Studies on Effect of CpG Suppression on Vertebrate Proteome

A Thesis

Submitted in the partial fulfillment of the requirement

for the award of the degree of

MASTER OF SCIENCE

IN

BIOTECHNOLOGY

Under the supervision of:

Dr. Vikas Handa

Assistant Professor

Submitted by:

Ravinder Kaur

Roll no. 301501013

DEPARTMENT OF BIOTECHNOLOGY

THAPAR UNIVERSITY, PATIALA -147004

DECLARATION

I hereby declare that the project work entitled "**Studies on Effect of CpG suppression on vertebrate proteome**" in the partial fulfillment of the requirement for the award of the degree of **Master of Science** in Biotechnology, Department of Biotechnology, Thapar University, Patiala, is an authentic record of my work during the period of six months from January 2017 to July 2017, under the guidance of **Dr. Vikas Handa**, Assistant Professor, Department of Biotechnology, Thapar University, Patiala. The matter embodied in this thesis has not been submitted in any part or full to any other University or Institute for the award of any degree in India or abroad.

Ravinder Kaur
M.sc (Biotechnology)
301501013
Thapar University

Date: 17th July 2017
Place: Patiala

CERTIFICATE

This is to certify that the thesis entitled "**Effect of CpG suppression on vertebrate proteome**" submitted by **Ms. Ravinder Kaur** in partial fulfillment of requirement for the award of degree of **Masters in Science** in the Department of Biotechnology, Thapar University, Patiala, is the record of the candidate's own independent original work carried out by her under my supervision and guidance. The matter embodied in this thesis has not been submitted in part of full to any other University or Institute for the award of any degree.

Ravinder Kaur
M.sc (Biotechnology)
301501013
Thapar University

Dr. Vikas Handa
Assistant Professor
Department of Biotechnology

ACKNOWLEDGEMENT

I thank the Almighty for showering his blessings throughout the preparation of my thesis. First and foremost, I would like to express my sincere and profound gratitude to my thesis supervisor, **Dr. Vikas Handa**, Assistant Professor and, Department of Biotechnology, Thapar University, Patiala, for her valuable guidance, undaunted motivation, encouragement, constant support and sound advice. His rare academic and professional insight, commitment and admirable dedication to the subject have always been a source of motivation for me. I would also like to thank him for providing the best laboratory facilities for conducting my research work.

I express my special gratitude to Dr. Moushumi Ghosh, Head of Department of Biotechnology, Thapar University, Patiala, for all her 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 would like to thank my lab-mates, **Ms. Japleen Kaur** and **Ms. Meera Sharma** for their support and help.

I shall retain my thankful indebtedness to my mother **Sarbjit Kaur** and family for giving me freedom and opportunity to pursue my own interest and for believing in me and enduring with me during difficult times.

Last but not the least I would like to thank all my batch mates for their support and companionship. I am deeply thankful to my friends (**Ms. Manpreet Kaur** and **Ms. Saruchi Kashyap**) for their encouragement and help whenever required.

Ravinder Kaur

Date: 17th July 2017
Place: Patiala

TABLE OF CONTENTS

ABBREVIATIONS

(O/E)	Observed / Expected ratios
A	Adenine
AdoMet	S-adenosyl-L- Methionine
C^5	Carbon at 5^{th} position
CpA	Phophodiester bond between Cytosine and Adenine
CpG	Phophodiester bond between Cytosine and Guanine
CpT	Phophodiester bond between cytosine and Thymine
DNA	Deoxyribonucleic acid
Dnmt	DNA methyltransferases
Dnmt1	DNA methyltransferase 1
Dnmt 2	DNA methyltransferase 2
Dnmt3a	DNA methyltransferase 3a
Dnmt3b	DNA methyltransferase 3b
G	Guanine
MTases	Methyltransferases
N^4	Nitrogen at 4^{th} position
N^6	Nitrogen at 6^{th} position
NCBI	National Centre for Biotechnology Information
RNA	Ribonucleic Acid
T	Thymine
TpG	Phophodiester bond between Thymine and Guanine
XCI	X-Chromosome Inactivation

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

DNA methylation is an epigenetic modification that occurs exclusively at C5 position of the cytosine in eukaryotes in context of CpG dinucleotide. CpG dinucleotides are the hotspots for mutation which causes deamination of cytosine and 5-methylcytosine. Due to deamination, CpG/CpG is converted into TpG/CpA. This mutation leads to the suppression of CpG and over expression of TpG and CpA. Suppression CpGs in coding regions of genome may affect the proteome of eukaryotes having methylated genomes. Codons consisting of CGs can be classified as CGN, NCG and NNC-GNN. The abundance of amino acids encoded by the three types of CG codons have been compared against their expected frequency in methylated genomes *(Homo sapiens and Mus musculus*) and unmethylated genomes as control(*Bacteriophage lambda, Haemophillus Influenza and E.coli.*). The amino acids analysis is not significant. The effects have been showed due to other evolutionary forces. But in di-amino acid analysis, Homo sapiens and mus musculus genome, showing (< 1 O/E) values lesser than *Bacteriophage lambda* in total genome and in coding sequence, higher number of di-amino acids showing (<1 O/E) values in methylated genomes which indicates the effect of CG suppression in proteomes.

Keywords: DNA methylation, CpG suppression, proteome, TpG, CpA.

CHAPTER 1

INTRODUCTION

Epigenetics is related to stably heritable phenotypes resulting from changes in a chromosome without alterations in the DNA sequence (Berger *et al.*, 2009). These changes may be induced by spontaneously, environmental factors or as a consequence of specific mutations. It involves three mechanisms: DNA methylation, histone modification and Non-coding RNAs. Epigenetic modifications include covalent modification of histone tails (such as acetylation, phosphorylation, ubiquitination and methylation). The epigenetic modifications are essential to many organism functions, such as gametogenesis, gene imprinting, X-chromosome inactivation, embryonic development, repression of transposable elements, ageing, silencing of transposons and tissue specific gene regulation activation etc. If these modifications occur improperly, this leads to adverse health effects (such as cancer & congenital diseases).

DNA methylation is the major epigenetic modification that affects DNA molecule in eukaryotes and control chromatin structure and gene expression. Changes in DNA methylation Profiles occur in physiological and pathological states, such as aging and metabolic diseases (Yan, J. et al., 2011). It is an enzymatic process by which methyl group is added to DNA at N6 position of adenine, and N4 and C5 position of cytosine (Jeltsch, A. 2002). However in eukaryotes methylation occurs only at 5 position of cytosine in context of CpG dinucleotide sequence (Rudolf Jaenisch and Adrian Bird, 2003). DNA methylation range occurs very low in arthropods, through intermediate in many non-arthropod invertebrates to high in the vertebrates (Bird 1980).

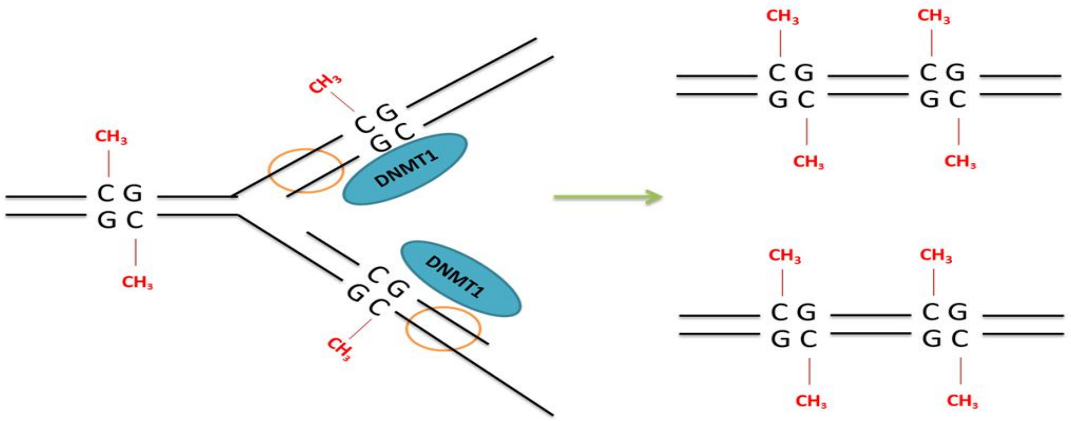
The other epigenetic modifications are histone modification and non- coding RNAs. Histone modification is a covalent post transcriptional modification that impinges on histone amino termini. Native cell within genome of human is packaged with histones & other proteins into chromatin. Histone proteins act to package DNA, which wraps around the eight histones, into chromosomes. The nucleosome is composed of an octamer of the 4 core histones (H3, H4, H2A & H2B) around which 147 base pairs of DNA are wrapped. Histone amino-terminal modification can produce antagonistic and synergistic affinities for chromatin associated proteins which dictate dynamic transitions between transcriptionally active and silent chromatin states (Jenuweinand and Allis, 2001).

Non-coding RNAs (ncRNA) is a functional RNA molecule that transcribed from DNA but not translated into protein. In general ncRNAs function to regulate gene expression at the transcriptional & post transcriptional level. These ncRNA divide into 2 main groups: short ncRNA (<30nts) & long ncRNA (>200nts). There are 3 major classes of non-coding RNAs are micro RNA (miRNA), short interfering RNA (siRNA) and piwi interacting RNAs (Inbar-Feigenberg, M. et al., 2013).

DNA Methyltransferases- DNA methylation is catalyzed by the enzymes called DNA methyltransferases (DNA MTase) which transfer the methyl group (-CH_3) to DNA. All the DNA methyltransferases use S-adenosyl L-Methionine (Ado Met) as the methyl donor because Ado Met is very effective donor for methyl groups (Jeltsch, A. 2002 and Singal, R. et al., 1999). In mammals, DNA methylation can be classified into 2 classes: Maintenance methylation and *de novo* methylation.

Maintenance methylation activity is important to preserve pattern of DNA methylation after every DNA replication cycle. Maintenance methylation methylates a hemi-methylated DNA that is methylated in one strand and un-methylated in other double strand DNA (M.Nakao, 2001). It gives the methylation pattern to newly synthesized daughter strand (that were also methylated in parent strand), and ignores un-methylated sites (that were also un-methylated in parent strand). Dnmt1 plays the role of maintenance methylation (Ramsahoye, B.H. et al, 1996).

Fig 1: Maintenance methylation

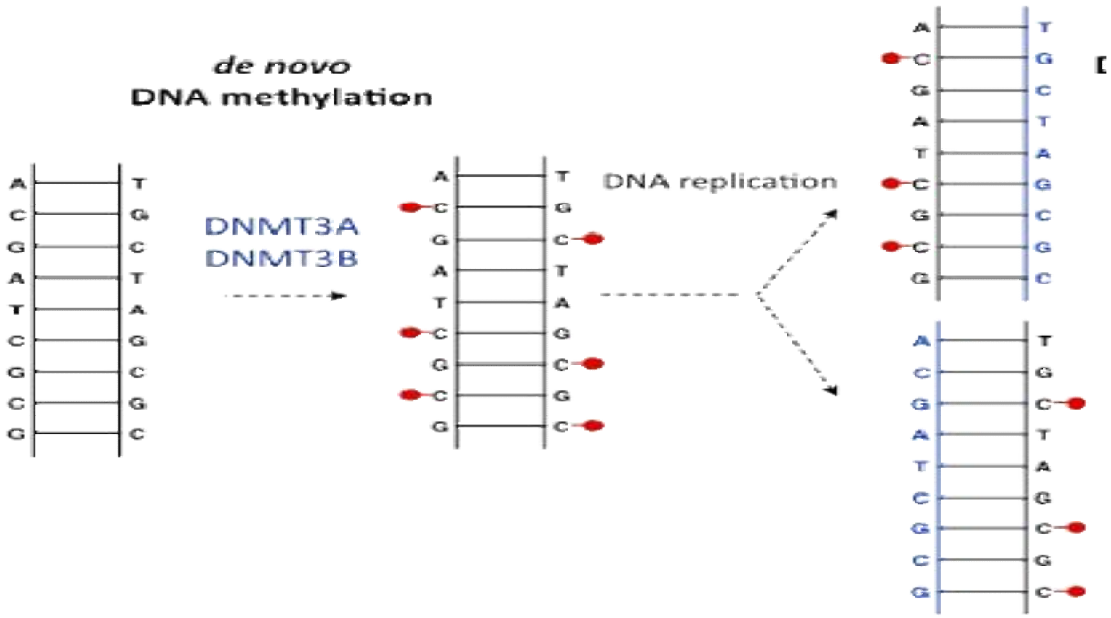


Source: https://sites.tufts.edu/epigenetics/maintenance-of-epigenetic-marks

De-novo methylation adds a methyl group to un-methylated base pair, resulting in the formation of a new hemi – methylated and then fully methylated base. It set up methylation pattern early in development. Dnmt3a and Dnmt3b play role for de-novo methylation (Klose, R.J. and Bird, A.P., 2006).

Dnmt2 has been identified as a DNA methyltransferases homolog. It lacks the large N-terminal regulatory domain. Therefore, the structure of the smaller domains also differs from known M Tases. Dnmt3l is a third member of Dnmt family. It is similar in related sequence to Dnmt3a/b but lacks functional catalytic activity (Auclair G. And weber M. 2012). Dnmt3l is abundant in early embryos and germ cells.

