Computational Micromodel for Epigenetic Mechanisms

Karthika Raghavan, B.Tech

A Dissertation submitted in part fulfillment of the

requirements for the award of

Doctor of Philosophy (Ph.D.)

to the

DUBLIN CITY UNIVERSITY

Supervisor:

Prof. Heather J. Ruskin

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of PhD is entirely my own work, that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed: .....

ID No.: .....

Date: .....

कर्मण्येवाधिकारस्ते मा फलेषु कदाचन।
मा कर्मफलहेतुर्भूर्मा ते सङ्गोऽस्त्वकर्मणि ॥ २-४७ ॥

Bhagavath Gita 2:47

TRANSLATION

"You have a right to perform your prescribed duty, but you are not entitled to the fruits of action. Never consider yourself the sole cause of the results of your activities, and never be attached to not doing your duty."

– Lord Krishna, Bhagavat Gita

Acknowledgement

"Science without religion is lame. Religion without science is blind." - Albert Einstein.

Given this conjecture, I am immensely thankful for the presence of divinity and guardian angels in my life. Firstly I am grateful for having the best parents – Nalini and Raghavan, in the whole wide world. I would like to secondly thank my brother, Aswin Kumar whose common sense and criticism at certain times kept me in line. I would like to simply mention about my lovable and affectionate husband Krishna Vijayaraghavan who always looked out for my future and also supported my decisions especially during the hectic times. My in-laws are also to be thanked for their endearing attitude towards my slow and steady progress in work. I would like to also acknowledge my good friends Claudio, Anita and my dearest Vindhya for being good listeners. I would like to thank my dear colleagues at Sci-Sym and Cloud Computing, (Marija, Jelena, Irina, Kabita, Oisin, Joachim and John amongst others) for working with me and also sharing a good laughter during times of stress. A special mention and sincere thanks to Dimitri, Ana and my ex-colleague Alina for helping me initially to swim through the challenges in work. Above all I would like to explicitly thank my supervisor Prof. Heather J. Ruskin for standing by me through all these valuable three years in Sci-Sym and DCU.

Contents

List of Figures

List of Tables

Abbreviations

bp - base pair

CDKN2A (P14/P16) - Cyclin-dependent kinase inhibitor 2A (gene[1])

ChIP - Chromatin Immuno Precipitation

CG - Cytosine Guanine

CRABP1 - Cellular retinoic acid binding protein 1

CTCF - 11-zinc finger protein or CCCTC-binding transcription factor

DM - DNA Methylation

DNA - Deoxy Ribonucleic Acid

HM - Histone Modifications

HP1 - Heterochromatin Protein 1

HPLC - High Performance Liquid Chromatography

IGF2 - Insulin like growth factor 2 (Gene)

iRNA - Interference Ribonucleic Acid

[1]The abbreviations of 9 specific test genes used in Chapter 6 have been provided here. For further information on other genes appearing in this document please refer to `http://www.genecards.org`

MAL - Myelin and lymphocyte protein

MBD - Methyl CpG binding domain protein 1

MCMC - Markov Chain Monte Carlo

miRNA - Micro Ribonucleic Acid

MGMT - Methylated-DNA-protein-cysteine methyltransferase (gene)

MLH1 - MutL homolog 1 (gene)

mRNA - Messenger Ribonucleic Acid

MSDK - Methyl Specific Digital Karyotype

MSI - Microsatellite instability

MSP - Sensitive Methyl Light

MSS - Microsatellite Stable

PCR - Polymerase Chain Reaction

RASSF1A - Ras association domain-containing protein 1

RLGS - Restriction Landmark Genomic Scanning

RNA - Ribonucleic Acid

RUNX3 - Runt-related transcription factor 3

TIMP3 - Metalloproteinase inhibitor 3 (gene)

Abstract

Definition and characterization of the role of Epigenetic mechanisms have gained immense momentum since the completion of the Human Genome Project. The human epigenetic layer, made up of DNA methylation and multiple histone protein modifications, (the key elements of epigenetic mechanisms), is known to act as a switchboard that regulates the occurrence of most cellular events. In multicellular organisms such as humans, all cells have identical genomic contents but vary in DNA Methylation (DM) profile with the result that different types of cells perform a spectrum of functions. DM within the genome is associated with tight control of gene expression, parental imprinting, X-chromosome inactivation, long term silencing of repetitive elements and chromatin condensation. Recently, considerable evidence has been put forward to demonstrate that environmental stress implicitly alters normal interactions among key epigenetic elements inside the genome. Aberrations in the spread of DM especially *hypo/hyper* methylation supported by an abnormal landscape of histone modifications have been strongly associated with Cancer initiation and development. While new findings on the impact of these key elements are reported regularly, precise information on how DM is controlled and its relation to networks of histone modifications is lacking.

This has motivated modelling of DNA methylation and histone modifications and their interdependence. We describe initial computational methods used to investigate these key elements

of epigenetic change, and to assess related information contained in DNA sequence patterns. We then describe attempts to develop a phenomenological epigenetic "micromodel", based on Markov Chain Monte Carlo principles. This theoretical micromodel ("EpiGMP") aims to explore the effect of histone modifications and gene expression for defined levels of DNA methylation. We apply this micromodel to (i) test networks of genes involved in colon cancer (extracted from an in-house database, *StatEpigen*) and (ii) to help define an agent-based modelling framework to explore chromatin remodelling (or the dynamics of physical rearrangements), inside the human genome. Parallelization techniques to address issues of scale during the application of this micromodel have been adopted as well. A generic tool of this kind can potentially be applied to predict molecular events that affect the state of expression of any gene during the onset or progress of cancer. Ultimately, the goal is to provide additional information on ways in which these low level molecular changes determine physical traits for normal and disease conditions in an organism.

Part I

Molecular Models

Chapter 1

Introduction - Epigenetics

1.1 What is Epigenetics?

"Epigenetics" was initially defined as a set of interactions between genes and the surrounding environment, which determines the phenotype or physical traits in an organism, [Murrell et al., 2005; Waddington, 1942]. A biological cell goes through different phases during its life cycle based on controlled signals from the nucleus, which in turn depend on the regulation of gene expression. The phrase "*epi*" (derived from the Greek for "over"/"upon") refers to the factors occurring above or around the broader events inside the genome. Epigenetics is specifically associated with the molecular mechanisms that occur without changing the actual DNA sequences, and strongly influences regulation of gene expression and the structural dynamics of Chromosomes, [Murrell et al., 2005]. The types of biological cells in multi-cellular organisms are capable of performing a spectrum of functions. These cells originate from a single pool of stem cells after differentiation and contain the same genetic contents but obtain functional diversity from differences in their epigenetic patterns, [Allis et al., 2007]. Similarly, at the organism level, identical twins originate from the same zygote but as they grow older they obtain different epigenetic profiles which strongly influence their re-

action to environmental stimuli and to resistance to diseases, [Choi, 2010]. Hence the epigenetic layer in human genome consists of molecular mechanisms that are involved in modifying gene expression without alteration of the actual DNA contents. These mechanisms are **DNA Methylation**, (DM) and **Histone Modifications**, (HM) which are discussed in detail in the sections that follow.

Epigenetic Molecular Events

Initial epigenetic research focused on genomic regions such as *heterochromatin* and *euchromatin* based on dense and relatively loose DNA packing, since these were known to contain inactive and active genes respectively, [Yasuhara et al., 2005]. Subsequently, key roles of DNA methylation, histone modifications, and other assistive proteins such as Methyl Binding Proteins (MBD) during gene expression and suppression were identified, [Baylin et al., 2006; Jenuwein et al., 2001]. DM refers to the modification of DNA by addition of a methyl group to the cytosine base (C), and is the most stable, heritable and well-conserved epigenetic change. A family of enzymes called DNA Methyl Transferases (DNMT) [Doerfler, Toth, et al., 1990; Riggs et al., 2004; Ushijima et al., 2003] introduce and maintain the methylation patterns inside the genome. Methyl-Cytosine or "mC" often referred to as the fifth type of nucleotide plays an extremely important role in gene expression and other cellular activities. Although DM is defined as a simple molecular modification, its effect can range from altering the state of a single gene to controlling a whole section of chromosome in the human genome. The genomic organization from a simple DNA sequence, nucleosome formation to a highly condensed chromosome is shown in Figure 1.1.

Histones are proteins that protect DNA sequences from being accessed by restriction enzymes and play a vital role in chromosome condensation, [Ito, 2007]. A "Histone Core", made of nine types of histone proteins (two copies of H2A, H2B, H3, H4 and a single H1), is attached to intervals of 146-148 base pairs (bp) of DNA molecules. A combination of modifications (such as *acetylation, methylation, phosphorylation, ubiquitination and sumoylation* carried out by many types of

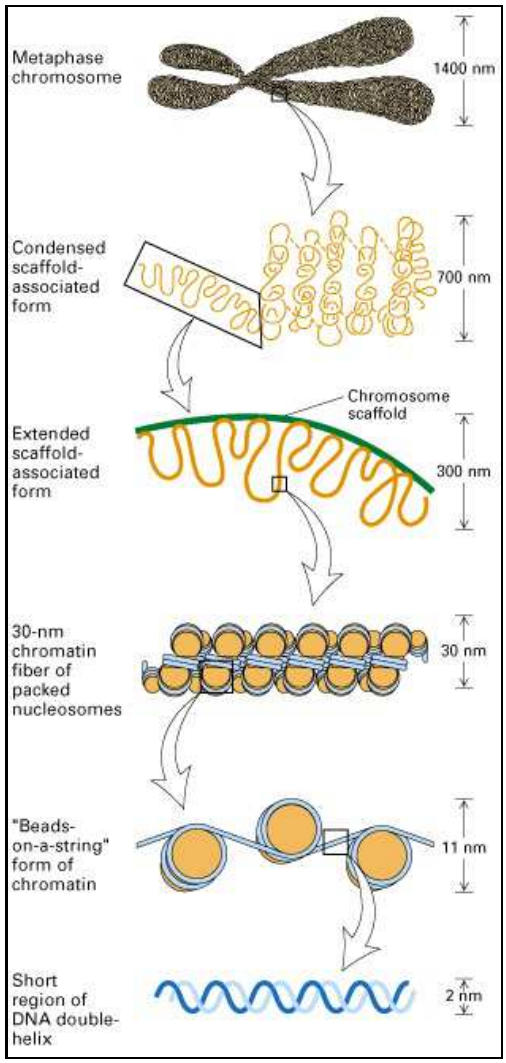


Figure 1.1: Genome Organization
DNA sequences along with histone proteins form a unit of nucleosomes. These nucleosomes that appear like “Beads on a String” rearrange to form a tightly arranged chromosome. Image adapted from
`http://bioweb.wku.edu/courses/biol566/Images/ChromatinF09-35.jpg`

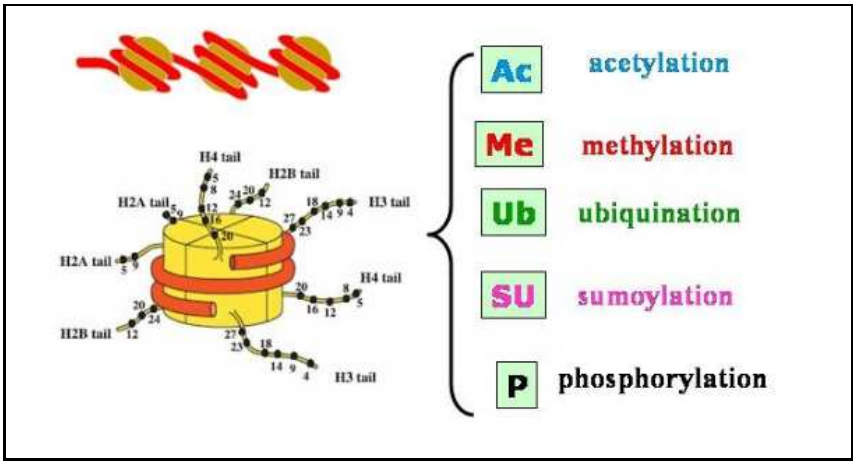


Figure 1.2: Nucleosome Modifications

About 146-148 bp of DNA molecules wind around a core of histone proteins. This image shows the type of amino acid modifications detected in histone core that form the nucleosome complex. Image adapted from (`http://chemistry.gsu.edu/Zheng.php`)

enzymes), within specific amino acids in each histone type leads to gene expression or inactivation [Kouzarides, 2007]. The process of histone modifications, unlike stable DNA methylation, is very dynamic in nature because the amount of modifications, (such as acetylation or methylation of Lysine or Arginine) change often in a biological cell, [Cedar et al., 2009]. Another fact about these changes is that activation of one modification affects the onset of modifications in other amino acids during cellular events, [Allis et al., 2007; Jung et al., 2009]. On the other hand it is also interesting to note that some histone modifications such as phosphorylation of serine amino acid are faithfully inherited across successive generations post DNA replication, [Kouzarides, 2007]. This unpredictable nature of such events increases the chance of assembling the wrong set of modifications to cause “Epigenetic Lesions”, [Allis et al., 2007], that lead to tumour formation.

1.2 Challenges and Motivation for Computational Modelling

Efficient epigenetic laboratory based techniques (Chapter 2), have lead to increased availability of data. Based on different experiments conducted so far, individual histone modifications [Cedar et al., 2009; Jenuwein et al., 2001], such as H3 lysine 36 (H3 K36) acetylation and H3 K9 methylation, have been identified as markers during transcription and gene suppression respectively. The collation of information from such experiments led to a conclusion about the behaviour of specific types of histone modifications. i.e. that the amount of global acetylation of histones is higher during gene transcription whereas global methylation is more likely to appear more during suppression, [Ajiro et al., 2010; Allis et al., 2007; Cedar et al., 2009; Kouzarides, 2007; Meng et al., 2009; Myers et al., 2003; Nakazawa et al., 2012; Turner, 2001; Wyrick et al., 2008]. Despite such strong conclusions, there are two major challenges: Firstly, information with regard to how histone modifications are orchestrated by DNA methylation during different stages of cell cycle and for different cell types is unclear, [Esteller, 2007]. Secondly, the dynamics of histone modifications and order in which these modifications appear are not fully understood.

Despite this lack of information, an abnormal interaction profile between histones modifications and DNA methylations with regard to cancer initiation has been reported [Esteller, 2007]. Recent evidences also indicate that epigenetic lesions along with nucleotide mutations trigger Cancer, [Brower, 2011; Pancione et al., 2012; Sell, 2004]. For example, in colon cancer during tumorigenesis*[1], aberrantly methylated genes such as MGMT or BRCA1 are associated with the occurrence mutations in the APC and TP53 genes, [Baylin et al., 2006; Esteller, 2007]. Following repeated cell divisions and faithful inheritance of genomic contents (both mutation and DNA methylation of genes) these low level events accumulate significantly to trigger malignant phenotypic changes at the top level in any organism. Ultimately, in an affected organism, (at the system or organ

[1] Phrases or words marked with * are explained in detail in the Appendix - C

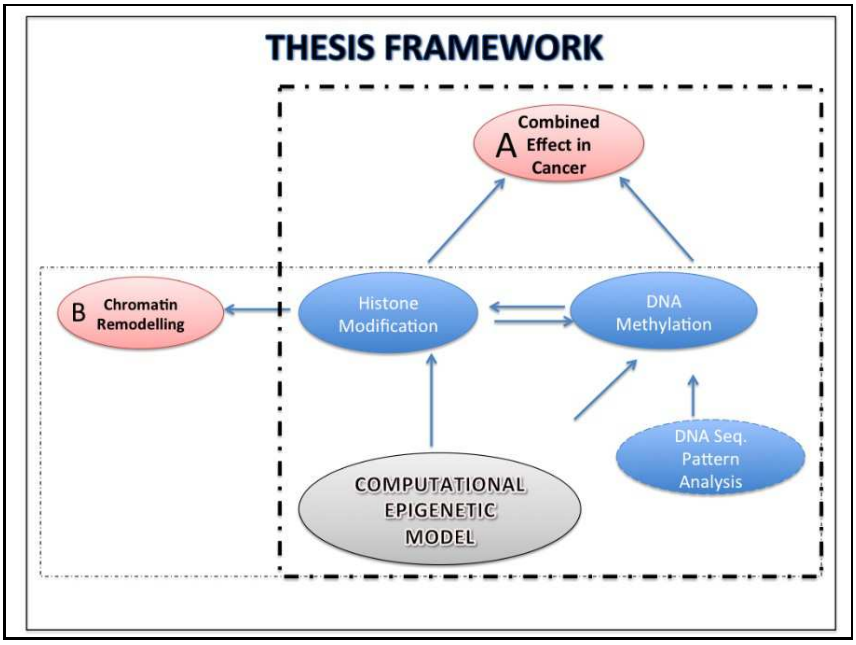


Figure 1.3: Overall Structure and stages in my thesis
The first stage in my thesis of epigenetic mechanisms involves defining an interdependency between the two key elements (indicated by blue colour) - "Histone Modifications and DNA Methylation". The second stage in my thesis consists of exploring the influence of these key elements in controlling (A) colon cancer progress and (B) chromatin structure, (both indicated by orange colour)

level), these molecular changes are known to spread quickly causing multiple failures which leads to death. Hence these facts make it a motivating and compulsive case to adopt computational modelling techniques to explore the epigenetic mechanisms.

1.3 Thesis Structure and Organization

A concise representation of the thesis is shown in Figure 1.3. Part I consists of developing models of key elements and their interdependencies in the epigenetic layer of the human genome. Part

II consist of modelling secondary events controlled by key epigenetic elements. The Appendices, (part III) consist of extended work and publication details. The contents in each chapter for Part I as follows,

- Chapter 2 details background on histone modifications and DNA methylation (DM). Laboratory-based experimental trials, DNA Sequence information and computational models developed so far are summarized. The last section introduces the concept of "Phenomenological Modelling" in epigenetics and a description of the set of assumptions, which form the basis of my research work.

- Chapter 3 discusses the DNA methylation process. This includes explanation of the significance of DNA sequences in establishing differential DNA methylation. Two methods are used to spot DNA sequence patterns, results obtained and their correlation with DNA methylation levels are dealt with in detail.

- Chapter 4 describes the development of our key stochastic tool, (EpiGMP[2]), for epigenetic molecular prediction. This micromodel incorporates expressions for the mathematical interdependency between levels of DM and HM using appropriate functional forms. With Markov Chain Monte Carlo algorithms, these have been applied to generate solutions of histone modification scenarios for specific levels of DNA methylation as input, ($\in$ [0,1]). The idea behind this **prototype** tool is to model the phenomenon and dependency between two main elements in epigenetic mechanisms. The chapter reports on characteristic histone modifications and expression levels for each gene.

While the first part of the thesis gives details about background, challenges and the unique computational micromodel development, the second half deals with applications of this micromodel. Three complex scenarios are explored by applying the tool. These include, (A) investigating the

[2]Epigenetic Molecular Predictor

epigenetic status of several genes across a chromosome in the human genome by combining the EpiGMP tool and information about DNA sequence patterns, (B) modelling the interactions among a group of either mutated or epigenetically altered genes during the carcinoma stage of colon cancer, and lastly (C) exploring the physical dynamics of chromatin structure based on the dynamics of histone modifications.

- Chapter 5 deals with improvements to EpiGMP framework. Efforts taken to combine the tool with pattern analysis from DNA sequences (Chapter 3), are explained. EpiGMP is applied to a large dataset from Chromosome 21 in humans. A Parallel version of EpiGMP tool is discussed, as this is required to expedite the time taken for the simulating such a large dataset. MPI and OpenMP routines (multi-processor and multi-thread), were added to the micromodel framework which was originally developed using C++.

- Chapter 6 involves investigation of molecular events associated with colon cancer and is intended as a prototype illustration of how the model framework can be augmented to draw inferences across different scales. This augmented model links levels or scales in overall by using, (i) data from a statistical epigenetic database called *StatEpigen* to decide occurrence of genetic *conditional* events at system or cell level (addressed as level 2), and (ii) EpiGMP, at molecular level to report the occurrence of histone modification and DNA methylation levels, (based on type of genetic event decided). A conclusive analysis of epigenetic aberrations that can possibly trigger Stage III in Colon cancer is presented.

- Chapter 7 presents a third type of application of the micromodel. Chromatin remodelling is a secondary epigenetic effect caused by the interactions between histone modifications and DNA sequences. It refers to the actual physical rearrangement of DNA and histone proteins that results in a compact human genome. The EpiGMP tool is again used to study molecular changes that characterize the chromatin structure in the human genome. For the first time

an agent-based modelling approach is proposed to study the chromatin packing. Preliminary results together with improvements currently being implemented in the model framework are discussed.

- Chapter 8 presents an overview of the principle findings and contributions of research achieved here. Description of the inherent limitations and discussion of possible future directions in computational modelling of epigenetic phenomenon is also given in the last section of this chapter.

Chapter 2

Background - Complexity & Modelling Initiatives

2.1 Cancer Epigenomics and Technological Approaches

DNA methylation was initially identified as one of the most primitive epigenetic mechanisms that organisms utilize to (a) protect genomic DNA and initiate the host resistance mechanism towards foreign DNA insertion and subsequently, (b) control gene expression, [Doerfler and Böhm, 2006]. From an evolutionary point of view as well, the catalytic domain in the structure of the methylation enzymes across all organisms has been preserved to perform methyl group addition. A major change however, in the level and functional utility of DNA methylation was noted in higher organisms such as eukaryotes, when its mechanism evolved from protecting the genomic contents to controlling their level of gene expression. In humans, there are two ways by which DNA methylation is established – (a) *de novo* methylation that establishes new landscapes of DNA methylation, (b) maintenance methylation responsible for inheriting existing DNA methylation profiles. Within the family of methylating enzymes (DNMT), two types namely DNMT3a/b/L and DNMT1 es-

tablish landscapes of DNA methylation in these two ways, [Doerfler and Böhm, 2006]. The *de novo* methylation process carried out by DNMT3a/b/L, is responsible for methylating embryonic cells which are devoid of any previous DNA methylation landscapes and methylated based on the DNA sequence contents. This process is also responsible for establishing parental imprinting[1] and X-chromosome inactivation that is permanently set within the organism enabling it to exhibit phenotypes, unique to itself as an individual from birth. In contrast, DNMT1 distribution is dynamic across a cell during its lifetime. This enzyme type is highly biased towards hemi-methylated[2] DNA sequences, making it responsible for propagating methylation landscapes after each cell cycle. While the DNA methylation process has been dealt with in detail, the recent evidences indicate that a new family of proteins called the TET proteins are involved in gene transcription and conversion of "mC" to "hmC" (*methyl* Cytosine to *hydroxy methyl* Cytosine) thereby *demethylating* the DNA sequences, [Williams et al., 2011; H. Wu et al., 2011]. The process of DNA demethylation is not described in detail in this thesis, although modelling the process is considered to be a part of the potential extensions to the model capability.

In the case of histone modifications, several families of enzymes carry out the addition or removal of chemical groups. Based on the type of amino acids and modifications, selective sets of enzymes are recruited to perform the removal or addition of the chemical groups. Among those enzymes, acetyl addition is done by a group called Histone Acetyl Transferases, (HAT), and removal by Histone Acetyl Deacetylases, (HDAC) on Lysine residues respectively, [Nakayama et al., 2001]. Methyl group addition is performed on arginine and lysine amino acids by Histone Methyl Transferases, (HMT), [Nakayama et al., 2001]. Some protein complexes with specific functional domain referred to as "JmjC", have also been recently discovered to perform histone demethylation, [Cho

[1] Genetic Imprinting is a process that permits faithful inheritance of the methylation status of genetic alleles from parents to the offspring.

[2] DNA sequences are made of double strands. A hemi-methylated DNA sequences refers to methylation of the sequence only in one strand.

et al., 2011; Klose et al., 2006]. The remaining significant, phosphate modification is managed by a broad type *kinase* enzymes in all mammals, [K. Zhang et al., 2005]. DNMT1, described above, is known to interact with histone deacetylases enzyme and some methyl adding proteins, (e.g. HP1), to remove acetyl and add methyl groups in histones, [Allis et al., 2007; Turner, 2001].

2.1.1 Aberrations caused by DNA Methylation and Histone Modifications

The significance of the role of DNA methylation in the cell has become most obvious following from the discovery of sporadic "methylation marks" during cancer progress, [Esteller, 2007]. Such changes in the methylation marks are mainly attributed to the abnormal function of DNMT enzyme complex. This abnormality results in gene imprinting disorders and malignancy formation due to *hyper/hypo methylation* of specific sections in the chromosomes, [Chahwan et al., 2011]. Among the most studied abnormalities recorded in connection to the failure of the DNMT enzyme complex, is *Immunodeficiency–Centromere instability–Facial anomalies (ICF) syndrome*. This is caused by mutations associated with coding for DNMT3B enzymes leading to global hypomethylation of repeat regions located in the pericentromere of human chromosomes, [Ehrlich. Et al., 2008]. Furthermore, *Prader-Willi* syndrome, *Angelman* syndrome and specific type of cancers such as *Wilm's tumour* have also been associated with imprinting disorders characterized by growth abnormalities, [Chahwan et al., 2011]. In these diseases, genetic mutations or altered DNA methylation cause improper imprinting patterns and lead to aberrant expression of the normally suppressed genes, [Chamberlain et al., 2010]. Cancer initiation and progress is mainly attributed to the sporadic occurrence of hypomethylation of oncogenes and hypermethylation of tumour suppressor genes, [Chahwan et al., 2011; Esteller, 2007]. Hence a combination of genetic abnormalities such as mutations and aberrant DNA methylation to specific genes trigger cancerous conditions leading to malignancies that spread across different systems in the human body, [Allis et al., 2007]. For example, in Wilm's tumour, the loss of imprinting of *IGF2* is associated with the spread of cancer

to the lungs, ovaries and colon. In general DNA methylation process when disrupted can lead to, (i) gene activation, promoting the over-expression of oncogenes, (ii) chromosomal instability, due to demethylation and movement of retrotransposons[3]/tandem repeat regions and consequently, (iii) acquired resistance of the human body to drugs, toxins or virus, [Chahwan et al., 2011]. Apart from failure in the control exercised by DM, there are certain protein "Onco-modifications" recently categorized as definitive signatures during occurrence of malignancies, [Fullgrabe et al., 2011]. Emerging evidence in the literature indicates the interface and control of histone modifications not only on gene expression but also on other nuclear processes such as DNA repair pathways, [Kurdistani, 2011; Sawan et al., 2010]. Some of the most frequently studied histone modifications, associated with DNA methylation and tumorigenesis are – acetylation of H3K18, H4K16 and H4K12, trimethylation of H3K4 and H4K20, acetylation/trimethylation of H3K9, trimethylation of H3K27, occurrence of histone variants[4] and also other external proteins such as MBD[5], HP1[6] and Polycomb that play role in chromosome rearrangement, [Chi et al., 2010; Fullgrabe et al., 2011]. Hence evidences on the role of these molecular modifications, as strong indicators of cancer, emphasizes the pivotal role of DNA methylation together with histone modifications in altering the "structure and integrity" of the human genome, [Esteller, 2007].

2.1.2 Data Acquisition

The advent of efficient technologies to detect DNA methylation and histone modifications, has significantly improved data quality over the last decade. Not all methods are reliable, but high

[3]The small DNA sequences that can amplify/replicate resulting in the formation of repeat regions. These repeat regions sequences are alternatively called as "genomic junk" because they are heavily methylated. More than 45% of the human genome is made of such retrotransposons.

[4]Histone proteins have been classified into 5 sub-classes namely H1,H2A,H2B,H3 and H4. Histone Variants are proteins, which belong to one of these classes but differ in a few amino acids along the peptide chain.

[5]MBD - Methyl binding domain protein 1 is a special type of protein that is capable of binding to methylated DNA sequences.

[6]HP1 - heterochromatin protein 1 is type of protein that is associated with gene suppression and heterochromatin formation.

throughput techniques are continually being improved upon rapidly. Some important technologies used most often to detect the epigenetic key elements are described below.

DNA Methylation Detection

Methods to detect DNA methylation can be classified into two categories namely gene specific and global methylation, based on the focus of study, [Esteller, 2007]. The names of various techniques falling within those two categories are summarized in Figure 2.1 and also Table 2.1. A breakthrough in DNA methylation analysis occurred with the invention of *Bisulphite Sequencing*, [Reed et al., 2010]. This method involves conversion of unmethylated Cytosine (C) into Uracil (U) base, which can be filtered away from the methylated cytosine. A common step followed in these detection methods is to combine Bisulphite sequencing with other DNA sequencing techniques such as Polymerase Chain Reaction (PCR)[7], citedbisulphite-comparison2. The only difference is after the Bisulphite sequencing step when, either enzymes, special proteins or fluorescent compounds are used to bind with DNA molecules in order to detect levels of DNA methylation, [Reed et al., 2010]. Table 2.1, gives a brief list of some methods invented for this purpose. A detailed description of each method is given in Appendix C. The most common disadvantages of all these approaches listed in Table 2.1 include the preparation of large number of samples for analysis and complex procedures, (especially the bisulphite-PCR combined techniques), [Shiraz et al., 2012]. Despite the limitations, these quantitative methylation techniques are being applied to study many biological processes such as methylation of imprinted genes, X-chromosome and also differentially expressed genes during cancer onset, [Esteller, 2007].

[7] A type of DNA sequencing technique which uses enzymes to trim the sequence and detect nucleotide bases.

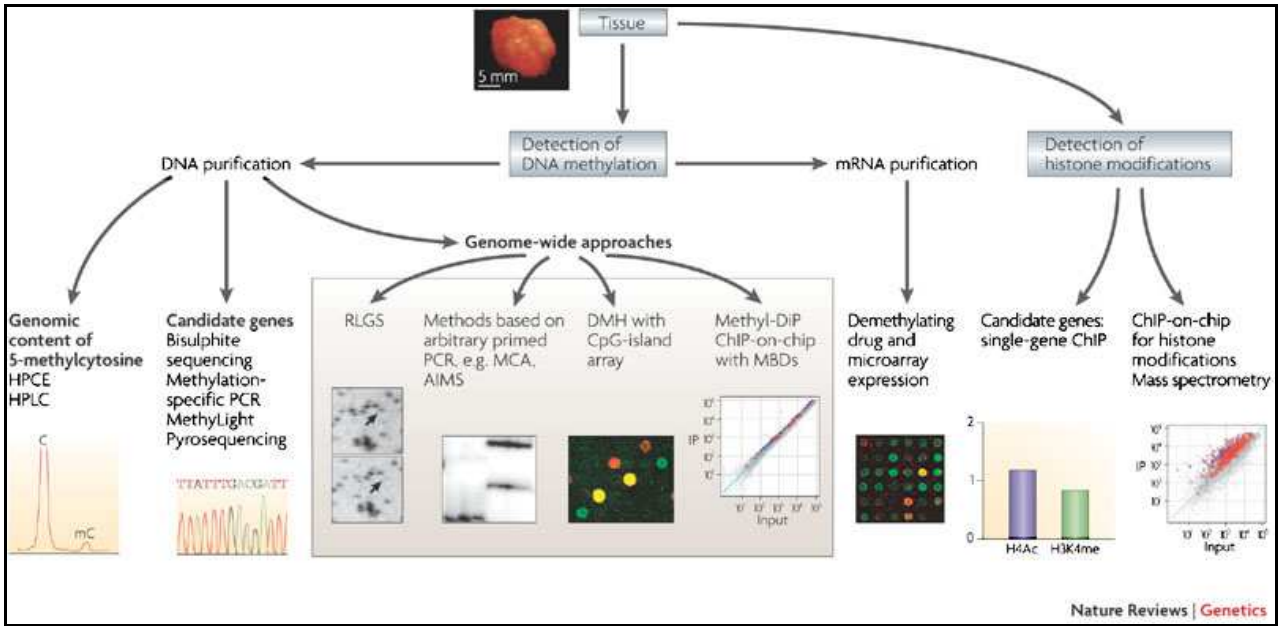


Figure 2.1: DNA methylation associated Techniques

Detection Procedure for DNA methylation and histone modifications across the Genome. Image adapted from [Esteller, 2007]

S.no	Key elements of Epigenetic Mechanisms	Detection Techniques
1	DNA Methylation (combined with Bisulphite or Bi-S)	*Global or Locus Specific* - a. Array based Bi-S with Antibody Hybridization* or Methyl Sensitive Restriction Enzyme Technique b. Non array based Bi-S with Restriction Landmark Genomic Scanning* c. Methyl Specific Digital Karyotype*
		Local or Gene Specific - a. Sensitive Methyle Light* (Qualitative) b. Cloning or Allele Specific* (Quantitative) c. Pyro and Manual Sequencing* (Qualitative)
2	Histone Modification	*Mass Spectrometry and Chromatography*
		Chromatin-Immuno Precipitation based - a. ChIP-on-Seq* (with DNA Sequencing) b. ChIP-on-Chip* (with Microarray Tecnology)

Table 2.1: Detection Techniques
A brief list of laboratory based techniques used to detect DNA Methylation and Histone Modifications. Phrases marked by "*" are defined in the Appendix section C.

Histone Modification Detection

Technologies to study histone modifications pose greater challenges than the ones that have been employed to study DNA methylation. If the golden standard for DNA methylation techniques was *Bisulphite Sequencing*, detection of post translational modifications in histone proteins is achieved by *Mass Spectrometry*[8]. Usually a combination of techniques which includes Chromatography, (HPLC)[9] and Mass Spectrometry is applied to analyse large samples, [Esteller, 2007]. Recent advancements in Mass Spectrometry, have enabled localization of specific histone modifications, and allowed unbiased quantification of the nucleoproteins from the biological samples. When combined with chromatography, the sample is initially separated by a column chromatography technique based on its constituent sizes, and then analysed by the mass spectrometer, [Moco et al., 2006]. The use of *Chromatin Immuno-Precipitation or ChIP* in combination with other quantification technologies to detect histone modifications have led to generation of new data in recent years, [Carey et al., 2009]. The standard procedure in ChIP, involves sonicating the samples with hybridizing agents to separate DNA molecules and nucleoproteins, followed using PCR to perform the DNA sequence analysis. The hybridizing agents utilized varies based on the protein of interest that needs to be extracted during the ChIP experiment, [Hoffman et al., 2009]. If a DNA sequencing method like PCR follows the ChIP procedure, the experimental technique is referred as ChIP-Seq and if a high throughput method such as DNA microarray is adopted, the method is called as ChIP-on-Chip, [Aparicio et al., 2001], (refer Table 2.1).