Fig 2: de-novo methylation

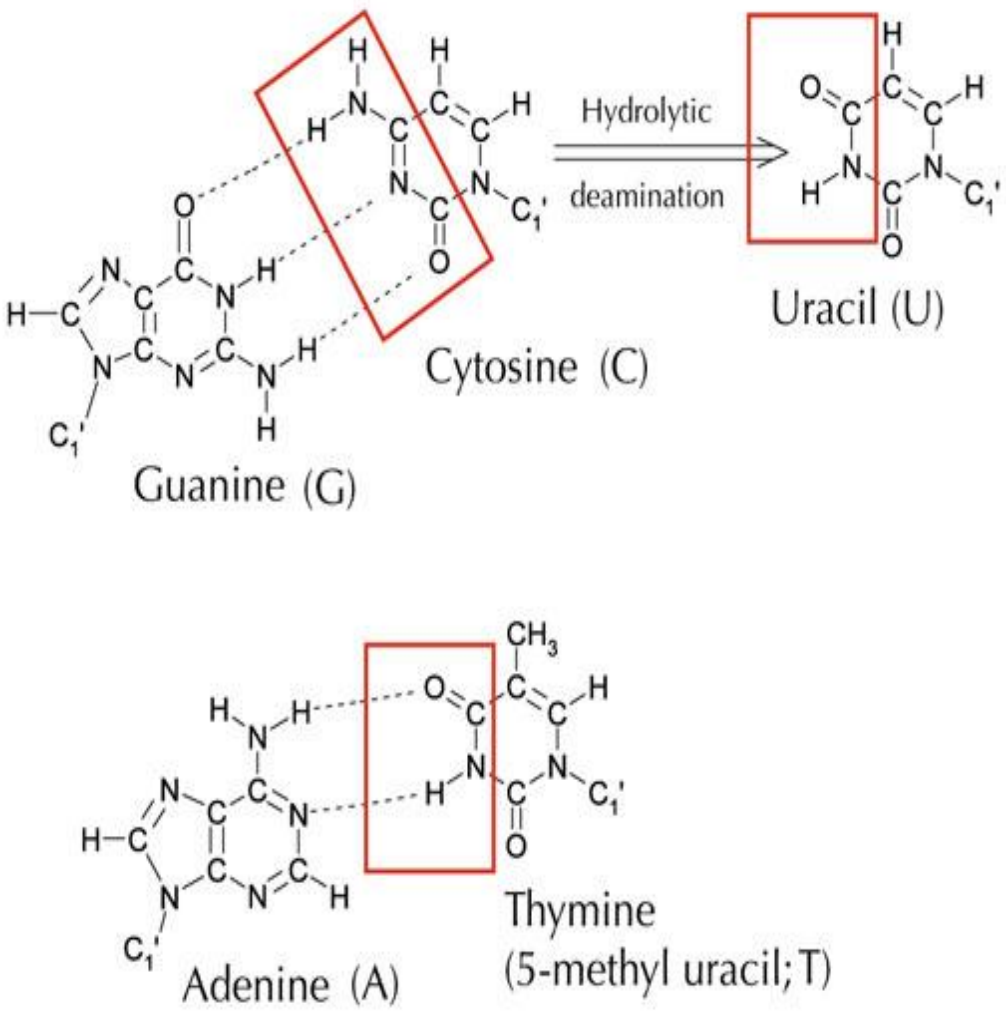


Source: Moison *et al.*, 2014.

The dinucleotide CpG, the methylation site is extremely underrepresented in vertebrate genome sequence. The observed CpGs in the vertebrate genomes are found to be five-fold fewer than the expected number (Bird and Tggart 1980). Approximately 80% of CpG's are methylated at the 5' cytosine. CpG dinucleotides are the mutational "hotspots" in the vertebrate genome. Mutations that takes place at CpG dinucleotide in the vertebrate genome cause spontaneous deamination of 5-methylcytosine to thymidine.

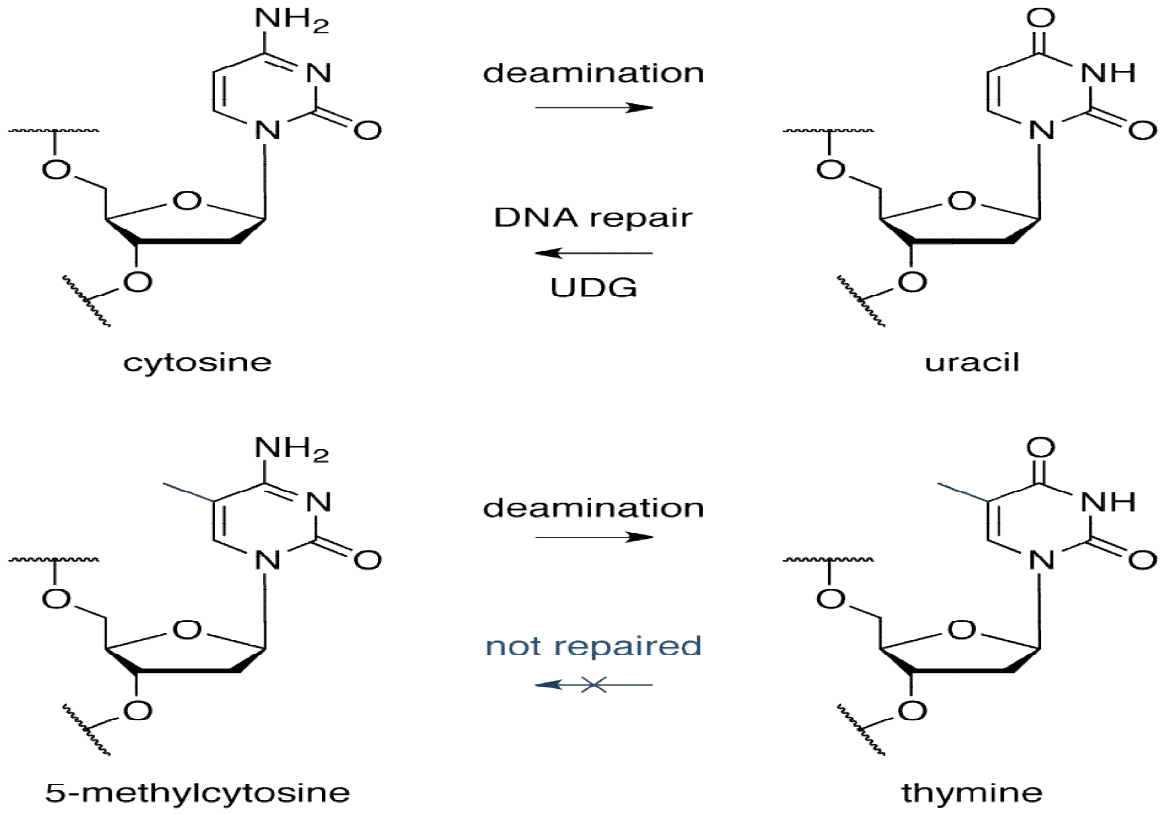
This result, the conversion of cytosine into uracil, through a process known as hydrolytic deamination (Adams, R. L. P. & Eason, R., 1984). 5-methylcytosine is deaminated 2-4 times more rapidly than cytosine. The rate of deamination of cytosine and 5-methylcytosine residues can be increases by DNA C-MTases, at low concentration of AdoMet After deamination of cytosine, when cell next replicates its DNA; the position opposite to the uracil would be taken by an adenine instead of guanine and change the message that this DNA encodes. Thus, CpG/CpG is converted into TpG/CpA.

Fig 3: AT &GC base pairing



Source: http://www.atdbio.com/content/56/Epigenetics.

Fig 4: Deamination of cytosine and 5-methylcytosine



Source: http://www.scienceinschool.org/2011/issue18/uracil.

.Spontaneous deamination causes Cytosine : Guanine pair (C:G) to mutated to Uracil: Guanine (U:G) mismatch. To maintain the integrity of genome these mutations repaired by DNA mismatch repair enzyme, Uracil DNA glycosylase. The enzyme detects a U:G mismatch and removes the Uracil replacing it with a Cytosine. If Cytosine is methylated at 5th position, its deamination results in Thymine (5-methyl Uracil). The G:T mispair resulting from deamination of 5-mCyt is more difficult for the cell to repair than G:U mispair, since thymine unlike uracil is a natural base of DNA (Mark L. Gonzalgo, & Peter A. Jones).

Due to deamination, CpG/CpG is converted into TpG/CpA and one 5-mC change would because loss of 2CpG and the gain of one TpG & one CpA (Bird, 1980). This mutation leads to the suppression of CpG and overrepresentation of TpG and CpA. Thus, DNA methylation causes the CpG deficiency in the methylated genomes of higher eukaryotes. The most extreme CpG deficiency (i.e. in vertebrates) have highest level of DNA methylation and poorly methylated genomes display no CpG deficiency while partially methylated genomes are intermediate deficient in CpG.

DNA methylation plays key role in genomic imprinting (which is extremely important in gene regulation) and in X-chromosome inactivation. Genomic imprinting is a process by which function of a gene is dependent on its parental origin. It involves that the two parental chromosomes are not equivalent and show either maternal or paternal expression in genome. When the imprinting is takes place, the maternal and paternal genomes contain specific alleles that will either express or repressed. If the allele is inherited from the father is imprinted, it means silenced then, only the allele from mother is expressed and if the allele from mother is imprinted, then father allele is expressed. Improper imprinting can lead 2 active copies or inactive copies leads to development of abnormalities or cancer. Over 40 imprinted genes are known in mammals.

In mammals, female carry 2 X-chromosomes while males have only one. For equal expression level of X-encoded genes in female XX and male XY cells, one of the copies of the X-chromosome present in female mammals is inactivated and this process is called X-chromosome inactivation (XCI). X-chromosome inactivation results in one inactive X-chromosome (Xi) and one active chromosome (Xa). Errors in DNA methylation cause multifactorial diseases and development of human cancer. DNA methylation also affects the gene transcription (Barakat & Gribnau, 2012).

In mammals, DNA methylation occurs at CpG dinucleotides. The GC rich (60-70%) sequences containing clustered unmethylated CpGs known as CpG islands. CpG islands are the regions having high number of CG dinucleotides (Handa & Jeltsch, 2005). These islands are 200 or more than 200 base pair longer with 50% G+C content (Takai & Jones, 2002). CpG islands having a high CpG frequency and they are clearly undermethylated. These islands are more ordinary among housekeeping genes (vinhinen, M. et al, 1996). They play important role in genome imprinting, gene silencing and X-chromosome inactivation. Approximately, 60% of genes contain CpG islands in their 5' (promoter) region.

CHAPTER 2

REVIEW OF LITERATURE

2.1 Epigenetics

It is a stably heritable phenotype resulting from changes in a chromosome without alterations in the DNA sequence. It involves three mechanisms: DNA methylation, Histone modification & non-coding RNAs. There are 3 types' signals that culminate in the formation of a stably heritable epigenetic state: "Epigenator", which emanates from the environment & activate an intracellular pathway (environmental variations), "Epigenetic Initiator" signal, which response to the epigenator & it is required to define the precise location of the epigenetic chromatin environment (non-coding RNAs), "Epigenetic Maintainers" signal, which support the chromatin environment in the first & following generations (histone/DNA modifiers) (Berger et al., 2009).

The stable alterations in gene expression are arising during development & cell proliferation. The epigenetic process is essential for differentiation & development, but they can also arise in mature humans and mice, either under the effect of environment or by random change (Jaenisch & Bird, 2003).

2.2 DNA methylation

It is the most abundant modification that affects the DNA molecule in eukaryotes. It occurs at the N6 position of adenine and the N4 & C5 positions of cytosine but only C5 position is observed in the eukaryotes (Hermann et al., 2004 and Bender, J., 2004).

DNA methylation is a modification that does not change the information which is coded by the DNA but participates in the modification of gene expression. In mammals, DNA methylation occurs at the C5 position of cytosine, mostly upstream of guanine in DNA double helix, therefore called CpG dinucleotides (Moison.C, et al., 2013).

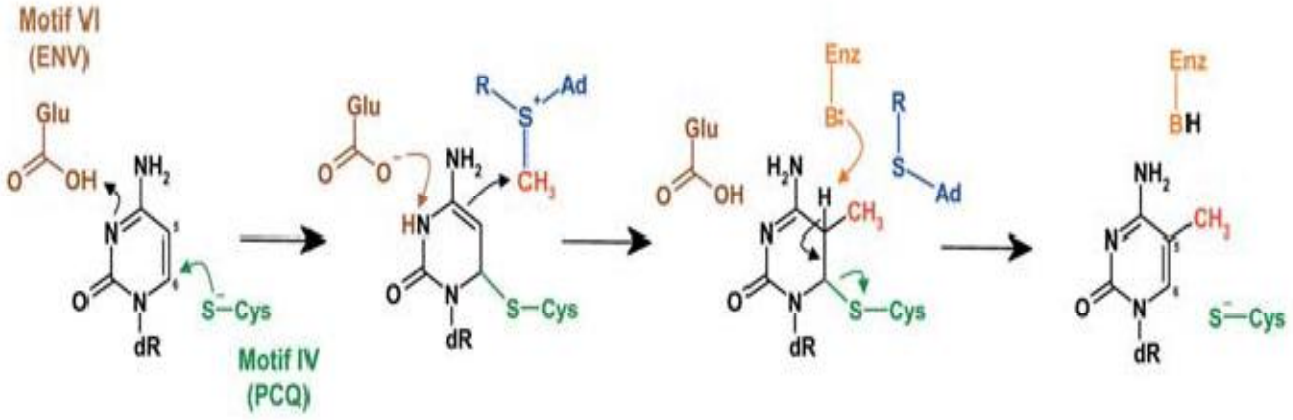
DNA methylation range occurs very low in arthropods, through intermediate in non-arthropod invertebrates to high in the vertebrates (Bird, 1980).