Data Limitations

One disadvantage is that although these methods have generated a lot of discrete data, they do not help to detect all modifications across the entire contents of the human genome at the same time,

[8] An analytical method which detects the mass-charge ratio of any particle passing through the detector.

[9] HPLC or High Performance Liquid Chromatography is a lab technique used to separate the different components from a mixture of liquids.

[Esteller, 2007]. Also, when it comes to ChIP based experiments; use of specific antibody proteins that hybridize and bind to the specific histone protein modifications is an essential protocol that needs to be followed. As a consequence, it is a challenge to design one antibody for the quantification of all possible histone modifications. ChIP-on-Chip experiments are very expensive and need to be performed at least thrice to ensure accuracy of results, [Esteller, 2007]. The combination of microarray technologies with ChIP, generates a huge amount of data, which needs to be statistically analysed, (involving normalization algorithms), in order to determine what is biological significant, [Malone et al., 2011]. In general, these technological approaches for both DNA methylation and histone modifications, require great expertise and high precision for implementation and are consequently expensive although sequencing is considered more cost effective than the alternatives. Laboratory-based measurements provide valuable information but the nature of experiments and in particular the tight focus on a single quantity or target means that data are both dispersed and incomplete for all questions that need to be answered, [Aparicio et al., 2001; Esteller, 2007]. In this situation, computational modelling has a role, and in particular can help, (i) explore fundamental patterns in DNA sequences across the human genome and (ii) to model molecular interactions among DNA methylation and histone modifications in order to understand how they may occur. These considerations provide strong motivations for model development to both address gaps in data and also incorporate diverse data as these become available from laboratory experimental trials.

2.2 Insights into the Human Epigenome and DNA Sequences

Histones are closely linked to DNA molecules and play a vital part in encoding information from them. Over time, histone proteins have diversified from a few ancestors into five distinct types of subunits in eukaryotes thus forming the octameric structure of a nucleosome, [Allis et al., 2007]. The histone octamer or core plays the most important role in condensing billions of DNA base pairs

compactly within 23 pairs of chromosomes in the human genome. Covalent post-translational histone modifications are mainly held responsible for chromatin architecture and propagation of many cellular events and, sometimes, disease onset. Thus, with more than one type of histone containing multiple types of modification in their tails present a potentially complex scenario, [Cedar et al., 2009; Jenuwein et al., 2001; Kouzarides, 2007; Zheng et al., 2003]. DM and HM most often have a mutual feedback influence hence maintaining a strong dependency over one another. A very interesting fact about histone modifications is that though the exact mechanisms are unknown, they are memorized by the cells post "DNA replication", especially those that aid in gene expression, methylation maintenance and chromosome structure stability, [Kouzarides, 2007]. Among all the histone modifications, methylation (mono/di/tri) and acetylation have been most studied in regard to their influence over gene expression and DNA methylation, [Cedar et al., 2009]. These modifications are quite often noted to compete for the same type of residues and are also known to recruit antagonistic regulatory complexes such as trithorax and polycomb proteins, [Allis et al., 2007]. For example, histone methylation was found to be important for DNA methylation maintenance at imprinted loci, which could lead to disorders such as Prader-Willi syndrome, [Chahwan et al., 2011]. Such individual experiments have helped unravel the connection step by step between levels of DM and specific histone modifications, [Barber et al., 2004; Ito, 2007; Meng et al., 2009; Sun et al., 2007; Taplick, 1998; Wyrick et al., 2008]. Irrespective of these elaborate measures, a complete picture of the molecular communications and their dynamics that control the cellular events is still lacking.

Relating information from DNA sequences to epigenetic alterations requires several stages. The role of principal events such as DNA methylation and histone modifications in controlling transcription have been discussed above. In general, transcription and translation processes provide information on the amino acid composition to encode protein sequences, which is in accordance with the well-known central dogma of molecular biology, (DNA$\rightarrow$RNA$\rightarrow$protein). The identifi-

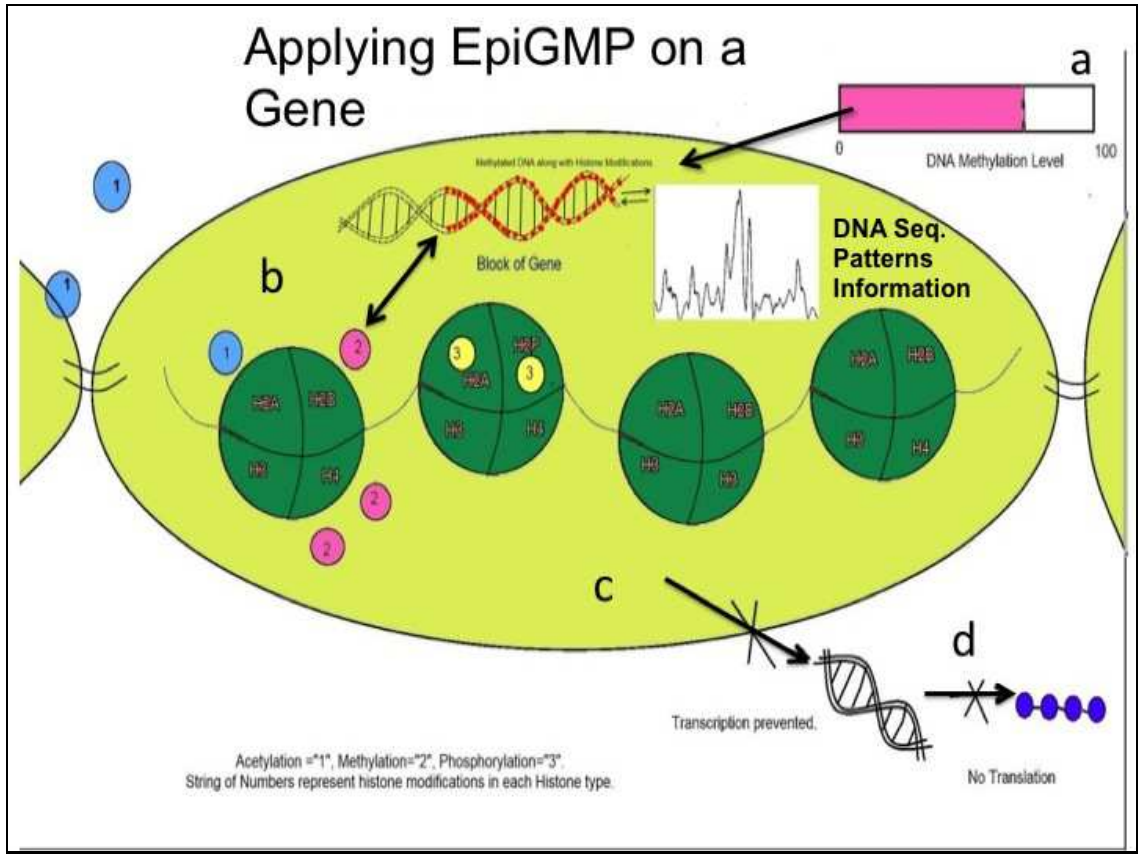


Figure 2.2: DNA Sequence Pattern Analysis & Epigenetics

Depiction of Epigenetic Control of gene expression and combining DNA seq. patterns information. Big yellow oval=gene, dark green circle=histone cores, blue circle=histone acetylation modification, pink circle=histone methylation modification, yellow circle=histone acetylation modification. Steps/stages in EpiGMP are,"a"= level of DNA methylation in gene, "b"=Histone modifications and DNA methylation interactions, "c"=transcription or gene expression level, "d"=translation or protein production. DNA methylation process also depends on DNA seq. patterns, when stage "b" occurs.

cation of any gene sequence by specialized binding proteins before inactivation or transcription, (followed by translation) is based on the presence of small oligonucleotides, [Ozers et al., 2009]. The same understanding applies to proteins associated with the establishment of DNA methylation. Hence it is necessary to investigate the nucleotide patterns and how they relate to epigenetic layer in the human genome.The schematic figure 2.2, represents the framework of the computational model and how effectively the role of DNA sequence patterns is incorporated along with epigenetic elements to investigate the process of gene expression. Given this argument, which reflects on the significance of DNA sequences, the following subsections provide details about nucleotide pattern analyses carried out so far.

Significance of DNA Sequence Patterns

Since the human genome consists of more than three billion base pairs, enormous efforts have gone into investigating its contents and organization, [Collins et al., 1998; Strachan et al., 1999]. The spread of DNA methylation in the genome is not randomly determined as it might be assumed, [Esteller, 2007]. Emerging evidence indicates that the underlying genotype or DNA sequence has a strong key role in enabling and propagating a spectrum of methylation landscapes, [Doerfler and Böhm, 2006; Gertz et al., 2011]. The nature of every biological cell is characterized by its preservation of the genetic contents and epigenetic landscape, also known as “dual inheritance”, [Holliday, 2006]. Consequently, it is of utmost importance to map the underlying nucleotide pattern in DNA sequences for further comprehension of the epigenetic phenomenon, [Gertz et al., 2011; Glass, Fazzari, et al., 2004; E. Segal et al., 2009]. Knowledge of the distribution and location of nucleotides, specifically dinucleotide “CG” can be utilized to understand the biological significance associated with determining the level of DM.

Note on CG dinucleotides

The DNA sequences in the human genome has been categorized as genes or coding regions and retrotransposons based the functionality. In relation to the spread of DNA methylation information, a special type of sequence located near genes which incorporates dinucleotide frequencies, are the CpG islands[10]. These islands are mostly found near the promoters, (5'-end), of genes and their methylation levels are closely monitored because they act as indicators of Cancer progress, [Esteller, 2007]. The CG dinucleotides are under-represented throughout the human genome, but are densely located in certain repeat regions, (retrotransposons) and CpG islands, [Esteller, 2007]. Previous analyses indicate that certain patterns of CG present in promoters and islands of specific genes are easily accessible by the DNMTs enzyme complexes, thus initiating methylation process. For example, studies conducted by Glass, Fazzari, et al., 2004, emphasize that the *de novo* methylating enzyme DNMT3a/L, is biased toward CG dinucleotides, appearing after every 8-10bp intervals [Glass, Fazzari, et al., 2004], (examples of spacing between CG dinucleotides are shown in Figure 2.3). Hence it is useful to perform a complete distribution or pattern analysis of nucleotides in human sequences; in particular of CG to understand how methylation is established and maintained. Although evidence about the nature of DNMT mechanisms in setting new methylation profiles is incomplete, analysing the global periodicities or distributions of CG dinucleotides will help to reveal at least a part of the hidden picture. A general overview of pattern analysis techniques is given together with a description of how time series analyses is applied in understanding CG dinucleotide occurrences in specific human sequences in the following subsections.

[10]DNA sequences are defined and classified as CpG islands if, (a) length of that DNA sequence $>$200 bp, (b) Total amount of Guanine and Cytosine nucleotides $>$50%, and, (c) the observed/expected ratio of CG dinucleotides for that given length of sequence, $>$60%, [Takai et al., 2002].

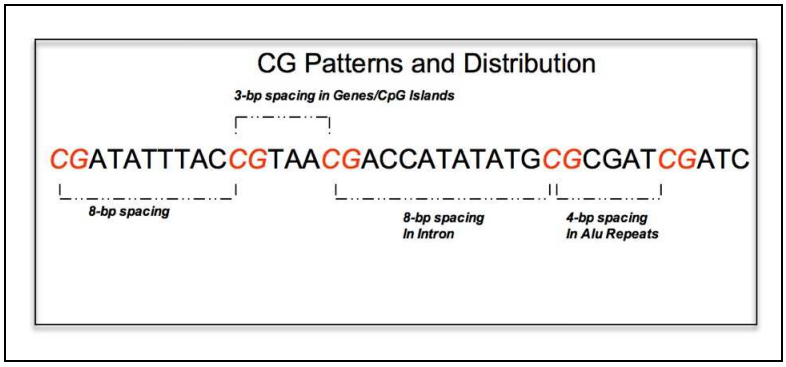


Figure 2.3: CG dinucleotide Spacing
Distribution of CG in Human DNA sequences.

2.3 Methods to analyse DNA Sequence Patterns

Several pattern recognition/analysis techniques or *time series analysis* methods, have been explored, ranging from simple statistical measures to sophisticated transformation and decomposition methods such as the Discrete Wavelet Transformation, (DWT). Calculation of the "Expected Frequency" based on the empirical probabilities of the occurrence of nucleotides was proposed by Whittle, and further developed on DNA sequences by Cowan, [Cowan, 1991; Whittle, 1955]. In the latter, transition probabilities (for all 16 types of dinucleotides - AA, AT, AC, AG, GG etc.) in the form of a matrix were constructed from known DNA sequences, to predict patterns along a new sequence. This particular analysis was performed on specific sequences containing the same starting and ending nucleotides. Another tool namely "GC-Profile" was used to visualize sequences and calculate nucleotide frequencies from the total amount of G and C nucleotides in small genomes, [Gao et al., 2006]. A standard pattern analysis can be conducted using the Fourier Transformation (FT), which allows decomposition of the time/spatial components in the data and construction of a frequency map, [Morrison, 1994]. Fields of application are wide in range with examples from – Physics (optics, acoustics and diffraction), Signal Processing and Communication Systems, Im-

age Processing, Astronomy, and DNA sequence analysis, amongst others, [A'Hearn et al., 1974; Goodman, 2005; Salz et al., 1969]. Early work by [Tiwari et al., 1997], involved use of Fourier Transformation technique to detect DNA patterns. In this method, small sequences from bacteria were first converted into four distinct sets of binary sequences, (each corresponding to location of a nucleotide), then analysed by applying Fourier Transformation technique. This was followed by a comparison between genes and non-coding regions, and identification of characteristic features/patterns such as 3bp periodicity in genes. This type of application gave rise to the phrase "Periodicity" of nucleotides i.e. count of appearance of specific patterns that appear in sequences, (sample of patterns with different base pair intervals/spacing among consecutive CG dinucleotides is shown in Figure 2.3). Subsequent research focused on these periodicities of small patterns (of length 10bp) in blocks of sequences. Thus the Fourier transformation was used to study frequency components of the sequences along a spatial axis where each nucleotide was represented by a directional vector. Periodicities in virus strains (SV40) were also studied to check for patterns of dinucleotides and their corresponding role in genome condensation, [Silverman et al., 1986]. The most prominent periodical pattern of 10-11bp, portrayed by pyridines (AA/TT/AT), which are involved in long-range interactions of up to 146 bp and aid in nucleosome alignment, was confirmed through these attempts. Refinement of this method through introduction of new parameters included calculation of *autocorrelation*[11], for specific patterns from DNA sequences. More recently, further improvements have been employed and tested on example sequences, [Epps, 2009]. Complete and significant analyses of patterns or biological markers on sequences were identified by, [Herzel et al., 1999] and [Hosid et al., 2004] from *E.coli* genome.

For the DNA sequences, the Fourier Transformation represents an initial analysis, which involves calculating the autocorrelation, profile for desired dinucleotide/ nucleotide and generating

[11] Autocorrelation is defined as correlation of a signal with itself after a time delay. Here the term addresses the intervals or distance in base pairs that separates consecutive CG dinucleotides. - Figure 2.3

frequency maps of the specific components. Further details following [Clay et al., 1995], with regard to its application to study nucleotide distribution in genes, non-coding regions and CpG islands are discussed in the next chapter. An extension to the Fourier analysis, *Discrete Wavelet Transformation* (DWT), is the application of a set of orthonormal vectors in space to localize and study both frequency and time/spatial components for a given dataset, [Kaiser, 1994]. The resulting coefficient matrix, a product of this family of vectors and input data helps to indicate regions of high and low frequencies along the spatial, (or sequential) axis based on an initial resolution factor, (e.g. Haar and Mortlet, [Kaiser, 1994]). Wavelets or specifically the method of DWT addressed here, have been quite extensively used to study financial markets, [Conlon et al., 2009], experimental data from Protein Mass Spectrometry and DNA sequence patterns amongst others, [Kwon et al., 2008]. Discrete Wavelet Transformation has been applied to visualise both frequency and location specific information of the DNA sequence patterns by the authors of [Tsonis et al., 1996; Zhao et al., 2001]. In our case, we have applied a modified version of DWT – Maximal Overlap Discrete Wavelet Transformation or MODWT, [Conlon et al., 2009] to study sequences from Chromosome 21 in the human genome. Elaborations on the methods to study patterns in DNA sequence and results thus obtained are reported in next chapter.

2.4 Modelling DNA Methylation and Histone Modification Interactions

Epigenetics is a relatively new field and formal models to study the associated phenomena are limited to date. The advent of favourable experimental techniques such as Protein Mass Spectroscopy, [Sundararajan et al., 2006], ChIP-Seq and ChIP-on-Chip, have led to new data and confirmed facts with regard to DNA-protein interactions and their role in cancer onset. Such experiments usually generate a large amount of data including measures such as direct count of modification detected

along the genome after specific intervals of DNA sequences, (standard intervals are 200 or 400 base pairs for histone modifications detection), [Sundararajan et al., 2006]. As discussed in detail, by [Bock et al., 2007], extracting comprehensible epigenetic information from experiments is a three-stage process. First, the biochemical interactions are stored as genetic information in DNA libraries, followed by applying DNA experimental protocols such as tiling microarray[12], (special type of microarray experiment) along with ChIP-on-Chip, and lastly applying computational algorithms to infer error-free epigenetic information from these experiments. These algorithms are mainly quantitative and help to establish a pipeline for prediction of probable epigenetic events. An initial coarse attempt to define the epigenetic, genetic and environmental interdependencies paved the way for an in depth study of the molecular factors that trigger these effects, [Cowley et al., 1992].

2.4.1 Quantitative Models

Among the computational attempts to model and analyse epigenetic mechanisms, some have successfully identified correlated histone signatures during gene expression using data from ChIP-on-Chip experiments and microarray-based gene expression measurements, [Karlić et al., 2010; Yu et al., 2008]. A Bayesian network model was constructed using the high-resolution maps[13] from laboratory experiments to establish causal and combinatorial relationships among histone modifications and gene expression, [Yu et al., 2008]. Quantitative measure of other proteins such as Polycomb, CTCF[14] and other transcription factors were also included to build these models. Based on Bayesian networks, conditional probabilities and joint probability distribution measures

[12] A subtype of microarray, tiling arrays are intensively applied to study sequences which form a part of contiguous regions in the genome. Hence it is useful in characterizing sequenced regions whose local functions that are largely unknown.

[13] The protocols in ChIP experiments, help to zoom and identify modifications in histones which sometimes saved as graphic objects, (like pictures). These are obtained by capturing the binding of target histone and a detectable chemical compound or another protein, [Hall et al., 2009].

[14] CTCF - 11-zinc finger protein or CCCTC-binding factor is an active transcription factor in the human genome

of datasets were calculated and a finely clustered molecular modification network was obtained. Repeated bootstrapping or random sampling verified the robustness of this Bayesian network. For initial analysis, datasets containing information from ChIP-on-Chip experiments ([Cuddapah et al., 2009] and [Boyer et al., 2006]) for histone protein modifications in human CD4+ immunity, cells and gene expression measurements from microarray experiments (obtained from [A. I. Su et al., 2004]), were extracted for clustering (using k-means), followed by construction of the Bayesian network.

Another quantitative model based on similiar information such as data from ChIP-on-Chip experiments, obtained from the literature, [Cuddapah et al., 2009], was developed using Linear Regression [Karlić et al., 2010]. In this case, the model was of the form: ($N_{i,j}$'=$N_{i,j}$+constant), where, $N_{i,j}$ = count of j_{th} modification in i_{th} gene in template samples. This equation was expanded to include more variables, to study multiple histone modifications, thus giving rise to more than one model type. Secondary information was also extracted and included in the model, namely, microarray expression data from the literature, [Schones et al., 2008] and promoter blocks information from Unigene database, (details of the website for this database is given in Appendix D). Here, loci of new sets of ChIP-on-Chip experimental results for histone modifications, were mapped on the human genome using annotation track information obtained from University of California Santa Cruz genome browser, (details of website for this online tool is given in Appendix D). These multivariate models were applied on different sequence datasets, which were based on Low CG or High CG dinucleotide concentration. The whole dataset thus obtained was divided into training and test sets namely – D1 and D2, where Pearson correlation coefficient values were used to confirm the accuracy of prediction, (D1) over the test set, (D2). This model was also extended over different cells, (with initial trials being conducted on CD4+ human cells), for nine histone modifications and for confirmation on CD36+ and CD133+ human immune cells respectively.

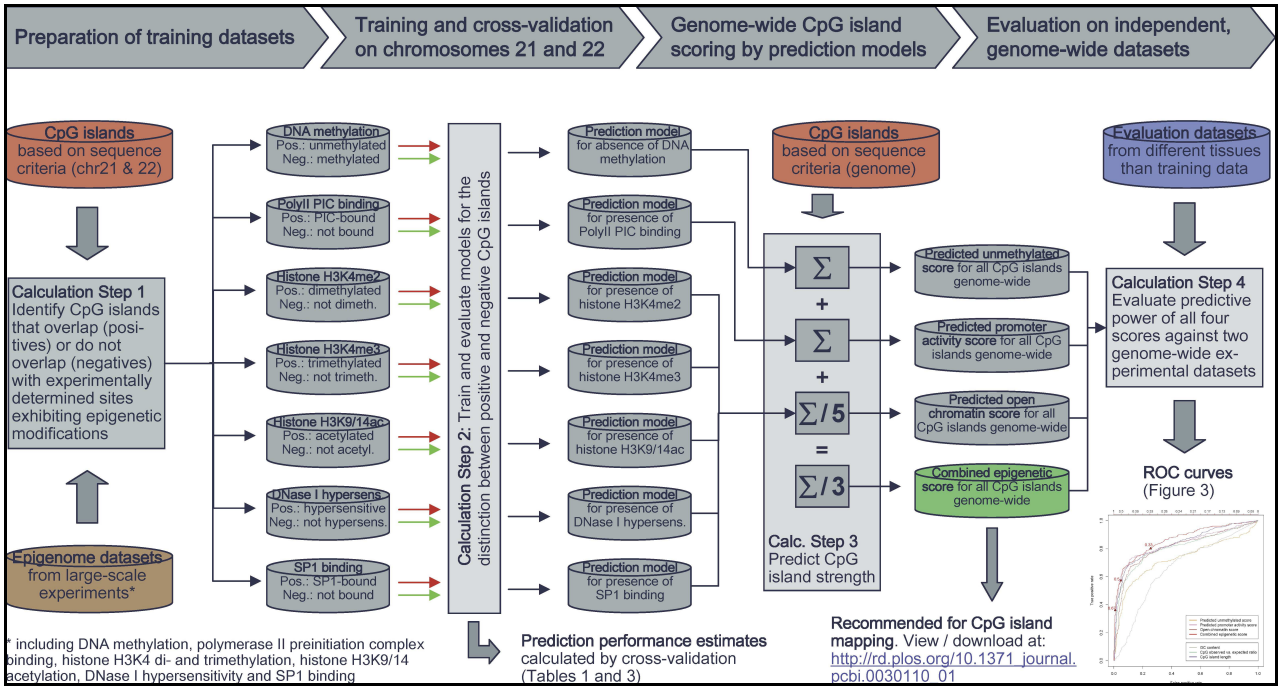


Figure 2.4: Quantitative Epigenetic Model

Previous attempt of computational modelling of epigenetic components by Bock et al. (Image adapted from [Bock et al., 2007]).

Bayesian Network Extensions

Other models based on Bayesian networks, have focused on developing tools to study DNA methylation and protein modifications, [Bock et al., 2007; Das et al., 2006; Jung et al., 2009; J. Su et al., 2010]. Among those, two models by J. Su et al. and Bock et al. have mainly focused on identifying the function of CpG islands using information on histone modifications. These type of "reverse" models explain the feedback connectivity between the two epigenetic events (HM and DM). The model by Bock et al. was an important initiative in computational epigenetics, since a clear pipeline for analysis of epigenetic data was proposed. The training model used several inputs from the experimental datasets to identify *bonafide* CpG islands. Inputs included – CpG islands that qualified based on criteria defined, [Takai et al., 2002] and epigenetic datasets from experiments (such as lysine modifications in histones, transcription binding factors, MBP, and SP1 proteins). This work consisted of three main steps which involved, (i) identification of predictive parameters from the datasets, (ii) cross validation and training of data using a linear support vector machine, and (iii) comparison of CpG islands previously identified in chromosome 21. These elaborate measures took into account the amount of histone modifications affecting the methylation status hence emphasizing on the strong connectivity between methylation levels and their corresponding epigenetic states. Similar to the model described, [Yu et al., 2008], another complementary attempt was made to construct regulatory patterns that appear in histones during high DNA methylation. A Bayesian network once again was used to predict a list of methylation modifications that leveraged the occurrence of DNA methylation (using the same datasets obtained from CD4+ cells[15] in humans), [Jung et al., 2009]. These independent and repeated attempts, on accumulation, helped to identify and confirm a definitive pattern and characteristic modifications that exist in epigenetic events in the human cells: for example, more acetylation modifications appear during gene expression and more methylation modifications are preferred during gene suppression.

[15] A special type of immunity cells in Humans that attack foreign bodies in the blood stream

2.4.2 Need for Phenomenological Model Framework

A major disadvantage in the development of these quantitative models, similar to performing the laboratory-based experiments is the restriction of obtaining results from a single source or studies performed to investigate a single disease onset, (as indicated in Section 2.1.2). Thus a general model framework able to link different epigenetic events is lacking. This has motivated a need to develop a "*phenomenological*" model that can report modifications occurring in genes associated with any type of cell or cancer, (provided there is evidence on the role of genes in diseases). As a consequence, in Chapter 4 we describe the development of a theoretical model incorporating information from several epigenetic events and test this on synthetic data. The novelty of this micromodel lies in accounting for the dynamics in the epigenetic mechanisms based on a stored library of possible histone modifications as well as methylation levels in the DNA sequences. The model allows sampling of possible solutions of histone modifications, using probabilities of transition. Based on the accumulative knowledge on the nature of modifications as mentioned above, probabilistic cost functions are used to set the interdependencies between variables (level of HM and DM) in this model.

2.5 Summary

The fundamental facts about Epigenetic Mechanisms, and their direct role in tumorigenesis[16] and Cancer progress was explained. The various experimental techniques utilized to investigate DNA methylation and histone modifications and their disadvantages have also been reviewed. This was followed by a detailed description of the computational algorithms developed so far to detect nucleotide patterns in human DNA sequences. The status of Computational Epigenetic models have been considered and the need for a more flexible framework has been argued. The next chapter in Part I elaborates on the analytical techniques used to extract epigenetic information from the DNA

[16]The production and development of tumours, usually cancerous in nature.

sequences, using decomposition methods introduced, (Fourier, Wavelet transformation). These are applied to real datasets, and results obtained are discussed.

Chapter 3

DNA Methylation Analysis

3.1 Introduction

Recent research has emphasized on the fact that DNA methylation is very much influenced by the nucleotide contents in the human genome, [Raghavan, Ruskin, and Perrin, 2011], which is known to contain hereditary information. Following the explanation of Fourier and Wavelet analyses techniques in the previous chapters, we apply these here to the study of DNA sequences from human chromosome 21. The main focus in this chapter is to investigate meaningful patterns of CG dinucleotide in DNA sequences as it strongly associated with DNA methylation across the human genome, [Glass, Fazzari, et al., 2004]. Other dinucleotides including GC, AA/TT are also of interest as they play a major role in gene expression and chromatin packing, respectively, [Tanaka et al., 2010]. From Fourier analysis we obtain the periodicity[1] information in terms of base pairs (bp) units for the distances between such repeats. Subsequently, we consider the wavelet transformation specifically the MODWT applied to *contiguous*[2] sequences obtained from laboratory based sequence experiments. The focus here is on the location of specific positions of dinucleotides and

[1] Periodicity and frequency, where the former is the inverse of the latter, are commonly used to describe signal components.

[2] Contiguous or contig sequences are overlapping DNA segments obtained from DNA sequencing experiments.

repeats along the length of the DNA sequences, as well as confirmation and elaboration on periodicities. The aim is to test the approach utilized to reconcile known biological features and explore new ones with a view of determining characteristic patterns and relating these to epigenetic changes of interest. Much of the work presented in this chapter has been published in [Raghavan, Ruskin, and Perrin, 2011], with some additional details explained here. Further remarks on the other types of dinucleotides are provided with supplementary materials given in Appendix A.

3.2 Methods

The first step in Fourier analysis involves calculating the autocorrelation of the CG distribution, so that background noise is removed (or e.g. contribution of other nucleotides in the sequence) and frequencies of the desired components are highlighted in the power (Fourier) spectrum. Positional Autocorrelation for each lag/interval "k[3]" and the power spectrum (or applying autocorrelation in Fourier maps), that provide information on global periodicities, were calculated as described, in equation 3.1 and [Hosid et al., 2004]. DNA sequences obtained using Map viewer, NCBI database and UCSC genome browser[4] were classified into 3 sets (20 genes, non-coding regions near the genes and all CpG islands in chromosome 21), for Fourier analysis (details in Table 3.1).

$$f_p = \frac{\sqrt{\sum_{i=1}^{m} sin\left((2\pi * \frac{i}{p}) * (X - X')\right)^2 + \sum_{i=1}^{m} cos\left((2\pi * \frac{i}{p}) * (X - X')\right)^2}}{\left(2\pi * \sum_{i=1}^{m} (X - X')^2\right)} \tag{3.1}$$

f_p = Normalized wave function amplitude at period - p

X = auto correlation profile of the dinucleotide

[3]The variable k used in Fourier Analysis and Wavelet Studies are different.

[4]Details of the website for these databases are given in Appendix D

X' = mean Auto Correlation

m=Maximum autocorrelation distance

Maximal Overlap Discrete Wavelet Transformation (MODWT), the second approach, was adapted to check for CG distribution along a given input sequence. This transformation indicates loci of high and low frequencies of CG, not detected by Fourier transformation [Conlon et al., 2009]. MODWT consists of the application of linear filters, (or vectors) that transform a given input into coefficients related to variations in the input signal over a set of scales (using non orthogonal basis vectors in this case). Similar to Discrete Wavelets, multi-scaling is achieved using mother and father wavelets, (Debaucies in our case). The wavelet representation of a discrete signal *f(t)* in $L^2(R)$ or Hilbert space [Kaiser, 1994] is given by:

$$f(t) = \sum s_{J,k} \ \phi_{J,k}(t) + \sum d_{J,k} \ \phi_{J,k}(t) + ... + \sum d_{1,k} \ \phi_{1,k}(t) \tag{3.2}$$

for, $1 < k < L$, with L, the length of filter used ($L = 4$ here,). The coefficients $s_{J,k}$ and $d_{J,k}$ are the smooth and detail components respectively [Conlon et al., 2009]. MODWT allows the input data to have any length, and also retains the downscaled (high frequency component) values for each level of decomposition (J in the above equation). For MODWT analysis, DNA strings were converted to numerical sequences, [Voss, 1992]. (For example, CGAATCG can be represented as 1000010, where 1 is recorded only for a position where C is followed by G; 0 otherwise). Raw and assembled contiguous DNA sequences, (i.e. a set of overlapping DNA segments derived from a single genetic source by DNA sequencing techniques), in chromosome 21 were utilized for this analysis using Map Viewer in NCBI database, (Table 3.1).

S.No	Method	Input Sequence Type
1.	Fourier Transformation	*Disease associated genes and introns located near them*– PRSS7, IFNGR2, KCNE1, MRAP, IFNAR2, SOD1, CRFB4, KCNE2, ITGB2, CBS, FTCD, PFKL, RUNX1, COL6A1, COL6A2, PCNT2, CSTB LIPI, TMPRSS3, APP
2.	Maximal Overlap Discrete aWavelet Transformation	*Contiguous Sequences with Accession No.*– NT_113952.1, NT_113954.1, NT_113958.2, NT_113953.1, NT_113955.2 and NT_029490.4

Table 3.1: Details on the type of sequence data used for each time series method

3.3 Results and Discussion

3.3.1 Fourier Analysis

The Fourier analysis calculated from auto correlation of a DNA segment, (of any length containing CG with lag or spacing "k" between occurrences), serves to detect the characteristic global frequencies for each of the three regions. The results are shown in Figure 3.1 and Table 3.2.

S. No.	Sequence Type	Characteristic Periodicity
1.	Genes	3
2.	CpG Islands	3,7
3.	Non-coding Regions	4,8,10-11, 24,25,26

Table 3.2: CG Characteristic Frequencies
Distribution of CG in defined regions of Chromosome 21, Human Genome Sequence No. & Regions & Characteristic Periodicities

CG Distribution

Figure 3.1 represents amplitudes of the power spectrum for all values of CG periodicities possible. Gene coding regions show an apparent peak at 3 bp, which might be expected due to the codon

bias, [Hosid et al., 2004]. CpG islands, throughout chromosome 21, also contribute to the peak at a periodicity of 3 bp since these are present near the promoter regions, (This related to coding theory triplets and translation to amino-acids). The islands also have 7 bp spacing possibly due to presence of repeats sequences containing CG, [Takai et al., 2002]. Most such islands (repeats) can be found near suppressed or silenced gene clusters, with high levels of methylation [Glass, Thompson, et al., 2007].

The authors of [Glass, Fazzari, et al., 2004], give a detailed list of possible periodicities and correlate their biological significance. Hence the initial step here was to identify the appearance of such previously defined frequencies. The placement of CG after 3 bp, in genes and in even more densely clustered CG islands prevents the DNMT complex from naturally methylating those regions. Hence spacing repeats of CG can be used in confirmation to define a CG island, over and above percentage CG content in any input sequence. Research has indicated that 8 bp intervals, (and also 4 bp), which correspond to satellite repeats between CG dinucleotides is a pattern that is, the most amenable to the attraction of DNA methylation complexes, (also indicated by [Glass, Fazzari, et al., 2004]). In fact, genes silenced in germ cells by the *de novo* methylation mechanism, have 8-bp spaced CG near their promoters, which is preferentially recognized by the methylating enzymes complexes of DNMT3a. Another peak, observed in Figure 3.1, between 10 to 11 bp periodicity, strongly known to support genomic structural condensation in lower organisms, is obvious but not large in wave amplitude, [Tanaka et al., 2010]

One of the more prominent and interesting features however is observed in the non-coding regions, which display patterns (between 24 and 26 bp). Other peaks, at periodicity of 15 and 20 bp, are less persistent and disappear when average wave amplitude versus periodicity over all 22 chromosomes for the 3 regions was plotted, with those at 8 bp and 24 to 26 bp only being consistent throughout. The hitherto unreported periodicity of an interval of length 24 to 26 bp, in the

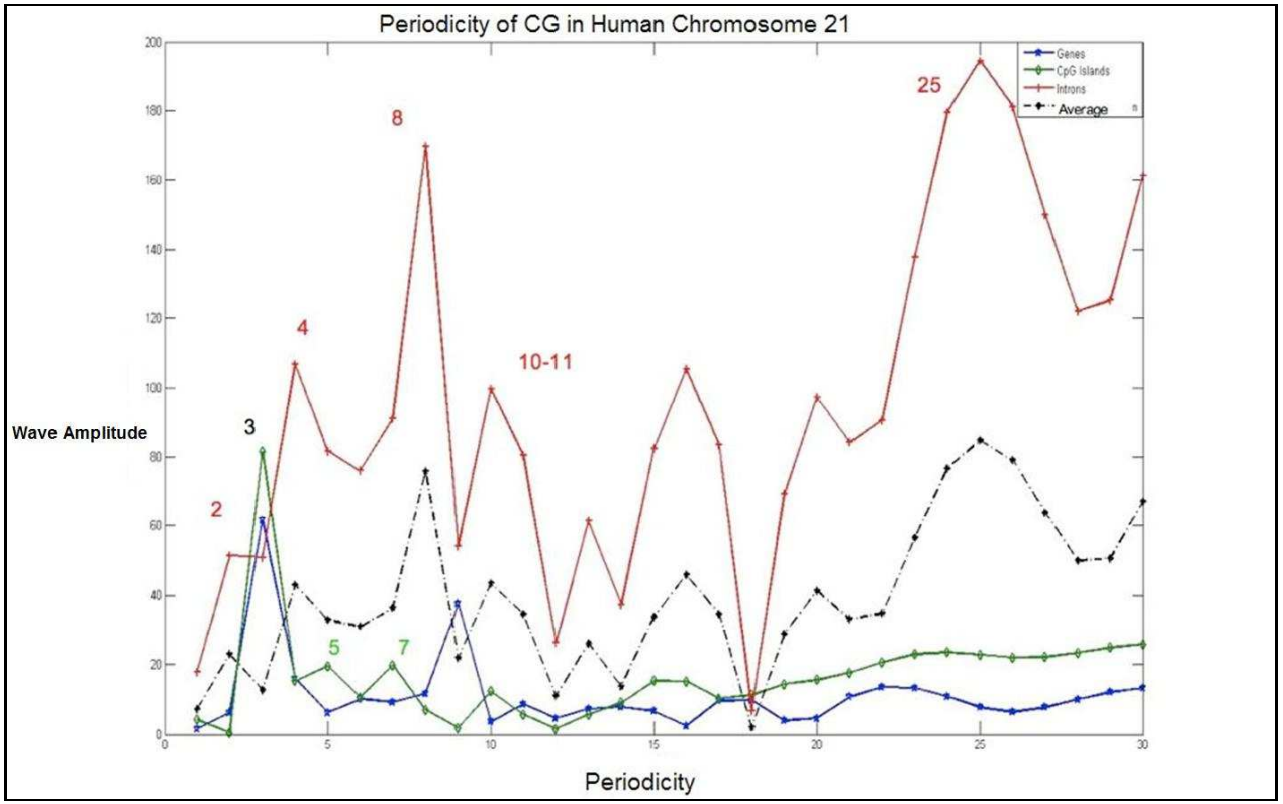


Figure 3.1: CG Fourier Map

Fourier analysis (Periodicity Vs Wave Amplitude) of global periodicities of CG dinucleotides in 20 genes (blue line), non-coding near them (red line) and all CpG Islands (green line) in chromosome 21. The average of the 3 region levels is shown as a dotted line.

non-coding region is less readily explained, but may be connected to DNA methylating mechanisms. One major clue, indicated in the work by [R. Li et al., 2010], is the presence of several million repetitive 25-mers in the human genome. Although not uniform throughout, this is known to be high, on average, in chromosome 21. Further, in [Yin et al., 2007], the authors explain that piRNA or Piwi protein associated iRNA or interference RNA, (which is significantly involved in cellular processes and propagation of *de novo* DNA methylation), is usually of length 24 to 26 nucleotides. This piRNA, (bound with small Interference RNA, i.e. components of RNA-induced silencing complex), is involved in silencing the retrotransposon regions in specific cells, through RNA interference mechanisms [Yin et al., 2007]. It appears that this specific periodicity of the CG dinucleotide as a human genome marker could help in understanding the role that iRNA associated DNA methylation plays in gene expression.

These interesting facts are only the tip of an iceberg in the sequence analysis of humans, especially with respect to differential gene expression, as controlled by epigenetics. The average wave amplitude over the 3 regions for each periodicity, (shown as the dotted line in Figure 3.1) also shows peaks at frequency 4, 8 and 24 to 26 bp. This is due to the fact that non-coding regions form more than 50% of Human genome, (with genes and CG Islands occurring in between) and hence the contribution of Intron regions to patterns and nucleotide periodicities is relatively high with respect to wave amplitude (contributing a large component to wave amplitude).

3.3.2 Wavelet Analysis

A summary of MODWT results is shown in Figures 3.2, 3.3 and 3.4. Contig sequences in chromosome 21 were analysed to help locate and interpret important regions that contained CG dinucleotide. We investigated each of the contig sequences (1, 2 and 3 out of 6 being reported here), for specific CG patterns using all values of scale, (J), allowed. We focus on high frequency components, denoted by $s_{J,K}$, for specific scales, in equation 3.2. (Maximum scale allowed is the log-

arithmic value of input length. Hence for all 3 contig sequences, J took values of range $\in$ [1,16], decided based on the length of input sequence).

In general, a scale range of 9 to 12 was found to be suitable to identify regions of interest, (genes, islands and introns) and to extract information on CG content. The most suitable scale however, appeared to be 11, which had a smoothed representation of the coefficients and reduced noise when a comparison of MODWT results and Fourier analysis was carried out. Figures 3.2, 3.3 and 3.4 show the MODWT coefficients, (blue line) and colored bar strips indicate presence of specific periodicities of possible interest along the sequences.

Several peaks in these figures, corresponding to MODWT coefficients in range ($>$ +0.03[5]), contain CG spacing patterns present near the genes, CpG islands and introns with CG separated by 8, 24, 25 and 26 bp, (understood from Fourier analysis). This emphasizes, as discussed in the previous subsection, that the long non-coding or retrotransposon regions, (with possible CG spacing at 24 to 26 bp), span or contain within themselves intermediate regions of gene and CG Island regions which are common features in the complex human genome, (dark/light blue and green bars).

Such peaks of large magnitude, relating to contributions from different component regions, in a given DNA sequence help to characterize features of the genome. Hence the coefficients ($>$ +0.03) indicate the presence (and overlap) of multiple CG patterns, associated strongly with the intron regions, (which predominate in the genome), but also reinforced by several genes and by island concentrations. In general however these specific distributions are indicated by peaks i.e. Wavelet coefficients greater than (+0.005). This suggests that wavelet analysis is extremely sensitive to CG distribution change for a given uncharacterized or contiguous sequence. The exploratory analysis, using the wavelet transformation, highlights regions of high CG and its distribution across DNA sequence loci. In particular, such patterns are seen to occur between loci, 20k-30k and 130k-140k

[5]The value of this wavelet coefficient was purely chosen based on the dataset chosen to analyse. On an average contigs from Chromosome 21 displayed this behaviour. The same explanation corresponds to coefficient of value 0.005

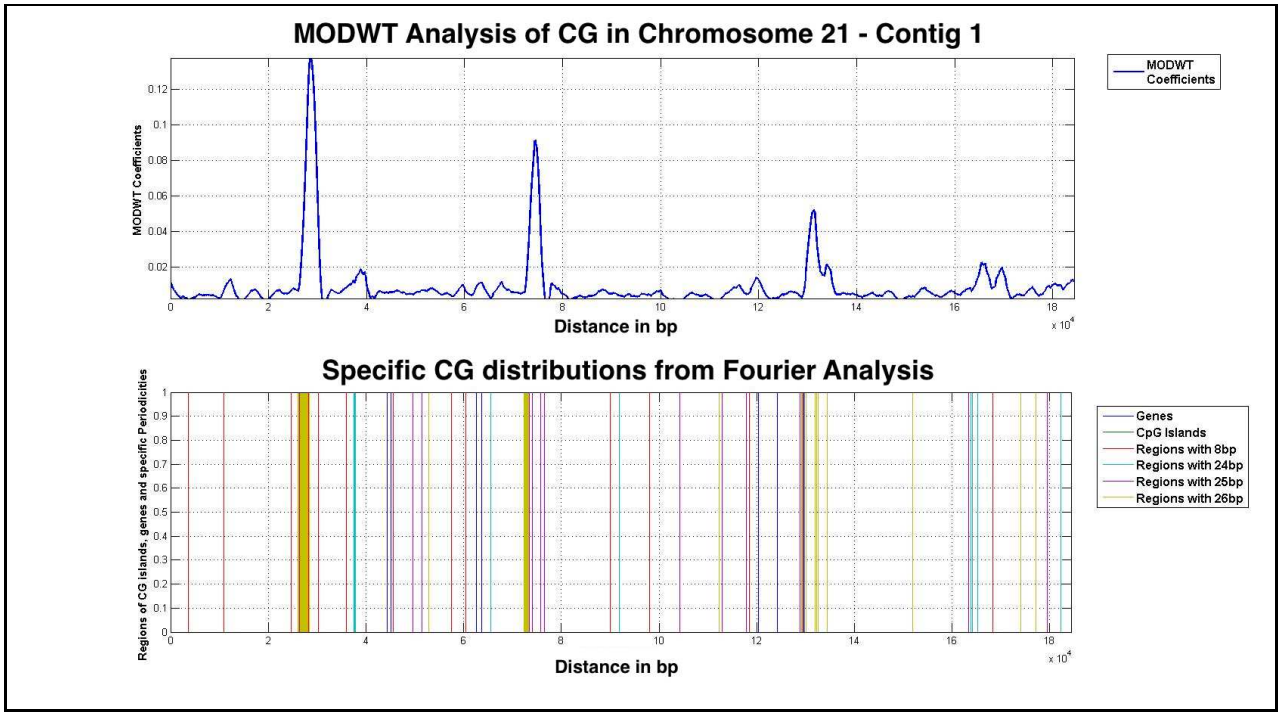


Figure 3.2: CG Wavelet Analysis - Contig 1

High frequency components from MODWT analysis (Distance in bp Vs MODWT coefficients) of Contig-1 with Accession No: NT_113952.1 at scale 11 - (thick blue line the graph 1). All types of vertical bars represent FA results in graph 2. blue lines=genes; dark green=CpG islands; red lines=8 bp interval among CG; light blue lines=24 bp interval among CG;magenta lines=25 bp interval among CG;yellow lines=26 bp interval among CG

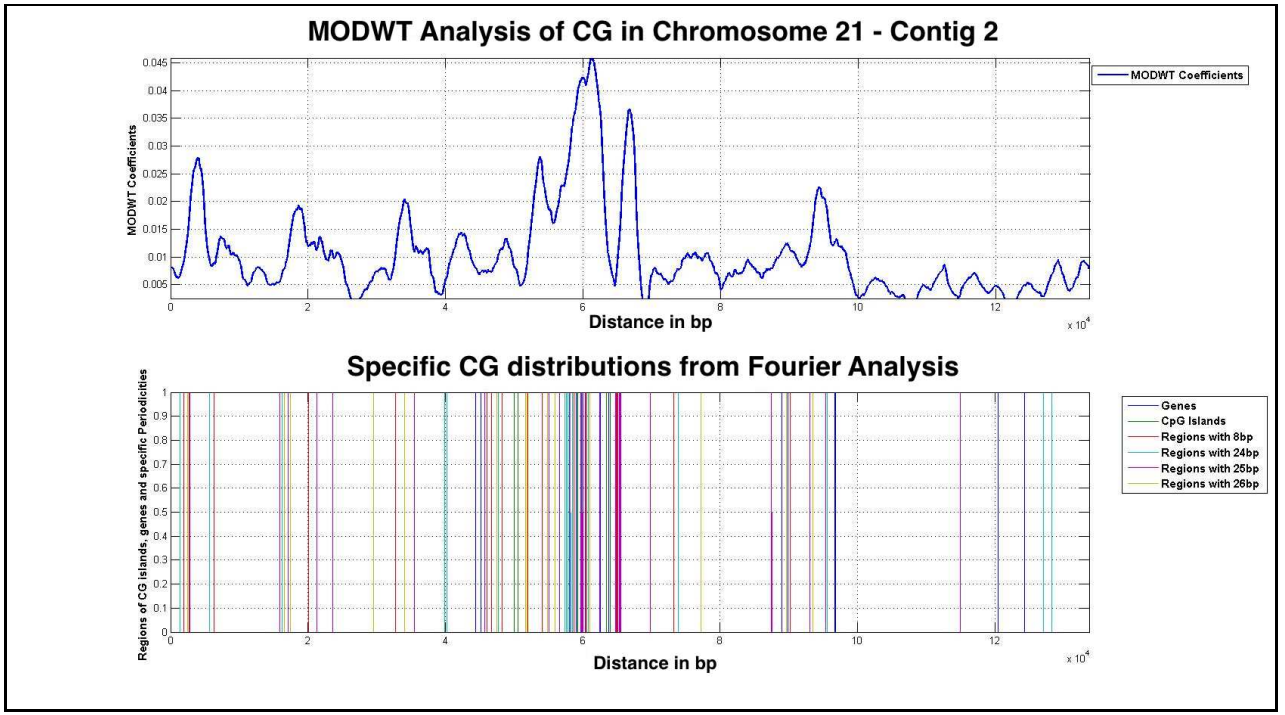


Figure 3.3: CG Wavelet Analysis - Contig 2

Representation of high frequency components from MODWT analysis, (Distance in bp Vs MODWT coefficients) of Contig-2 with Accession No: NT_ NT_113954.1 at scale 11 - Thick blue line in graph 1. All types of vertical bars represent FA results in graph 2. blue lines=genes; dark green=CpG islands; red lines=8 bp interval among CG; light blue lines=24 bp interval among CG;magenta lines=25 bp interval among CG;yellow lines=26 bp interval among CG

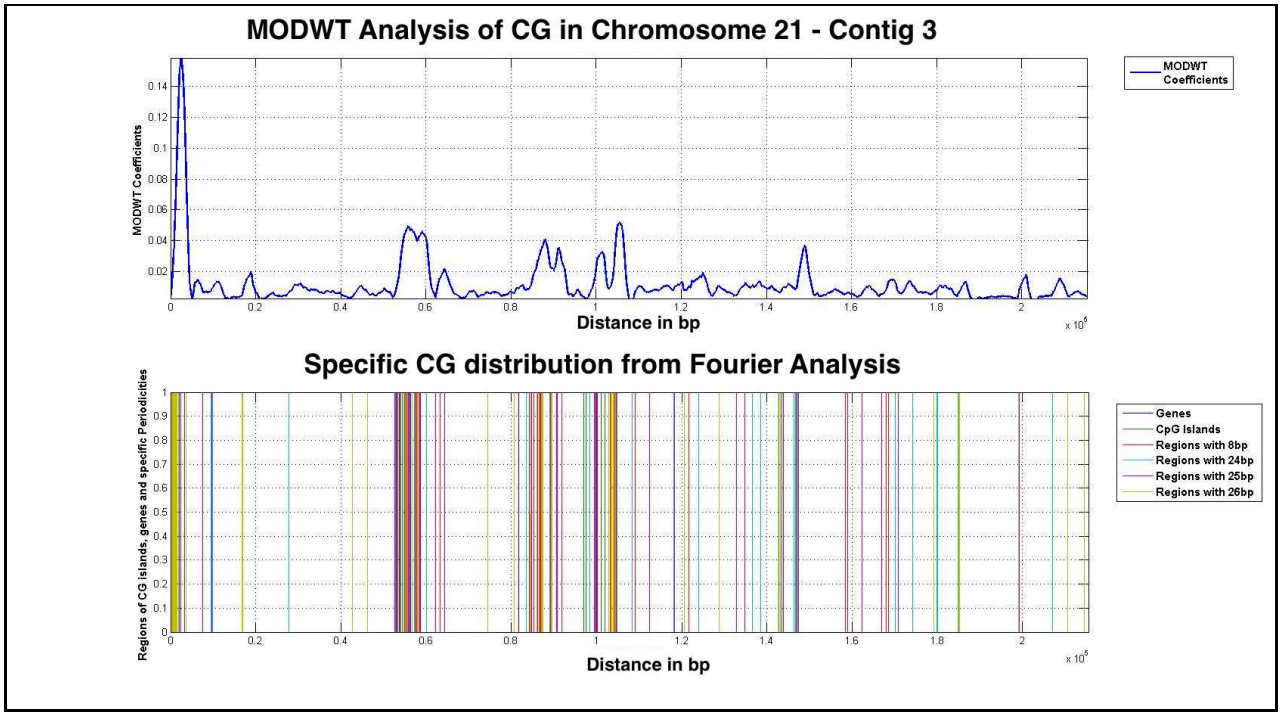


Figure 3.4: CG Wavelet Analysis - Contig 3

Representation of high frequency components from MODWT analysis, (Distance in bp Vs MODWT coefficients) of Contig-3 with Accession No: NT_ NT_113958.2 at scale 11 - (Thick blue line in graph 1). All types of vertical bars represent FA results in graph 2. blue lines=genes; dark green=CpG islands; red lines=8 bp interval among CG; light blue lines=24 bp interval among CG;magenta lines=25 bp interval among CG;yellow lines=26 bp interval among CG;

in Contig 1 (Fig. 3.2), between 55k-65k loci in Contig 2, (Fig. 3.3) and 0-5k, 50k-60k, 80k-90k in Contig 3, (Fig. 3.4). Methods to analyse significance of wavelet coefficients computed for multi-component (and overlapping) data have been reported before for other applications before, [Kwon et al., 2008]. One such method is the calculation of Variance of Wavelet Coefficients, which has been applied further to understand difference in the distributions among CG and other types (GC, AA/TT), and elaborated more in Appendix A.

3.4 Conclusion

Two signal decomposition approaches have been applied to DNA sequences, with a view to investigating patterns of CG dinucleotide in relation to epigenetic mechanisms. Fourier analysis highlighted characteristic frequencies, present in CpG islands and in non-coding regions, but provided no location information. Wavelet analysis gave further detail on frequencies, but also on loci along the DNA sequence at which major effects occurred. Periodicities, (content repeats in bp) were related to known biological features in some instances, (notably bp spacing) corresponding to codons, to *de novo* methylation and others, but also to other unconfirmed potential methylation mechanisms at 24-26 bp spacing. The concentration of CG dinucleotide patterns at specific locations was also reported, although discriminating between contributions from gene, intron and non-coding regions using MODWT was not possible here. Improved understanding of how pattern components, and their sequence locations affect methylation may help us understand how epigenetic changes are stimulated.

This chapter concludes with discussion on the significance of DNA sequence patterns. In the next chapter we consider the second set of major epigenetic changes namely histone modifications and describe attempts to model these by reporting the development of a micromodel called "EpiGMP". This model is also considered as a tool, which predicts the type of modifications asso-

ciated with histone nucleoprotein during various levels of DNA methylation. Fourier analysis was implemented using Matlab, and wavelet transformation using Matlab MODWT package, (details for downloading MODWT package is given in Appendix D).

Chapter 4

EpiGMP - Histone Modification Tool

4.1 Introduction

The aim of this chapter is to build a computational model based on inter-relations between key epigenetic elements. This inevitably required considerable simplification in order to abstract and incorporate individual events and the inter-dependencies between them. Ideally, we wanted to demonstrate that the model formulated was capable of reproducing known histone modifications for given initial DNA methylation values as well as exploring possible new modifications. In the long term this type of investigation would enable the establishment of a comprehensive framework to investigate different epigenetic changes and in the end influence the phenotype of an organism. The initial model rationale and formulation is described here and its various improvements in the next chapter. It is to be noted that much of this chapter has been discussed in [Raghavan, Ruskin, Perrin, et al., 2010] with further details presented here about model features.

4.2 Methods

4.2.1 Conceptualization

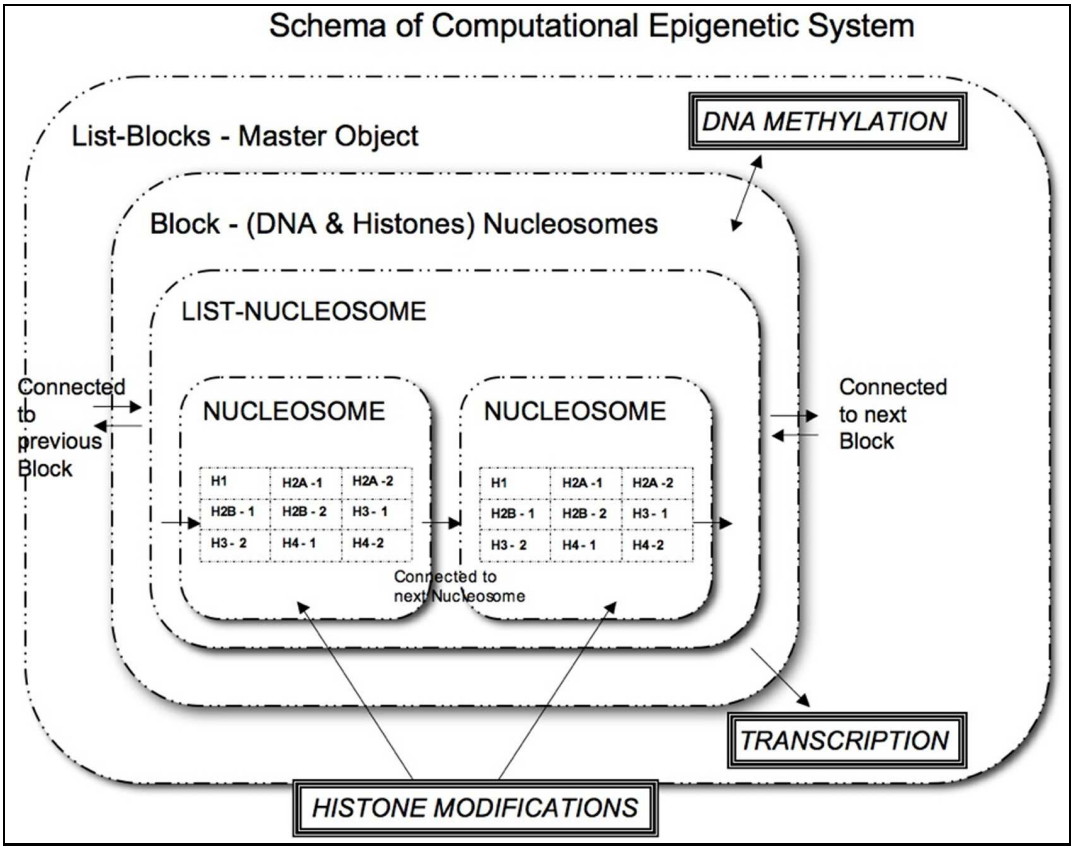


Figure 4.1: Computational Epigenetic Micromodel - Schema

Structure and layers of our computational model closely represent known epigenetic mechanisms. The master object is `List-Block` that generates a `Block` of genes, and contained inside each `Block` are the DNA and `histone` objects forming

textttNucleosome unit. There are 8 histone objects - pairs of H2A, H2B, H3 and H4 along with one H1 object. Each of the histone objects are updated with the modifications over each time-step during the simulation. Hence during each time step, the model is aware of the histone modifications and DNA methylation which defines the system evolution.

In our model, each object represents a natural entity (such as a histone, nucleosome or gene) in the epigenetic layer. Consequently, the model execution starts with a master object that generates a chain of gene Blocks. Each gene `Block` has access to its own set of DNA sequences and `histone` objects (forming a Nucleosome Unit). With respect to `histone` objects, each has a set of tables, updated constantly in terms of the chemical modifications that appear after each time step. Although the objects provide a good mimic of the natural system construction, the major focus is on simulating epigenetic events. This is done by allowing the model to move between possible histone states, (containing one combination of possible chemical modifications at a time) over several time-steps, (explained in detail in the next section) using a stochastic approach. This method as a result is used to define the interdependencies between histone modifications, DNA methylation and transcription progress as closely to natural system as possible.

4.2.2 Evolution of Histone Modifications

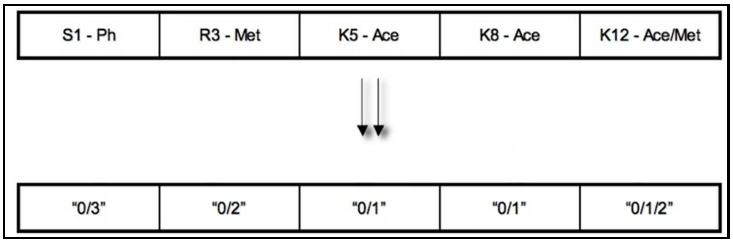


Figure 4.2: General representation of histone *states*
The number of modifiable amino acids chosen for each histone type differs. S, R and K refer to Serine, Arginine and Lysine amino acids respectively. In general, each modification is encoded as a number - Acetylation, (Ace) as "1", Methylation, (Met) as "2", Phosphorylation, (Ph) as "3" and no modifications as "0". The string of numbers or the histone *state*, represents the possible combination of modifications within that particular histone type. (H4 type is shown here).

To observe how modifications are handled dynamically, in nature, information (extracted from the literature [Kouzarides, 2007]) on the number and type of amino acids for each histone type is

fed into the model before the simulation. So when a given type of modification occurs during a particular time step, the corresponding table is updated within that histone. This encoded information is used to define the intrinsic interdependencies of histone modifications, i.e. (how these affect and are affected by the level of DNA methylation), and their combined effect on the output parameter "Transcription".

Histone Modification Data – Collection and Representation

As a part of collating information in order to study the complex interdependency among several histone modifications and DNA methylation, the first step involved tabulating key epigenetic modifications that triggered apparent cellular events. As previously mentioned, *a priori* knowledge about the general behaviour of histone modifications were the main focus at this stage. An interesting fact to be noted is that the literature contains a lot of information about the role of modifications associated most frequently with H3 and H4 types among all histones, [Cedar et al., 2009; Jenuwein et al., 2001; Kouzarides, 2007]. For example, a review on translating the histone code was published by the authors of [Jenuwein et al., 2001], in 2001, gives a clear definition on how to understand the influence of all Lysine amino acid acetylations and methylations. The article elaborates on the antagonistic nature of Lysine 9 or K9 methylation in suppressing gene expression and that of Lysine or K4 methylation during transcription. Another informal yet very detailed discussion on the role of histone modifications including acetylation, methylation and phosphorylation was given in a book written by the author of [Turner, 2001]. A few years later, in 2007, more definitive roles of the same set of modifications and other types were reported, [Kouzarides, 2007]. Here the significance of "cross-talk information" among different modifications was described. The presence of more than one type and how they interact to control the local chromatin structure were explained in detail. Few among many examples include presence of H3K4 methylation, H3S10 phosphorylation, H4K5 acetylation amongst others during transcription, the negative effect of H3S10 on H3K9 methylation

and ubiquitinilation of H2B amino acids before H3K4 trimethylation, [Kouzarides, 2007]. Also recently, the role and influence of DNA methylation on histone "non-acetylation" modifications were reviewed, [Cedar et al., 2009]. Here the authors related the formation of long-term repression of genes and X-chromosome inactivation through co-ordinated interactions between DNA methylation and histone modifications such as H3K27 and H3K9 methylation.

As a consequence of recording the presence of diverse sets of modifications during different genetic events, a tabulation of possible modifications in the 4 types of histones was carried out. This table numbered as 4.1 in this chapter, gives the details on the number of amino acids chosen per histone types and the corresponding types of modifications associated with them. Further information on amino acid positions (relative to their order in the histone peptide sequence), and the possible combinations of modification states in each histone type are also provided. This information is encoded as *numerical strings* and stored in the model during each simulation as a set of nodes, which must be visited or chosen (i.e. possible combinations of histone modifications that are preferred in nature). To represent the amino acid modifications in histones more conveniently, each histone has numerical strings of specific length associated with the possible number of modifiable amino acids. Each of the chemical modifications is encoded as a number, (Acetylation as "1", Methylation as "2", Phosphorylation as "3" and no modifications as "0") and appears in specific positions of the string to mark the choice of modification preferred by the model.

In the case of the H4 type histone (as shown in Figure 4.2), an example of H4 histone "state" has only 5 amino acids and each amino acid has a particular modification associated with it. If the current combination of modifications or (*histone state*) is "3-0-0-0-0" it can be interpreted as, the first amino acid (S1) is phosphorylated and the other amino acids (Table 4.1) are not modified. This process generates many combinations of the possible states in each histone type.

S.No.	Histone Type	No. of Amino acids	Amino Acid & Position	Corresponding Modification	No. of States
1.	H1	0	-	-	-
2.	H2A	4	S1-R3-K5-K9	Ph-Met-Ace-Ace	16
3.	H2B	10	K5-S10-K11-K12-S14-K15-K16-K20-K23-K24	Ace/Met-Ph-Ace-Ace-Ph-Ace-Ace-Ace-Met-Ace	1536
4.	*H3	6	R2-T3-K4-R8-K9-S10-T11-K14-R17-K18-T22-K23-R26-K27-S28-T32-K36-K37	Met-Ph-Met-Met-Ace/Met-Ph-Ph-Ace/Met-Met-Ace/Met-Ph-Ace/Met-Met-Ace/Met-Ph-Ph-Ace/Met-Met	6300
5.	H4	5	S1-R3-K5-K8-K12	Ph-Met-Ace-Ace-Ace/Met	48

Table 4.1: Amino Acid Positions & Modifications

Details of specific amino acids and their corresponding modifications in all histone types. * - H3 has a special type of representation based on amino acid type and the corresponding modification. K - Lysine, S - Serine, T - Threonine, R - Arginine, Ace - Acetylation, Met - Methylation, Ph - Phosphorylation

Application of Markov Chain Monte Carlo (MCMC)

A stochastic process (also referred to as a random process), contains a set of random values, which allow the next state of the system to be selected. A Markov chain is a mathematical model for stochastic process governed by a "transition probability" in order to choose the successive states of the system, [Doubleday et al., 2011]. The choice of developing any stochastic process to solve a problem arises from the fact more than one solution or state of existence is possible for a system. A Markov process can be defined to have no memory, i.e. the next state of the system depends on only the preceding state. In simple words, a stochastic process with *markov* property randomly traverses through a target set or distribution without any memory of previous solutions, to choose the next state based on the current state using a value of transition probability, [Doubleday et al., 2011].

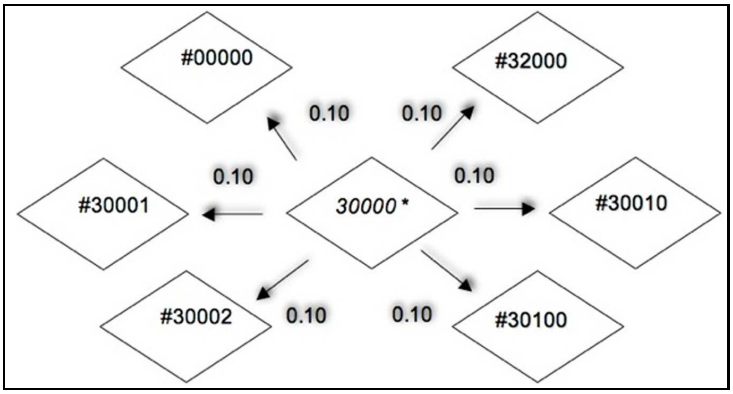


Figure 4.3: Probability of shift among histone states
Only one change is possible at each time step hence each histone state can potentially shift to only one of its specific neighbours. * = Current state. # - neighboring state. Probabilities of shift ($\in$ [0,1]) can be given by the user initially.

Mathematically, a sequence of random variables X^0, X^1, ... belonging to χ, (represents collection of random variables), has markov property or is a markov chain if,

$$Pr(X^{t+1} = x | X^1 = x^1, ..., X^t = x^t) = Pr(X^{t+1} = x | X^t = x^t) \tag{4.1}$$

for all t = 1,2, ... and $x \in \chi$.

The Markov Chain Monte Carlo technique originated from the field of Statistical Physics in order to study the movement of particles and henceforth has been repeatedly applied in the fields of Finance, Image Analysis, Bioinformatics amongst others, [Doubleday et al., 2011; Miklós, 2003; Tu et al., 2002]. The Monte Carlo integration is used to approximate the expectation or prediction pattern obtained by applying Markov chain model. To elaborate more, when Markov Chain Monte Carlo is applied, a system is allowed to evolve and choose among the possible states depending on the transition probability over a repeated number of times. The repetition allows average behaviour

of the system over a period of time to be understood. MCMC is one of the most successful types of Bayesian computing until today. The other derivations of the Bayesian computing techniques are Metropolis-Hastings Algorithms and Gibbs Sampler, [Dubes et al., 1989], which are not dealt in detail in this thesis. In our case, the representation of several histone modifications gave rise to 4 distinct libraries of histone modifications. Among the types, H2A has the smallest library with 16 possible states, followed by H4 with 48, H2B with 1536 and H3 with 6300 histone states as mentioned in Table 4.1. In this EpiGMP tool, MCMC is applied to randomly choose among possible histone states, which in turn depends on probabilities of transition, (when states are chosen).

Each time-step or *iteration* of the model corresponds to the addition or removal of a modification group from the possible combination of histone states. Equivalently, this step resembles the action of enzymes that are involved in chemical modification of histone proteins. In the computational model, only one change or modification is made during every iteration when the model moves between the possible histone states, based on probability of transitions/shifts. The potential shift to a "neighbouring state" from the current histone state is calculated during each iteration of the model. Probabilities of shift also provide a window of control to introduce stress into the system so as to see how the output parameters and the modifications fluctuate over several time-steps. When there is a shift between states, based on the given probability, the corresponding modification graph in each histone type is updated with the changes. In this way, the model can keep track of the dynamic changes easily and use these to describe the resulting output parameters. Our model can also handle multiple additions of the same modification in an amino acid (Mono/di/tri acetylation, methylation or phosphorylation [Kouzarides, 2007]). Although this is invisible to the user, it is taken into account during calculation of global modification levels in each `Nucleosome`. The actual transition that occurs between possible histone states is decided *randomly*, unless the user wishes to input a revised probability distribution (i.e. based on known or desired experiments). This random function, which decides the next state, is based on a uniform distribution, and returns

the index of next random state chosen. The algorithm assumes each histone state to be a node and its probability of shift to a neighbouring state as an edge, (Figure 4.3)

Dijkstra's Algorithm

If further inputs by the user are necessary during the model run, the probable path to achieve the user-desired histone state for a specific time step is calculated by Dijkstra's algorithm [Dijkstra, 1959]. Dijkstra's algorithm was founded by a Dutch computer scientist called Edsger Dijkstra in 1959, [Dijkstra, 1959]. The main idea here is to find the shortest path that connects two distinct nodes in a graph, G, made of vertices and nodes, $([V,E] \in G)$. Dijkstra's algorithm has been applied widely to study GIS networks for water flow and transport, public transportation development and perturbed control systems, [Djokic et al., 1993; Lossow, 2007; Schulz et al., 2000]. In our case, the algorithm is allowed to choose the shortest path from the current state to the choice specified by the user. Since the model is limited by allowing only one change at a time, the user-specified histone state is reached within a minimum number of iterations.