DNA methylation is a common factor of vertebrate genome which is predominately occurs at cytosine in CpG dinucleotide & converts 5-methylcytosine (Bird & Taggart, 1980 and Mugal, C.F., et al., 2015). It occurs at different sequence contexts although at much lower frequency. It is known that dinucleotide CpG is extremely underrepresented in genome sequence of organism that methylates their nuclear DNA. These 5-methylcytosine or CpG dinucleotides are the mutational "hotspots" in the vertebrate genome (Tom S. Shimizu, etal., 1997).

2.2.1 DNA Methyltransferases

DNA methylation occurs by enzymes known as DNA methyltransferases (N Tases). These all methyltransferases use Ado Met (S-adenosyl-L-Methionine) as the source of methyl group. The methyl group of the AdoMet is bound to the sulphonium atom which destabilizes the molecule. The reaction mechanism initiate with nucleophillic attack of the enzyme on the target cytosine ring at C6 to makes a covalent complex between the DNA & enzyme. The formation of covalent bond activate the C5 atom towards electrophillic attack & this leads to the addition of methyl group at C5 position of cytosine followed by elimination of proton at 5 position & resolution of the covalent intermediate .

Fig 5: Methylation reaction of DNA (cytosine –C5) MTases



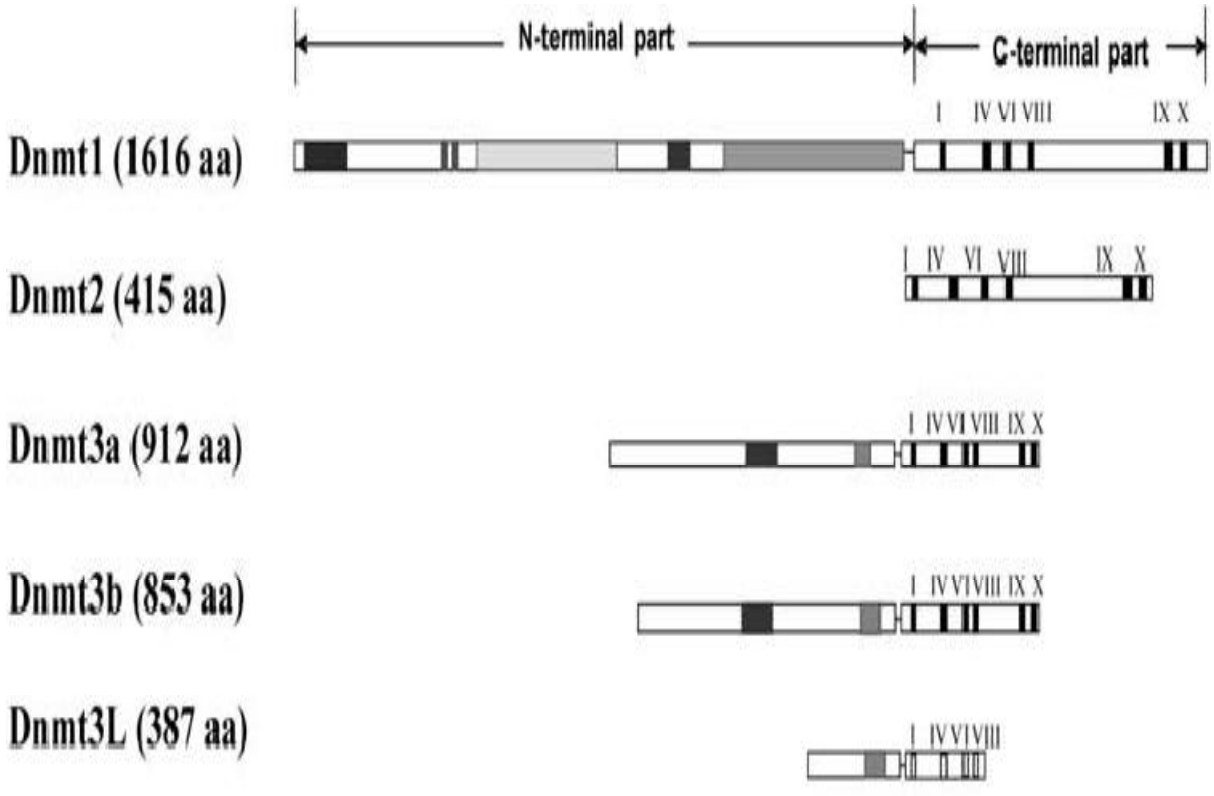
Source: Hermann *et al.*, 2004.

.There are two methylation process occurs: de-novo methylation (which establishes the methylation state by Dnmt1) & Maintenance methylation (which copies it into daughter DNA strand by Dnmt 3a & Dnmt 3b) (A. Hermann et al., 2004).

Dnmt1: Dnmt1 exhibits preference for hemi-methylated DNA sites. Dnmt 1 contains large N- terminal domain with regulatory function and a small C- terminal catalytic domain. Dnmt1 is a large protein which contains 1620 amino acid residues. It has 5-30 fold preference for hemi-methylated substrates & as a result has been allow function in the maintenance of methylation pattern (Timothy H. Bester, 2000).

Dnmt3: This family consists of Dnmt3a & Dnmt3b, which are highly related to each other but encoded by different genes. These were found to methylate CG dinucleotides without preference for hemi-methylate sites (A. Jeltsch, 2002).

Fig 6: Domain organization of mammalian Dnmts

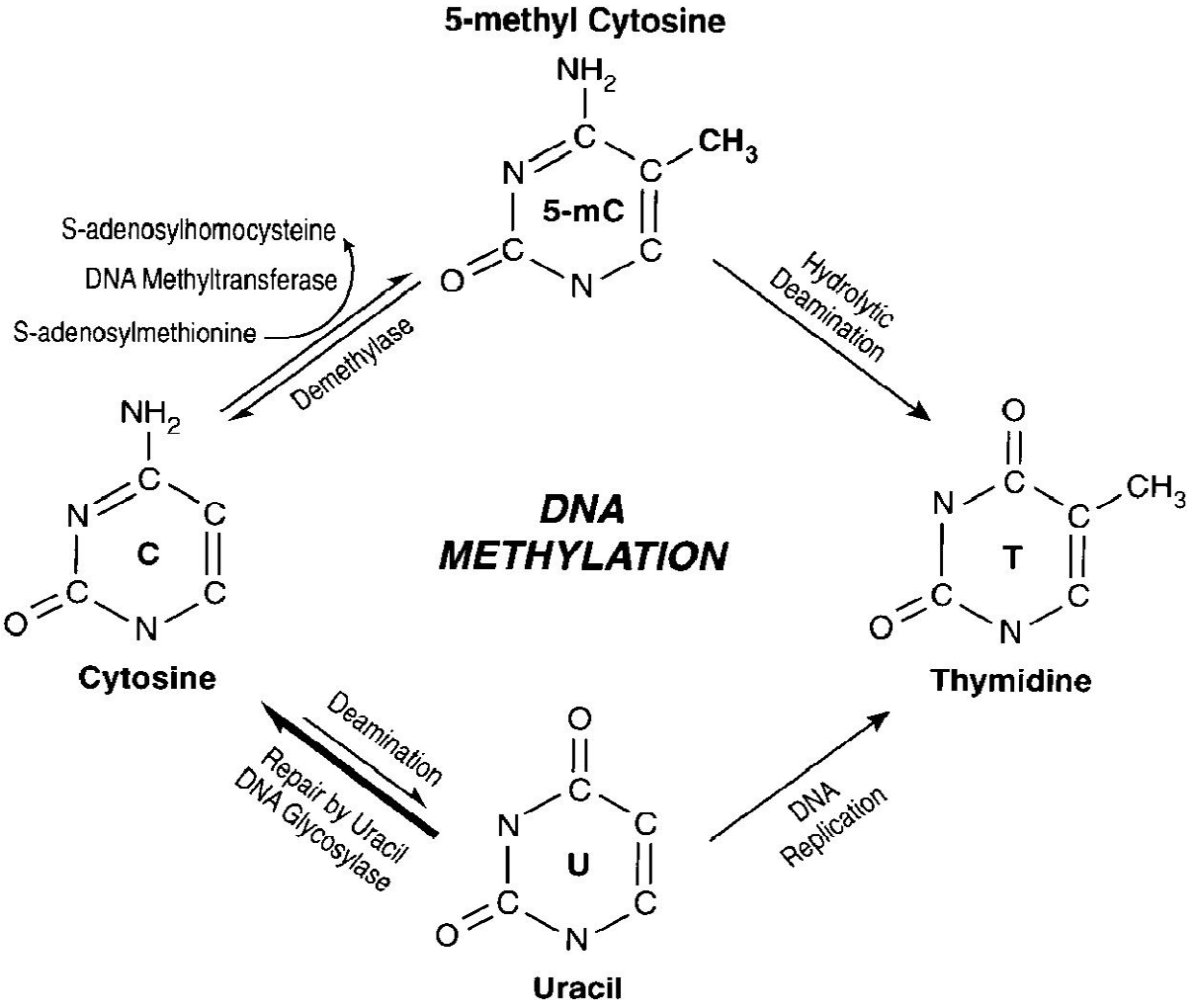


Source: Hermann *et al.*, 2004.

2.2.2 Deamination

In higher eukaryotes, the presence of 5-methylcytosine in the DNA has the disadvantage i.e. potentially mutagenic, because it causes spontaneous deamination of 5-methycytosine to thymidine & this results, the hydrolytic deamination of cytosine to uracil. It is initiated by hydroxy ion attack at the C4 of cytosine base that is protonated at N3 position. The rate of deamination of 5-methylcytosine and cytosine is increase by DNA C-MTases, at low concentration of AdoMet After this, when cell next replicates its DNA, the position opposite to uracil would be taken by adenine instead of guanine & change the message that encoded by DNA. Thus, CpG/CpG is converted into TpG/CpA (A. Jeltsch, 2002).

Fig 7: DNA methylation



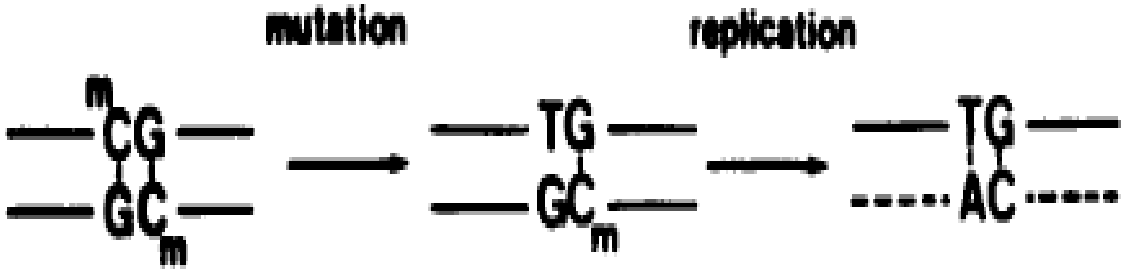
Source: Singal, R. *et al.*, 2017.

To maintain the integrity of genome these mutations repaired by DNA mismatch repair enzymes. Uracil is a non-natural base in DNA & the uracil bases are removed by enzyme, Uracil DNA glycosylase. But in case of thymidine, is a natural base of the DNA that cannot be repaired by any mechanism or pathway.

2.2.3 CpG suppression

DNA methylation causes CpG deficiency because CpG is the only one in genomes that are heavily or partially methylated. The organisms with high level of methylation are extremely CpG deficient but poorly methylated genomes display no significant CpG deficiency and partially methylated genomes are intermediately deficient in CpG. Since, one 5mc change would cause loss of 2 CpG and the gain of one TpG & one CpA. This mutation leads to the suppression of CpG & over expression of TpG & CpA.

Fig 8: Mutation from methylated mCpG to TpG & CpA



Source: Bird, P.1980.

Mutation of 5mc causes mismatch T-G pair, if this mismatch is not repaired, it give rise to a T-A pair after replication (Bird, 1980).

In vertebrate DNA low CpG is the result of deamination which occurs in mCpG dinucleotide. DNA methylation occurs in high amount than invertebrate DNA. The analysis was done on the number of nucleotide sequences & the analysis supported the variance that the vertebrate genome is deficient in CpG dinucleotide. In addition, the deamination is less extensive in the more stable & high G+C rich regions but if G+C content is lower than it is also deficient in CpGs. The Genome region with 70% G+C content have high number of CpG but region with 40% G+C completely devoid of CpGs (R.L.P. Adams & R. Eason, 1984).

Due to high mutation rate from C to T in methylated CpGs causes CpG deficiency and this CpG deficiency varies according to the G+C content of the sequence (Duret & Galtier, 2000).

All animals without exception (vertebrates & invertebrates) mitochondrial sequences show significant CpG suppression. The methylation-deamination-mutation scenario (CpG to TpG-CpA mutations) with associated excess of TpG & CpA dinucleotide may not apply to mitochondrial genome. Since, in invertebrates the associated methylase is absent and the methylase cannot access the mitochondrial organelle in vertebrates. Some experiments have been unable to detect methylase activity in vertebrates mitochondrial DNA. Furthermore, contradictory to expectations under CpG methylation the methylase mutations or CpG suppression is not overrepresented in mitochondrial sequences (Karlin et al., 1993).