H3 Modification

The possible number of amino acid modifications for H3 histone obtained from the literature was prohibitively large. In consequence these are stored in a different manner to permit compression. A one dimensional array of size *six*, based on the importance of six specific types of modifications and their corresponding amino acids is considered. For example, given a coded representation "4-0-0-0-0-0" the first position corresponds to all Arginines that could be methylated (see Table 4.2). This allows the system to choose and modify any arginine from its population. i.e. this could be one among R2/R8/R17/R26 or all of them together. A value "V" from the closed range [0,4] is chosen randomly to show the number of Arginines modified (based on a uniform distribution random function that returns a random amino acid and the number to be modified in each array

position).

Position in Compressed Array	No. of Amino acids	Corresponding Modification	Amino Acids in H3 Modified
1.	4	Methylation	R2,R8,R17,R26
2.	4	Phosphorylation	T3,T11,T22,T32
3.	2	Phosphorylation	S10,S28
4.	2	Methylation	K4,K37
5.	6	Methylation	K9, K14, K18, K23, K27, K36
6.	6	Acetylation	K9, K14, K18, K23, K27, K36

Table 4.2: Compression of H3 type histone

Details on the Compression of H3 histone states. Content of amino acids is classified based on the type of amino acid type and modification that applies, into 6 groups. K - Lysine, S - Serine, T - Threonine, R - Arginine

In the H3 array,while the first position corresponds to methylation of any or all arginines ($V \in [0,4]$), the second position to threonine phosphorylation ($V \in [0,4]$), the third to Serine phosphorylations ($V \in [0,4]$), [Cedar et al., 2009; Kouzarides, 2007; Turner, 2001] and the fourth position in particular corresponds to lysine methylation ($V \in [0,2]$) that could relate to high transcription. In contrast, the fifth position relates to lysine methylation ($V \in [0,6]$) that potentially encourages more DNA methylation and position six ($V \in [0,6]$) represents acetylation modifications that appear during transcription. As a side effect of compression, the user cannot choose any specific amino acid (such as R2 or R8 etc) to be modified during the iterations since the model deliberately only permits random choice. The details on grouping of amino acids and the compression is given in Table 4.2.

4.2.3 Epigenetic Interdependency

The system has a simple yet strong and well-defined inter-dependency amongst evolution of histone states, transcription rate and level of DNA methylation inside each `Block`. There are 3 main interactions in our model.

a. Histone Modifications $\longrightarrow$ Transcription.

Equations (4.2), (4.3) and (4.4) define how transcription (T) is calculated after each time interval in our model. The user can set the time-interval to 1, 5, 25 or 100 time-steps, since the biological rate of change of all types of modifications is elusive. Here, the transcription variable is affected by the number of modifications in all nucleosomes in a `Block`. The choice of an exponential function in any application is based on expressing an output that depends on variables that are continuously changing. In our case, the histone modifications are dynamic and used to define the instantaneous state of the model at any time point of the simulation.

$$T_{per\ time\text{-}interval} = P_T * ((\prod_{i=1}^{m} e^{(2Ace\text{-}1)}) * (\prod_{i=1}^{m} e^{(1\text{-}2Met)})) \tag{4.2}$$

m = No of time-steps set by the user in a time-interval

P_T = Probability of transcription occurring (by default this value is set to 50 % or 0.5 – unbiased)

$$Ace = \sum \text{Average of the no. of Acetylation Modifications in all histone types in “n” Nucleosomes} \tag{4.3}$$

$$Met = \sum \text{Average of the no. of Methylation Modifications in all histone types in “n” Nucleosomes} \tag{4.4}$$

Here the probability of transcription to occur is initially 50% (or 0.50), which is altered subsequently by the histone modifications, hence making the transcription event a function of the modifications within this stochastic model. Also, the system ensures that if a promoter type `Block` has high levels of DNA methylation, transcription is blocked for all the gene Blocks that follow the promoter, again in agreement with the literature [Allis et al., 2007]. This step is implemented so that only the promoter decides transcription of the genes, as occurs in nature [Turner, 2001].

The second interaction is:

b. Histone Modifications $\longleftrightarrow$ DNA Methylation.

As mentioned above, based on information from the literature, the system allows H3 and H4 type histones alone to influence DNA methylation and vice versa, [Cedar et al., 2009; Jenuwein et al., 2001; Kouzarides, 2007].

The influence of DNA methylation on direction of histone evolution is as follows,

(i) Probability Values for histone states containing more acetylation modification –

$$P'_{a\text{-}b} = P_{a\text{-}b}/e^{(2(D-k))} \tag{4.5}$$

(ii) Probability Values for histone states containing more methylation modification –

$$P'_{a\text{-}b} = P_{a\text{-}b} * e^{(2(D-k))} \tag{4.6}$$

k = Mean DNA Methylation Value (set to 50 % or 0.50)

a, b = current and neighbouring histone states (H3 and H4 types) respectively.

P_{a-b} = Initial Probability of shift from state a to state b.

P'_{a-b} = Probability of shift from state a to state b in the successive iteration.

D = DNA Methylation level in initial iteration.

The system maintains probability values within a closed range of 0 and 1 with the help of a *scaling factor* or the mean DNA methylation value. During each time-step, the probability of shift of every histone state (H3 and H4), is altered by the DNA methylation level (as given in equations 4.5 and 4.6). The user must set the initial probability for the first iteration. Conversely, histone states (acetylation and methylation modifications) can be used to express the level of DNA methylation, which is calculated in two ways at the start of each iteration.

1. The user can specify a value between 0 and 1, else

2. Based on the initial histone states chosen, the system calculates the DNA methylation value.

DNA methylation for one `Block` is calculated in a very simple way in the model.

$$D' = D - R * A \tag{4.7}$$

D' = DNA methylation for current iteration.
D = DNA methylation from previous iteration.
R = Random Value ($\in$ [0,1.0]) generated by the system based on uniform distribution.
A = Average of the ratios of the current level to the maximum level possible in methylation and acetylation modification in a `Block`.

The DNA methylation for the first iteration, if not provided by the user, is assumed to be zero. The model randomly initiates stochastic behaviour and similarly incorporates interactions of histone modifications with DNA methylation. Equation 4.7 is implemented within the system (for successive iterations) based on a *conditional probability*. The system generates another random value (between 0 and 1 based on uniform distribution) and if this value is less than 5% of DNA methylation value (from previous iteration), the equation 4.7 is implemented and methylation is

updated. This threshold step is very important since it controls the system evolution and does not allow all modifications to have a uniform effect on DNA methylation. It should be noted that the transcription rate is calculated based on the time-interval set by the user and DNA methylation values are set after each iteration or time-step. The third type of interaction, a consequence of the two mentioned above, is between DNA methylation and transcription. In this last type, the former is inverse to the later and evidence for this is presented in Section 4.3. Hence through these inter-dependencies, we try to mimic the mechanisms in a simple manner that control gene expression.

4.2.4 Model Simulation

The steps given below explain the simulation in a simple and concise manner.

1. *Read and Store Inputs*

   (a) Histone Data -The possible combinations of histone modification as described above are read and stored in the model. These include *string of histone states* and the *probabilities of shift* between the states. (The possible types of modifications are given in Tables 4.1 and 4.2)

   (b) User Selected Values are provided –

      i. Default Parameters: Number of Blocks and number of nucleosomes per `Block`. The model also requires the total number of iterations, (or time-steps) and time-intervals after which output must be reported. (Figure 4.1), for 1 run of the simulation.

      ii. Optional Parameters: DNA methylation and histone states preferred by the user (in which `Block`, `Nucleosome` and at what iteration/time-step)

2. *Create Objects*

(a) Create the number of Blocks (promoters/genes/ isolator/Introns/silencer), nucleosomes, nine histone types (default) and modification tables for each histone, as specified by the user.

3. *Simulate*

(a) If the user has not chosen to explore a preferred histone state, start with zero modifications. Based on the DNA methylation value (either mentioned by user or calculated based on those histone states in the current iteration), and the probabilities of shift for each state, a random selection of states is made. Simultaneously the modification tables are updated based on the current state. For example if state 02002 in H4 is chosen, update methylation tables for H4 histone.

(b) For specific time-intervals the transcription rate, (using equations 4.2, 4.3 and 4.4), is recorded and after each time-step the DNA methylation value (based on the modification tables as mentioned above or by taking the value specified by the user in a desired time-step) is calculated. Also, the probabilities of shift based on the DNA value from previous time-step, (using Equations, 4.5 and 4.6), is subsequently altered.

(c) Simulation is continued till the maximum value of iteration is reached.

(d) The pseudo code of the algorithm is as follows,

Algorithm 4.1 Algorithm of the EpiGMP tool - Version 1

```
1: for Block a=1:max do                                   ▷ no. of Blocks (max=64*4 here)
2:     for Iteration b=1:max do                           ▷ max=5000 iterations or time-steps
3:         for Nucleosome d=1:max do                      ▷ max=no of nucleosomes per Block
4:             for Histone Type j=1:8 do                  ▷ 8=default no of histones per textttNucleosome
5:                 if User has specified a choice then
6:                     DESIRED HISTONE STATE SET
7:                 else
8:                     RANDOM HISTONE STATE SELECTION
9:                 end if
10:            end for
11:        end for
12:        CALCULATE DNA METHYLATION
13:        if a==user mentioned time-step then
14:            CALCULATE TRANSCRIPTION
15:        end if
16:    end for
17: end for
```

4. *Store Outputs*

   (a) Results for the specified time interval, inside each `Block` are –

      i. Transcription rate

      ii. DNA methylation level

      iii. Global modification levels for each `Block` (Methylation, Phosphorylation and Acetylation)

      iv. Count of the number of times each state is visited in all 8 histones for each `Nucleosome`.

4.2.5 Simplifications and Model Restrictions

As the major focus is on histone evolution, a few simplifications are made here to test the system reliability.

1. The model currently handles only three modifications i.e. acetylation, methylation and phosphorylation as their biological significance is known from the literature [Kouzarides, 2007].

2. Although our model can handle several hundred nucleosomes per `Block` (as in reality), we illustrate with only one `Nucleosome` per `Block` to track and analyse the evolution of histones over several time-steps. This ensures that the evolution of modifications in a single `Nucleosome` is clearly identified and changes are demonstrated. The `Nucleosome` number will be increased for further investigations to improve realism, (implemented in Chapter 5).

3. The system is initialised with "Zero" modifications, which are slowly increased over several iterations. Intervention by the user to permit input of desired histone states in any of the time steps is implemented but not tested in the results here.

4. DNA methylation during the first iteration for the promoter `Block` was specified by the user (such as high and low conditions of DNA methylation values as specified in results section). For subsequent iterations of the simulation, the system evolution determines the values.

5. The model is set to remove or add only **one** value in a time-step. It is to be noted that this is the only disadvantage of using EpiGMP because allowing one change slows the system dynamics but helps to record the number of modifications most efficiently.

4.3 Results and Discussion

In order to investigate the system behaviour, a set up of 64 objects of `Block` type with a single `Nucleosome` per `Block` was implemented and evolution of histone states was observed over 5000 iterations. Sixteen promoters, each of which controlled 3 subsequent genes, were added within the chain, to form 64 `Blocks`[1], (refer Figure 4.1 for the multi-layers of objects in EpiGMP).

[1] the 64 `Blocks` here refer to a set of synthetic data and not any real information on genes or promoter

Histone states, transcription progression, DNA methylation and global histone modification levels for every `Block` were recorded every 25 iterations.

4.3.1 Transcription Progression

The third type of interaction in our model between DNA methylation and transcription is analysed below is,

c. DNA Methylation $\longleftrightarrow$ Transcription.

Given the first two interactions between the two epigenetic key elements as mentioned in methods section, the model is able to efficiently simulate an inverse relation between the components of the third interaction, *Transcription(T)* and *DNA methylation*, (as reported in the literature). Transcription values (T $\in$ [0,1.0], represented in figure 4.4), for increasing DNA methylation levels (specified by the user, in this case) were observed during 3 different simulation runs. Here, an inverse relation between transcription and DNA methylation levels is consistently prominent. A high rate of transcription is observed when the DNA methylation is approximately 0.35 or less, while at the midpoint of range (within intervals $\in$ [0.4,0.7]), transcription decreases sharply. Higher values, ($>$0.75) of DM evidently prevent any increase in transcription. This behaviour is a reflection of the model selection of specific histone modification over several iterations. For DNA methylation in the range 0.47 to 0.7, (hemi-methylation state), the rate of transcription is severely affected. This observation could relate to how transcription is blocked under methylation of the CpG islands within the promoter, (broadly in agreement with the real system as reported in the literature [Turner, 2001]).

4.3.2 Evolution of Histone States

Here, we analyse histone modifications for only two cases, i.e. *high* DNA methylation ($>$ 0.85) and *low* DNA methylation ($<$0.15), solely for the *Promoter* type `Block` as any changes to this

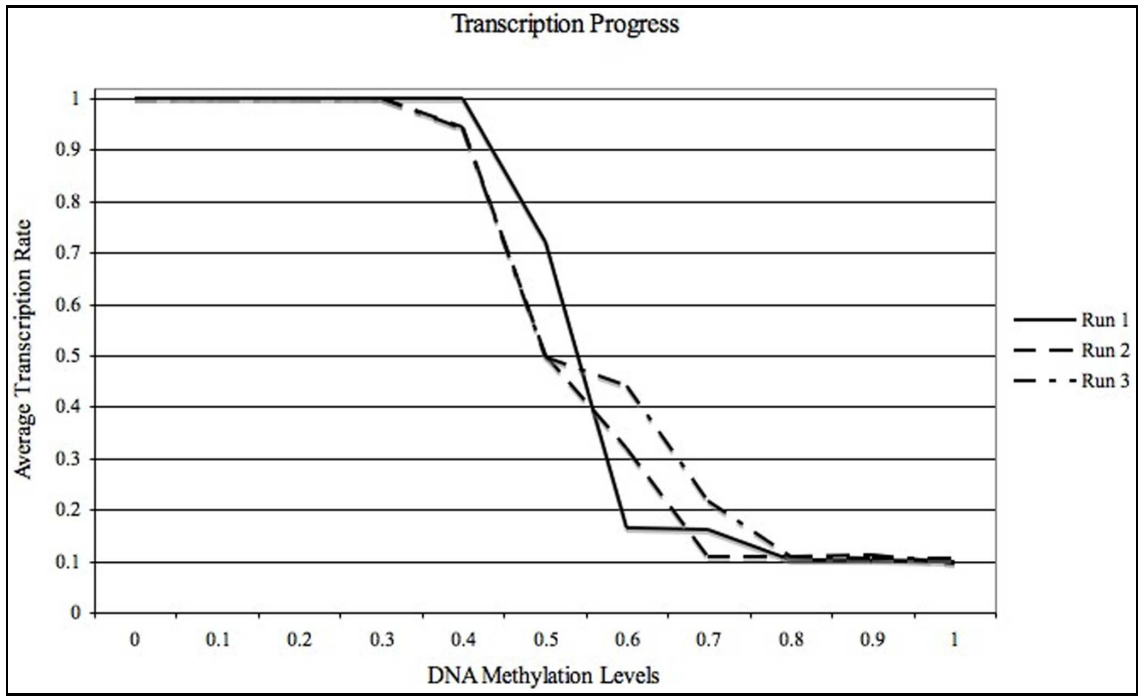


Figure 4.4: Average Transcription Progress

Derivation for 16 promoters over 5000 iterations during 3 different simulation runs. The third type of interaction between transcription rate and DNA methylation level (or percentage) was observed here. Transcription rate (25 time-steps = 1 time-interval) is inversely proportional to DNA methylation level (decided by user in this case, for testing purposes).

`Block` affect the succeeding genes. These conditions are analysed to study biological cases such as those, which apply when an unexpressed oncogene is activated, or when a tumour suppressor gene is inactivated.

4.3.2.1 Case 1: Low DNA Methylation

For small fixed levels of DNA methylation, (refer to Figure 4.4) acetylated histone states are preferentially chosen, which in turn lead to a stable and high transcription rate. These simulations are carried out to show how the system effectively emulates the biological process of transcription of genes for low DNA methylation levels. As discussed above, this version of the EpiGMP tool considers the role of H3 and H4 type histones alone and hence evolution of histone states of those types are discussed in the following subsections, [Jung et al., 2009; Sun et al., 2007; Turner, 2001; Yoo et al., 2006].

H4 Analysis

Figure 4.5 shows the system preference on an average (%) of H4 histone states in 16 promoters for 10 different simulation trials[2]. We tested the consistency and robustness of the system by initially assigning 10 datasets with various "probabilities of shift" for H4 (H4-1 and H4-2) type histone. These probabilities (of a move from the current state to any of its neighbors) were generated randomly by a system-defined function (based on a pseudo random number generator - Mersenne Twister, which is robust, has a large range of period and a high order of dimensional equidistribution [Matsumoto et al., 1998]). Acetylated amino acids states, such as the 11^{th}, 35^{th} and 47^{th} predominated in more than 75% of the datasets i.e. states containing acetylated amino acids such as K5, K8 and K12 (see Table 4.1) were highly visited. Even when the probability for one of the three preferred states was decreased during any test set, the system preferred the other two states

[2]Each trial was initiated with a different set of transition probabilities

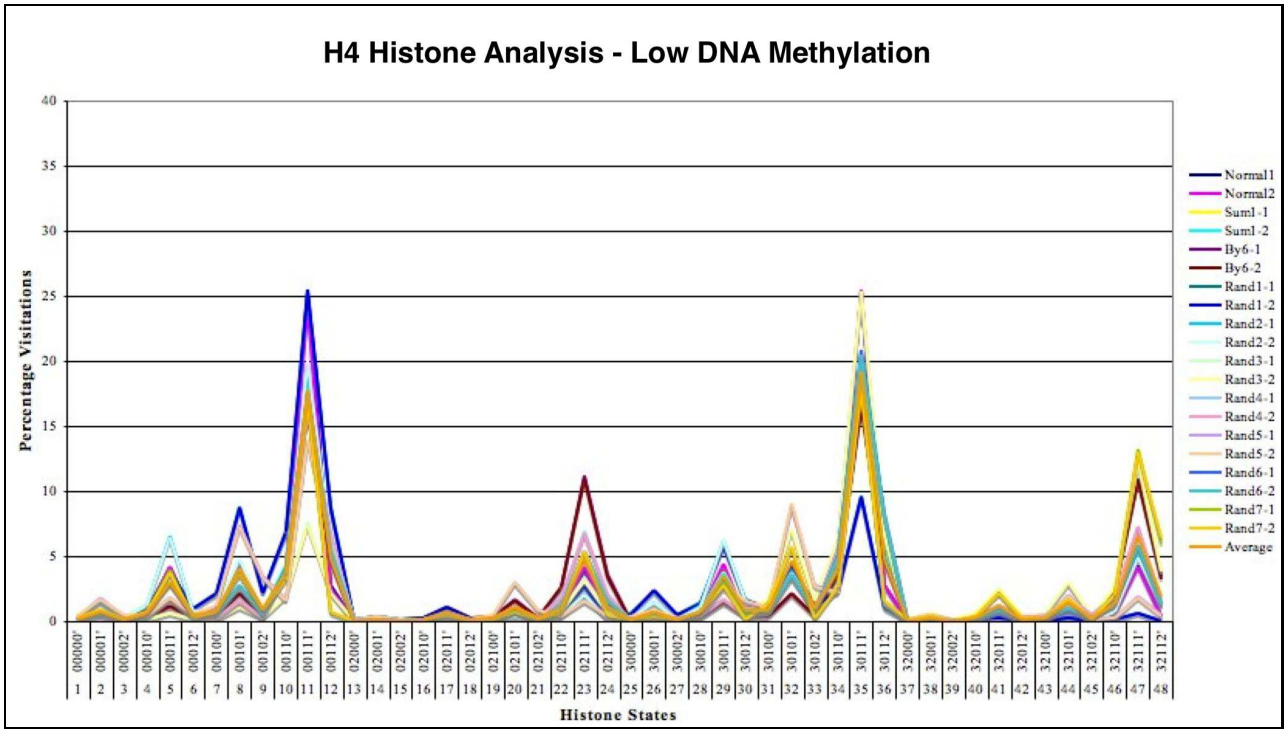


Figure 4.5: H4 Acetylation Analysis

Evolution of H4 (H4-1 and H4-2) histone states in the 16 promoters for 10 different datasets during low DNA methylation levels (<0.15 or 15%). H4-1 and H4-2 histone states were tested with 10 datasets of random probability values assigned, (represented by colors in the graph).

containing lysine acetylation. Such consistent results demonstrate the ability of our model to reproduce the presence of these modifications during transcription, as reported by authors in [Taplick, 1998; L. Zhang, X. Su, et al., 2007], during expression of oncogenes.

H3 Analysis

The depiction and interpretation of H3 results reflect the way these are addressed in the model. In each of the H3 associated figures, (Figures 4.6 and 4.8), a unit on the X-axis represents an expansion of what the linear array of H3 histone stands for (Amino acid, Position in H3 Array, Number, Modification). The Y-axis gives an average percentage of visitations of the states containing the modifications described by each unit in the X-axis. Since the number of H3 histone states generated, even after compression, is the largest among all histone types (refer to Table 4.1), we report and analyse specific and prominent H3 states that are significant based on the published information.

H3 histone states that contain the maximum number of lysine acetylation modifications (refer Table 4.2) such as K6 are only visited during high level of transcription. Hence, we analyse the modifications within those states in particular. These states, contain least or no Lysine methylation, (corresponding to position K5(0-6) in Figure 4.6 and in H3 array- of amino acids K9, K14, K18, K23, K27, K36). Also, in these states, phosphorylation of Serines is higher, (i.e position 3 in H3 array depicts S10 and S28 phosphorylation - refer to Table 4.1), and as substantiated also by the literature, [Sun et al., 2007]. In general, however it was found that that preference given to other amino acids positions (R1, T2 and K4 series) and their corresponding modifications is very similiar. This means that apart from serine, other modifications could be neutral or default modification during transcription.

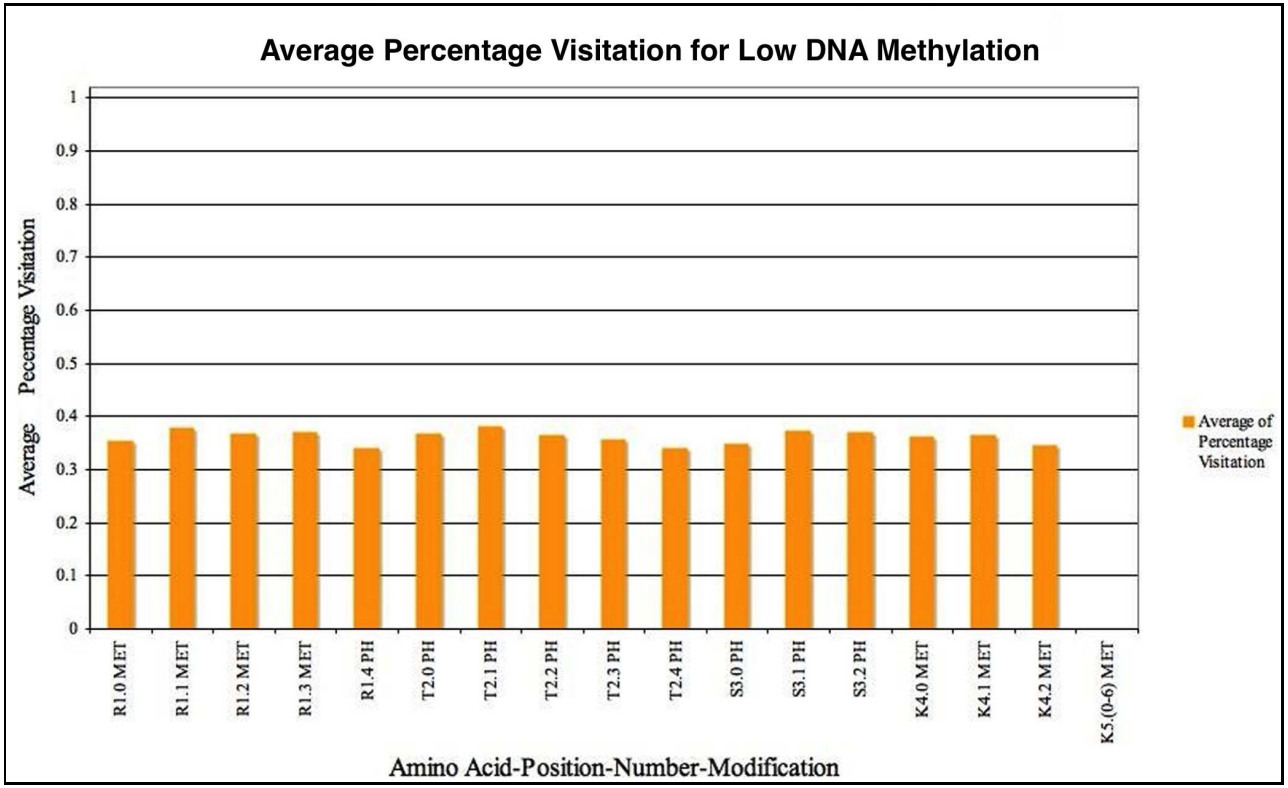


Figure 4.6: H3 Acetylation Analysis

Analysis of Average percentage visits of H3 histone states containing *Lysine acetylation .i.e K6* in 16 promoters after 5000 iterations (for low levels of DNA Methylation (<0.15 or 15%)).States containing Lysine acetylation are visited the most. Hence we analyse the average of percentage visitation of model to all other modifications (except Lysine acetylation - K6 in Figure) during the simulation. Here each unit in the X-axis represents an amino acid–position in H3 array–number of Amino acids–Modification possible. The Y-axis elaborates on the average of percentage visitation of H3 states that contain the modification depicted in each unit of X axis.

4.3.2.2 Case 2: High DNA Methylation

For higher levels of DNA methylation (>0.85, Figure 4.7) during the simulation, the preference is more towards choosing *methylated* histone states. This biased behaviour of the system leads to reduced transcription rate.

H4 Analysis

Figure 4.7 shows the average percentage occupation of H4 type histone states for 16 promoters. The system was again tested with 10 datasets with various probabilities assigned to the histone states in H4 (H4-1 and H4-2). The system was persistently found to occupy methylated amino acids, states such as the 15^{th}, 39^{th} and 45^{th} in more than 8 out of 10 datasets i.e. methylation of K12 was predominantly high. Such strong evidence, (during histone deacetylation and methylation) of modification to a crucial lysine position in H4, is a potential indicator of transcription repression and initiation of DNA methylation. Figure 4.7 hence indicates the possible presence of this modification during real gene repression. Another interesting observation is the appearance of serine phosphorylation (state 39, Figure 4.7 and state 35, Figure 4.5) which show the importance of this specific modification during expression or otherwise. This suggests that the modification could be present from the time that the H4 histone complex was formed, [Barber et al., 2004].

H3 Analysis

Interpretation of Figure 4.8 is similar to Figure 4.6. We analyse specific H3 states so as to aid in comprehension of the results. Figure 4.8 shows the modifications that were preferred during high DNA methylation (>0.85). Only states which contained lysine methylation (amino acid positions such as K9, K14, K18, K23, K27, K36 as in Table 4.2 - position K5) were visited. Hence we analyse the preference of other modifications within H3 states that contain Lysine methylation. Here, in contrast to Figure 4.6, within those states, lysine acetylation was negligible (acetylation

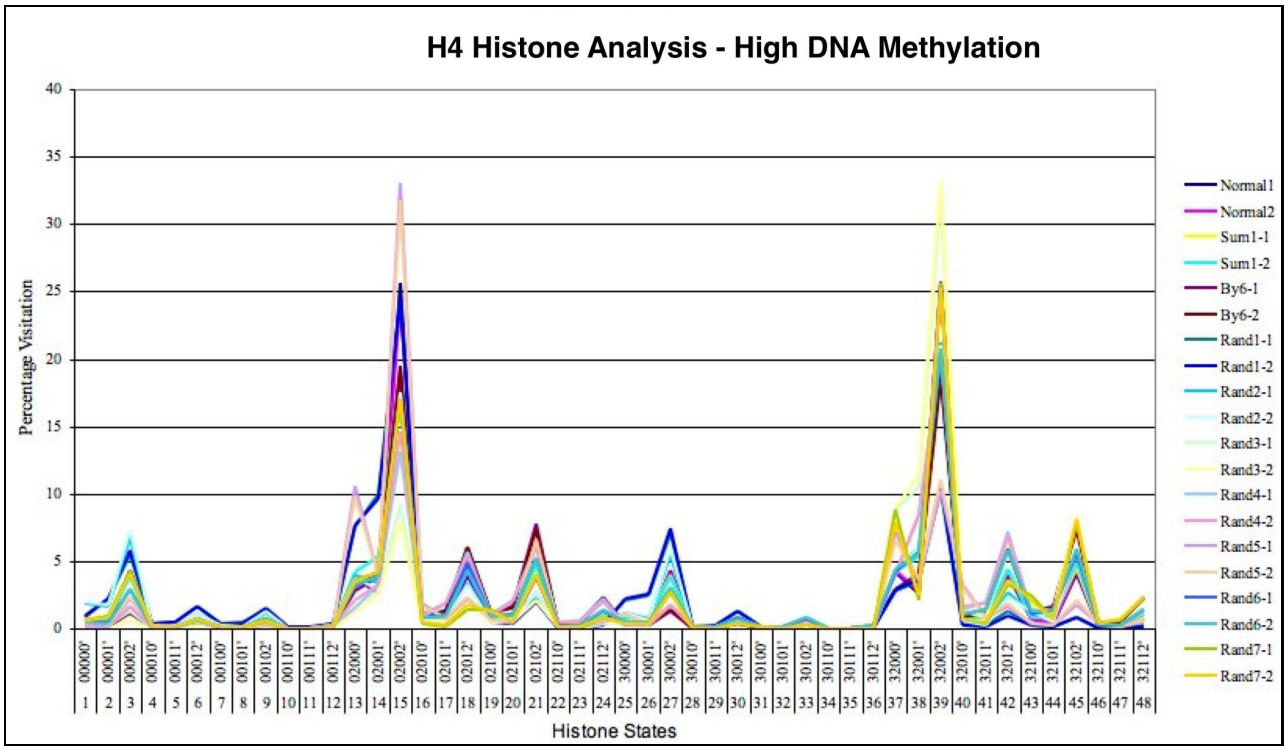


Figure 4.7: H4 Methylation Analysis

Evolution of H4 (H4-1 and H4-2) histone states in the 16 promoters for 10 different datasets during high DNA methylation levels (>0.85 or 85%). H4-1 and H4-2 histone states were tested with 10 dataset of random probability values (represented by colors in the graph).

of K9, K14, K18, K23, K27, K36 - Table 4.2) as these are preferentially methylated.

This is shown by the least number of times the system visited those states that contain Lysine acetylation, (position K6(0-6) in Figure 4.8). Also, recruitment of states containing highly phosphorylated Serine was low. Such observations, on the content of lysine acetylations and serine phosphorylations during high DNA methylation suggest that our model can successfully reproduce results from laboratory studies [Sun et al., 2007] and also indicate presence of other modifications as yet unexplored in the literature.

4.3.2.3 Comparative Study

Figure 4.9 contrasts percentage visitation for H4 histone states under high (>0.85) and low (<0.15) DNA methylation levels. As DNA methylation controls the direction of histone evolution, the states visited for high levels of DNA methylation are not visited for low levels and vice versa.

Standard deviations, shown as error bars, are calculated from the results containing the number of visits for each state. The deviation is high for less visited states and low for highly visited states. This means that the system tolerance to initial selection (determined by random selection using PRNG Mersenne Twister) is also good with specific states consistently chosen over several iterations. This consistency in predicting characteristic histone modifications under defined DNA methylation levels leverages our models capability to mimic the real system to an accurate level. Hence, we expect to obtain similar histone patterns under stable DNA methylation values, for corresponding experimental observations.

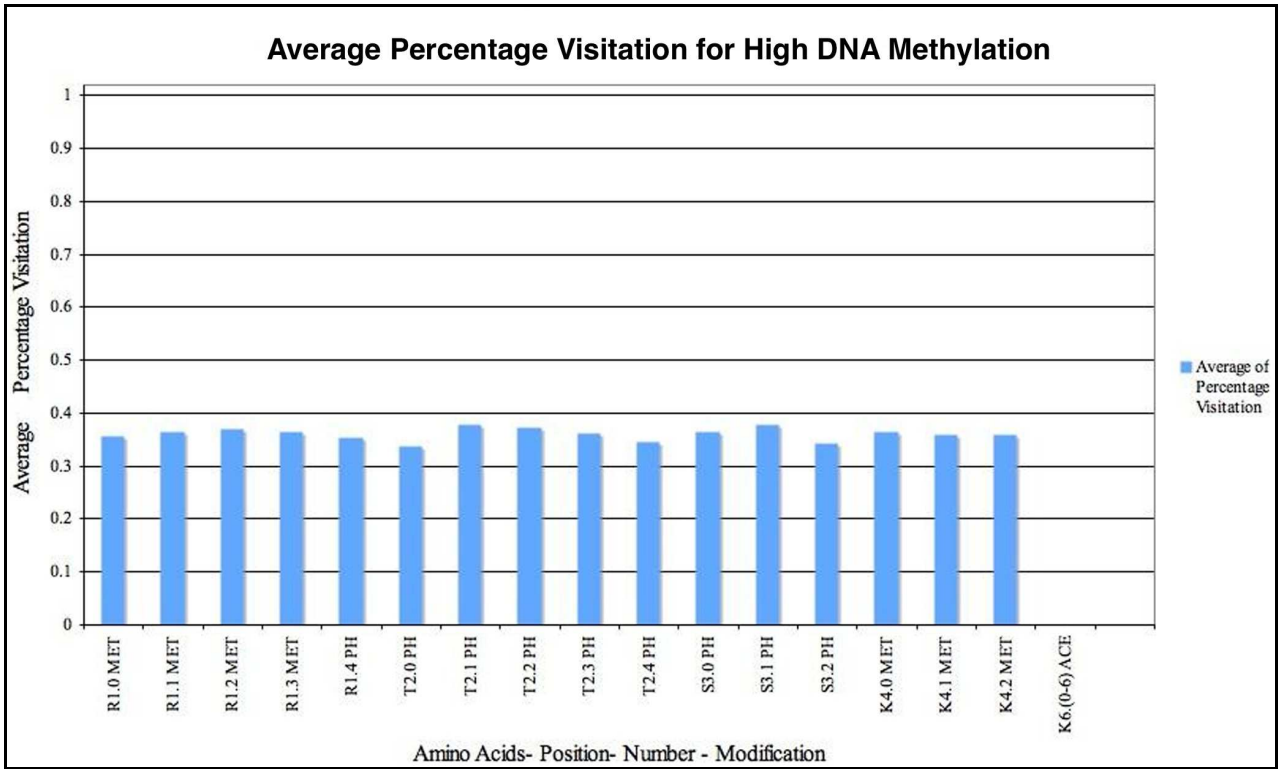


Figure 4.8: H3 Methylation Analysis

Analysis of Average percentage visits of H3 histone states containing *Lysine methylation i.e. K5* in 16 promoters after 5000 iterations (for high levels of DNA methylation (<0.85 or 85%)).States containing Lysine methylation are visited the most. Hence we analyse the average of percentage visitation of model to all other modifications (except Lysine methylation K5) during the simulation. Each abbreviated unit in the X-axis can be expanded to represent (amino acid (short form)–position in H3 array–number of Amino acids changeable–Modification). The Y-axis elaborates on the average of percentage visitation of H3 states that contain the modification given on the X axis.

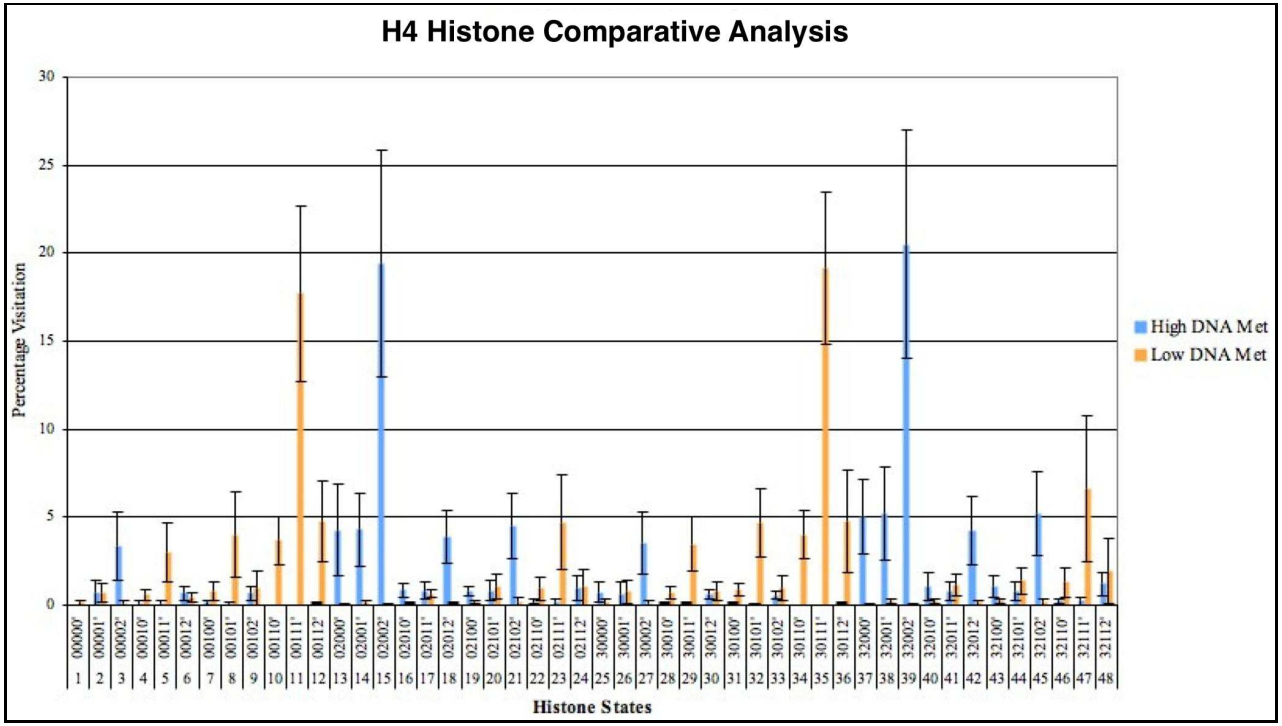


Figure 4.9: H4 States Comparison

A Comparison between the average (of all 20 test results obtained for H4-1 and H4-2) preferences of H4 states for high, (blue bars) and low (yellow bars) DNA methylation Levels. Error bars represent the standard deviation calculated from the total number of visits, for every H4 histone state (occupancy) during the simulation.

4.4 Conclusion

The current version of the model, has been demonstrated to be capable of reproducing known histone modification under stipulated DNA methylation levels, and also report unexplored modifications such as K12 methylation in Figure 4.7. Preference of histone states containing Lysine acetylation during high transcription, and increased number of methylation modifications in H3 and H4 states for higher values of DNA methylation confirms this. Further analysis of the additional modification - (phosphorylation), reveals that for H3 type histone it supports transcription (Serine phosphorylation, in Figure 4.6), while this simultaneously stays neutral in H4 type histone (see Figures 4.5 and 4.7). Such results demonstrate the model capability and its' potential as a tool to simultaneously trace the evolution of modifications in histone states for different histone types and to investigate, how the epigenetic profile is affected overall. While individual results from laboratory experiments in epigenetics and methods to analyse them, have been reported [Barber et al., 2004; Taplick, 1998; L. Zhang, X. Su, et al., 2007], our model is the first of its kind to determine the occurrence of several modifications at one time-step. This provides a basis for further investigations of abnormal conditions such as cancer and other genetic disorders. Extensions to the current framework which include, (i) contribution of H2A and H2B [Barber et al., 2004] modifications, on the model evolution and output parameters, (ii) calculations based on CG patterns in CpG islands and other regions of the human genome, (not just simple DNA methylation values as described here) and omission of model restrictions (described in subsection 4.2.5 in this chapter) are reported and discussed in detail in the next chapter.

Part II

Applications

Chapter 5

Model Improvements and Parallelization Strategies

5.1 Introduction

The focus so far has been to individually model the primary components of Epigenetic mechanisms. In Chapter 2 the basic mechanisms of DNA methylation – background and related factors which influence it were considered. Subsequently, the DNA sequence patterns in human chromosome were analysed to correlate them with the spread of DNA methylation. In the last chapter, a novel stochastic micromodel/tool that could predict the level of histone protein modifications based on simple values of DNA methylation was introduced. In this chapter, the improvements to the EpiGMP micromodel that includes new equations to calculate transcription and information about DNA methylation learnt from nucleotide pattern analysis in DNA sequences are discussed elaborately. A large dataset, (namely all CpG islands throughout chromosome 21), is used to test the combined improved model and this necessitates parallelization of the code which is also discussed.

5.2 Improvements – Epigenetic Interdependency

In the last chapter we defined the mutual influence of DNA methylation values over the level of histone modifications and vice verse. Although the inverse relation between transcription and DNA methylation was established there, the emulation of DNA methylation mechanisms and the effect of those on gene expression require more careful handling for larger iterations, ($>$ 10,000). In this section the improved model structure and refined outputs are presented without changing the dynamics and algorithm of the micromodel. The mathematical relation among key elements is well described in Figure 5.1. Exponential equations were used to define the interdependency among the epigenetic events, (equations 4.2 to 4.7) from previous chapter. The same relations are adopted here except for the calculation of the level of transcription. The improvement in defining transcription calculation, (from Equation 4.2 in Chapter 4) includes allowing DNA methylation levels to also directly affect this output and is given in Equation 5.1.

$$T_{total} = P_T * ((e^{2Ace\text{-}1}) * (e^{1\text{-}2Met}) * (e^{1\text{-}2D})) \tag{5.1}$$

T_{total} = Total level of transcription.
P_T = Probability of transcription occurring (by default this value is set to 50 % or 0.5 – unbiased).
Ace = Average of acetylation modification across all histone types.
Met = Average of methylation modification across all histone types.
D = DNA methylation value $\in$ [0.1,0.9]

Another improvement to be noted is that the calculation of average acetylation and methylation modifications as indicated in Equations (4.4) and (4.3) is altered to accommodate contributions of H2A and H2B types as well. The main idea behind implementing this step is to introduce and test performance by allowing other histone types to contribute to the evolution of histone states in the computational micromodel. In the following sections, the model reports histone modifications for

methylation levels in CpG islands that contain sequences with specific CG dinucleotide patterns extracted from the human genome, (based on the analysis explained in Chapter 3).

5.3 Application of EpiGMP to study CpG Islands

The two main model components, (DM and HM[1]) are key to gene expression and the framework here is designed to address and simulate the natural behaviour of these constituent elements. The overall model structure connecting these two components has been split into two implicit modules in this chapter - *EP1* and *EP2* (epigenetic modules 1 and 2). EP1, made of EpiGMP tool explains modelling histone modifications, which are concisely represented as active "nodes", and randomly sampled/visited. This sampling, which is influenced by DM levels accounts for the stochastic selection of all dynamic histone modifications. This type of modelling, used as a prediction tool differs from a deterministic approach, (for example, a differential equation system [Nelson et al., 2002]), in providing more than one possible solution, for specific attributes or values. Results from previous analyses stated that unique CG patterns were associated with specific levels of DM across the genome [Glass, Fazzari, et al., 2004; Raghavan, Ruskin, and Perrin, 2011]. The direct correlation between DNA sequences and information about associated methylation levels was combined in a simple manner. Hence as consequence, in EP2, all CpG islands from human Chromosome 21[2] were analysed using Fourier Transformations to check for inherited and stable CG patterns that decide methylation levels. Figure 2.3 in Chapter 2 gives a concise representation of these patterns. The islands are categorized into 'groups' with different DNA pattern sequences, following the Fourier analysis. For example, group 3 in the model consisted of CpG islands which contained a maximum number of CG dinucleotides separated an interval of 3bp, (thus generating 31 distinct groups of islands). These groups were separately checked for histone protein modifications on application

[1] DNA methylation and histone modification

[2] Chromosome 21 was chosen because it is the smallest chromosome in the human genome

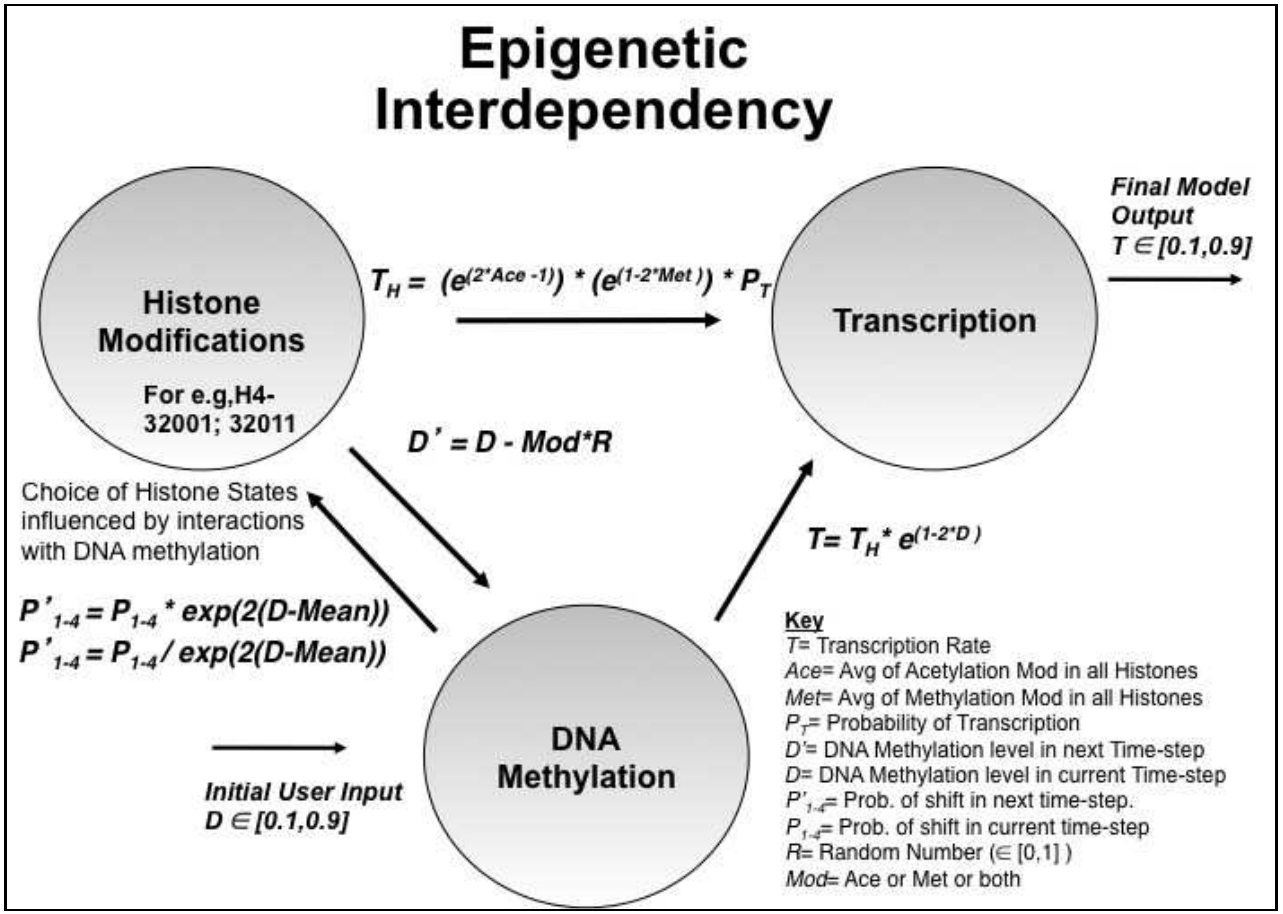


Figure 5.1: Epigenetic Interdependency

Interactions between Epigenetic Elements in the Complex System. DM, associated with CG patterns in the DNA sequences and HM influence each other over each time step. *Transcription,* the output based on both parameters is calculated at regular intervals. Initial input to EpiGMP micromodel is a value of DNA methylation.

of the micromodel. The stochastic model is thus designed to explore how the dynamics associated with histone modifications can control DNA methylation, as well as utilizing additional sequence information to that end. Real system analogies may be found in cancer-associated conditions, (such as silencing or activation of CpG islands located near tumour suppressor and oncogenes), [Esteller, 2007]. In this version of the model, the number of nucleosomes totally depends on the length of the DNA sequence, unlike the previous model version. Provision for deciding a specific histone modification before the start of simulation has also been implemented in order to add more features that improve the overall efficiency. Conversely, the model only handles 3 modifications as mentioned in the previous chapter because acetylation, methylation and phosphorylation are directly involved in gene expression, unlike other types whose role is still under investigation.

5.4 Parallelization

Parallel computing is a form of simultaneous computation that is used to handle large problems, (in terms of high complexity/data or both), using more than one processor [Grama et al., 2003]. Here, the data-driven nature of the model makes parallelization a natural choice, with load distributed across processors. Message Passing Interface or MPI is one of the well known and powerful libraries designed to initiate and maintain communications between processors or nodes, [Gropp et al., 2004]. This API, (Application Programming Interface) was utilized here to execute the simulation for each group of blocks (with common sequence patterns) in a cluster node, i.e. each group of blocks assigned to one computation node (with 8 processors per node). In general this allowed distribution and hence reduced the problem of data overloading. The main disadvantage of MPI libraries is that it does not permit excess inter-node communication, [Gropp et al., 2004]. Hence in this version of model, there is no inter-node communication, unless more than 2 groups share the same CpG island. On the other hand, OpenMP is a form of implementation of "multithreading" concept where a master thread (set of serial instructions), divides into a number of slave threads

and assigns a task to each thread, [Chapman et al., 2007]. The runtime environment later associates each thread to a processor available in the hardware. The purpose of using OpenMP routines along with MPI was to expedite simulation inside a node by sharing limited number of computational objects and yet computes results for large data efficiently. Many real time applications allow hybrid computing, (use of MPI and OpenMP libraries), to exploit the advantages of both approaches, [Rabenseifner et al., 2009]. In our case the huge dataset that contained more than four hundred genes were categorized into 31 groups based on CG spacing. As a consequence, a hybrid version that allotted data to 31 distinct nodes and shared the assigned data among the 8 processors inside a node, was adopted.

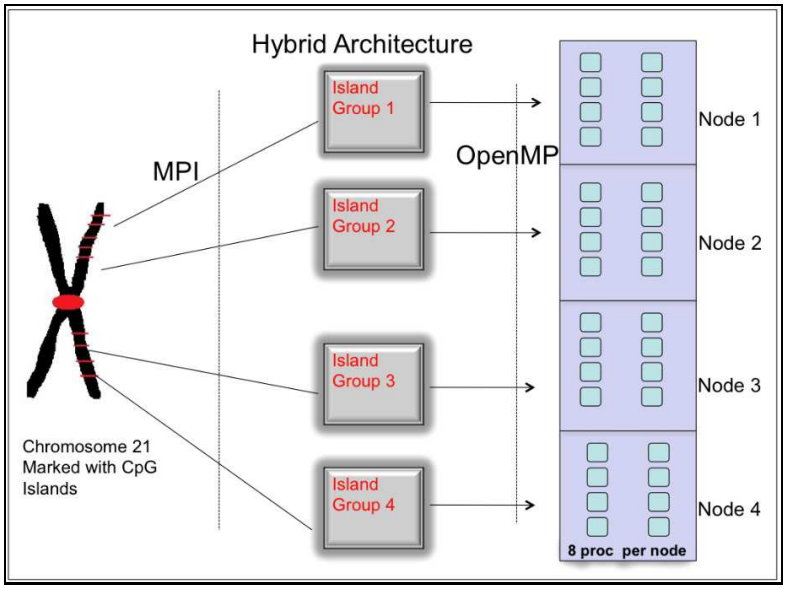


Figure 5.2: Parallelization Framework(PF)
Hybrid Architecture Layout for the Parallel Version of EpiGMP micromodel. 31 groups of CpG islands are derived from 464 DNA sequence files. Each group was assigned to a computational node, (using MPI routines), and objects in each group were shared inside a node among processors equally, (using OpenMP routines.)

The HPC platform, (Dual Intel Xeon Quad core of ram 2.66GHz) was used to perform simulations for model hybrid version. The number of nucleosome objects per block of island varies based on the length of DNA sequences. While MPI routines play a major role in optimization and reduction of the total time taken for simulation, OpenMP functions help to efficiently distribute the data by sharing memory among all processors in a node, (as described in Figure 5.2). The simulation steps and short pseudo code for the hybrid version of EpiGMP are given below.

5.4.1 EpiGMP Hybrid Version Psuedocode

1. *Read and Store Inputs*

   (a) Histone Data - possible number of histone "string representations" (or modification data), for each histone type

   (b) DNA sequences and information on CG positions throughout sequences.

   (c) User Selected Values – DNA methylation per a Block in a specific time-step, total number of iterations (or time-steps).

2. *Create Objects*

   (a) Each CpG island is assigned to a group based on the CG pattern found in it. The length DNA sequence of each island decides the number of nucleosomes and histone objects to be created.

3. *Simulate*

   (a) For each CpG island irrespective of the group, evolution of each histone type is simulated and the new information of modification levels is sent to the master processor. Inside each node the master processor receives the updated modification tables and re-computes output parameters, such as new DM and transcription. Simultaneously,

probabilities of shift are amended, based on the DNA methylation value (see Figure 5.1 for interdependencies), which in turn is associated with methylation in the CpG islands. The load balancing of CpG islands is achieved as described by the algorithm given below.

Algorithm 5.1 Algorithm of the EpiGMP tool - Version 2

```
1: Each Group of Islands or genes recruits one Computational Node
2: DQUARD=8                          ▷ universal constant represent no of processors in an node
3: Group 3                           ▷ Showing the steps to simulate group 3 CpG island
4: MPI_Init()
5: #pragma omp for (Block b=1:max)
6: schedule (static,total-blocks/DQUARD)
7: for Iteration a=1:max do                          ▷ max= maximum no of iterations
8:     for Nucleosome d=1:max do                     ▷ max=maximum no if nucleosome objects
9:         for Histone Type j=1:8 do                 ▷ 8=total no of histone types
10:            RANDOM HISTONE STATE SELECTION
11:        end for
12:    end for
13: end for
                     ▷ Check for common islands in diff. groups to exchange output values
14: if group3.Block1==group5.Block3 then             ▷ an example of exchange
15:    MPI_Send/receive DM values
16: elseTake No Action
17: end if
18: Calculate Output parameters and Synchronize
19: MPI_Finalize()
```

(b) Continue simulation, memory sharing (using OpenMP routines) and MPI communications, if applicable, till maximum no. of iterations are reached.

4. *Store Outputs*

(a) Results for the specified time interval, inside each Block –

i. Transcription rate

ii. DM value (assumed to be methylation of each CG dinucleotide)

iii. Count of the number of times each state is visited in all 4 histones for each nucleosome, (H2A and H2B reported here).

(b) Time taken for Simulation of each group of CpG islands, in Minutes.

5.4.2 Details about Dataset & Simulation

In order to investigate the system behaviour, all contiguous sequences in chromosome 21 were combined and analysed (total sequence length being 35,660,412 bp). Chromosome 21 has 464 CpG islands, present in all contig sequences, which were extracted based on criteria described by [Takai et al., 2002], using a Perl script before the micromodel was executed. These islands were segregated into 31 groups based on the CG patterns observed in each island as described above. Histone evolution was observed for 10,000 iterations and outputs were recorded after every 1000 time-steps. The DNA methylation was set to high and low values, (0.9 or 90% for high and 0.1 or 10% low DM) for each island in all groups. Outputs namely, the time taken by each CpG island group in minutes and evolution of histone states of type H2A and H2B are reported below.

5.5 Results

5.5.1 Hybrid Version – Performance and Time Analysis

The simulation was carried for each group of CpG islands separately, with each group taking 2.5 computational hours on average to complete. The total time for this hybrid version took about 7 hours unlike the serial version, which required a total of 23 hours for the whole simulation to complete. The total time per group in minutes is shown in Figure 5.3. Here the amount of time taken for simulation is directly related to the number of objects (number of CpG islands or nucleosomes per island), in each island group. Group 2, made of 110 CpG islands took the maximum time while groups 19, 27-29 and 31 took the least simulation time as they contained

only 1 island each. Other groups such as 17, 18, 21 and 23 contained islands that were less in number but huge in DNA sequence length. As a result, the number of nucleosomes per island was more and this increased the time taken for simulation, ($>$ 100 minutes).

The result, (Figure 5.3) reported here, is an analysis of islands from Chromosome 21. The same categorization based on CG patterns would result in different sizes of CG-groups for other chromosomes. This hybrid framework is very efficient in terms of reduction of simulation time and hence will be adopted for all future simulations and applications of EpiGMP tool. The human genome consists of a significant number of CpG islands that contain the maximum number of CG dinucleotides separated by 3bp. The specific pattern is strongly associated with differentially expressed genes. Hence histone modifications (H2A and H2B), for this specific group of islands are reported in the next two subsections

5.5.2 H2A Analysis

Figure 5.4 contrasts the different modifications observed in H2A during high and low methylation conditions averaged over 3 simulation trials. During high methylation condition (DM level $> 85\%$), selective histone states such as the 5^{th} and 13^{th}, (number or order of H2A histones are displayed in X-axis of Figure 5.4) were most preferred i.e Arginine was methylated in H2A the most under this condition. The Literature [Eckert et al., 2008] indicates that specific cell types, do not contain this modification and hence develop into tumorous cells, (this is an explicit evidence of down regulation of methylation modification leading to tumour growth). Under lower DM conditions ($< 15\%$), 4^{th} and 12^{th} states were most visited implying high priority to Lysine 5 and 9 modifications.

Acetylation of K5 is notably found more during gene expression while that of K9 is an unexplored modification [Cuddapah et al., 2009; Wyrick et al., 2008]. This unreported K9 acetylation in H2A, could be a potential modification that supports gene expression based on the model results

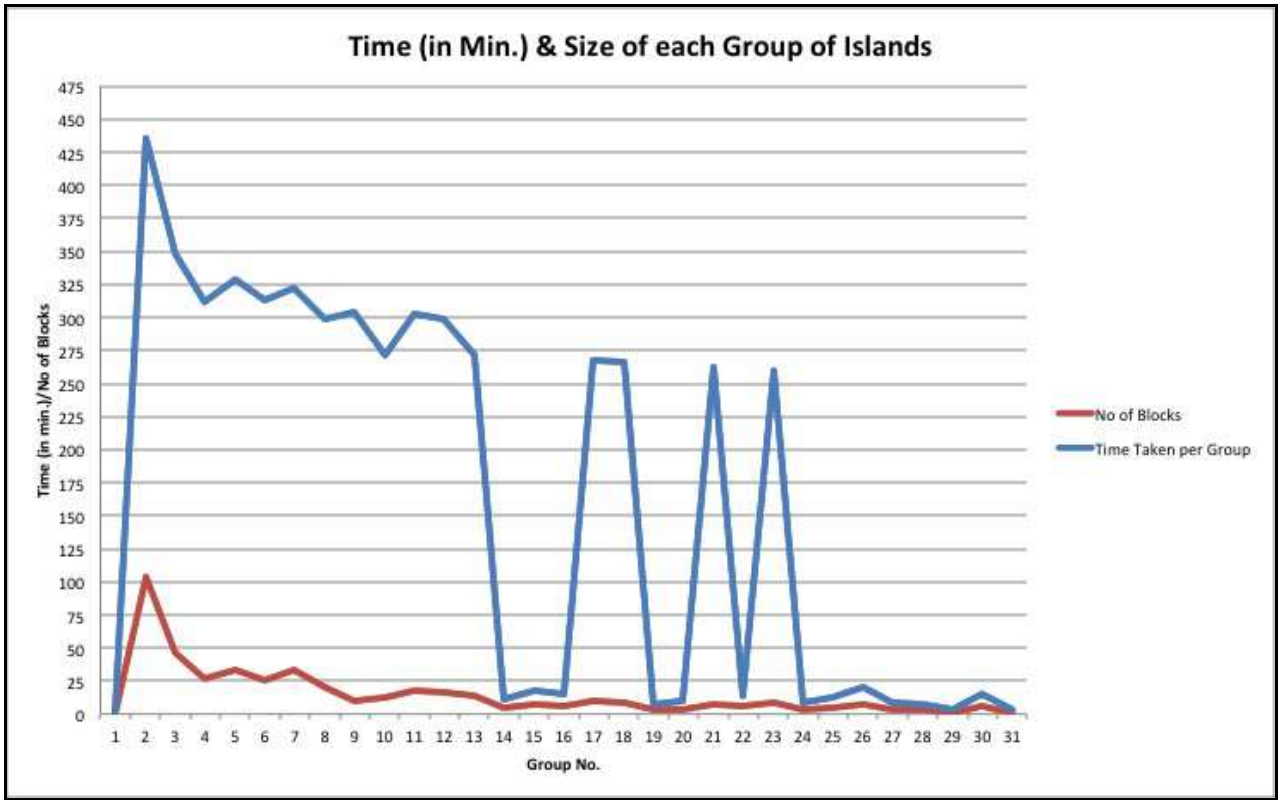


Figure 5.3: Time Analysis of Parallel Framework

Details on the time taken and size of each group of CpG Islands. The number of islands per group is indicated by the red graph line and the time of simulation per group, (in minutes) is represented by blue graph line.

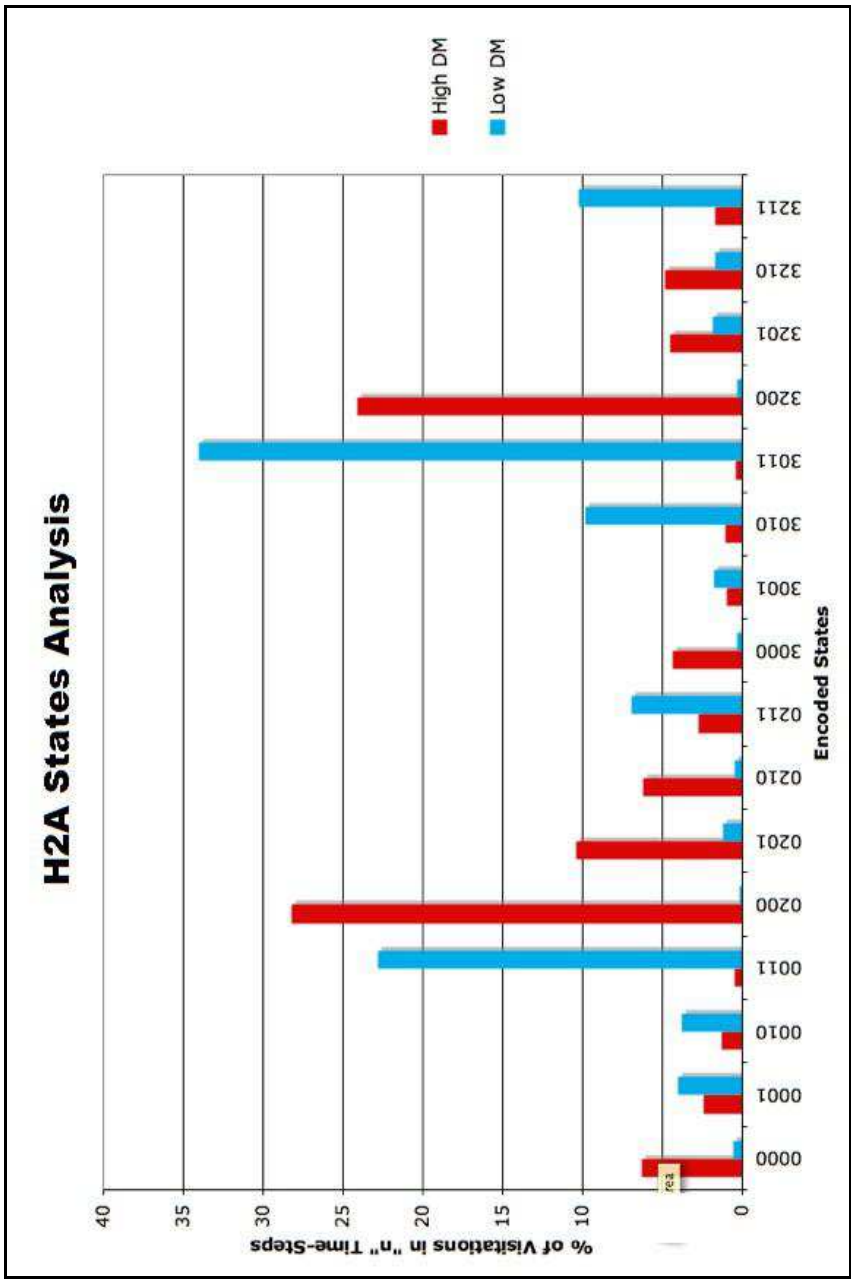


Figure 5.4: H2A Analysis
A Comparison between the average (over 3 Simulation runs) preferences of H2A states for high (red) and low(blue) DNA methylation Levels.

derived through repeated simulations.

5.5.3 H2B Analysis

Figure 5.5 shows highly visited states during high and low DM for H2B type histone. Since the number of states possible (refer Table 4.1) for H2B is more than 1000, Figure 5.5 shows those 15 states visited the maximum number of times over all trials. In case of low DM, several states containing acetylation modifications are most often visited. i.e. Acetylation of K5, K12, K15, K16, K20 and K24. The role of K15 acetylation has been extensively studied in connection to cell death in the literature [Ajiro et al., 2010]. In specific cells this modification is compulsively retained to keep the cell alive and removed if cell apoptosis (self destruction) is carried out [Myers et al., 2003]. However acetylation of K5, K12, K16, K20 and K24 is known to occur in H2B in general, though the biological significance is still under investigation [Wyrick et al., 2008].

Serine phosphorylation quite interestingly appears (S10 and S14) for both levels of DM in our simulation, but is known to play a supporting role in cell apoptosis (antagonistic to K15 acetylation) and structure balance. During high methylation, few modifications such as methylation of K5 and K23 were consistently noted or preferred despite the indication that the former is required for transcription [Yoo et al., 2006]. In the latter case the biological significance is unknown. Agreement of our model findings with known effects reported in the literature are consequently encouraging [L. Zhang, Eugeni, et al., 2003]. In case of H2B, fewer modifications were reported in the literature with known effect, but the others reported by the micromodel have been efficiently categorized here through repeated testing.