The dinucleotide relative abundances in 1657 human genes and in non- coding genes sequences shows consistant strong biases that reflecting avoidance of CG and TA at all codon sites and in non-coding regions. Specifically, CG at first and second codon site is approximately 0.70, second and third codon site is approximately 0.36 and at third and fourth codon site is approximately 0.44. But in the non-coding sequence CG is approximately 0.36. The higher value of CG at codon site first and second probably reflects requirements of Arginine usage. The TA values are about equally low at codon positions first and second (0.54), second and third (0.56) and third and fourth (0.61) in human introns and intergenic regions.

A paradox related to the phenomenon of G as most frequent nucleotide (about 32.3%) at codon sites and C is most frequent nucleotide (29.3%) at codon site 3. Whereas the dinucleotide CG frequency is significantly deficient. To explain this inequality recognizes the methylation/deamination/mutation pathway and it was assume that methylation/deamination/ mutation creates mutations at nucleotide C at much more then at nucleotide G (Karlin et al, 1996).

2.3 CpG islands

The GC rich (60-70%) sequences containing clustered unmethylated CpGs known as CpG islands. They do not show any suppression of CpG. CpG islands founds at the 5' ends of all housekeeping genes. DNA of CpG Island is found in shorter regions of 1-2 kb, which together account for approx. 2% of the genome, & it has distinctive properties when compared with the DNA in the rest of the genome (Cross & Bird, 1995).

In 40% mammalian genes, CpG islands are present in the exonic & promoter region. They plays important role in genome imprinting, gene silencing, X-chromosome inactivation, carcinogenesis and the silencing of intragenomic parasites. CpG Island is a 200 bp region of DNA with a high G+C content (greater than 50%) (Takai & Jones, 2001).

2.4 Amino Acid Composition of Proteins

The analysed correlation between third and first + second Codon positions leads to change the composition of amino acids. First and second Codon positions are related with specific amino acids. Thus, third Codon position increase or decrease in GC levels, amino acid composition tends to change towards amino acids relative to codons characterized by lower or higher GC levels in first and second Codon positions. This difference affects the relative abundance of Alanine, Glycine, Proline and Arginine on the one hand and Asparagine, isoleucine, isoleucine, lysine, leucine Methionine and tyrosine on the other hand, in the proteins that are encoded by genes that are GC rich or GC poor in their first and second position (Bernardi et al, 1991).

The G+C content of the coding DNA sequences creates an unexpected bias in amino acid composition. The amino acid present in proteins are forced by the nucleotide composition of the related gene. Because there is frequently a strong bias in the composition of the leading and lagging strand of the chromosomes, prominent from each other by enrichment in C+A and G+C , Thus, proteins coded from leading strand are enriched in valine comparatively to those coded from lagging strand which are enriched in isoleucine and threonine. Additionally, there is a general bias in GNN codons that could affect the overall amino acid composition of the proteomes. This bias creates inequity between Asn, Cys, Ile, Lys, Met, Phe, Ser, Thr, Trp, Tyr on one hand ala asp Gln, Glu, Gly, His, Pro, Val on the other hand (Pascal et al, 2006).

2.5 Codon bias frequency

Codon bias refers to the difference in the frequency of occurrence of synonymous codons in a coding DNA. In human DNA, there is strong dinucleotide relative abundance bias, at the different Codon sites. At Codon site 1, the dinucleotide G is the most frequent and the frequency of the nucleotide A at Codon site 2 is at the same level. These reflect the amino acid usages, mostly those acidic residues whose average is 12% in human proteins. The lowest frequency at site 1 is about 16% of T. Overall, frequencies at codon site 3, shows the order C > G > T > A with frequency C= 29% and frequency A=19%.

The frequency of GA at Codon sites (1 &2) is usually high evidently reflecting high usage of acidic amino acids whereas TA is the lowest, probably reflecting stop Codon avoidance (Karlin and Mrazek, 1996).

2.6 Correlation of Codon usage for different amino acids

A strong negative correlation within protein sets of vertebrate species is observed regarding amino acid usage of strong verses weak Codon types. In this respect, the frequency of strong at Codon site 1and 2 has a correlation of about -0.4 with weak. This negative correlation reflects high G+C region alternating with low G+C region of the gnome probably conforming the isochore phenomenon in mammalian species (Karlin and Mrazek, 1996).

CHAPTER 3

SCOPE OF STUDY

Study methylated genomes to see the effect of CG suppression. In our case we try to indicate the effect of CG suppression on proteomes. But our results are not significant in amino acids. In di-amino acids the effect of CG loss is reflected in the proteomes of methylated genomes when compared with un-methylated genome. Such comparison is expected to reveal about the effect of CG suppression on proteomes.

CHAPTER 4

OBJECTIVES

- To determine the frequency of amino acids and dipeptide sequences in polypeptide sequences of different organisms.
- To compare the abundance of amino acids and dipeptide sequences encoded by sequences consisting of CG dinucleotides among methylated and unmethylated genomes.

CHAPTER 5

MATERIALS AND METHODS

5.1 Data source

The DNA sequences were downloaded from the National Centre for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov/). To study the effect of CpG suppression on vertebrate proteome five genome sequences was selected. The genome sequences that are selected are:

Table 1: Genome sequences taken from Gen Bank

Genome sequence name	**Accession number**
Bacteriophage Lambda	NC_001416.1
Haemophillus Influenza	NC_000907.1
Escherichia coli	NC_002695.1
Homosapiens Chromosome 21	NC_000021.9
Mus musculus chromosome 19	NC_000085.6

5.2 Sequence analysis tools

Microsoft excel: The data was analysed on the Microsoft Excel Sheet by using applications of the sheet Excel sheet used for statistical and computational analysis. The tools which are used:

Table 2: Microsoft Excel Functions

Tools	Function
COUNTIF	Statistical function, to count cells that meet single criteria. It is used to count numbers, dates & text that match specific criteria.
CONCATENATE	Used to combine text from different cells into one cell.
SUM	Used to sum a column or row of numbers.
IF	One of the logical functions, to return one value if a condition is true and another value if its false

Notepad++: Notepad++ used for macro recording (that records the work for playback at a later time) to analyze or manipulate DNA sequences. Macro recording manipulate genome sequence such as convert sequence into one single string. Macro recording was applied on those sequences which are required to be repeated several times.

Computation of mono- and di-nucleotide frequencies in DNA sequence: http://WWW.biohp.org/minitools/chaos game representation/demo.php?FCGR). This tool is used to determining mono and di-nucleotide frequency of DNA sequences. Chaos game representation of genome sequences has been used for visual representation of genome sequence patterns as well as alignment free comparisons of sequences based on oligonucleotide frequencies.

5.3 Methods

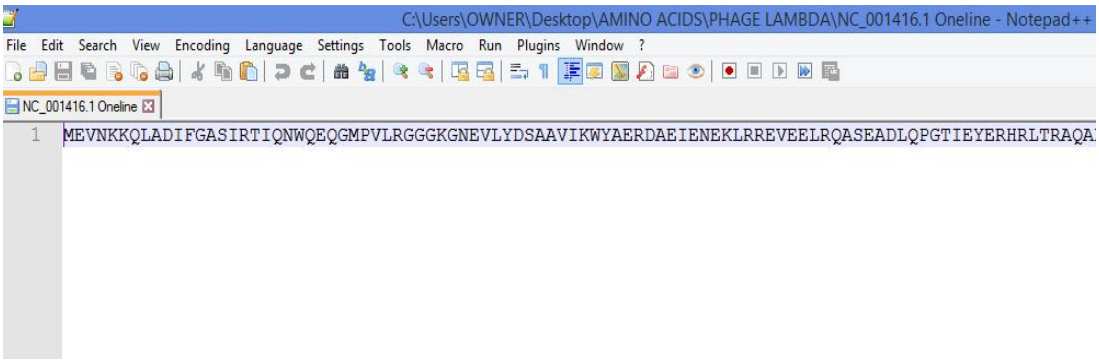
5.3.1 Analysis for amino acid and di-peptide frequency

The five genome sequences are downloading from the NCBI (National Centre for Biotechnology Information) site by using the appropriate accession number of each genome

sequence. The genome sequence is downloading into the fasta format. For further analysis, downloaded sequence is paste into Notepad++ & using following steps:

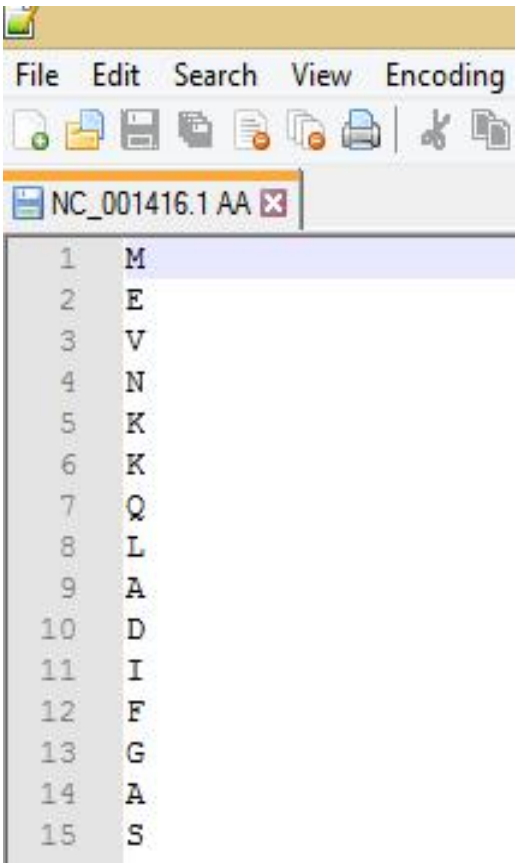
1) In Notepad++, with the help of Macro recording the sequence is converted into a single line.

Fig 9: Conversion of genome sequence into single line in notepad++



2) Recorded macro again for amines and diamines in single string.

Fig 10: Recorded macro for Amines **Fig 11 (11.1): Recorded macro for di-amines**



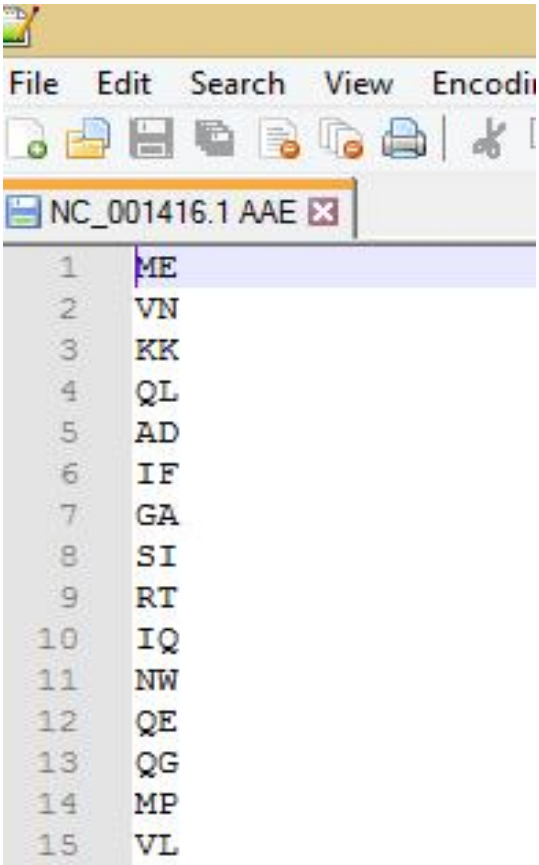
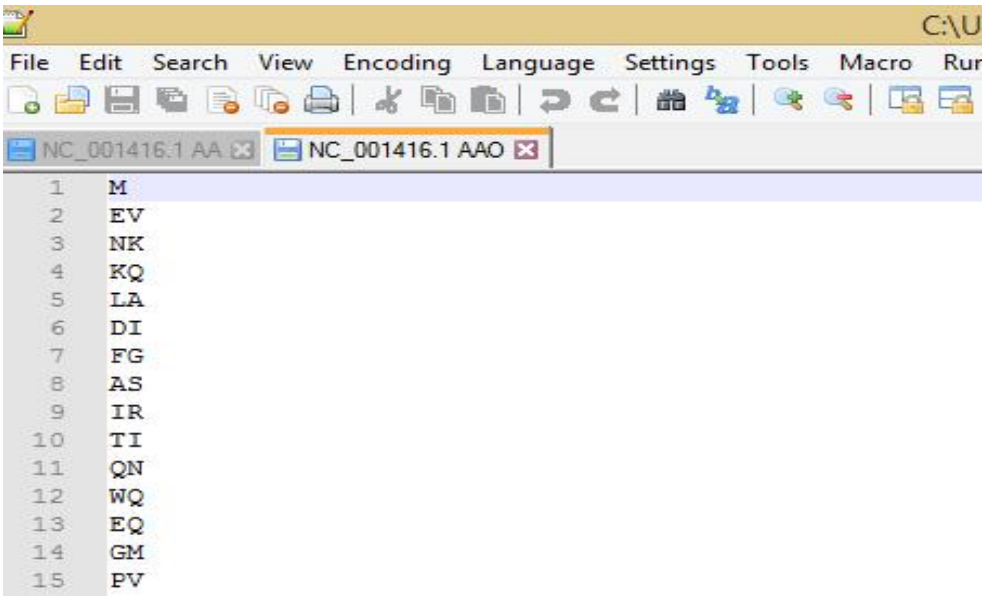
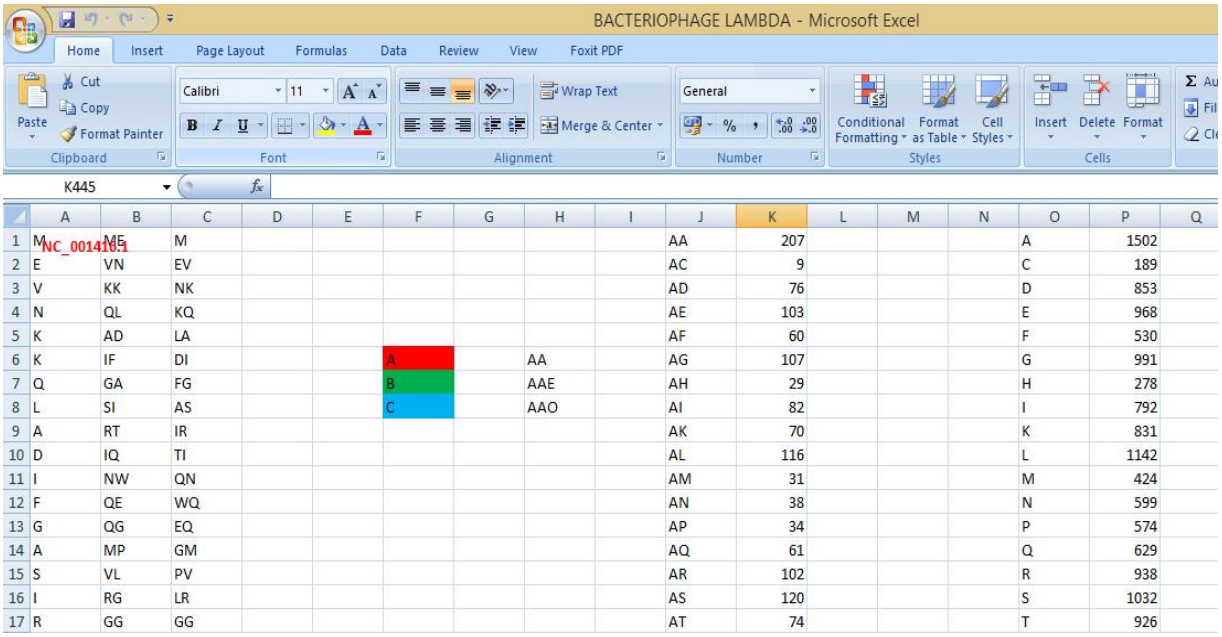


Fig 11.2: Recorded macro for di-amines



3) Calculated the frequency of amino acids and di-peptide of all five genome sequences by using the COUNTIF & CONCATENATE formula.

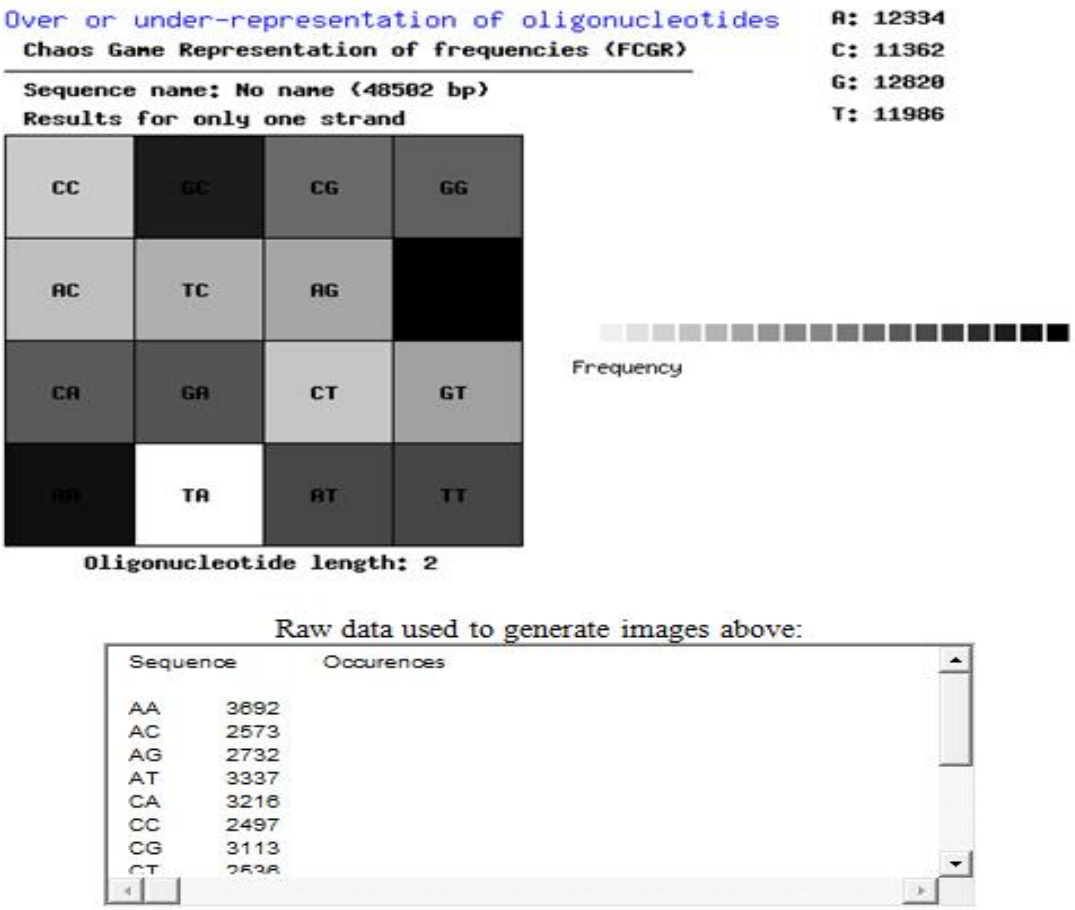
Fig 12: Calculated frequencies of amines and di-amines in MS Excel.



5.3.2 Analysis for base composition of genomes

The base compositions of genomes are calculated by the FCGR (Chaos Game Representation of Frequencies). The sequence is downloaded from the NCBI and then paste into FCGR and calculates for only upper strand.

Fig 13: Analysis for base composition



Raw data used to generate images above:

Sequence	Occurences
AA	3692
AC	2573
AG	2732
AT	3337
CA	3216
CC	2497
CG	3113
CT	2538

CHAPTER 6

RESULTS

In higher eukaryotes, only cytosines are methylated at position 5 in context of CpG dinucleotides. The CpG dinucleotides are hotspots for mutation in the vertebrate genome. When 5-mC is spontaneously deaminated, it gets converted into thymine it produces a T:G mismatch. Repair of the mismatch may produce TpG/CpA mutant in half of the cases. Such mutations in the course of evolution have resulted in suppression of CpG and overrepresentation of TpG and CpA in methylated eukaryotic genomes. One 5-mC change would cause loss of two CpG and the gain of one TpG and one CpA.

6.1 Effect of CG suppression on proteomes

Due to CG methylation induced mutation, CG/TG is converted into TG/CA. Many such mutations might take place in the coding regions. Some of the mutations in the coding region might lead to amino acid substitution in the encoded polypeptide sequences. There are two types of codons that contain CG, CGN & NCG and those where CG is split in two successive codons as NNC-GNN. When CG contain codon is converted into TG and CA then it may affect the amino acids which are coded by them.

CGN & NCG: There are two possibilities i.e. NCG is converted into NTG or NCA for example, GCG is a codon which codes for Alanine (A) is converted into GTG which codes for Valine (V) or GCA which codes for Alanine (A). In this case, Alanine loss occurs if it converts into valine.

NCG	⟶	NTG	Or	NCA
(A) G**CG**	⟶	G**TG** (V)	Or	G**CA** (A)
(T) A**CG**	⟶	A**TG** (M)	Or	A**CA** (T)

Another possibility is CGN converted into TGN or CAN for example CGG codon (codes for Arginine) converts into TGG (codes for Tryptophan) or CAG (codes for Glutamine) codon. In this, Tryptophan (W) or Glutamine (Q) is gain and Arginine (R) is lost because of CpG suppression.

CGN ⟶ TGN Or CAN

(R) **CG**G ⟶ **TG**G (W) Or **CA**G (Q)

(R) **CG**A ⟶ **TG**A (Stop Codon) Or **CA**A (Q)

CG split in two successive codons: Same occur in case of the codons that partially contain CG as NNC GNN is converted into NNT GNN or NNC ANN and affects the amino acids.

NNC GNN ⟶ NNT GNN NNC ANN

CAC GAT ⟶ CAT GAT or CAC AAT

H**D** H**D** H**N**

GTC GGT ⟶ GTT GGT GTC AGT

V**G** V**G** V**S**

6.2 Determining the frequency of all the amino acids in 5x genomes

The frequency of all amino acids is calculated by using the "COUNTIF" formula in MS Excel.

Table 3: Frequency of all amino acids

	Bacteriophage lambda	*Haemophilis influenza*	*Escherichia Coli*	*H.sapiens chr.21*	*M.musculus chr.19*
A	1502	42144	148245	50858	100471
C	189	5256	18284	16886	31029
D	853	25546	82019	35473	73828
E	968	33130	91791	54563	104699
F	530	22524	59649	26708	51948
G	991	34043	115219	46169	94592
H	278	10416	34808	19525	38384
I	792	36040	92588	33287	62669
K	831	32017	70879	45805	82396

L	1142	53443	162877	74010	146756
M	424	12340	43300	15278	32335
N	599	24709	63707	27328	52244
P	574	18912	68996	45524	100169
Q	629	23657	69624	37093	68247
R	938	22756	88192	41611	85250
S	1032	29672	94096	63788	130575
T	926	26540	86586	39845	76936
V	973	34187	109831	46689	87017
W	243	5713	24023	9363	17044
Y	452	15881	45235	19242	38957
Sum	14866	508926	1569949	749045	1475546

6.3 Determined base composition of 5x genome:

Table 4: Base composition of Genome

Genome	**Size (480bp)**	**Accession Number**	**Coding Sequence**		**Complete Genome Sequence**	
Bacteriophage Lambda	48502	NC_001416.1	A C G T	12058 10441 12386 9932	A C G T	12334 11362 12820 11986
Haemophillus Influenza	1830138	NC_000907.1	A C G T	480603 272652 323062 457026	A C G T	567623 350723 347436 564241
E.coli	5498450	NC_002695.1	A C G T	1158534 1148762 1295858 1129934	A C G T	1361495 1386237 1392518 1358200
Homosapiens Chromosome 21	46709983	NC_000021.9	A C G T	640796 605137 628511 525240	A C G T	11820664 8185244 8226381 11856330

Mus musculus Chromosome 19	61431566	NC_000085.6	A C G T	1289416 1314017 1299976 1075235	A C G T	16732680 12449343 12440880 16602953

Base composition of genomes of coding and complete genome sequence is determined by FCGR (Chaos Game Representation of Frequencies). The sequence is downloading from the National Centre for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov/) and then paste into the FCGR. Then calculate it only for upper strand and create the FCGR and frequencies for the four bases are calculated.

6.4 Calculate probability of each codon based on base composition of amino acids:

Probability of each codon is calculated on the basis of base composition by calculating the P(N) value of each base. The P(N) value is calculated by dividing the frequency of each base by total frequencies of 4 bases.

Probability of each codon = P(N) * P(N) *P(N)

6.5 Calculate expected counts of amino acids:

The probabilities of each codon were used to calculate the probability of finding their corresponding amino acids. For example, Proline is encoded by four codons (CCG, CCA, CCT and CCG). The probability of Proline is P(CCG) + P(CCA) + P(CCT) + P (CCC). Expected count of every amino acid was determined by multiplying its probability with total number of residues in the proteome.

Table 5: Expected counts of Amino acids of complete genomic sequence

	Bacteriophage lambda	*Haemophilis influenza*	*Escherichia coli*	*H.sapiens chr.21*	*M.musculus chr.19*
A	920	18517	100241	31384	67347
C	467	14895	49021	22727	44830
D	481	14984	49140	22658	45181
E	518	14985	49311	22664	45338

F	437	24189	47813	32755	59924
G	1039	18344	100695	31542	67193
H	426	15126	48918	22545	45284
I	687	39430	71706	51917	95176
K	499	24482	48213	32567	61077
L	1331	54264	145750	77996	150156
M	247	9240	24319	13404	25820
N	463	24480	48045	32558	60865
P	816	18693	99788	31227	67501
Q	459	15127	49089	22551	45442
R	1439	33503	149552	54048	112685
S	1342	45056	146910	67890	135203
T	886	30253	98007	45096	90726
V	971	29791	98213	45460	89817
W	257	5656	24873	9329	19166
Y	450	24334	47929	32656	60393

Table 6: Expected counts of Amino acids of Coding sequence

	Bacteriophage lambda	*Haemophilis influenza*	*Escherichia coli*	*H.sapiens chr.21*	*M.musculus chr.19*
A	957	19066	104324	49473	101687
C	414	15209	49402	20227	39932
D	502	15993	50653	24678	47886
E	603	17615	54558	27711	51897
F	332	21515	43077	16904	33028
G	1135	22592	117682	51384	100601
H	424	13498	44903	23760	48403
I	641	37528	66623	32314	60983
K	587	26205	48777	28252	51476
L	1166	50670	137364	60325	119903
M	245	10017	25117	11467	21550
N	489	23793	45285	25160	47497
P	807	16091	92482	47633	102786
Q	508	14866	48365	26680	52458
R	1560	36682	158882	77183	153585
S	1270	42966	141619	66021	131994
T	932	28364	93269	50440	100861
V	910	31960	102614	42941	83209
W	252	6734	28094	11247	21727
Y	403	22625	44167	20623	39608

6.6 Analysis based on observed/expected frequency of Amino acids.

The observed counts of amino acids in different proteomes are divided by expected counts of amino acids in different proteomes. The values in the table are the result of observed/expected.