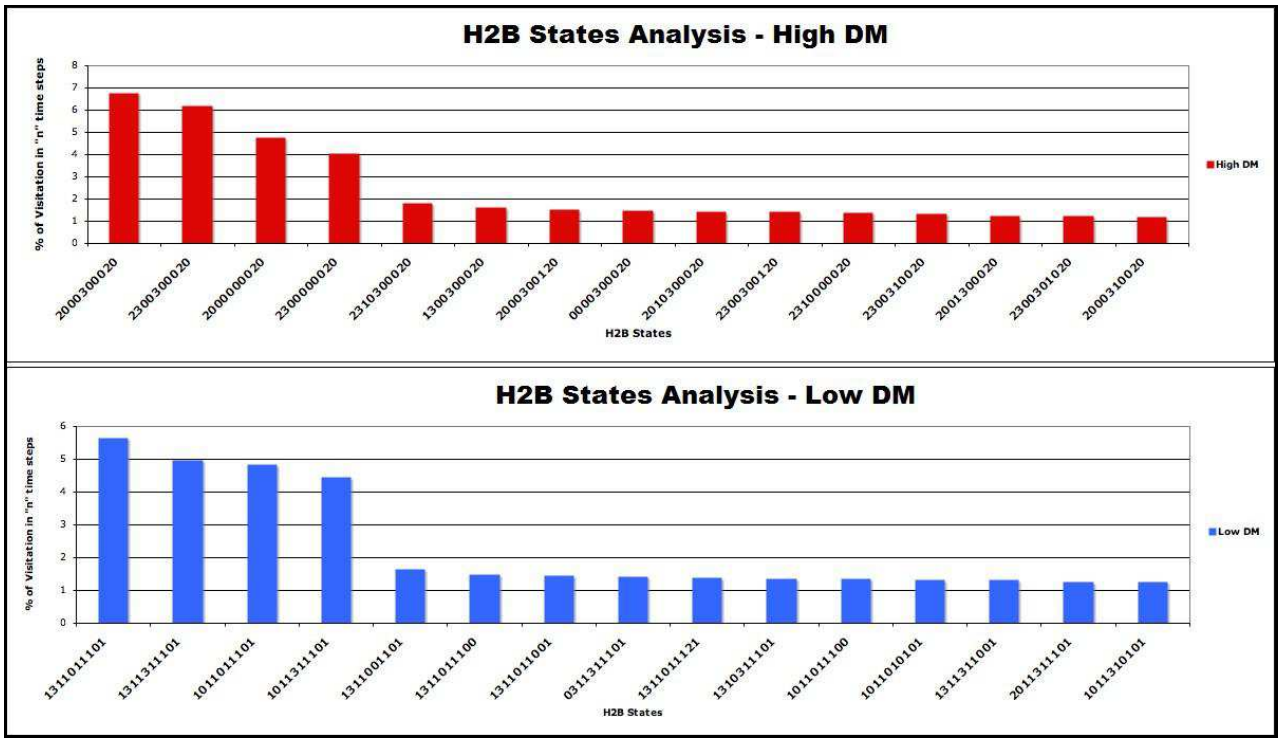


Figure 5.5: H2B Analysis

A Comparison between the average (over 3 Simulation runs) preferences of the most visited 15 states of H2B histone type for high (red) and low (blue) DNA methylation Levels.

5.6 Conclusion

Our EpiGMP micromodel iterates between possible histone modifications that can be associated with a specific level of DM. DM fluctuations over specific time intervals are associated with distinct CG dinucleotides in the DNA sequence. A Hybrid parallel version of micromodel, combines the advantages of both MPI & OpenMP libraries evidently and has been applied here to study CpG islands in Chromosome 21 since its the most optimized approach. This enables investigation and analysis of a network of "gene groups" simultaneously as it occurs realistically in nature. For example, evolution of common genes shared by 2 groups involved in Prostrate and Colon cancer can be studied. As DNA methylation controls the direction of evolution of histone states, high and low DM levels lead to different preferred levels of histone state selection. Nevertheless, system tolerance to initial selection of histone states is good, with specific "marker" states consistently chosen during execution. This consistency in predicting characteristic histone modifications under defined DNA methylation levels emphasizes the model's ability to mimic the real system accurately and provides an effective tool for investigating epigenetic events.

Ultimately, the aim of the predictive model must be association of the molecular events and epigenetic mechanisms with aberrant effects at the macro-scale, such as onset of disease. This framework can be amended to determine the interactions among many groups with functionally different genes and ultimately identify alterations in the several systems inside an organism. In the next chapter, we apply EpiGMP to a network of genes that play a significant role in Carcinoma or stage III of Colon Cancer. The first step is to apply the tool to every gene in the network to predict epigenetic molecular events, followed carrying out by an analysis of how the expression level of each gene iteratively, (over many cellular generations) affects the whole network during the progress of Carcinoma.

Chapter 6

Colon Cancer: A Case Study

6.1 Introduction

Cancer, long considered to be one of the most serious life threatening disease conditions, is currently the most investigated topic with regard to human health, [Miyamoto et al., 2012; Vogelstein et al., 2002]. This group of diseases is characterized based on its origin in the human body, leading to one or more types of malignant or non-aggressive cancer. The principal feature is uncontrolled growth of cells and tissues, with the initiation period consisting of tumour growth, which further develops by spreading across different systems, through blood and lymph fluids. Colon cancer[1], is one of the most highly incident type of cancer in humans that consumes more than half a million lives worldwide every year, [Campbell et al., 2012]. The high incidence of this cancer type is mainly attributed to the consistent accumulation of anomalies such as genetic mutations over several cell divisions. The epigenetic layer consisting of DNA methylation and histone modifications is also responsible for the tumour development. A tumorous genome contains random point mutations on specific genes, local *hypermethylation* of tumour suppressor genes and global *hypomethylation* of repeat regions causing abnormal phenotypic traits such as Micro Satellite In-

[1] Alternatively addressed as colorectal cancer; originates in the human colon or in the rectum

stability (MSI)[2], [Loeb, 1994]. Such major changes affect important pathways controlling the cell cycle and others such as the WNT pathway[3], DNA repair pathways and mainly apoptosis pathway, [Lind, Thorstensen, et al., 2004; Netzer et al., 1998]. The development of colon cancer is a multistep process and progresses from benign polyps[4] to malignant tumours that have the potential to spread across the body, [Barat et al., 2010; Migliore et al., 2011]. In more than half the subjects diagnosed with an advanced colon cancer, "adenoma" polyps have been previously detected in the bowel and known to be associated with initial stages of malignancy, [Netzer et al., 1998].

6.1.1 Epigenetic basis of Tumorigenesis in Colon Cancer

Hereditary and somatic molecular changes in many genes have been reported during the development of the adenoma polyps which eventually lead to invasive cancer or metastasis, [Migliore et al., 2011]. These aberrations can be further divided into – (a) hypermethylation of tumour suppressor genes and DNA repair or damage control genes, (b) hypomethylation of oncogenes, (can disrupt normal pathways and cause uncontrolled growth of cells) and lastly, (c) mutation of one or both types of these genes disrupts DNA methylation status leading to dysfunction of cellular events. Considerable research has focussed on analysing the mutated and epigenetically altered "marker" genes such as – APC, MLH1, MGMT, KRAS and TP53. For example, a study conducted by [Lind, Thorstensen, et al., 2004] explored the expression states of these genes from different cell lines and reported on overall frequencies of these states being affected. These included methylated

[2]The human genome consist of simple sequence repeats, (SSR) formed by dinucleotides or short oligonucleotides, (e.g. CACACACA). The absence of specific marker SSR within coding regions of tumour suppressor genes during cancer is called Microsatellite Instability or MSI

[3]A intermediate pathway that signals many proteins, and consequently allow expression of genes responsible of cell growth, differentiation and surviva.

[4]A polyp is a an extended mass that grows from the intestinal walls

tumour suppressor genes recorded during MSI and MSS[5], such as MLH1, CDKN2AP14(Arf) and CDKN2AP16(Ink). Another case of disruption to DNA repair genes is that of MGMT methylation which encourages or leads to mutation and inactivation of TP53 tumour suppressor gene, [Esteller et al., 2001]. One of the most prominently studied genes in colon cancer is a tumour suppressor called "APC" (or *Adenomatous Polyposis Coli*) which is known to be highly methylated and mutated, although the order of these two events can vary depending on the cell line affected, [Lind, Thorstensen, et al., 2004]. One among many genes known to be methylated along with APC mutation is E-Cadherin. APC is involved in WNT pathway, which leads to increased transcription levels of genes like MYC and CCND1 and consequently "stimulates cell proliferation", [Lind, Thorstensen, et al., 2004; Lustig et al., 2003]. Such events are further cemented by changes in levels of marker histone modifications, (H3K9 or Lysine9) in these genes, [Nakazawa et al., 2012]. These consolidated facts reveal that colon cancer is a disease triggered by a sequential order of events at different levels (molecular to system level) over a prolonged period of time. While a number of steps are involved, the aim of this chapter is to propose and test the model framework to see how events at a system level, (like a network of genes), are linked to the triggering molecular events. This is one of the first and highly focussed attempt at extending the model across several scales and is carried out as a feasibility study, in part to determine the where the difficulties lie and what key questions must be addressed.

6.1.2 Cross Scale Information for Colon Cancer

In the previous chapters we established the mathematical interdependency among epigenetic events and considered the nature of genetic information contained in the nucleotide sequences and its association with epigenetic information. Questions raised about the progress of cancer and malignancy

[5]Human genome consist of simple sequence repeats, (SSR) formed by dinucleotides or short oligonucleotides, (e.g. CACACACA). The presence of specific marker SSR within coding regions of tumour suppressor genes during cancer is called Microsatellite Stability or MSS

development, [Esteller, 2008], include whether these non-genetic factors are responsible for the progress of the disease. It seems clear that epigenetic information augments that obtained from gene expression measurements, proteins and mRNA[6] stability. For example, mRNA correlation with protein levels for transcription factors, varies from gene to gene and is not entirely sufficient to understand differential gene expression especially during cancer [Guo et al., 2008]. Further alterations in *miRNA*[7] or micro RNA occurs commonly in human cancers and interest has grown on how epigenetic mechanisms contribute to miRNA dysregulation. In a research article exploring this question for colon cancer, the authors, [Suzuki et al., 2009], used ChIP-on-Chip analysis and concluded that recorded patterns can offer insight on the associations of this dysregulation with chromatin state. Studies on the determination of chromatin state and prediction of genome wide enhancers have been reported, using computational approach based on Genetic algorithm, [Fernández et al., 2012]. Such biological models help to confirm the facts observed through laboratory-based experiments to investigate the indicators of the spread of colon cancer.

Computational Models associated with Colon Cancer

There have been other extensive computational models that relied on discrete information to investigate the lethality of colon cancer and consequently aid in designing preventive measures and healthcare policies. Based on results from biological experiments, a "gastric crypt" model by [Perrin and Ruskin, 2010], explored the spread of methylation and differentiation of cancerous stem cells in gastric crypts of the human colon. The major advantage of these simulation trials is that they permit a major reduction in the time taken to investigate specifically the spread of malignancy, (where biological experiments for this specific study took up to 18 months of trials). Other computational multi-scale models were developed to understand – (i) malignant events that propelled

[6] messenger RNA or mRNA is a type of RNA that helps to translate genetic code into a string of amino acids or protein sequences

[7] miRNA or micro RNA is a type of small RNA, usually 22bp long that is involved in binding to mRNA transcripts and prevention of protein translation.

colon cancer across different stages, (from benign polyp, adenomas and carcinogenic tissues to a metastatic system), [Kuntz et al., 2011], (ii) advantages of different treatments including radiation therapy and other control strategies, [Rutter, Knudsen, et al., 2011] and (iii) evolution and adaptation of improved health policies for cancer patients, [Rutter, Miglioretti, et al., 2011; Rutter, Zaslavsky, et al., 2011].

These previously designed models reveal the complexity of signals and cascade of events that instigate malignancy progress in the human colon, and consequently the systematic procedures to be followed in the development of a possible cure for the same. Hence our focus is to understand epigenetic events at the molecular level and how different levels, (from molecules, genes, systems to organisms), integrate and scale up to influence disease dynamics inside the human body. Data on genetic and epigenetic events and their interactions are becoming increasingly available but drawing these together towards a plausible multi-level model is far from trivial. A very elaborate and useful resource for DNA methylation alone called "MethDB", was recently created [Grunau et al., 2002], followed by review of the state-of-art epigenetic resources has been discussed by the authors in [Kaja et al., 2011; Shakya et al., 2012]. In addition, a detailed, though targeted, resource on statistics of epigenetic events for colon cancer called "StatEpigen", [Barat et al., 2010] is available, (website of the database given in Appendix D). The following subsections explain the data extracted from StatEpigen and how the original algorithm of epigenetic micromodel was modified to investigate the complex signalling in colon cancer pathways.

6.1.3 Statepigen Database

As information about mutated and methylated genes has increasingly become available, [Barat et al., 2010], a systematic consolidation of this information for different stages of colon cancer was needed. This motivated the creation and development of StatEpigen, which is a statistical epigenetic database, relying on manually curated epigenetic data and providing information on the

frequency of gene associated events during the different stages of colon cancer. These include somatic *hyper* and *hypo* methylation, mutation, gene expression, Multiple Satellite Instability, CpG Island Methylator Phenotype or CIMP, occurrence of special histone modifications and other cellular events. The main aim of this database is to gather fragmented information on genetic-epigenetic events so as to aid in building a platform for the development of computational models.

Typical information extracted from this database includes– (i) Simple frequency values of genetic/epigenetic events for a given cancerous phenotype together with source of cell lines, (ii) Conditional frequencies of other secondary events given a primary event occurring for one or more genes, (i.e. groups of molecular events with combinations of event types, which are found together in the same samples). Since the data available in StatEpigen is fairly extensive and diverse, our dataset focussed on the following characteristics –

1. Conditional events during the carcinoma stage of colon cancer only are considered for the model. For example the frequency of hypermethylation of any gene (event 2), given the occurrence of hypermethylation of MGMT gene (event 1) from sporadic samples is considered.

2. Three major events of hypermethylation, gene expression and mutation only are considered.

3. The first event is fixed to be hypermethylation of a gene. Hypomethylation and gene expression are strictly omitted as a first event to simplify our modelling framework.

4. The dataset does not include conditional events if both event 1 and 2 are associated with the same gene. For example, if Event 1 and 2 were hypermethylation and mutation of APC gene, this entry was eliminated from the list.

The carcinoma stage in colon cancer is a well-documented disease condition because it is a highly incident cancer type and hence was chosen for further investigation of epigenetic lesions

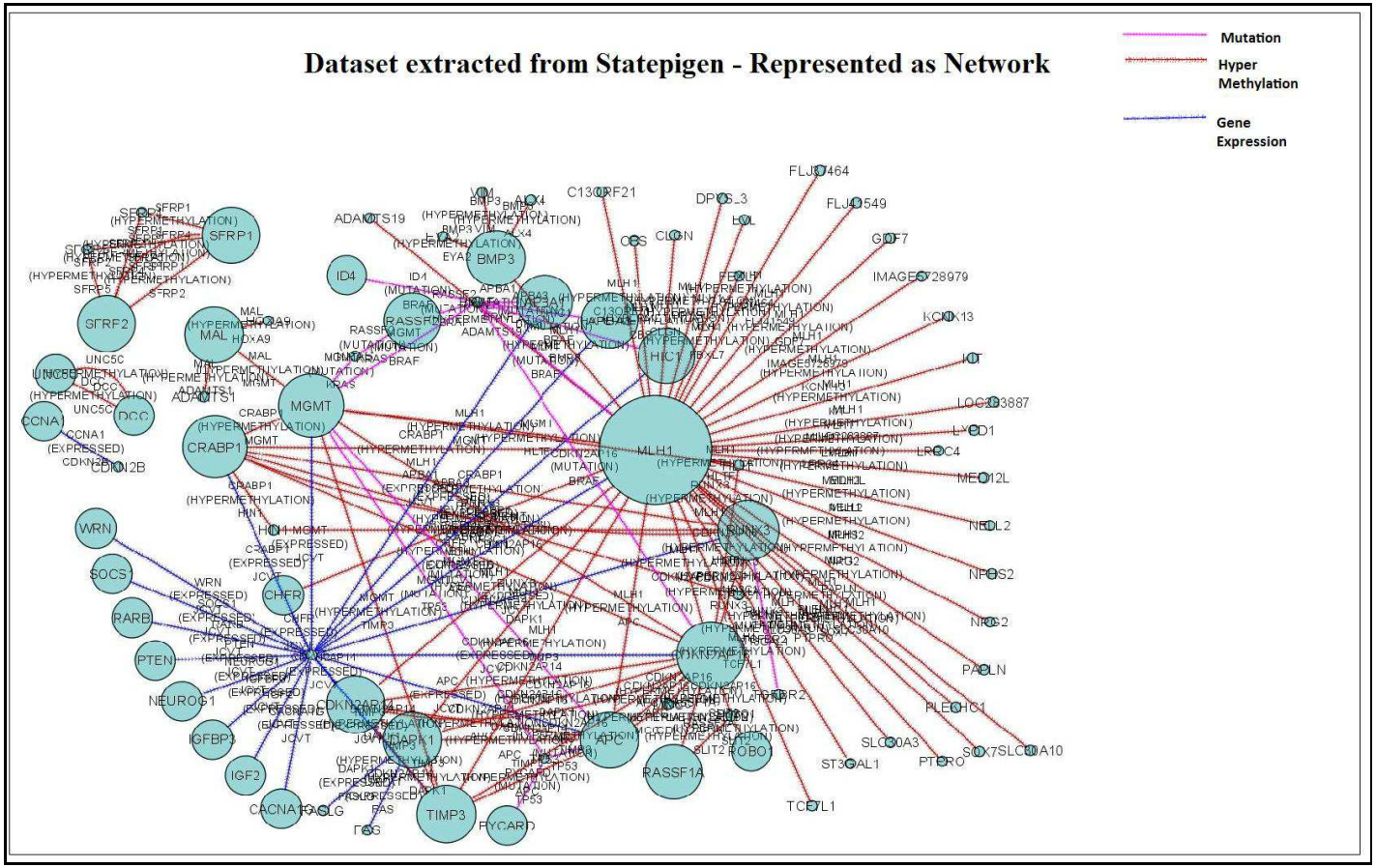


Figure 6.1: Colon Cancer Gene Network

Network representation of Dataset extracted from StatEpigen. The nodes represent genes, and edges represent the conditional events. Each type of event or edge - gene expression, mutation and methylation is assigned blue, pink and red color respectively. The network has 85 nodes, 121 edges in total and amongst which 33 nodes have atleast 1 outgoing edge. Node size depends on the no of outdegree for each node. All the genes appearing in the network here undergo a genetic/epigenetic event, which is associated with colon/colorectal cancer. (Source: StatEpigen Database)

that trigger its progress. The hypermethylation of genes was chosen as event 1 because it provided a window for testing the colon cancer network when EpiGMP was applied, (as EpiGMP is known to take DNA methylation values as inputs). Two events associated with the same gene were not allowed, as this would reduce the chance of studying interrelations among genes during cancer. After removal of duplicates and information that did not satisfy the conditions made above, 121 relations among 85 genes involved in colon cancer from sporadic cell lines were extracted. Details of all the genes and their interrelations are given in Appendix D. The Figure, 6.1 shows the dataset redrawn as a network, using Cytoscape software, [Shannon et al., 2003], for better visualization of the level of connectivity among gene sets. An example obtained from the database contains the following information – Gene 1: MGMT, Event 1: Hypermethylation, Gene 2: APC, Event 2: Mutation, Frequency of Event 2 given that Event 1 occurs: (0.21). Such a network may be described in terms of weights of connecting edges clustering coefficients or other network parameters. In the current case we are interested in those genes, which may be considered most influential in terms of their observed involvement and potential to be involved or associated with other genes. For this reason the figure shows most frequently occurring genes or those with highest association as the nodes with largest, number of outgoing edges. Edges represent conditional relations extracted from the database and their weights are the frequency of the conditional events occurring. Ideally a single step in the model, when applied to the gene network, corresponds to applying the EpiGMP tool to each gene, (biologically this correlates to completion of cell cycle and initiation of next generation). The occurrence of each conditional event was tested through simple stochastic tests, and results at this level reported the status of the network and count of appearance of the possible number of edges in it. This simulation trial does not include information about DNA sequence patterns, (EP2 compartment from the previous Chapter) to contribute to the final outputs of the simulation trials carried out.

6.2 Methods

The epigenetic prediction tool, (described in Chapters 4 and 5), provides a framework to account for and report possible molecular changes occurring during abnormal epigenetic events. These include over-expression of oncogenes and silencing of tumour suppressor genes. The idea was to apply it to an extended set of real data in order to determine the probable chain of events leading to metastasis and instability within the patient affected. The precise and qualitative nature of empirical data obtained from StatEpigen, provided a perfect opportunity to investigate relations among genetic/epigenetic events and simultaneously report the "driving molecular changes" inside each gene that causes cancer. Hence the final model developed for this purpose consists of two distinct inner and outer layers as explained below, (as also shown in Figure 6.2).

6.2.1 Application of EpiGMP – Layer 1

Layer 1 (or the inner layer) consists of the application of the EpiGMP tool on each gene in the network obtained from StatEpigen, (Figure 6.1). While the main input required at this stage is DNA methylation levels, (DM $\in$ [0,1]), the consequent outputs reported are the recalibrated DNA methylation, Transcription levels and histone modifications, (as described in Chapters 4 and 5). The inner core of the 2-part cancer model proposed here, focuses on investigating changes at the molecular level, and hence the micromodel's dynamics is quickest. This is because, each iteration here corresponds to selecting a possible histone modification among a library of possible solutions, where aggregated changes feed back to methylation status. To elaborate more, a hyper-methylated gene would essentially have methylation values set to 90% (or 0.90) initially and in turn generate levels of histone modifications. Correspondingly, an over-expressed gene would set least DNA methylation and consequently report more acetylation changes. The ideal number of time-steps or iterations for the EpiGMP if applied to a gene was set to 10,000, in order to choose possible modifications, (refer Chapter 4 for definition of time-step). Since the tool was applied to a gene

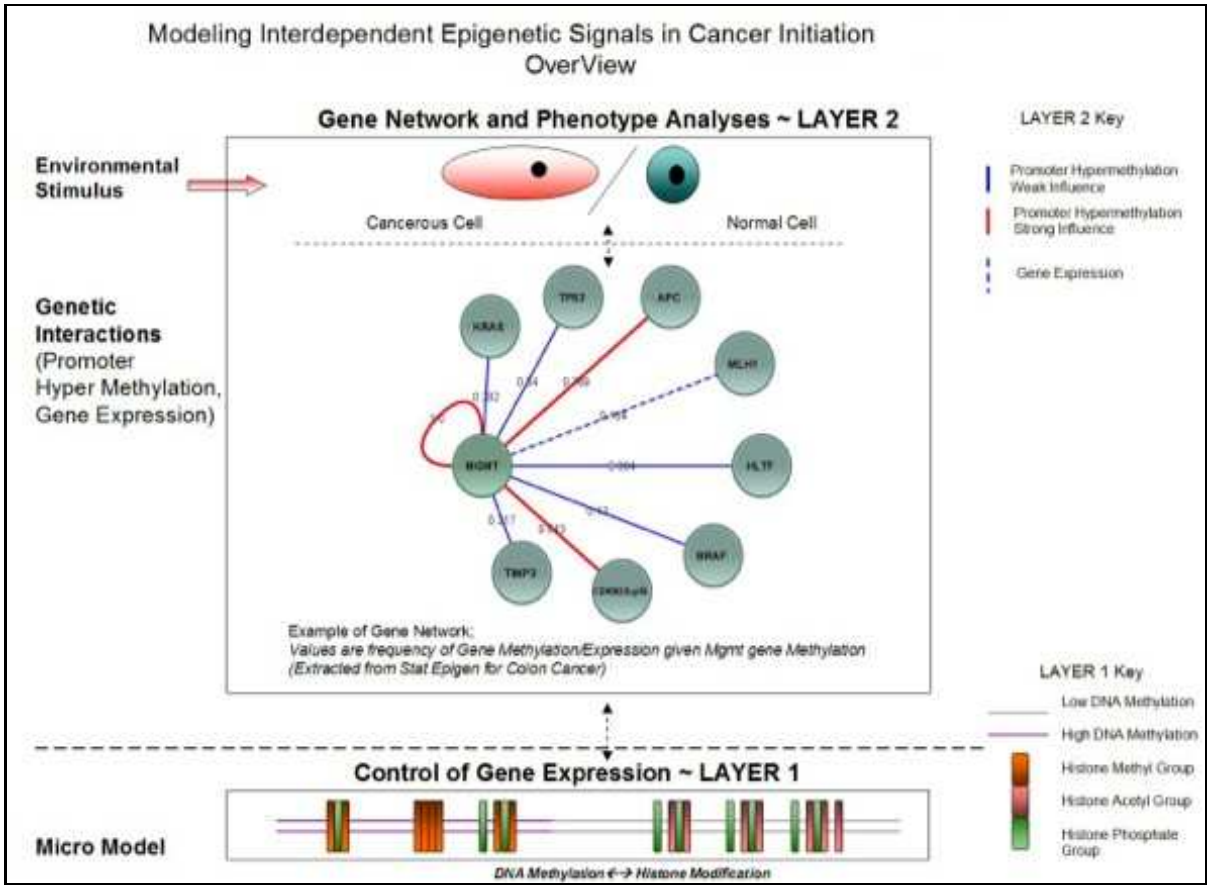


Figure 6.2: Two layers in the model framework for colon cancer

The two layers which form a part of the our computational model here. Bottom layer (layer 1 or molecular layer) consist of EpiGMP application. Top layer (layer 2 or gene layer) consist of network of genes; whose information is extracted from StatEpigen Database, [Raghavan, Roznovat, et al., 2011]

network extracted from the database, the user was allowed to set methylation level of only 1 gene, while information from the database and random variables were utilized for the remaining 84 genes in the network. This decision forms the second layer in this model version as explained in the next subsection.

6.2.2 Utilization of Information from StatEpigen – Layer 2

At Layer 2 (or the outer layer), the relations and state of each gene is decided by combining information from StatEpigen and applying RGT or Random Graph Theory. RGT is one of the most popular methods to study large graphs such as social, tele-communication and internet based networks [Newman et al., 2002]. The main idea behind this theory is that the nodes, edges or both are randomly created and sampled to study the network topology and infer interactions among nodes. In this model, the edges that represent associated events among different genes, are allowed to *randomly* occur, (from the true pool of edges that exist in the StatEpigen network file). The dynamics of this layer are much slower than the dynamics of the molecular changes inside a gene at layer 1. An iteration at this layer corresponds to the biological cell division, where genetic/epigenetic events that appeared in the previous generation are faithfully inherited in the current generation. Hence within each loop/iteration at layer 2, the DNA methylation levels are decided and followed by applying EpiGMP model to each of the 85 genes to report molecular changes. The inputs for this layer are the name of a single gene preferred by the user and its level of methylation, based on which further decisions regarding the methylation of other genes, (the occurrence of a given edge in the network), are taken by using random variables or random graph theory. It is very important to note that (i) the inputs for layer 1 and 2 remain the same as mentioned above, and (ii) the methylation/mutation status or edge appearance at layer 2 with regard to all genes are faithfully retained for the next iteration/generation, to resemble the real phenomenon of epigenetic inheritance. Since there are 3 main types of events or edges, (DNA methylation, mutation or gene expression),

two sets of random variables are used to decide their occurrence, which are explained in the next subsections.

Local RandomVariables

For each of the 121 conditional events, a specific or local random variable is created to decide its state of expression. An example of a conditional event involving two genes CDKN2AP16, (gn1), and MGMT, (gn2) is handled as follows. Frequency ($F_{gn1\text{-}gn2}$) of gn1 (hypermethylation) given that gn2 (hypermethylation) occurs is (0.543). A corresponding random variable is generated, $R_{gn1\text{-}gn2}$ within the range $\in$[0,1]. If this random variable satisfies a condition namely, ($\in$ [0, $F_{gn1\text{-}gn2}$]), methylation level of gn1 is set to 100%, but if $R_{gn1\text{-}gn2} \in$ [1-$F_{gn1\text{-}gn2}$, 1.0], the methylation level remains unaltered. For the case of a mutation event, the Boolean value of a mutation variable for the associated gene is changed to 1 if the random variable satisfies the condition and remains 0 otherwise. This process is repeated once for all 121 conditional events within one loop at the layer 2.

Global Mutation Variable

While it is possible to change the methylation profile of genes explicitly, mutation events are further controlled using a global variable, called "*GM*" whose boolean, (values assumed to be 1 or 0), status determines the occurrence of mutation of genes during each generation cycle or layer 2. This variable permits genetic mutation among possible genes if it has a value of 1 and doesn't allow if otherwise. To calibrate the value of *GM*, a slightly deviant approach that utilizes the concept of Shannon index, is used to make the decision for the whole network. A Shannon index, is a quantitative measure that helps to account for the diversity in a given population of entities, [Kaneko, 2009]. It has been applied in ecology (for a collection of organisms), and demography, (diverse sets of population). Equation 6.1 gives the usual form of the Shannon index. The number

of subspecies assumed in this case here is two, (assuming genes are considered to be mutable or non-mutable). Two separate arrays of Shannon indices, (total number of sub-species or R=2), where probabilities in set 1 $\in$ [0, 0.5] and those in set-2 $\in$[0.51,1.0]. Random values are generated using Mersenne Twister to choose which category the genes in the network fall under, namely mutable (GM=1) or non-mutable, (GM=0). It is to be noted that the value of GM is checked once for every generation in the model simulation.

$$H_i = -\sum_{i=1}^{R} \left[p_i * log(p_i) \right] \tag{6.1}$$

where,

H_i = Shannon diversity index for a specific subset in a big population,

R = Total number of species in the community (richness),

p_i = frequency/probability of belonging to i*th* species type.

This double-layered model requires the DNA methylation value of a selected gene as an initial input as described previously. Based on the methylation level of that gene, further decisions are made using two types of random variables described above. The first type is a global variable, which sets the permission to allow any mutation event to occur, and the other is an array of randomly generated variables, that is associated with the 121 events extracted from the database. In other words, this array contains answers for the occurrence of each edge in the network. This is followed by applying EpiGMP tool based on the value of DNA methylation set for each gene, (given the network decisions as described), and reporting histone modification levels. The dynamics of the two layers in the model are different and hence, maximum number of iterations for layer 2 is 150 and the micromodel or layer 1, is 10,000 time-steps.

6.2.3 Algorithm

The following points below enumerate the simulation steps –

1. *Read and Store Inputs*

   (a) The list of genes involved in the colon cancer network and relations among the possible genes is read, (in other words conditional events connecting two genes and the frequency of the events).

   (b) DNA sequence of the genes specified in network files

   (c) Information is read and stored on histone modification data from input files, (read data corresponding to 4 types of histones from files), in association with each gene in the colon cancer network.

   (d) Specific User defined Values –

      i. A user selected "test-gene" from the network and

      ii. DNA methylation value$\in$ [0,1]) for the test-gene.

2. *Create Objects*

   (a) Numerical vectors to store methylation levels and mutation status of each gene in the network.

   (b) Adjacency matrix (A)[8] to represent the occurrence of events in the network.

   (c) Based on the length of DNA sequence of each gene, corresponding number of nucleosomes (DNA sequence length/148 bp per nucleosome = no of nucleosome), are created.

   (d) For each nucleosome object, histone data structures for each histone type, to store modification levels.

[8] An adjacency matrix is a square matrix that represents the presence of edges among different nodes are 1 and zero otherwise in a network.

3. *Simulate*

   (a) If the "test-gene" matches, one of the genes in the list, reset its DNA methylation value as specified by user.

   (b) The pseudo code of the model is given below,

4. *Store Outputs*

   - Results after each cell cycle/generation inside the network include (at Layer 2) –

     (a) GM value based on random sampling of Shannon indices

     (b) Edge value: M - Mutated or D - Hyper methylated or G - Expressed or N - None

     (c) Corresponding mutation status of associated genes in the network

     (d) Adjacency matrix of network (A).

     (e) Results of EpiGMP tool for each gene (at Layer 1) including –

        (i) Transcription level

        (ii) Frequency of each histone modification.

        (iii) Re-calibrated DNA methylation values

The resulting adjacency matrix – (A), is used for analysing and interpreting the network evolution in the following section.

6.3 Results

As reported earlier, colon cancer exhibits progressive stages which are characterized by the occurrence of different genetic events, [Lind, Thorstensen, et al., 2004]. In this simulation, which consists of two layers – layer 1 for molecular changes in genes and layer 2 for changes in gene expression levels, the analysis of the genetic network involved in colon cancer or layer 2 is the

Algorithm 6.1 Algorithm of the EpiGMP tool (Application) - Version 3

```
1: DQUARD=8  ▷ no of proc per node  ▷ For this simulation only one cluster node was used as
   we only test for 1 group of genes involved in carcinoma stage of colon cancer.
2: MPI_Init()
3: for g=1: Max (Generation) Iterations (max = 150) do
                                        ▷ Set DNA methylation of test gene to specified value.
4:     if g!=1 then
5:         Read data from prev Generations
6:     end if
7:     1. Calculate value of GM
8:     if GM==1 then
9:         Allow Mutation
10:    else
11:        Do Not Allow Mutation
12:    end if
13:    2. Now check for occurrence of all conditional events using RGT
14:    3. Reset DNA methylation values and mutation status accordingly
15:    4. Apply EpiGMP here
16:    #pragma omp for (Block i=1:max) schedule (static,total-blocks/DQUARD) ▷ Inside each
   block of gene
17:    for m=1:max do                                   ▷ Max no. of iterations set to 10^4
18:        for d=1:max do                          ▷ parse through each nucleosome object
19:            for j=1:8 do                          ▷ parse through 8 types of histones
20:                RANDOM HISTONE STATE SELECTION
21:            end for
22:        end for
23:    end for
24:    5.For each Block, Calculate Output parameters
25:    5.1 Global histone mod. levels
26:    5.2 Transcription
27:    5.3 new DNA methylation
28:    6.Save the following variable of current generation
29:    6.1 GM variable value for network
30:    6.2 DNA methylation levels and
31:    Mutation variable for each gene
32: end for
33: MPI_Finalize()
```

principal focus of interest, since, the results of EpiGMP have been explained already, in Chapters 4 and 5.

S.no	Genes	Out Degree	Path length to MLH1
1	MLH1‡	39	0
2	CDKN2AP16‡	10	1
3	MGMT‡	8	1
4	CRABP1‡	7	1
5	RUNX3‡	6	1
6	CDKN2AP14‡	4	2
7	MAL‡	3	2
8	RASSF1A‡	2	2
9	TIMP3‡	4	3
10	APC,BMP3,DAPK1, SFRP1,SFRP2	3	inf
11	APBA1, HIC1	2	inf
12	CACNA1G,CCNA1,CHFR, DCC, ID4, IGF2, IGFBP3,NEUROG,PTEN, PYCARD,RARB,ROBO1, SOCS1,UNC5C, WRN	1	inf

Table 6.1: Colon Cancer Gene Details

Details of the genes with at least one outgoing edge. ‡ - refers to genes connected to MLH1 genes which have highest outdegree here. Dotted line separate genes connected and not connected to MLH1. "inf generally denotes that the gene is not connected to MLH1. Please refer the StatEpigen dataset presented in Appendix D for more information about the genes and the associated conditional events.