Table 7: Observed/Expected frequency of Amino acids of complete genome sequence

	Bacteriophage lambda	*Haemophilis influenza*	*Escherichia coli*	*H.sapiens chr.21*	*M.musculus chr.19*
A	1.63	2.28	1.48	1.62	1.49
C	0.40	0.35	0.37	0.74	0.69
D	1.77	1.70	1.67	1.57	1.63
E	1.87	2.21	1.86	2.41	2.31
F	1.21	0.93	1.25	0.82	0.87
G	0.95	1.86	1.14	1.46	1.41
H	0.65	0.69	0.71	0.87	0.85
I	1.15	0.91	1.29	0.64	0.66
K	1.67	1.31	1.47	1.41	1.35
L	0.86	0.98	1.12	0.95	0.98
M	1.72	1.34	1.78	1.14	1.25
N	1.29	1.01	1.33	0.84	0.86
P	0.70	1.01	0.69	1.46	1.48
Q	1.37	1.56	1.42	1.64	1.50
R	0.65	0.68	0.59	0.77	0.76
S	0.77	0.66	0.64	0.94	0.97
T	1.05	0.88	0.88	0.88	0.85
V	1.00	1.15	1.12	1.03	0.97
W	0.95	1.01	0.97	1.00	0.89
Y	1.01	0.65	0.94	0.59	0.65

Table 8: Observed/Expected frequency of Amino acids of coding sequence

	Bacteriophage lambda	*Haemophilis influenza*	*Escherichia coli*	*H.sapiens chr.21*	*M.musculus chr.19*
A	1.57	2.21	1.42	1.03	0.99
C	0.46	0.35	0.37	0.83	0.78
D	1.70	1.60	1.62	1.44	1.54
E	1.61	1.88	1.68	1.97	2.02
F	1.60	1.05	1.38	1.58	1.57
G	0.87	1.51	0.98	0.90	0.94
H	0.66	0.77	0.78	0.82	0.79
I	1.23	0.96	1.39	1.03	1.03
K	1.42	1.22	1.45	1.62	1.60
L	0.98	1.05	1.19	1.23	1.22
M	1.73	1.23	1.72	1.33	1.50
N	1.22	1.04	1.41	1.09	1.10

P	0.71	1.18	0.75	0.96	0.97
Q	1.24	1.59	1.44	1.39	1.30
R	0.60	0.62	0.56	0.54	0.56
S	0.81	0.69	0.66	0.97	0.99
T	0.99	0.94	0.93	0.79	0.76
V	1.07	1.07	1.07	1.09	1.05
W	0.97	0.85	0.86	0.83	0.78
Y	1.12	0.70	1.02	0.93	0.98

Amino acids encoded by CGN and NCG codons as well as those encoded by TGN, CAN, NTG and NCA were determined. Based on this information it was determined what fraction of which amino acids will be lost and which amino acids will be gained due to CG/CG mutation to TG/CA. In addition to the twice weight of CGN and NCG encoded amino acids in comparison with TGN, CAN, NTG and NCA encoded amino acids, the total number of codons encoding the lost or gained amino acids (based on codon degeneracy) were also taken into consideration while assigning score to the amino acids that may get affected. The list of affected amino acids and their scores are shown in the following table. CGN and NCG encoded R, A, T, S & P may be lost and thereby are given negative sign to their score while TGN, CAN, NTG and NCA encoded W, C, Q, H, V, M & L may in turn be gained. We apply "IF" function on the observed / expected to determine the effect mCG on amino acids in our 5x genomes.

Table 9: Percentage of amino acids

Percentage of amino acids may get affected due to methylated CG			
R	-66.67	W	+50
A	-12.5	C	+50
T	-12.5	Q	+50
S	-8.33	H	+50
P	-12.5	V	+12.5
		M	+50
		L	+16.66

Arginine (R) has been shown expected loss in *Homo sapiens* and *Mus musculus* in comparison with *Bacteriophage lambda, E.coli and Haemophillus influenza.* Also,

Methionine (M) and Glutamine (Q) have shown expected gain in *Homo sapiens and Mus musculus* in comparison with *Bacteriophage lambda, E.coli and Haemophillus influenza.*

Based on table, Loss of R > A = T = P > S

Gain of M = W = C = Q = H > L > V

But in our results, marginal affect was observed where R is lost on the basis of (O/E << 1) in *Homo sapiens and Mus musculus* but not much more than *Bacteriophage lambda, E.coli and Haemophillus influenza.*

A is not lost. Also, O/E is comparable to *Bacteriophage lambda, E.coli and Haemophillus influenza.*

T is lost a little but comparable to *Bacteriophage lambda, E.coli and Haemophillus influenza.*

P is not lost and comparable to *Bacteriophage lambda, E.coli* and *Haemophillus influenza.*

S is also not lost and is even higher than *Bacteriophage lambda, E.coli* and *Haemophillus influenza.*

M is gained less than *Bacteriophage lambda*, *E.coli* and *Haemophillus influenza*.

W is not gained and comparable *to Bacteriophage lambda, E.coli* and *Haemophillus influenza.*

C is not gained but more than *Bacteriophage lambda, E.coli* and *Haemophillus influenza.*

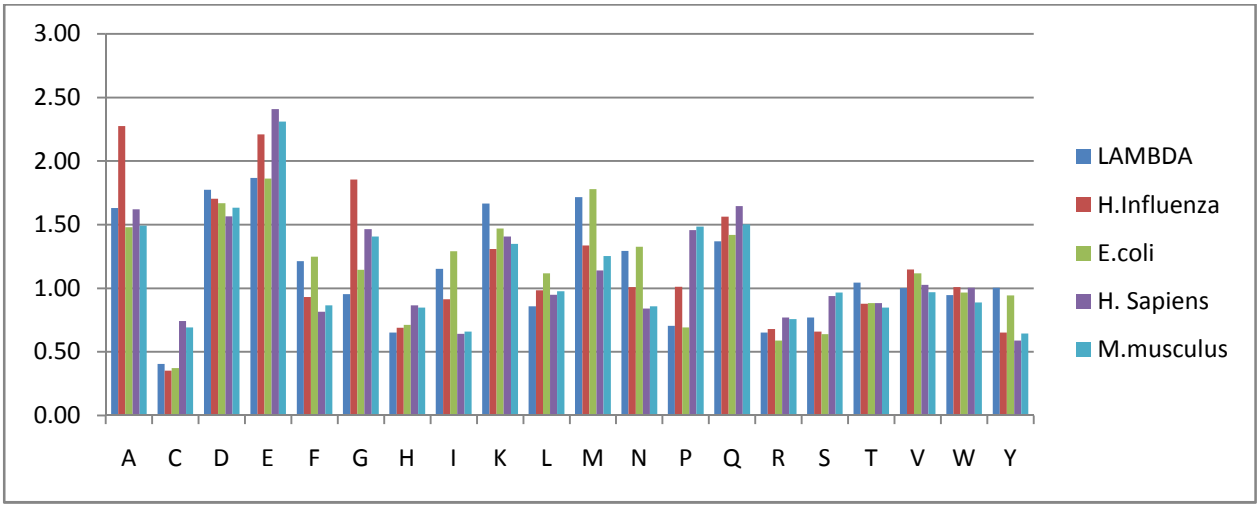
Q is gained but comparable to *Bacteriophage lambda, E.coli and Haemophillus influenza*.

H is not gained but more than *Bacteriophage lambda, E.coli and Haemophillus influenza*.

L is not gained and comparable to *Bacteriophage lambda, E.coli and Haemophillus influenza*.

V is not gained and comparable to *Bacteriophage lambda, E.coli and Haemophillus influenza*.

Fig 14: Graph observed/expected frequency of Amino acids of complete genomic sequence.



Since no expected gain of W, C, Q, H, V, M, L and loss of R, A, T, S, P was observed in methylated genomes sample sequences (*Homo sapiens* and *Mus musculus)* when compared to genomes not showing methylation related CG suppression (*Bacteriophage lambda, E.coli and Haemophillus influenza*), the expected count of amino acids was calculated based on base composition of coding regions of the genomes only.

In coding sequence, marginal affect was observed where R is lost on the basis of (O/E <<1) in *Homo sapiens* and *Mus musculus* but comparable to *Bacteriophage lambda, E.coli* and *Haemophillus influenza*.

A is not lost but lesser than *Bacteriophage lambda, E.coli and Haemophillus influenza*.

T is lost but lesser than *Bacteriophage lambda, E.coli and Haemophillus influenza*.

P is lost but less as not as much as than *Bacteriophage lambda, E.coli and Haemophillus influenza*.

S is not lost but lesser effect than *Bacteriophage lambda, E.coli and Haemophillus influenza*.

M is gained but comparable to *Bacteriophage lambda, E.coli and Haemophillus influenza*.

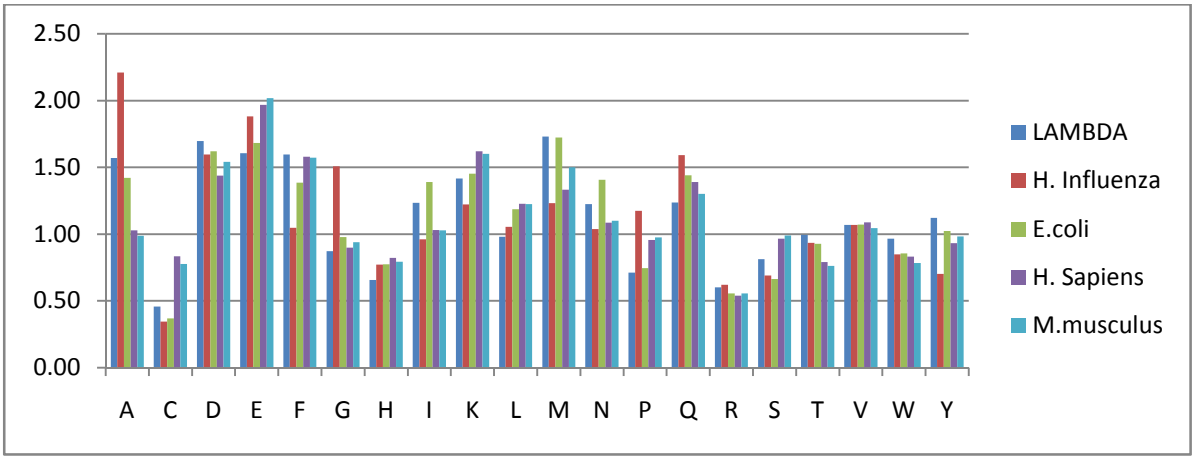
W is not gained and comparable to *Bacteriophage lambda, E.coli and Haemophillus influenza*.

C is not gained but O/E of is greater than *Bacteriophage lambda, E.coli and Haemophillus influenza*.

Q is gained but comparable to *Bacteriophage lambda, E.coli and Haemophillus influenza*.

H, L and V are not gained but comparable *Bacteriophage lambda, E.coli and Haemophillus influenza*.

Fig 15: Graph observed/expected frequency of Amino acids of coding sequence.



The effects of CG/CG mutation to TG/CA mutation in methylated genomes might have been over showed due to other evolutionary forces.

6.7 Calculate Probability of each codon based on base composition of di-amino acids

CG can be a part of codon as CGN& NCG. Another possibility is CG split into 2 successive codons as NNC GNN.

NNC can be G, D,V,A, S, N, I, T, C, Y, F, R, H, L, P and NNG can be G, E, D, V, A.

NNC GNN ⟶ NNT GNN OR NNC ANN

There are total 75 permutations of di amino acids.

We calculated the probability of di amino acids of coding sequence and complete genome sequence for all the NNC GNN permutations of codons using base composition of the respective genome.

6.8 Calculate expected counts of di-amino acids:

Expected count of every diamino acid was determined by multiplying its probability with total number of residues in the proteome.