The network in Figure 6.1 is directed, has 33 nodes with at least one outgoing edge and an overall average outdegree of 3.66. As previously mentioned, the double-layered model requires user to give a gene and its methylation level. Table 6.1 gives details of all genes which have at least 1 out going edge. MLH1 gene is one of the most documented in colon cancer, and has the

highest outdegree in the StatEpigen dataset, (Table 6.1). Hence based on the connectivity to this gene, 8 other genes were also chosen (marked by ‡), to act as an input to the model, (as mentioned in methods section, the model framework requires a gene as one of the inputs to the model).

Information about the role of "Test Genes" in Carcinoma

Although the nine genes were chosen initially based on their connectivity in the cancer network, the role of most of these genes are associated with tumour suppression during colon cancer. MLH1 is a very vital gene involved in DNA Mismatch Repair system, (MMR), which helps DNA polymerases to rectify the defects in the double helix following DNA replication. Mutations or methylation of this gene is commonly found in many types of cancer. The CDKN2A family of genes, coding for P14 and P16, acts as a negative regulator of cell proliferation and hence preventing tumorous growth. The encoded proteins of this gene bind to cell receptors to prevent cell growth and initiate apoptosis. MGMT on the other hand is one of the most frequently focussed genes in colon cancer research. The main role of MGMT is to suppress tumour and defend against oxidation effects of specific compounds that prevent the repair of DNA. The CRABP1 gene encodes for proteins involved in retinoic acid or vitamin A mediated differentiation of cells. Methylation of this gene prohibits the differentiation process itself, hence making it another important gene associated with colon cancer. The other genes acting as tumour suppressors include RUNX3 and RASSF1A, where the former controls gene expression process and the latter is required for apoptosis, mediated through receptor signalling pathways. While the protein encoded by the MAL gene is required in T-Cell signal transduction, (methylation leading to wrong signals in pathways) protein encoded by the TIMP3 gene is involved in inhibition of *metalloproteinases*, and hence cell migration, (defects allow cells to migrate to different systems the body and spread malignancies during cancer). ***The details about the function of these genes and others mentioned in this thesis was obtained from a database called Genecards, [Rebhan et al., 1998], which stores functional information about***

genes, proteins and transcription factors, (website link to Genecards is given in Appendix D)

A high DNA methylation value of 100% for each gene was provided as the second input because the restrictions on simulation only allow hypermethylation of genes as a first event. Each of the 9 simulation trials (based on providing names of 8 genes and MLH1 as inputs to the model) evolved over 150 generations and this was repeated over a 100 times. Hence a total of 135,000, (9*150*100), matrices were generated to check the occurrence of events/edges in the cancer network. The results section consists of 3 parts with each subsection summarizing simple tests applied on the network parameters. Subsection 6.3.1 (or Edge Analysis I) gives a simple count of the average number of edges recorded in all matrices. Subsection 6.3.2 (or Edge Analysis II) uses a metric called Hamming distance to further explain the order of edges appearance. Subsection 6.3.3 (or Edge Analysis III) reports an analysis of small motifs or subgraphs of size 3 and 4 in all the matrices, using a tool called, *mfinder*, [Kashtan et al., 2002].

6.3.1 Edge Analysis I

Figure 6.3 shows a count of the possible events or edges in the cancer network during the first 30 generations. In general, the total edge count did not increase beyond generation 30 hence Figure 6.3 and 6.4 show results recorded up to the 30th generation or cell cycle only. The top graph, in Figure 6.3, shows the number of edges that appear in more than 50% of the total simulation runs. The second shows count of edges that appear in less than 50% of the simulation trials and the last shows count of all edges that appear at least once in all generations. From these, it is interesting to note that, although MLH1 had the highest outdegree, MAL activated the largest number of events (89 edges or events in the network), followed by RASSF1A, which influenced 88 edges or events. This is probably due to the fact that although MLH1 had a large outdegree, the probability of a few events associated with it were least, despite repeated trials, (see StatEpigen dataset in Appendix D). All input genes excluding MAL and RASSF1A activated 86 events in the network

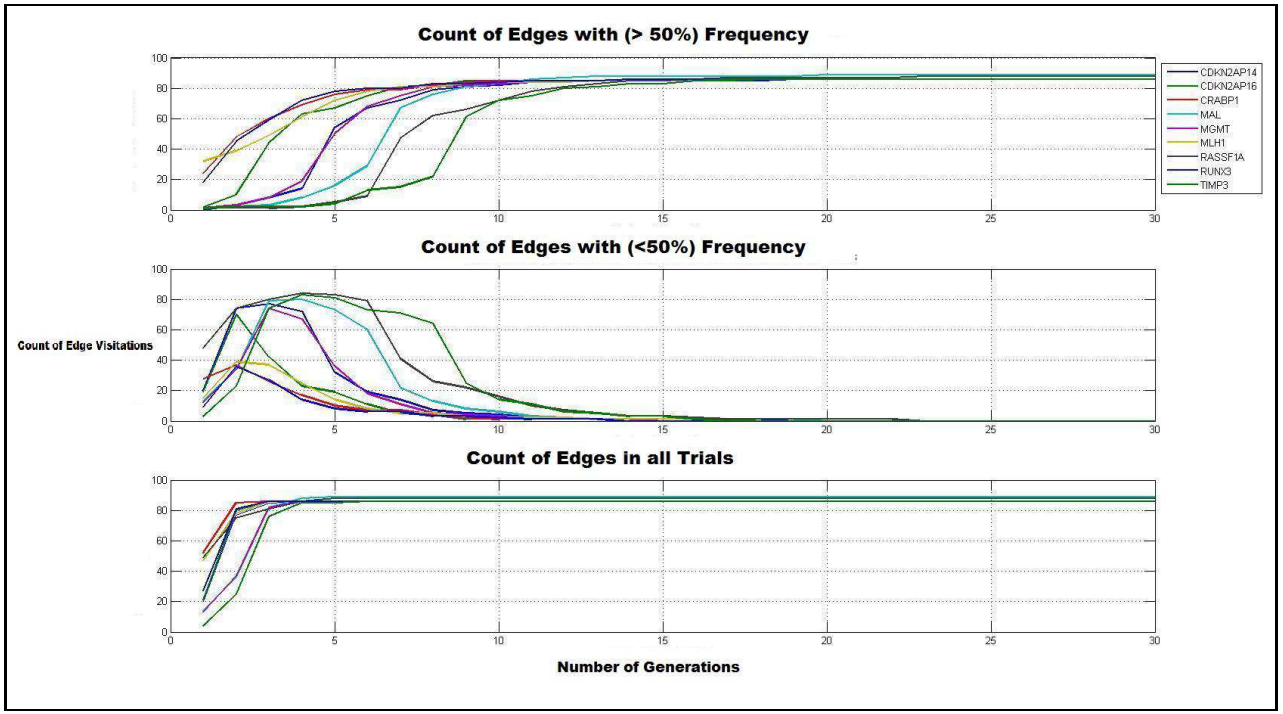


Figure 6.3: Edge Analysis 1

The top graph shows count of number of edges, (expressed as %), whose average frequency of appearance is >50% in the 9 different trials. The middle graph shows count of number of edges, (expressed as %), whose average frequency of appearance is <50% in the 9 different trials. The bottom or last graph shows count of number of edges, (expressed as %), in the 9 different trials using 9 genes in the simulation trials. Each of the 9 colors in the 3 graphs represent the 9 genes marked by ‡ in Table 6.1.

including MLH1. MGMT and RUNX3 based simulations showed a marked increase in the number of edges appearing after the fifth to sixth generation (or cell cycle), while for MAL and RASSF1A, a similiar growth behaviour occurs around the 7^{th} generation. An increase in the number of edges in the network appears at a later stage during simulation trials initiated by hypermethylation of CDKN2AP16 (test gene). The rest of the input sets, which include hypermethylation of MLH1, CDKN2AP14, RUNX3, CRABP1 and TIMP3, follow a general pattern with an increase in the number of edges reaching (more than 60) between third and fifth generation. The middle graph has an inverse behaviour where the number of edges appearing in less than 50% of the simulation trials is high until the fifth generation but slowly decreases to zero, which means that most edges are completely visited after 30th generation. Of the possible 121 edges, a maximum of 89 only were activated possibly due to two reasons, (i) presence of cliques or closed pathways, (e.g. genes such as SFRP-1, 2 4 and 5) that are not connected to any of the 9 genes or (ii) some conditional events have a "zero" probability of occurring which leads to a permanent inactive state for the associated genes.

Subsequently an average of the 100 simulation trials for all 9 types of input genes, for each generation was calculated. Each matrix of averages was further converted to a binary matrix where average values $> 50\%$ were assigned a value of 1 and 0 otherwise. Only 30 generations were considered for the 9 inputs (9*30=270), as no further major changes in the count of edges appearing after this generation were observed. This observation is mainly because, the changing the epigenetic states of specific nodes/genes allows affecting a majority of the edges in the network, (more than 86 out of 121). Hence it can be understood that those specific set of "test genes" given in the Table 6.1 (marked by ‡) play some significant role in carcinoma stage of colon cancer.

6.3.2 Edge Analysis II

The Hamming distance of dissimilarity, [Wijk et al., 2010] is written as:

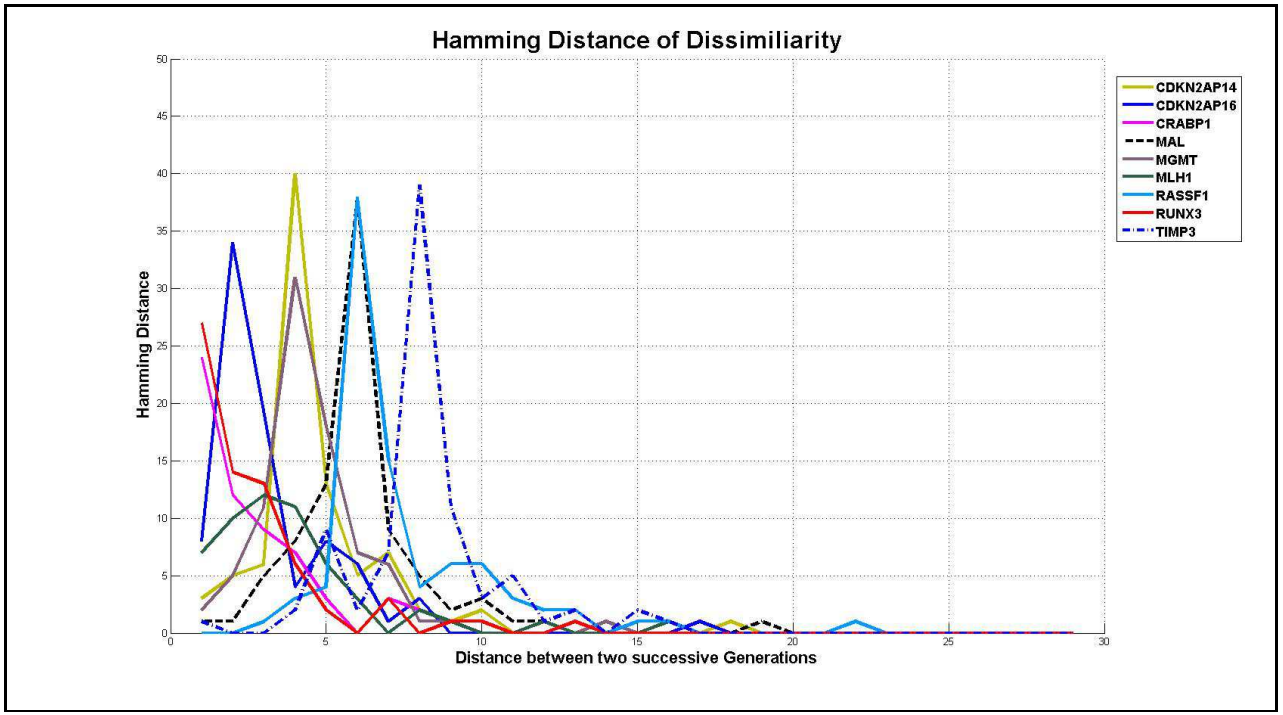


Figure 6.4: Edge Analysis 2

Hamming Distance of Dissimilarity calculated between any two successive generations for all 9 type of Simulation Trials with 9 different inputs. Each of the 9 colors in the graph represent the 9 genes marked by ‡ in Table 6.1.

$$d_{Hamming}(g_1,g_2) = \sum_{i \neq j}^{N} \left[A^{(1)}{}_{ij} \neq A^{(2)}{}_{ij} \right]; \tag{6.2}$$

where,

$A_{ij}{}^{(1)}$ = adjacency matrix for graph g_1.

N= Number of edges.

The Hamming distance as a crude measure was used to comprehend the order of genetic events. Figure 6.4 shows the Hamming measure of dissimilarity obtained for every pair matrices representing successive generation, (using Equation 6.2). The hamming distance is a discrete metric measure and every adjacency matrix produced after each generation is considered as an independent variable for the analysis. Large peaks in Figure 6.4, imply the appearance of a large number of events during that generation. These include the RUNX3 and CRABP1 genes during generation 1, MGMT and CDKN2AP14 during generation 4 and lastly, RASSF1A and MAL, during generation 7. CDKN2AP16 initiates rapid evolution of the network as a majority of events occur during generation 3, while TIMP3 (with a large distance or path length=3 from MLH1), among those 9 input genes, exhibit a large increase in the number of edges only after generation 7. It is also interesting to note that RASSF1A shows some activity after generation 21 unlike the other genes, which are quick to activate hypermethylation, mutation and gene expression across the network, (within the first 15 generations). Information from the literature indicates that independent hypermethylation of RASSF1A occurs in 20% of sporadically occurring Colorectal cancer, [Engeland et al., 2002]. This event has been closely linked not only with colon cancer, [Gonzalo et al., 2010] but also Chronic Myeloid Leukaemia, Uveal Melanoma and Thyroid cancer amongst others, [Avramouli et al., 2009; Merhavi et al., 2007; Xing et al., 2004]. These test simulations do suggest the effects of

hypermethylation of RASSF1A and its role in RAS-pathway[9] to mediate uncontrolled cell growth, [Xing et al., 2004]. The influence of MAL hypermethylation in this cancer network is in agreement with evidences from the literature, which indicate this epigenetic event plays a major role in early stage of breast and colon cancer development, [Horne et al., 2009; Lind, Ahlquist, et al., 2008].

6.3.3 Edge Analysis III (Motif Count)

Counting the number of edges and calculating the Hamming distance of dissimilarity helped to identify the number and order of events that occurred over many generations. Consequently, the final and biggest step is to investigate the most preferred *group of events* for sporadic colon cancer. A motif, (alternatively addressed as subgraph) is a small subset forming a part of any whole graph or network. Understanding the preferred patterns or connections in a network is very essential, as it helps to identify the significant pathways in colon cancer. This subsection reports on the statistics of incidence of possible sub-graphs or motifs based on several simulations of the colon cancer network. Subgraph or motif detection is a separate and complex algorithm, which is far from trivial. Commonly used tools to detect motifs or sub-graphs in a network are *mfinder*, Fanmod and MAVisto, [Kashtan et al., 2002; Schreiber. Et al., 2005; Wernicke et al., 2006]. Among those, mfinder is preferred due to its reliability, ability to compute motifs of up to size 8 and also its fast in-depth search of the possible sub-graphs for a given input matrix or relational data. The mfinder tool is written in C++ and utilizes a tree search algorithm and random sampling to identify motifs of a specified size. Here, the most commonly occurring motifs of sizes 3 and 4 are investigated in all the simulated matrices. The aim is to help identify the most affected pathways leading to alterations in the state of gene expression in colon cancer. The motif analysis in this case involves 3 stages – (i) identify the possible connections or patterns in motifs of size 3 and 4 that can be detected in the whole network of colon cancer (Figure 6.1), (ii) make a list of the nodes/genes involved in forming

[9]The Ras-pathway in humans involves activation of certain genes to produce corresponding proteins which in turn activate genes responsible for cell growth differentiation and survival.

those patterns detected, and finally (iii) report corresponding statistics on frequency of the possible configurations[10] in all 270 simulated network samples.

Note: Steps 1 and 3 are discussed below, while step 2 involving a detailed list of the nodes that follow the connections possible is given in a file called "countmotif.xls" along with this thesis document.

S.no	Motif Size	No. of Patterns	Pattern Number–Count of Configurations following the pattern
1	3	10	1-804, 2-137, 3-7, 4-206, 5-53, 6-9, 7-3,8-1, 9-1, 10-1
2	4	66	1-2542,2-1374,3-10,4-661,5-990,6-747, 7-1,8-205,9-19,10-13,11-30,12-7, 13-20,14-11,15-42,16-35,17-5,18-36, 19-1,20-2,21-9,22-90,23-39,24-7, 25-4,26-27,27-6,28-3, 29-7,30-5, 31-30,32-11,33-2,34-2,35-39,36-17, 37-2, 38-874,39-251,40-10,41-27,42-18, 43-2,44-1,45-32,46-3,47-3,48-1, 49-4,50-3,51-2,52-1,53-2,54-1, 55-1,56-1,57-1,58-1,59-1,60-1, 61-1,62-1,63-1,64-1,65-1,66-11

Table 6.2: Genetic Subgraphs/Motifs Details
This table summarizes information about motifs with sizes 3 and 4. Total types of patterns (connections) were found to be 10 and 66 for each size. Number of configurations following each pattern is reported here.

In order to simplify the analysis of the results in this section, definition of some formal terms used henceforth is given as follows.

1. **Motif**: A motif *(m)* is a subsection or part of a whole network or graph (*[V,E]* $\in$ *G*), whose

[10] where different genes represent the same node in a motif of a specific size to form a configuration

number of nodes and edges, ($[v,e] \in [V,E]$) are less than the total number of nodes and edges of the whole graph. A motif is alternatively referred to as a *subgraph* here. In this section, we only handle motifs of size 3 and 4, (totally number of nodes = 3 or 4; number of edges in the motif can vary but within the range = [2,6] for size 3 and [4,12] for size 4).

2. **Patterns**: The term pattern in this chapter refers to the direction of edges that connect each node in a motif of size 3 or 4. (Refer to Figure 6.5 in this chapter and Figure D.2 in Appendix D to see patterns with a motif size of 3 and 4 respectively.)

3. **Configuration**: A configuration contains 3 or 4 nodes based on the motif of focus and belongs to only one pattern type. More than 1 Configuration can belong to a specific pattern but differ from each other based on the gene that forms the node in the pattern. The reverse of more than 1 pattern associated with 1 configuration is possible but NOT considered in this analysis. Example of genes that form two configurations to follow pattern 1, (connection of pattern 1 given in Figure 6.5) are; configuration 1 - MLH1 (gene 1), APC (gene 2) and BMP3 (gene 3), configuration 2 - MLH1 (gene 1), CDKN2AP14 (gene 2) and HLTF (gene 3).

The first step consisted of converting the Table D.1 in Appendix D, to an appropriate input file and analysing possible patterns or connections that could exist in a motifs of size 3 and 4 using mfinder. Consequently the mfinder tool reported 10 types of patterns among motifs with size 3, and 66 types of patterns among motifs with size 4, (Figure 6.5 and Figure D.2). The total number of configurations (of motif size 3) that followed one of 10 patterns was 1223 and those that followed one of the 66 patterns was 8297 (of motif size 4), as reported in Table 6.2. In the following subsections, we report and analyse, (i) the genes that form a node in specific configurations, (highest and least incident), reported by mfinder for the StatEpigen Network file, (given in Figure 6.1), (ii) frequency of appearance of each possible configuration in all our simulated matrices, (of

the network). The frequency of configurations of motif size 3 and 4 is reported in Figures 6.6 and 6.7. The configurations in these figures are arranged in the increasing order of display the patterns they follow, (**for example, the configurations that follow pattern 1 come first followed by configurations that follow pattern two, etc.**).

Possible types of Patterns in Motifs with Size 3

The size of the top three largest groups of configurations associated with pattern types 1, 4 and 2 are 804, 206 and 137 respectively, (Table 6.2 and Figure 6.5). Hence we explore and report the common genes that occupy specific nodes in the configurations following these patterns alone. MLH1 gene at node 1 constitutes 94% of the first 804 configurations, which follow pattern type 1. The number of configurations that follow pattern 1 and either contain CDKN2AP16, MGMT or CRABP1 at node 1 are 24, 19 and 9 respectively.

Of the given 137 configurations that belong to pattern 2, the most of them contain genes RUNX3 and CRABP1 as node 1. In the 204 configurations that follow pattern 4, genes at node 1 are APBA1, APBA2 and CACNA1G each forming 26, 24 and 16 configurations respectively.

Frequency of Configurations based on Motif size 3

Figure 6.6 shows the average of the occurrence, (or frequency expressed as %) of all 1223 configurations of size 3. This average is calculated over all 270 matrices, (9 types of input genes sampled for 30 generations). A general trend observed, (although individual results for all 9 inputs not shown here), was that some configurations were consistently chosen over others in all samples irrespective of input genes, (maintaining a high average of 88% for frequency of occurrence for some configurations), provided to test the changes or appearances of edges in the network. Samples with input genes such as MLH1, CRABP1 and RUNX3 generated networks, (frequency of configurations during specific input genes are not shown here), containing the same set of configurations

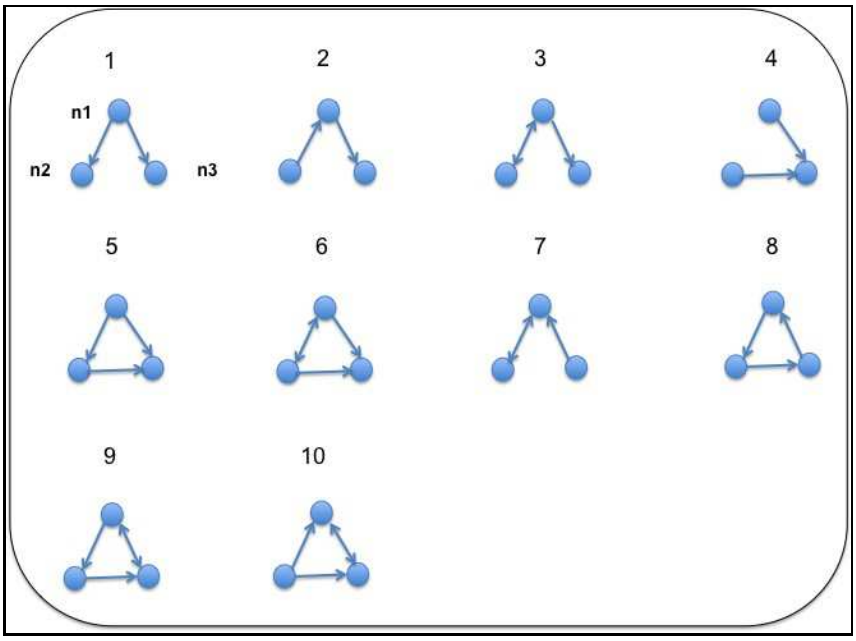


Figure 6.5: Motif Dictionary

Type of patterns formed with 3 nodes found in our colon cancer Network. The total number of configurations with different nodes that follow one of these patterns is 1223.

chosen repeatedly hence normalizing their frequencies to a value above 90% in Figure 6.6. These configurations mostly contained MLH1 as node 1 or node 2. In other trials with input gene TIMP3, those configurations had a slightly lower frequency of only 75%. This is because TIMP3 has the longest path length to the MLH1 gene, (refer Table 6.1) not allowing the configurations containing MLH1 to appear so frequently. Hence it is obvious that configurations containing MLH1 as node 1, especially from analysing those that follow pattern 1, have a very high incidence across all sampled matrices (average close to 90%).

This is because, MLH1 as mentioned before, has a high outdegree, which allows it to be a part of a lot patterns, detected by mfinder. This also reflects on the overlapping fact obtained from the literature that hypermethylation of MLH1 plays a crucial role in colon cancer progress. For example, configuration 7 which following pattern 1, consist of genes MLH1, BMP3 and CDKN2AP16 (Arf) as nodes 1, 2 and 3 respectively has been most frequently detected in all the model simulations. Results from these simulation agree with evidences in the literature that hypermethylation of MHL1 appears along with hypermethylation of BMP3 and CDKN2AP16 during colon cancer, [Zou et al., 2007]. We discuss general characteristics of configuration following each pattern for size 3, but more information about the configurations and their connectivity pattern is given in the support file, "countmotif.xls", as mentioned previously.

1. (Configurations with Pattern 1):

   The first 280, among all possible 1223 Configurations, which follow pattern 1, are formed by key genes such as MLH1 and BMP3 and are highly frequent and visited. The third node in these Configurations was represented either by CDKN2AP14, P16, or CHFR. To supplement the observations recorded so far, information from the literature states that methylation of MLH1 and BMP3 occur together during carcinoma stage of colon cancer, [Koinuma et al.,

Figure 6.6: Edge Analysis IIIA

Frequency of all possible configurations based on Motifs of size 3. The average over all 9 types of Input genes in 30 generations is reported here. X-axis is number of configurations (total=1223) and Y-axis average of frequency of visits for each of the configurations, (expressed in percentage). The configurations in these figures are arranged in the increasing order of the patterns they follow, (**for example, the configurations that follow pattern 1 come first followed by configurations that follow pattern two, etc.**)

2005]. Other configurations that count from 370 and 392, which follow pattern 1 and are highly visited as well. This is because all of them containing CDKN2AP16 at node 1. This specific gene was previously established as highly influential "marker" gene in the colon cancer network, [Zou et al., 2007]. However, an unstable set of frequency values among consecutive configurations (resembling sinusoidal graph) is observed especially from 423rd to 801st configuration. This is because configurations with PTPRO gene as node 2 or 3 were not preferred in any of our simulation samples. Although the probability of hypermethylation of PTPRO is recorded to be quite high (0.833 from Table D.1), the model did not favour these types of configurations.

2. (Configurations with Pattern 2):

   Of all the 140 configurations, counting from 804 to 945, that followed pattern 2, none were visited in any of our simulations. Among them, the genes that formed node 1 were CRABP1 and RUNX3.

3. (Configurations with Pattern 3):

   Configurations following pattern 3 that contained CDKN2AP16 at node 1 and CDKN2AP14 at node 2, were observed to have high occurrences. This observation was recorded in all simulated networks tested with different initial input genes. This Configuration (following pattern 3) reported by the model, has already been recorded in the literature where methylation of P14 and P16 are given as probable indicators of invasive colon cancer, [Hibi et al., 2002].

4. (Configurations with Pattern 4):

Configurations that followed pattern 4, although second highest in number, had a very sparse occurrence in sampled network/matrices. Only 2 configurations following that pattern, containing MGMT as node 2 and APC as node 3, were incident in all result datasets. The reason for the model to prefer those 2 configurations following pattern 4 alone was because of high frequency values associated with hypermethylation and mutation of MGMT and APC genes.

5. (Configurations with Pattern 5):

   Configurations, following pattern 5 had a high frequency in all samples if they had MLH1 as node 1 and APC as node 3, or just CRABP1 at node 1. An interesting fact from the literature indicates that correlation between MLH1 hypermethylation and APC mutation is less which contradicts our simulated results, [Vogel et al., 2009]. The reason could be because the authors of this article, [Vogel et al., 2009] reported from 1 type of experiment, (experiments conducted on a single line of cells). Contrastingly, the hypermethylation of CRABP1 gene is a common event recorded during colorectal cancer as also indicated in our simulated results, [Ahlquist et al., 2008].

6. (Configurations with Pattern 6):

   The configurations following type 6 were again least preferred when they containing JCVT as node 3. This unexpected behaviour is probably due to the fact that JCVT is expressed through alternative pathways, (activation of genetic/epigenetic events connecting JCVT and other genes in the network), and hence edges in these motifs do not appear at all.

7. (Configurations with Pattern 7):

Configurations that followed pattern 7 were again not detected at all in our model simulations, inspite of the fact that nodes 2 and 3 were formed by P14 and P16 respectively. This is another contradictory observation involving these two genes in the model simulation as the literature indicates that P14 and P16 are actively involved in colon cancer progress, [Hibi et al., 2002].

8. (Configurations with Pattern 8 & 9):

   Edges of Configurations that follow these patterns are highly visited, as there is only one motif under each type. The genes involved in configurations following both patterns 8 and 9 are - MLH1, P16, RUNX3, CRABP1, TIMP3 and P14. All these genes are known to have a very important role in colon cancer, as indicated from these simulation trials and also information from the literature, [Ahlquist et al., 2008; Hibi et al., 2002; Zou et al., 2007].

9. (Configurations with Pattern 10):

   Only 1 configuration follows pattern 10 and this was not visited in the samples. This is only because, there are many edges connecting the 3 nodes, (bi-directional edges), which reduces the chance of observing all the edges together to form this type of pattern in simulated random networks.

Summary of Motifs with size 3

In summary, Configurations, (irrespective of the pattern they follow), least and highly preferred in the simulated networks, had their own set of common features which is listed as follows,

1. Features of least preferred configurations:

   (a) JCVT as node 2 or 3.

(b) BRAF as node 1.

(c) PTPRO as node 3.

(d) ADAMS19 as node 2.

(e) APC as node 1 or 2 with TP53 as node 3

(f) combination of MLH1 as node 1 and DAPK1 as node 2.

2. Features of highly preferred configurations:

(a) MLH1 as node 1 or 2.

(b) CDKN2AP14 and P12 as either node 1 or node 2.

(c) MLH1 as node 1 and MGMT as node 2, and vice verse.

The relation between genes as common characteristics of configurations highly preferred and the probability of occurrence associated with those genes have been discussed above. The configurations that are least preferred sometimes do not appear because they contain edges that have least probability of occurring. For example, JCVT has one of the highest number of incoming edges, but the probability of most those edges is less than 30%. This reduces the chance of observing the configurations with JCVT at node 3 present in configurations with pattern 1. Similarly, probability of APC and TP53 mutation occurring together is higher that 50% according to StatEpigen, but is not observed at all in any of our simulation. It is also interesting to note that the authors of a research experiment reported in 1998, [Smits et al., 1997] reported that TP53 mutation did not occur when a specific type of mutation associated with APC was recorded during colon cancer. This fact is in concord with our simulated results, although the nature of this evidence is very specific with respect to the type of mutation in APC gene,[Smits et al., 1997].

Possible types of Patters in Motifs with Size 4

As previously mentioned, there are 66 patterns and 8297 configurations that follow those patterns. The details about each pattern is given in the Figure D.2 in Appendix D. Figure 6.7 contains the average of visitation of these configurations, (expressed in percentage), in all 270 adjacency matrices. In general, it was found that configurations that follow patterns 1, 2 4, 5, 6 and 38 are highest in number, (as indicated in Table 6.2), although they do not directly correlate to their frequency of appearance in our simulated matrices. In order to simplify the results, we discuss about only those genes forming configurations that are most and least preferred in the sampled matrices.

Frequency of Configurations based on Motif size 4

Configurations that were detected based on motif with size 4 are similiar to the configurations based on motifs with size 3 detected previously. In other words, configurations of size 3 could be detected within larger configurations that followed patterns involving 4 nodes. Since it is not possible to explain the statistics of configurations following each of the 66 patterns, the Figure 6.7 has been marked with 5 regions of highly incident configurations, (among all 8297) whose average frequency of occurrence is $> 75\%$. Hence information about the incidence of configurations and which patterns they follow, (located within each of the 5 dense regions) are explained as follows.

The configurations that have a high frequency irrespective of the patterns they fall under, do share some genes like MLH1 at node 1. Similarly, a list of genes was commonly found at nodes in configurations that were least preferred in sampled matrices. The dense regions contain configurations counting from – 1 to 500, 1090 to 1550, 2510 to 4500, 5600 to 7000 (with a brief dip around 6000-6200) and 8210 to 8300. Configurations found in one of the dense regions were also found to belong to one of these pattern types (Large Occurrence Types or LOT) ~1-4, 6, 8, 9, 15, 16, 18, 21, 23, 27-30, 31, 34, 37, 46, 47, 49-50, 56-57, 59-66. It was very interesting to note that configu-

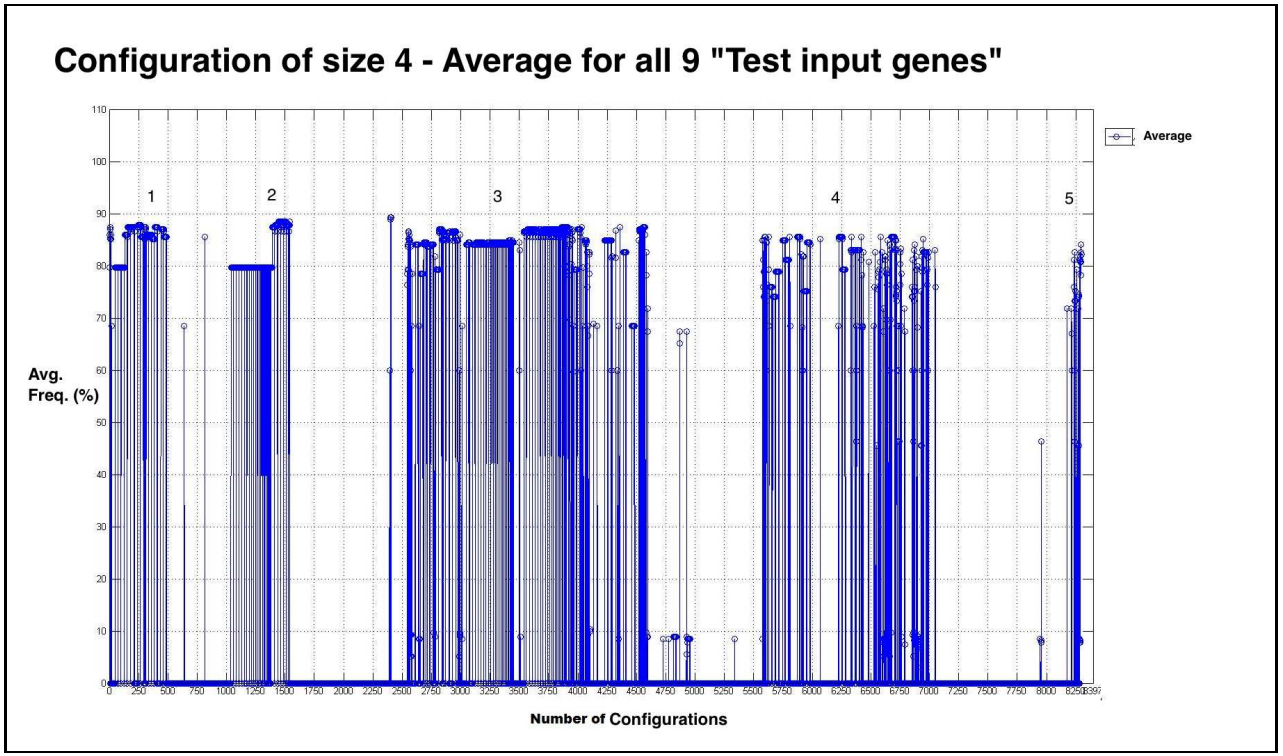


Figure 6.7: Edge Analysis IIIB

Frequency of all possible configurations based on Motifs of size 4. The average over all 9 types of Input genes in 30 generations is reported here. X-axis is number of configurations (total=8297) and y axis is the of average visits over all sampled matrices, (expressed in percentage). The configurations in these figures are arranged in the increasing order of the patterns they follow, **(for example, the configurations that follow pattern 1 come first followed by configurations that follow pattern two, etc.)**

rations following rest of the patterns, (those not listed in LOT), had the lowest or zero frequency of appearance in all trials. The description of genes forming highly visited configurations, (following their corresponding patterns) are listed below.