Table 10: Expected counts of di-amino acids of complete genomic sequence

	Bacteriophage lambda	*Haemophilis influenza*	*Escherichia coli*	*H.sapiens chr.21*	*M.musculus chr.19*
GG	17.08	127.11	1633.65	271.64	655.48
DG	16.43	207.66	1597.26	390.32	883.02
VG	15.97	206.42	1593.39	391.50	876.18
AG	15.14	128.31	1626.28	270.28	656.98
SG	30.58	416.03	3183.46	779.86	1761.20
NG	15.81	339.26	1561.67	560.86	1189.55
IG	15.36	337.24	1557.89	562.55	1180.33
TG	14.56	209.62	1590.05	388.37	885.04
CG	15.97	206.42	1593.39	391.50	876.18
YG	15.36	337.24	1557.89	562.55	1180.33
FG	14.93	335.23	1554.12	564.25	1171.18
RG	15.14	128.31	1626.28	270.28	656.98
HG	14.56	209.62	1590.05	388.37	885.04
LG	14.15	208.38	1586.21	389.54	878.18
PG	13.42	129.52	1618.95	268.93	658.49
GE	8.52	103.84	800.02	195.19	442.28
DE	8.20	169.64	782.20	280.47	595.81
VE	7.97	168.63	780.30	281.32	591.19
AE	7.55	104.82	796.41	194.21	443.29
SE	15.26	339.87	1558.98	560.38	1188.36
NE	7.89	277.15	764.77	403.01	802.64
IE	7.67	275.50	762.92	404.23	796.42
TE	7.27	171.25	778.67	279.07	597.18
CE	7.97	168.63	780.30	281.32	591.19
YE	7.67	275.50	762.92	404.23	796.42
FE	7.45	273.86	761.07	405.45	790.25
RE	7.55	104.82	796.41	194.21	443.29
HE	7.27	171.25	778.67	279.07	597.18
LE	7.06	170.23	776.78	279.91	592.55
PE	6.69	105.81	792.82	193.24	444.31
GD	7.91	103.82	797.24	195.13	440.74
DD	7.61	169.62	779.48	280.39	593.74

VD	7.40	168.61	777.59	281.24	589.14
AD	7.01	104.81	793.64	194.16	441.75
SD	14.17	339.83	1553.56	560.23	1184.23
ND	7.32	277.12	762.11	402.90	799.86
ID	7.12	275.47	760.27	404.12	793.65
TD	6.75	171.23	775.96	278.99	595.10
CD	7.40	168.61	777.59	281.24	589.14
YD	7.12	275.47	760.27	404.12	793.65
FD	6.91	273.83	758.43	405.34	787.50
RD	7.01	104.81	793.64	194.16	441.75
HD	6.75	171.23	775.96	278.99	595.10
LD	6.55	170.21	774.08	279.83	590.49
PD	6.21	105.80	790.06	193.19	442.77
GV	15.97	206.42	1593.39	391.50	876.18
DV	15.36	337.24	1557.89	562.55	1180.33
VV	14.93	335.23	1554.12	564.25	1171.18
AV	14.15	208.38	1586.21	389.54	878.18
SV	28.60	675.64	3105.01	1123.98	2354.20
NV	14.78	550.97	1523.19	808.35	1590.07
IV	14.36	547.68	1519.50	810.78	1577.75
TV	13.62	340.43	1550.87	559.74	1183.04
CV	14.93	335.23	1554.12	564.25	1171.18
YV	14.36	547.68	1519.50	810.78	1577.75
FV	13.96	544.42	1515.82	813.23	1565.51
RV	14.15	208.38	1586.21	389.54	878.18
HV	13.62	340.43	1550.87	559.74	1183.04
LV	13.23	338.40	1547.11	561.43	1173.86
PV	12.54	210.35	1579.05	387.59	880.20
GA	15.14	128.31	1626.28	270.28	656.98
DA	14.56	209.62	1590.05	388.37	885.04
VA	14.15	208.38	1586.21	389.54	878.18
AA	13.42	129.52	1618.95	268.93	658.49
SA	27.11	419.97	3169.10	775.96	1765.24
NA	14.01	342.47	1554.63	558.06	1192.28
IA	13.62	340.43	1550.87	559.74	1183.04
TA	12.91	211.61	1582.88	386.43	887.07
CA	14.15	208.38	1586.21	389.54	878.18
YA	13.62	340.43	1550.87	559.74	1183.04
FA	13.23	338.40	1547.11	561.43	1173.86
RA	13.42	129.52	1618.95	268.93	658.49
HA	12.91	211.61	1582.88	386.43	887.07

LA	12.54	210.35	1579.05	387.59	880.20
PA	11.89	130.75	1611.65	267.58	659.99

Table 11: Expected counts of di-amino acids of coding sequence

	Bacteriophage lambda	*Haemophilis influenza*	*Escherichia coli*	*H.sapiens chr.21*	*M.musculus chr.19*
GG	20.30	178.87	2148.11	890.36	1813.10
DG	19.76	266.10	1920.47	907.76	1798.37
VG	16.28	253.04	1873.07	744.06	1499.65
AG	17.11	150.96	1904.28	857.24	1832.69
SG	33.49	479.66	3580.92	1624.15	3314.22
NG	19.24	395.86	1716.96	925.50	1783.77
IG	15.85	376.44	1674.57	758.60	1487.47
TG	16.66	224.58	1702.48	874.00	1817.80
CG	16.28	253.04	1873.07	744.06	1499.65
YG	15.85	376.44	1674.57	758.60	1487.47
FG	13.05	357.97	1633.23	621.80	1240.39
RG	17.11	150.96	1904.28	857.24	1832.69
HG	16.66	224.58	1702.48	874.00	1817.80
LG	13.72	213.56	1660.45	716.39	1515.85
PG	14.43	127.41	1688.12	825.36	1852.48
GE	10.78	139.47	995.88	480.16	935.33
DE	10.49	207.48	890.35	489.54	927.74
VE	8.64	197.30	868.37	401.26	773.63
AE	9.09	117.71	882.84	462.30	945.44
SE	17.78	374.00	1660.14	875.88	1709.73
NE	10.22	308.66	796.00	499.11	920.20
IE	8.42	293.52	776.35	409.11	767.35
TE	8.85	175.11	789.28	471.34	937.76
CE	8.64	197.30	868.37	401.26	773.63
YE	8.42	293.52	776.35	409.11	767.35
FE	6.93	279.12	757.18	335.33	639.89
RE	9.09	117.71	882.84	462.30	945.44
HE	8.85	175.11	789.28	471.34	937.76
LE	7.29	166.52	769.80	386.34	781.99
PE	7.66	99.34	782.62	445.11	955.65
GD	8.98	126.63	924.59	427.60	863.04
DD	8.75	188.38	826.61	435.96	856.03
VD	7.20	179.14	806.21	357.34	713.84

AD	7.57	106.87	819.64	411.70	872.36
SD	14.82	339.57	1541.30	780.01	1577.58
ND	8.52	280.24	739.02	444.48	849.08
ID	7.01	266.50	720.77	364.33	708.04
TD	7.37	158.99	732.78	419.75	865.28
CD	7.20	179.14	806.21	357.34	713.84
YD	7.01	266.50	720.77	364.33	708.04
FD	5.78	253.42	702.98	298.63	590.43
RD	7.57	106.87	819.64	411.70	872.36
HD	7.37	158.99	732.78	419.75	865.28
LD	6.07	151.19	714.69	344.05	721.55
PD	6.38	90.19	726.60	396.39	881.78
GV	16.28	253.04	1873.07	744.06	1499.65
DV	15.85	376.44	1674.57	758.60	1487.47
VV	13.05	357.97	1633.23	621.80	1240.39
AV	13.72	213.56	1660.45	716.39	1515.85
SV	26.85	678.56	3122.42	1357.28	2741.26
NV	15.43	560.01	1497.12	773.43	1475.39
IV	12.71	532.54	1460.16	633.96	1230.32
TV	13.36	317.70	1484.49	730.39	1503.54
CV	13.05	357.97	1633.23	621.80	1240.39
YV	12.71	532.54	1460.16	633.96	1230.32
FV	10.47	506.41	1424.11	519.63	1025.95
RV	13.72	213.56	1660.45	716.39	1515.85
HV	13.36	317.70	1484.49	730.39	1503.54
LV	11.00	302.12	1447.84	598.68	1253.79
PV	11.57	180.24	1471.97	689.75	1532.22
GA	17.11	150.96	1904.28	857.24	1832.69
DA	16.66	224.58	1702.48	874.00	1817.80
VA	13.72	213.56	1660.45	716.39	1515.85
AA	14.43	127.41	1688.12	825.36	1852.48
SA	28.23	404.81	3174.44	1563.75	3350.02
NA	16.22	334.09	1522.06	891.08	1803.03
IA	13.36	317.70	1484.49	730.39	1503.54
TA	14.04	189.53	1509.22	841.50	1837.43
CA	13.72	213.56	1660.45	716.39	1515.85
YA	13.36	317.70	1484.49	730.39	1503.54
FA	11.00	302.12	1447.84	598.68	1253.79
RA	14.43	127.41	1688.12	825.36	1852.48
HA	14.04	189.53	1509.22	841.50	1837.43
LA	11.57	180.24	1471.97	689.75	1532.22

PA	12.16	107.53	1496.49	794.67	1872.49

6.9 Analysis based on observed/expected frequency of di-amino acids:

In this analysis, the observed counts of di amino acids are divided by expected counts of di amino acids.

Table 12: observed /expected frequency of diamino acids of complete genome sequence

	Bacteriophage lambda	*Haemophilis influenza*	*Escherichia coli*	*H.sapiens chr.21*	*M.musculus chr.19*
GG	1.02	4.72	1.30	2.92	2.69
DG	2.28	3.92	1.95	3.00	2.57
VG	0.85	2.82	1.15	1.52	1.50
AG	1.77	5.52	1.77	3.04	2.47
SG	1.11	1.82	0.89	1.75	1.52
NG	1.77	2.52	1.73	1.50	1.61
IG	1.22	2.67	1.47	0.95	0.82
TG	1.08	2.28	1.16	1.80	1.49
CG	0.47	1.33	0.54	1.78	1.24
YG	0.98	1.63	1.12	1.21	1.12
FG	1.11	2.53	1.52	1.34	1.44
RG	0.68	1.79	0.55	1.43	1.39
HG	0.62	1.46	0.84	1.29	1.47
LG	0.72	2.95	1.11	1.83	1.59
PG	0.82	1.44	0.84	3.63	3.08
GE	1.67	5.20	2.28	3.73	3.27
DE	4.09	5.94	3.63	3.95	4.17
VE	1.79	3.56	1.99	2.57	2.22
AE	3.41	7.42	2.96	4.96	4.02
SE	1.59	1.77	1.14	2.55	2.42
NE	2.73	2.80	2.25	2.07	2.05
IE	2.57	3.04	2.24	1.86	1.65
TE	2.00	2.48	1.44	2.38	1.96
CE	0.75	0.96	0.69	2.13	1.43
YE	2.09	1.53	1.42	1.60	1.46
FE	1.61	2.46	1.79	1.94	1.73
RE	1.70	2.63	1.31	2.42	2.20

HE	0.89	1.39	1.16	1.65	1.51
LE	1.53	3.11	1.81	3.54	3.04
PE	1.46	3.68	1.78	4.33	4.11
GD	1.93	3.84	1.86	3.17	2.78
DD	3.94	3.65	3.07	2.99	3.01
VD	1.89	2.68	1.88	2.05	2.13
AD	2.71	5.17	2.46	2.62	2.56
SD	1.53	1.48	1.07	1.75	1.74
ND	2.05	2.14	2.25	1.57	1.34
ID	2.34	2.41	2.44	1.35	1.33
TD	1.89	1.96	1.40	1.80	1.50
CD	0.47	0.92	0.65	1.31	1.34
YD	2.11	1.51	1.74	1.07	1.14
FD	3.18	2.65	2.34	1.32	1.39
RD	1.38	1.87	1.02	1.73	1.69
HD	0.89	1.12	1.19	1.17	1.02
LD	1.35	2.77	1.72	1.88	2.03
PD	1.93	2.12	1.41	2.81	2.86
GV	1.13	3.17	1.42	1.81	1.39
DV	1.89	2.76	1.95	2.05	2.34
VV	1.16	1.84	1.39	1.39	1.18
AV	1.75	3.63	1.61	2.15	1.90
SV	0.93	1.00	0.73	1.15	1.04
NV	1.05	1.52	1.41	1.17	0.89
IV	1.00	1.44	1.34	0.80	0.82
TV	1.34	1.30	1.06	1.29	1.09
CV	0.30	0.57	0.40	1.02	0.90
YV	0.87	0.91	0.88	0.76	0.69
FV	1.43	1.40	1.31	1.07	0.97
RV	0.60	1.20	0.62	0.95	0.87
HV	0.55	0.66	0.66	1.15	0.81
LV	0.87	1.64	1.13	1.11	1.12
PV	0.92	1.61	0.96	2.12	1.65
GA	1.34	5.14	1.46	2.53	2.34
DA	3.06	4.58	2.49	3.00	2.19
VA	1.57	3.51	1.68	2.26	1.56
AA	3.86	6.41	2.34	4.74	3.72
SA	1.56	2.09	0.93	1.70	1.60
NA	2.00	2.95	1.95	1.41	1.29
IA	1.86	3.35	2.10	1.17	1.02
TA	2.07	3.02	1.26	2.03	1.49

CA	0.49	0.97	0.50	1.16	1.04
YA	1.29	1.87	1.26	0.73	0.80
FA	1.59	2.72	1.81	1.21	1.15
RA	0.88	2.15	0.70	1.76	1.55
HA	0.77	1.48	0.92	1.32	1.04
LA	1.70	3.77	1.79	2.19	1.92
PA	1.28	2.61	1.04	3.22	2.83

Table 13: Observed / Expected frequency of di-amino acids of coding sequence

	Bacteriophage lambda	*Haemophilis influenza*	*Escherichia coli*	*H.sapiens chr.21*	*M.musculus chr.19*
GG	0.86	3.35	0.99	0.89	0.97
DG	1.90	3.06	1.62	1.29	1.26
VG	0.83	2.30	0.98	0.80	0.88
AG	1.56	4.69	1.51	0.96	0.88
SG	1.02	1.58	0.79	0.84	0.81
NG	1.46	2.16	1.57	0.91	1.08
IG	1.18	2.39	1.37	0.70	0.65
TG	0.95	2.13	1.08	0.80	0.72
CG	0.46	1.08	0.46	0.94	0.72
YG	0.95	1.46	1.04	0.90	0.89
FG	1.26	2.37	1.45	1.22	1.36
RG	0.60	1.52	0.47	0.45	0.50
HG	0.54	1.36	0.78	0.57	0.71
LG	0.74	2.88	1.06	0.99	0.92
PG	0.76	1.46	0.80	1.18	1.09
GE	1.32	3.87	1.83	1.52	1.54
DE	3.19	4.85	3.19	2.26	2.68
VE	1.65	3.04	1.79	1.80	1.69
AE	2.83	6.61	2.67	2.08	1.88
SE	1.37	1.60	1.07	1.63	1.68
NE	2.10	2.51	2.16	1.67	1.79
IE	2.34	2.85	2.20	1.84	1.71
TE	1.64	2.42	1.42	1.41	1.25
CE	0.69	0.82	0.62	1.49	1.10
YE	1.90	1.44	1.40	1.58	1.52
FE	1.73	2.42	1.80	2.35	2.13
RE	1.41	2.35	1.18	1.02	1.03
HE	0.73	1.36	1.14	0.97	0.96

LE	1.49	3.17	1.83	2.57	2.30
PE	1.27	3.92	1.81	1.88	1.91
GD	1.70	3.15	1.61	1.45	1.42
DD	3.43	3.28	2.89	1.92	2.09
VD	1.94	2.53	1.82	1.61	1.76
AD	2.51	5.07	2.38	1.23	1.29
SD	1.46	1.48	1.07	1.26	1.31
ND	1.76	2.12	2.32	1.42	1.26
ID	2.38	2.49	2.57	1.49	1.49
TD	1.73	2.11	1.49	1.20	1.03
CD	0.49	0.87	0.63	1.03	1.11
YD	2.14	1.56	1.83	1.18	1.27
FD	3.81	2.86	2.53	1.80	1.85
RD	1.28	1.83	0.99	0.82	0.86
HD	0.81	1.20	1.25	0.78	0.70
LD	1.45	3.12	1.87	1.53	1.66
PD	1.88	2.49	1.53	1.37	1.44
GV	1.11	2.59	1.20	0.95	0.81
DV	1.83	2.47	1.81	1.52	1.86
VV	1.32	1.72	1.32	1.26	1.11
AV	1.80	3.55	1.54	1.17	1.10
SV	0.99	1.00	0.72	0.95	0.89
NV	1.00	1.50	1.43	1.23	0.96
IV	1.13	1.48	1.40	1.03	1.05
TV	1.37	1.39	1.11	0.99	0.86
CV	0.34	0.53	0.38	0.93	0.85
YV	0.98	0.93	0.92	0.97	0.88
FV	1.91	1.51	1.39	1.68	1.48
RV	0.62	1.17	0.60	0.52	0.50
HV	0.56	0.70	0.69	0.88	0.64
LV	1.05	1.84	1.21	1.04	1.05
PV	0.99	1.88	1.03	1.19	0.95
GA	1.18	4.37	1.24	0.80	0.84
DA	2.67	4.28	2.32	1.34	1.07
VA	1.62	3.42	1.60	1.23	0.90
AA	3.59	6.52	2.24	1.54	1.32
SA	1.50	2.17	0.93	0.84	0.84
NA	1.73	3.03	1.99	0.88	0.85
IA	1.90	3.59	2.19	0.89	0.80
TA	1.90	3.37	1.32	0.93	0.72
CA	0.51	0.94	0.47	0.63	0.60

YA	1.31	2.01	1.31	0.56	0.63
FA	1.91	3.05	1.93	1.13	1.08
RA	0.82	2.19	0.67	0.57	0.55
HA	0.71	1.65	0.97	0.60	0.50
LA	1.84	4.40	1.92	1.23	1.10
PA	1.25	3.17	1.12	1.08	1.00

Table 14: No. Of NNC GNN diAA exhibiting observed /expected values

Genome	No. Of NNC GNN diAA exhibiting observed /expected Value <1 (out of 75)	
	Total genome	Coding sequence
Bacteriophage Lambda	22	22
Haemophillus Influenza	8	7
E.coli	17	19
Homo sapiens Chromosome 21	**5**	**31**
Mus musculus Chromosome 19	**9**	**34**

From this table, when we compare human and mouse total genome based di-amino acids then number of amino acids encoded by NNC GNN showing <1 Obs/Exp value are comparable or fewer than *Bacteriophage lambda* total genome. However when similar analysis was performed using expected value based on base composition of only coding regions of the genomes, there were much higher number of di amino acids showing <1 Obs/Exp value in human and mouse. This indicates that effect of CG loss is reflected in the proteomes of methylated genomes when compared against the control genomes.

CHAPTER 7

DISCUSSION

DNA methylation causes CpG deficiency in higher eukaryotes. The organisms with high level of methylation (i.e. the vertebrates) are extremely CpG deficient conversely, poorly methylated genomes display no significant CPG deficiency while, partially methylated genomes are deficient in CpG to an intermediate extent. DNA methylation in vertebrate genomes mostly occurs at cytosines in CpG dinucleotides. These dinucleotides are known as mutation hotspot which results in the deamination of cytosine and 5-methylcytosine. This deamination leads to the suppression of CpG and over-representation of TpG and CpA.

The DNA methylation occurs in high amount in vertebrates then invertebrates. The methylated species are markedly suppressed while the un-methylated forms show an excess of CpGs. The CpG deficiencies vary according to the G+C contents of sequences.

In the present study, it has been attempted to study the effect of under representation of CpGs on sequence of amino acid in protein encoded by a vertebrate genome. For example, Arginine (which is coded by CGG) is converted into Tryptophan (which is coded by TGG) or Glutamine (which is coded by CAQ). Arginine is lost by CPG suppression and Tryptophan (W) and Glutamine (Q) are gained.

We analyse the effect of CpG suppression on vertebrate proteome. Since, there is no significant result was observed in the amino acid in the coding and complete genome sequence when compared non-methylated genomes with methylated genomes related to CpG suppression. The effect of mutation in methylated genomes on the relative amino acid abundance protein seqeunces has been found to be minimal or totally absent. It appears that most of the effect of CpG mutations due to methylation is in non-coding regions. The effect in coding regions is not effectively translated into protein sequences due to degeneracy of codon system. Moreover evolutionary pressure also is expected to have very profound effect against CpG methylation based mutations as the protein sequences are functionally active part of genome and thereby essential.

But in di-amino acids, the *Homosapiens* and *Mus musculus* show<< 1 obs. / exp value for many di-peptides. This number is much more than in prokaryotic systems. Values are comparable or lesser than *Bacteriophage lambda* in total genome. However, when this analysis was performed with coding sequences of the genomes; there were much higher number of di-amino acids showing < 1 obs. /exp. values in *Homo sapiens* and *Mus musculus*. This indicates that the effect of CG loss is reflected, though minimally, in the proteomes of methylated genomes (*Homo sapiens and Mus musculus*) when compared *with Bacteriophage lambda, E.coli and Haemophillus influenza.*

CHAPTER 8

CONCLUSION

Due to DNA methylation, in vertebrate genomes the CG is under-representative and TG and CA is over-representative. We investigate the effect of CG suppression on vertebrate proteome.

There is no significant result was observed into the amino acid sequence in methylated genome. But in di-amino acids, *Homo sapiens and Mus musculus* showing (< 1 Obs. / Exp.) values lesser than to *Bacteriophage lambda* in total genome and in coding sequence, there were much higher number of di-amino acids showing (< 1 Obs. / Exp.) values in *Homo sapiens* and *Mus musculus* . This indicates the effect of CG suppression in the proteomes.

CHAPTER 9

REFERENCES

1) Adams, R.L.P. and Eason, R. (1984). "Increased G+C content of DNA stabilises methyl CpG dinucleotides." Nucleic Acid Research, **12**, 5869-5877.
2) Auclair, G. and Weber, M (2012). "Mechanisms of DNA methylation and demethylation in mammals." Biochimie, **94**, 2202-2211.
3) Barakat and Gribnau (2012). "X chromosome inactivation in the cycle of life." Development, **139**, 2085-2089.
4) Berger, L., Kouzarides, T., Shiekhattar, R. (2009). "An operational definition of Epigenetics." Genes and Development, **23**, 781-783.
5) Bestor, T. H. (2000). "The DNA methyltransferases of mammals." Human Molecular Genetics, **9**, 2395-2402.
6) Bestor, T. H. (2000). "The DNA methyltransferases of mammals." Human Molecular Genetics, **9**, 2395-2402.
7) Bird, P. (1980). "DNA methylation and frequency of CpG in animal DNA." Nucleic Acid Research, **8,** 1499-1504**.**
8) Bird, P. and Taggart, H. (1980). "Variable pattern of total DNA and rDNA methylation in animals." Nucleic Acid Research, **8**, 1485-1498.
9) Cardon, L. R., Burge, C., Clayton, D. A., Karlin, S. (1993). "Pervasive CpG suppression in animal mitochondrial genomes." Proceedings of the National Academy of Sciences, **91**, 3799-3803.
10) Cross, S. H. and Bird, A. P. (1995). "CpG Islands and genes." Genetics and Development, **5**, 309-314.
11) Duret, L. and Galtier, (2003). "The covariation between TpA deficiency, CpG deficiency, and G+C content of human isochors is due to a mathematical artefact." Mol. Biol., **17**, 1620-1625.
12) Gonzalgo, M. L. and Jones, P. A. (1997). "Mutagenic and epigenetic effects of DNA methylation." Mutation Research, **386**, 107-118.

13) Handa, V. and Jeltsch, A. (2005). "Profound flanking sequence preferences of Dnmt3a and Dnmt3b mammalian DNA methyltransferases shape the human epigenome." J Mol Biol, **348**, 1103-1112.
14) Hermann, A., Gowher, H., Jeltsch, A. (2004). "Biochemistry and biology of mammalian DNA methyltransferases." Cellular and Molecular Life Sciences CMLS, **61**, 2571-2587.
15) Inbar-Feigenberg, M., Choufani, S., Butcher, D. T., Roifman, M., Weksberg, R. (2013). "Basic concepts of Epigenetics." Fertility and Sterility, **99**, 606-616.
16) Jaenisch and Bird (2003). "Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals." Nature Genetics Supplement, **33**, 245-254.
17) Jeltsch, A. (2002). "Beyond Watson and Crick: DNA methylation and molecular enzymology of DNA methyltransferases." ChemBioChem, **3**, 274-293.
18) Jenuwein and Allis (2001). "Translating the histone code." Science, **293**, 1074-1080.
19) Karlin, S. and Mrazek, J. (1996). "What Drives Codon choices in Human genes." J. Mol. Biol, **262**, 459-472.
20) Klose, R. J. and Bird, P. (2006). "Genomic DNA methylation: the mark and its mediators." Trends in Biochemical Science, **31**, 89-97.
21) Moison, C. and Arimondo, P. B. (2013). "DNA methylation in cancer." Atlas Genet Cytogenet Oncol Haematol, **18**, 285-292.
22) Mugal, F. C., Arndt, P.F., Holm, L., Eleegren, H. (2005). "Evolutionary consequences of DNA methylation on the GC content in vertebrate genomes." Genes Genome Genetics, **5**, 441-447.
23) Nakao, M. (2001). "Epigenetics: interaction of DNA methylation and chromatin." Gene, **278**, 25-31.
24) Onofrio, G. D., Mouchiroud, D., Aissani, B., Gautier, C., Bernardi, G. (1991). "Correlations between the compositional properties of Human genes, Codon Usage and amino acid composition of proteins." J Mol Evol, **32**, 504-510.
25) Pascal, G., Medigue, C. and Danchin, A. (2006). "Persistant biases in the amino acid composition of prokaryotic proteins." Functional Genomics and Bioinformatics, **28**, 726-738.
26) Ramsahoye, B. H., Davis C. S., Mills, K. I. (1996). "DNA methylation: biology and significance, Blood Reviews, **10**, 249-261.

27) Shimizu, T. S., Takahashi, K., Tomita, M. (1997). “CpG distribution patterns in methylated and non-methylated species.” Gene, **205**, 103-107.
28) Singal, R. and Ginder, G. D. (1999). “DNA methylation.” Blood Reviews, **93**, 4059-4070.
29) Takai and Jones (2002). “Comprehensive analysis of CpG islands in human chromosome 21 and 22.” Biochemistry, 99, 3740-3745.
30) Yan, J., Zierath, J. R., Barres, R. (2011). “Evidence for non-CpG methylation in mammals.” Experimental Cell Research, **317**, 2555-2561.

CpG Suppression

ORIGINALITY REPORT

%15	%5	%13	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Karlin, S.. "What Drives Codon Choices in Human Genes?", Journal of Molecular Biology, 19961004 Publication	%2
2	Adrian P. Bird. "DNA methylation and the frequency of CpG in animal DNA", Nucleic Acids Research, 1980 Publication	%1
3	A. Hermann. "Biochemistry and biology of mammalian DNA methyltransferases", Cellular and Molecular Life Sciences CMLS, 10/2004 Publication	%1
4	www.pasteur.fr Internet Source	%1
5	Albert Jeltsch. "Beyond Watson and Crick: DNA Methylation and Molecular Enzymology of DNA Methyltransferases", ChemBioChem, 04/02/2002 Publication	%1
6	Giuseppe D'Onofrio. "Correlations between the	