1. (Dense Regions 1 & 2):

   The Dense regions 1 and 2 contained only configurations following pattern 1, and amongst these, 97% of the configurations highly incident in the samples contain MLH1 as node 1 followed by APC as node 2 (from 1^{st} to 2542^{nd} motif in Figure 6.7).

2. (Dense Region 3):

   This region included configurations following patterns 2, 3 and 4. Among those following pattern 2, which were counted from 2543^{rd} to 3916^{th}, again contained mostly MLH1 as node 1 and either RUNX3 or CRABP1 as node 2 in Figure 6.7. Similiarly configurations based on pattern 4 were numerous but the ones that contained MLH1 as node 1 were alone detected in the result samples.

3. (Dense Regions 4):

   This dense region contained configurations following patterns– 6 ,8, 9, 15, 16, 18, 21, 23, 27-30, 31, and 34. The configurations based on pattern 6 was similiar to that of pattern 4 with more than 50% of the detected configurations containing MLH1 as node 1 and MGMT, CDKN2AP14, P16, CRABP1, DAPK1 or TIMP3 assigned as node 2 amongst all other configurations detected. Although more than 800 possible configurations were detected to follow pattern 38, these motifs hardly appeared during the simulation similiar to that of configurations following pattern 5.

4. (Dense Regions 5):

This region contained highly occurring configurations, which belonged to patterns 46, 47, 49, 50, 56, 57 and 59-66. Configurations following patterns 53 to 55 are less in number, but again have a high incidence. Irrespective of the directions of edges associated with all 4 nodes found in these motifs, most of them have CDKN2AP16 as a node 1. This implies that, given the structure of these configurations, the effect of p16 in many pathways is crucial as mentioned in the literature, [Goto et al., 2009]. There is a strong fluctuation in the level of frequency of these configurations in all 9 datasets, indicating that a gene can be affected by more than one pathway. Calculation of visitation statistics of all configurations based on motifs of size 4 are not elaborately explained but provided in Appendix.

Frequency of Motif Occurrence - Size 4

The brief regions of gaps in the middle of dense regions 3 and 4 are very apparent. On comparing configurations based on motifs with size 3 and 4, we found that both types shared some common features. The results indicate that the configurations (size 4) did not strictly appear if they had one or more of the following genes present in their structures. Hence the list of these genes that form the least preferred configurations is presented here.

1. JCVT gene as node 3 or 4,
2. JCVT and DAPK1 genes as node 2 and node 3,
3. BRAF as gene node 4
4. ADAMTS19 gene as node 3
5. PTPRO as node 4, (similiar to the observations in sub-graphs of size 3.)
6. HIN1 gene as node 4,
7. MLH1 as node 1 and DAPK1 as node 2 (together)

While some common genes listed above have been already found in configurations of size 3, there are other exclusive genes present in configurations, of size 4 which were least preferred. The reason for the appearance of these common genes is still not clear, as the probability information in relation to the events associated to these genes did not directly correlate well with these findings. Hence we strongly believe that during colon cancer, pathways or occurrence of genetic events involving these genes are least preferred to initiate cancer growth under sporadic cases. The number of patterns and configuration for this network that can be derived if their size is more than 4 nodes is prohibitively large hence they are not discussed here.

Genetic/Epigenetic Events associated with Adenoma Stage of Colon Cancer

Adenoma or stage II in colon cancer precedes carcinoma and is characterized by malignant polyps that grow in the intestinal walls. The cancer condition spreads across the entire wall of human intestine but not to the lymph nodes. While the role of MLH1, MGMT and other tumour suppressors have been explored through this 2-part modelling framework, the focus of study was only carcinoma stage. Further analysis on the role of these genes observed during adenoma stage, (alternatively addressed as adenocarcinoma) from StatEpigen database gives a slightly different picture. The carcinoma stage was characterized by hypermethylation of genes such as MLH1, MGMT and P14/P16 although only a few among these genes were aberrantly modified during the adenoma stage. A major observation in adenoma stage however was the interactions among the TP53 and APC genes, which were prominently noted in carcinoma too. The other genes that were mutated during adenoma following the hypermethylation of MGMT and CDKN2AP16 were KRAS, BRAF and HLTF. The literature indicates that mutation of both KRAS and BRAF is one of the most common events through all stages of colon cancer, [Weisenberger et al., 2006]. Hypermethylation of BMP3 was also another notable epigenetic event in adenoma, which persisted in Carcinoma. Hence it is obvious that some epigenetic changes occur in the early stages of colon, which con-

tinue to drive the aberrations across the genome and spread the malignancy within the colon and to different tissues in the body. While the network was tested for epigenetic events associated with important tumour suppressor genes and those which protected the cell from proliferation and growth, it is also possible to confirm the occurrence of these events during adenoma stage of colon cancer. The immediate set of steps as part of the future work involved in application of EpiGMP to study cancer networks are elaborated in next section.

6.4 Conclusion & Future Work

In this Chapter, an analysis was performed which sought to comprehend not only the molecular changes but also predict the possible occurrences of specific genetic/epigenetic events, for carcinoma stage of colon cancer. Layer 1 consisted of applying a previously developed epigenetic tool to predict histone modifications and DNA methylation level. Outputs of the inner layer consisted of deciding the events across network of genes actively known to be involved in colon cancer. The aim was to explore the possible number of events, patterns and chain of genetic/epigenetic events at cell level for a sporadic condition of colon cancer. The number of events that were known to surely occur were reported in Edge Analysis I, followed by Edge Analysis II involving the calculation of Hamming distance of dissimilarity to understand when the edges appeared, (with the simulation being carried out for 150 generations).

Finally, (in Edge Analysis III), an attempt was made to investigate the preferred combinations of genes expressed, methylated or mutated by reporting patterns among motifs or sub-graphs (of size 3 and 4), in all the simulated samples. The configurations with nodes represented by genes such as MLH1 (node 1), CDKN2AP14, CDKN2AP16, RUNX3, CRABP1 (as node 1 or 2) and APC (as node 2 or 3) were most preferred irrespective of the pattern they followed. The reason for the strong association of these genes in such pattern formation is that they all have a comparatively large

outdegree. The significance of these genes, especially a crucial marker such as mutation of APC to trigger malignancy in colon is also demonstrated in our results, [Goto et al., 2009; Hibi et al., 2002; Koinuma et al., 2005]. In fact the most frequent marker in any type of cancer is the mutation of TP53 gene. Although this specific event appears in our simulation, the pathways chosen to trigger this event is more than 1, (as indicated in a number of conditional events from StatEpigen). It is also interesting to note that motifs associated with TP53 mutations and APC mutation were least preferred by the model, in contrast to the information from the literature. We also reported a list of possible genes and associated events that had the least probability of occurring during carcinoma stage. It is possible that these features could appear for a different stage of colon cancer or another cancer type. A model for the prediction of probable events for colon cancer has not been developed before, especially in terms of recurring gene relation structures or chain of important events in the cancer scenario. This is one of the first crude attempts aimed to analyse both molecular and also cellular changes happening over several generations of cell division.

Cancer is not a simple disease, with many complex factors involved in propagating it to different stages of development. In attempting to extend the model analysis and interpretation of the molecular and gene level changes, other factors known to affect the colon system can be implemented, (theoretically) – including lifestyle, eating habits, age and cancer history of a person. This initial attempt aims to show how information from different levels may be integrated to highlight and interpret abnormal events during Colon Tumorigenesis. The genes that were extracted from the epigenetic database are also known to be involved in other types of cancer, (for example, APC is known to be involved with BRCA1 in advanced stages of breast cancer [Bahar et al., 2001]). The same framework can also be applied in future to study other stages of colon cancer and malignancies originating from other systems such as - circulatory system, breast, pancreas and liver. Our

final aim in the near future is to combine and model each of the actual complex biological layers, (involving – genetic molecules, gene expression data, proteins involvement, cellular behaviour, and consequently phenotypic changes in tissues and organs) using the tools we developed so far along with further information from the literature, and hence investigate the dynamics of cancer progress and spread in the human body.

Chapter 7

Chromatin Remodelling

7.1 Introduction

In mammalian cells, evolution has led to the efficient packing of a large number of DNA molecules around several nucleoprotein units (or nucleosomes) to form the structure of chromatin. A Nucleosome, which is the basic unit of this chromatin, consists of a core protein complex, (2 copies of 4 histones), along with a variable number of DNA molecules wound around the core and linking to the next nucleosome unit. Based on factors such as number of DNA base pairs bound to each nucleosome and modifications in the histone proteins, the chromatin nucleosomes that are either widely spaced to form an open structure or more closely aligned to produce a compact chromatin structure, [Jiang et al., 2003]. This structural condensation not only helps to compactly pack the billions of DNA molecules within the human genome but also has three more advantages. Firstly, protection from radiation induced damage and genotoxic stress, secondly, induction of a stable "kinetochore"[1] in centric and pericentric regions made of satellite sequences and lastly provide "an additional level of control over the process of transcription", [Gilbert et al., 2005]. Although a lot of research has been carried out to investigate the modifications and positioning of nucleosome

[1] kinetochore are chromatid proteins which help to form spindle fibres during cell division.

in the chromatin, little is known about the preferred higher order structures or factors influencing the chromatin as a whole. In this chapter, the relation between histone modifications, nucleosome positioning, and their combined effect on the overall chromatin structure is explored. Using the Epigenetic micromodel, a simple agent-based modelling framework is proposed in order to account for the epigenetic molecular signatures that are responsible for controlling structural dynamics of the human genome.

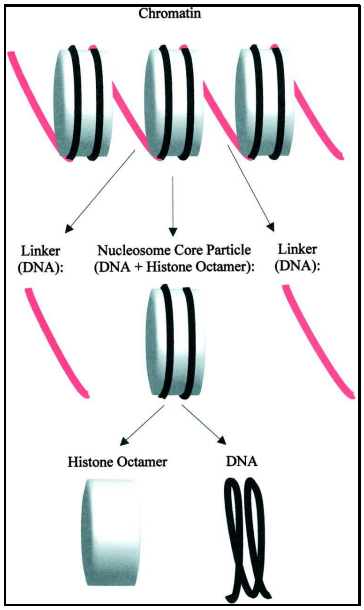


Figure 7.1: Nucleosome Linkage

Nucleosome forms the basic unit of chromatin. Each unit consist of 200bp in human genome, and has a core of 147bp wound around the histone protein complex, the rest forming a linker DNA which could vary in length under different circumstances. Image adapted from [Schrem et al., 2002]

7.1.1 Chromatin Organization

The histone core complex has 147 bp of DNA molecules wound around itself by establishing 14 non-covalent bonds, [Alberts et al., 2002; Paponov et al., 1980]. The linker DNA, which connects

to other nucleosome units, varies in length for different situations such as, – (i) during different events associated with chromatin packing in a cell, (ii) in different cell types of the same organism and, (iii) between different types of species. In humans it is known that each nucleosome unit has access to 200 base pairs, (147bp + variable linker DNA length) of DNA molecules. The linker histone (one copy of H1 type histone in humans), controls the exit and entry of the DNA double strand around the nucleosome core by accessing an additional 20 bp, of the linker DNA molecules (apart from the core 147 bp), during chromatin packing. Figure 7.2 shows the components that form the basic unit of Nucleosome inside the chromatin. Under low salt concentrations, the chromatin appears like a "beads on a string" linear structure, which is often associated with an open or a lenient packing. The wrapping of DNA molecules, allows condensing the sequence 1.65 times, (when one nucleic acid is approximately 0.34 nm in length and wraps around the nucleosome with a diameter of $\sim$ 11 nm). In the presence of favorable conditions, this linear chromatin condenses into a helical rearrangement of nucleosomes to form the 30 nm chromatin fibre, [Woodcock et al., 1993]. The strength and compactness in the fibre is determined by "DNA-histones" protein interactions between nucleosomes. The dynamics of this 30 nm fibre is of specific interest to researchers as recent findings indicate that this fibre has "physiological relevance" in local regulation of DNA associated metabolic pathways, [Szerlong et al., 2011]. The complex structure of the fibre is often referred to as "ubiquitous" and secondary in nature [Routh et al., 2008; Szerlong et al., 2011; Woodcock et al., 1993]. Consequently, the validation of this structure remains controversial, despite the fact that many models representing the structure and orientation of this chromatin fibre have been proposed, [Szerlong et al., 2011]. Figure 7.2 depicts the components of chromatin that condense and form a closely linked and tightly packed structure. Consequently the following subsections describe in detail about the chromatin assembling process, formation of secondary fibres and models proposed to date.

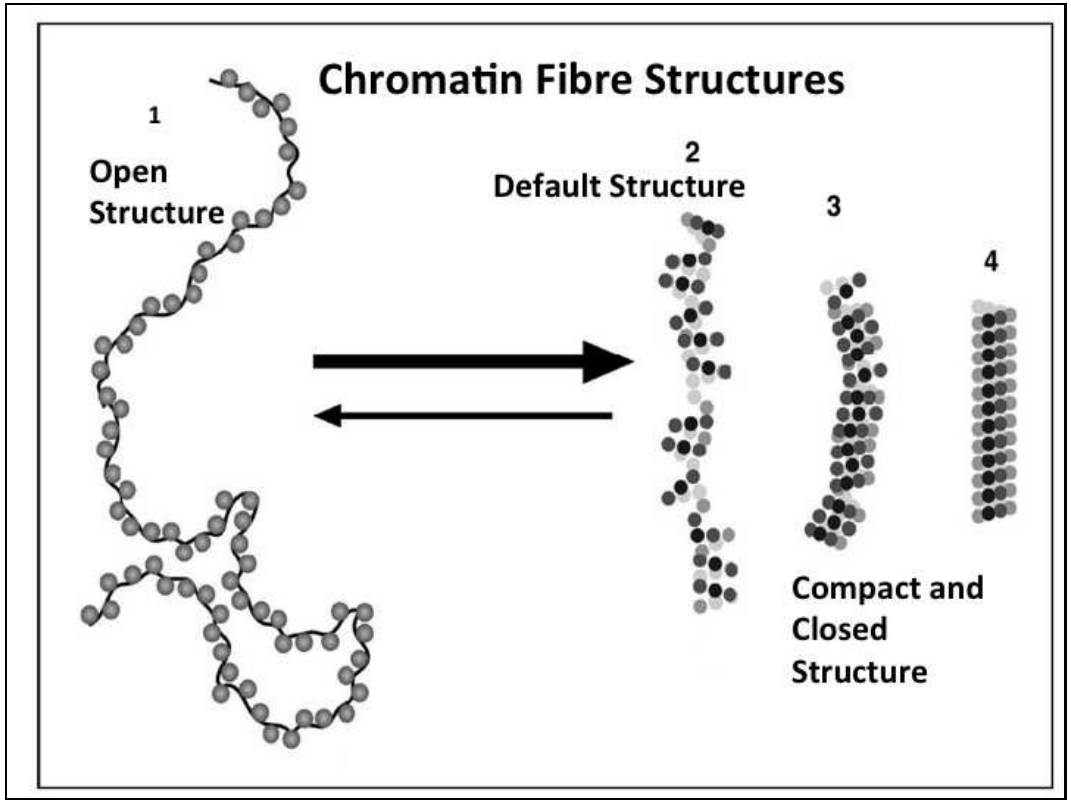


Figure 7.2: Chromatin Condensation

Process of condensation of Nucleosome divided into 4 parts, (Left to Right) - (1)Open "beads on a string structure", (2)Open chromatin structure forming a 30 nm fibre with discontinued condensation. (3) & (4) Refers to the formation of high order chromatin structure that is closed and compact in appearance, Image adapted from [Gilbert et al., 2005].

Role of DNA molecules and Histone Modifications

The DNA sequence content plays a major role in determining the position or placement of nucleoproteins in the chromosome. Previously conducted studies on yeast genome, reported that 10 bp alignment of alternating AA/TT and GC dinucleotides facilitate sequence-dependent positioning of the nucleosome complex, [Szerlong et al., 2011]. These authors also mention that sequence-dependent curvature of DNA can determine interactions between the DNA and the histone complex. In fact the loci of non-covalent bonds between DNA sequences and histone complex are the main targets for chromatin remodelling proteins which cause nucleosome repositioning[2], [Jiang et al., 2003]. Apart from the effect of DNA molecules (which are generally known to be negatively charged), some marker histone modifications can also help stabilize the corresponding open or closed chromatin structure. Laboratory-based experimental evidences have proved that a systematic and methodical assembly of chromatin remodelling components can help to unwind the compact fibre structure in order to allow the transcription of genes, [Ma et al., 2004; Strahl et al., 2000]. Modifications that aid in chromatin relaxing, during transcription initiation include – H3R3, R17 and R26 methylation, H3K4 methylation and H3S10 phosphorylation and H3K19 acetylation, [Ma et al., 2004]. Some researchers consider that chromatin remodelling factors are intermediate translators which work to change the secondary structure based on the histone modification observed, [Volpe et al., 2003]. While acetylation helps to relax the structure, histone methylation signatures and DNA methylation are known to pull the nucleosomes closer to form the 30 nm fibres. The methylation process in DNA sequences along with H3 K9 histone methylation modification instigate chromatin remodelling complexes such as MePC, (methyl-CpGbinding protein) that can induce a closed local structure in the chromatin, [Martinowich et al., 2003]. These facts make it a compelling case to investigate the process of chromatin structural dynamics and factors influencing it such as – sequence content and epigenetic events, (histone modifications and DNA

[2]repositioning refers to the sliding movement of a nucleosome along a DNA sequence

methylation). A brief review of theoretical models proposed so far to understand the structure of chromatin fibres is discussed below.

Chromatin Fibre Models

The folded 30 nm chromatin fibre, initially observed through X-ray diffraction experiments, is known to have a superhelical structure, [Ma et al., 2004]. Among the many models developed for the nucleosome packing in the fibre, the most widely accepted models belong to two classes such as – (i) zig-zag and, (ii) solenoid. The first suggests that the *helical conformation* is a "two-start" with a zig-zag nucleosome arrangement, each of which is connected to the next by a relatively straight linker DNA, [Woodcock et al., 1993]. The diameter of this chromatin is within the range of 25-30 nm, with a nucleosome packing density of 5-6 units after every 11 nm of length. The nucleosomes are allowed to interact closely with every alternative $(i+2)^{nd}$ that stacks on top of each other. A further refinement was proposed to the zig-zag model resulting in a variant called "cross-linked" model. The most prominent difference between the two sub-type models is the orientation angle of linker DNA which is inclined between 0-50 degrees along the length of the chromatin axis in the former, and approximately perpendicular in the cross-linked model, [C. Wu et al., 2007]. The second class is a one-start solenoid model, which differs from the former in topology, dimension and nucleosome packing density. This model has a diameter ($>$ 30 nm) and stacks up to 6 nucleosomes in 1 "gyre" or twist. This stacking allows nucleosomes to become closer to every $(i+6)^{th}$ unit, [Routh et al., 2008] due to bending of linker DNA initiated by linker histone in the solenoid structure. This packing in a simple manner allows "nucleosomes of adjacent gyres become interdigitated", [Routh et al., 2008]. In the framework proposed in this chapter, the agents or nucleosomes adopt the two-start zig-zag model because experimental evidence based on the properties and structure of the 30-nm fibre is more inclined to support this type of model. Hence in order to link the spatial movement of nucleosomes and epigenetic changes, EpiGMP tool

is applied to a group of nucleosomes, (forming a gene), to assess the interactions between HM and DM and consequently decide if the zig-zag structure should be adopted or not. Information about evolution of temporal component (application of EpiGMP), and spatial component, (adaptation of zig-zag model), and how these influence one another inside the model framework is given in the following subsections.

7.2 Proposal: Agent-Based Model to study Chromatin Dynamics

Agent-based Modelling is a class of modelling techniques that attempts to mimic behaviour and interactions among a group of autonomous agents in order to understand their effect on a system as a whole. This type of modelling helps in understanding the underlying complexity in any reducible system made of smaller units. Agents in a system are recognized to have attributes such as *autonomy, Social behaviour, Reactivity* and *Proactivity*, [Jennings et al., 1998; Perrin, 2008]. If agents exist in a closed system, they undergo restricted temporal and spatial changes based on interactions with their peers alone. If in an open system, the agents behaviour is also affected by environmental influences. This is the first time an agent-based modelling approach is proposed and explored in order to study chromatin structure. The structure of the 30 nm chromatin as mentioned before is an important determinant of the human genome transcription process. Since nucleosomes, made of core proteins and a variable number of DNA molecules, are the basic units of this fibre, each nucleosome is assigned as an agent. These undergo epigenetic changes over time, (or temporally), and as a result rearrange their loci with respect to their peers to form an open- or closely-packed chromatin fibre. Chapter 4 in Part I gives details on the working of the EpiGMP tool. In this chapter, the same tool is applied to report the expression levels of genes and hence predict the movement of nucleosomes. To elaborate more, the EpiGMP decides histone modifications and Transcription levels, which is allowed to influence the consequent movement of nucleosomes in a Euclidean 2-dimensional plane. Since the dynamics of chromatin resembles a spring (during expansion and

compression movement), the Kamada-Kawai, [Kamada et al., 1989] algorithm is applied to decide the resultant spatial parameters, (displacement along the 2-D plane). This force-directed algorithm has been the most widely accepted and applied method in graph or network drawing. Here, a graph theoretic approach is applied to calculate an energy function, (or Potential Energy of Spring) which is a product of the spring strength and the relative displacement of a node in Euclidean plane, (or agent in our case). Here, the strength of the spring is calculated by considering each nucleosome as an individual contributor to the dynamics of the whole chromatin. The strength is derived from utilizing DNA methylation or average acetylation levels for each nucleosome, (reported by the EpiGMP model).

Agents in a system follow a specific set of rules or instructions to closely resemble the real system being simulated. The rules or assumptions followed to model the chromatin structure are given as follows,

1. The number of DNA molecules per nucleosome is assumed to be 200 bp as indicated for human genome, [Tanaka et al., 2010]. Hence based on the length of the DNA sequence of an input gene, the number of nucleosomes is dynamically decided.

2. The linker DNA can vary but the number of core DNA molecules wound around the nucleosome is fixed at 147 bp.

3. Each nucleotide (A/T/G/C) on an average is assumed to have a length of 0.34 nm.

4. The chromatin structural dynamics are only tested for specific genes which are known to contain properly-phased nucleosome units.

5. In this primary version of the model, effects of either DNA methylation or Global acetylation modifications are alone considered to calculate the spatial movements.

6. The structural changes of nucleosomes along the 2-Dimensional Euclidean plane alone is reported here.

7. Inputs to the simulation require - DNA methylation levels ($\in$ [0,1.0]), Name of gene and Number of Linker DNA molecules to be preset.

The following section explains in details, the formulas for energy calculations and Kamada-Kawai algorithm used to control the structural dynamics.

7.3 Method

A very efficient graph theoretic-based approach was proposed by Kamada and Kawai in 1989, [Kamada et al., 1989]. The approach in this algorithm depends on setting a spring system scenario, where altering energy correlates to changing the Euclidean distances between nodes or agents or nucleosomes in this case. This model does not propose a force or energy of repulsion or attraction, but rather the change in distance between nodes relative to a default position assigned initially.

$$l_{i,j} = L * d_{i,j} \tag{7.1}$$

$$L = L_0 * max_{i<j} d_{i,j} \tag{7.2}$$

Here, in Equations 7.1, 7.2 and 7.3, $d_{i,j}$, corresponds to shortest distance between any two nucleosomes. The shortest distance is a product of length of a nucleotides (0.34 nm) and number of linker DNA molecules connecting the nucleosomes. The variable in Equation 7.1 - $l_{i,j}$ is the ideal length of a spring between i^{th} and j^{th} nucleosome where L is the desirable length. According to Equation 7.2, L_0 is the length of the side of the plane area which is takes the value of $d_{i,j}$ and $\max_{i<j} d_{i,j}$ is the maximum distance occupied by the total number of base pairs in nanometers.

$$k_{i,j} = K / d^2_{i,j} (or K')$$ (7.3)

$$K' = \text{Average Acetylation level} * (2 * (\text{DNA Methylation - Mean DNA Methylation}))$$ (7.4)

The most important substitution here is in Equation 7.3, where *K* refers to the DNA methylation level of a gene and *K'* refers to the average acetylation levels per nucleosome. The system permits the use of either *K* or *K'* during a simulation to calculate the energy values. If *K'* is preferred, values are obtained using the equation 7.4.

$$E_{i,j} = 1/2 * k_{i,j} * (|p_i - p_j| - l_{i,j})$$ (7.5)

$$E_{i,j} = 1/2 * k_{i,j} * \left((x_i - x_j)^2 + (y_i - y_j)^2 - 2 * l_{i,j} * \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}\right)$$ (7.6)

The energy value is calculated for individual nucleosomes in terms of epigenetic change in every unit which independently drive them closer or further from their neighbour. One also has to note that the energy values and consequent spatial movement of a nucleosome only depends on its previous neighbor and not any other structure. x_i and x_j refer to the x-coordinate values of i^{th} and j^{th} nucleosome while, y_i and y_j refers to the y-coordinate values of i^{th} and j^{th} nucleosome respectively, (given that j^{th} nucleosome is the predecessor lying to the left of i^{th} nucleosome).

7.3.1 Kamada-Kawai Algorithm

The Kamada-Kawai algorithm is slightly altered to allow a curtailed movement of the agents within the 2-Dimensional space. The system is initially assigned to a relaxed chromatin state with corresponding co-ordinates to the nucleosomes. The following steps explain the algorithm adopted in our model, for "n" nucleosomes.

Algorithm 7.1 Algorithm of the EpiGMP tool - Version 1

```
1: Apply EpiGMP model here
2: Calculate energy each nucleosome
3: for j=1:n-1 do                                   ▷ where n is the total no of nucl. objects
4:     i=j+1
5:     if i ≠ 1 then
6:         Compute d_{i,j} for 1 ≤ i ≠ j ≤ n;
7:         Compute l_{i,j} for 1 ≤ i ≠ j ≤ n;
8:         Compute k_{i,j} for 1 ≤ i ≠ j ≤ n;
9:         if Energy of i^th Nucl > Medium Level Energy then
10:            Nucl_i(x) += CLOSED HORIZONTAL DISPLACEMENT CONSTANT
11:            Nucl_i(y) += CLOSED VERTICAL DISPLACEMENT CONSTANT   ▷ Allow i^th
   nucl to move closer to its prev, (j)^th nucl  ▷ j^th nucl, (the predecessor) lies to the left of i^th nucl
   ▷ Position of 1^st Nucl is kept immobile and constant
12:        else
13:            Nucl_i(y) += OPEN VERTICAL DISPLACEMENT CONSTANT        ▷ Form a
   "Beads on String Structure"
14:        end if
15:    end if
16: end for
```

Unlike the original Kamada-Kawai algorithm, in this case, the displacement values along both axes are not derived from a pair of differential equations but assigned to be a constant. This step is adopted to control the movement of nucleosomes to form a helical zigzag arrangement or have a open "beads on string" conformation.

7.4 Preliminary Results

Information from the literature states that nucleosomes are not regularly placed to form a chromatin structure, [C. Wu et al., 2007]. Based on the gene expression frequency and DNA sequence contents, some parts in the genome are either devoid or contain completely phased[3] nucleosome structure. The three main inputs for this agent-based model are (i) name of gene, (ii) level of DNA methylation and (iii) length of linker DNA. Hence in the simulation reported here, we study the structural dynamics of the MGMT gene, based on the fact that differentially expressed genes, (such as MGMT) have nucleosomes regularly placed after a fixed number of linker DNA molecules along the sequence length. The length of linker DNA was set to 53 bp (or equivalently 18.02 nm, given that each bp is 0.34 nm). Eight nucleosomes agents were created depending on the length of the DNA sequences, (and assumption that each nucleosome has access to 200bp in human genome). Equations (7.1) to (7.6), help to understand how the required energy values are calculated given that K is the level of DNA methylation input. The change in this energy value for different inputs of DNA methylation levels and linker DNA is shown in Figures 7.4 and 7.3. Here, an attempt to only mimic the open or "Beads on a String", structure was achieved although explicit results of agent/nucleosome movement is not shown.

7.4.1 Energy with more Linker DNA

Figure 7.3, shows the progress in energy values calculated in the system for increasing levels DNA methylation (substituted for "K", as shown in Equation 7.3). This same energy was found to be decreasing when acetylation levels, (derived from equation (7.4)) were substituted for methylation, in Equation 7.3. The DM-based energy can be assumed as the attraction force that tried to bring the agents close to one another. On the other hand the energy calculated using acetylation levels, acts in the opposite direction in order to pull apart the closed packing and allow lenient

[3]The term phase refers to alignment of nucleosomes after regular and fixed intervals of base pairs.

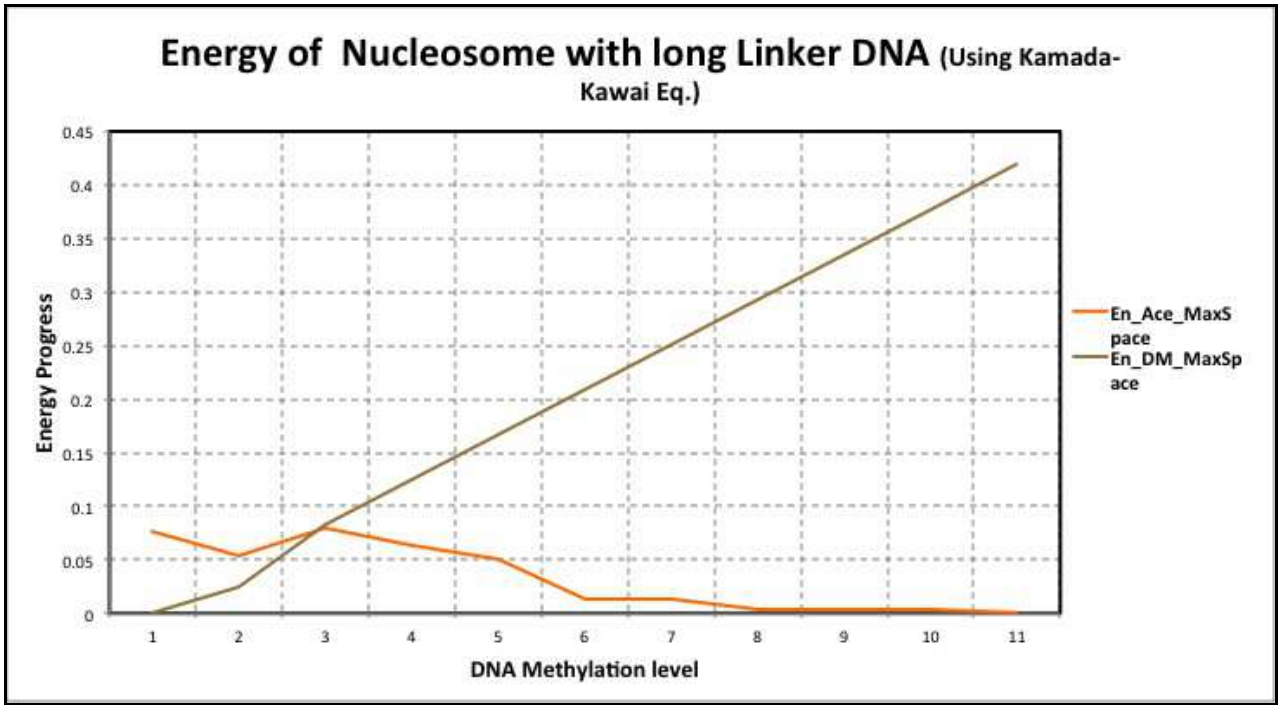


Figure 7.3: Nucleosome Energy (More Linker DNA)

Change in Spring Potential Energy for different values of DNA methylation and Avg. Acetylation for more Linker DNA. Green colour graph represents energy values calculated based on DNA methylation and yellow colour graph represents energy calculated using Average acetylation levels, extracted from micromodel outputs.

spacing among nucleosome agents. These contrasting energy values for different modifications is because of the theoretical interrelations established by EpiGMP tool, (refer Equation 4.5 and 4.6 in Chapter 4). i.e. The required behaviour such as increasing levels of DNA methylations, ($\in$ [0, 1]) reduced the global acetylation levels profoundly. The same observations have been reported from experiments conducted by groups investigating chromatin associated epigenetic changes, [Ma et al., 2004; Patra et al., 2008]. For example, The authors of [Ma et al., 2004], studied the nucleosome dynamics of the human MMP-9 gene, and reported that low levels of histone acetylation modifications were found in the closed or inactive form chromatin. However the onset of HAT, (*Histone acetyle transferases* - enzymes that add acetyl molecules), created an imbalance in the electric charge around the nucleosome forcing it to unwind and open so as to allow transcription factors to bind.

7.4.2 Energy with less Linker DNA

Although the length of linker DNA does not affect the compactness of the chromatin fibre it is known to affect the angle of nucleosome orientation in the fibre to suit the tight packing, [C. Wu et al., 2007]. In our successive simulation, this idea was tested by setting the linker DNA length to 0 bp. In Figure 7.4 changes to the energy for different values of DM are alone shown. It should be noted that the slope for this graph is constant relating the energy values directly to different DNA methylation values, (except that the range of resulting energy values are quite different.) Energy values are not calculated using acetylation modifications as such a state does not exist for an open chromatin structure.

7.5 Conclusion & Future Work

This chapter explored the relationship between epigenetic changes and chromatin dynamics in human genome. An efficient framework to model nucleosome movements in Euclidean space and

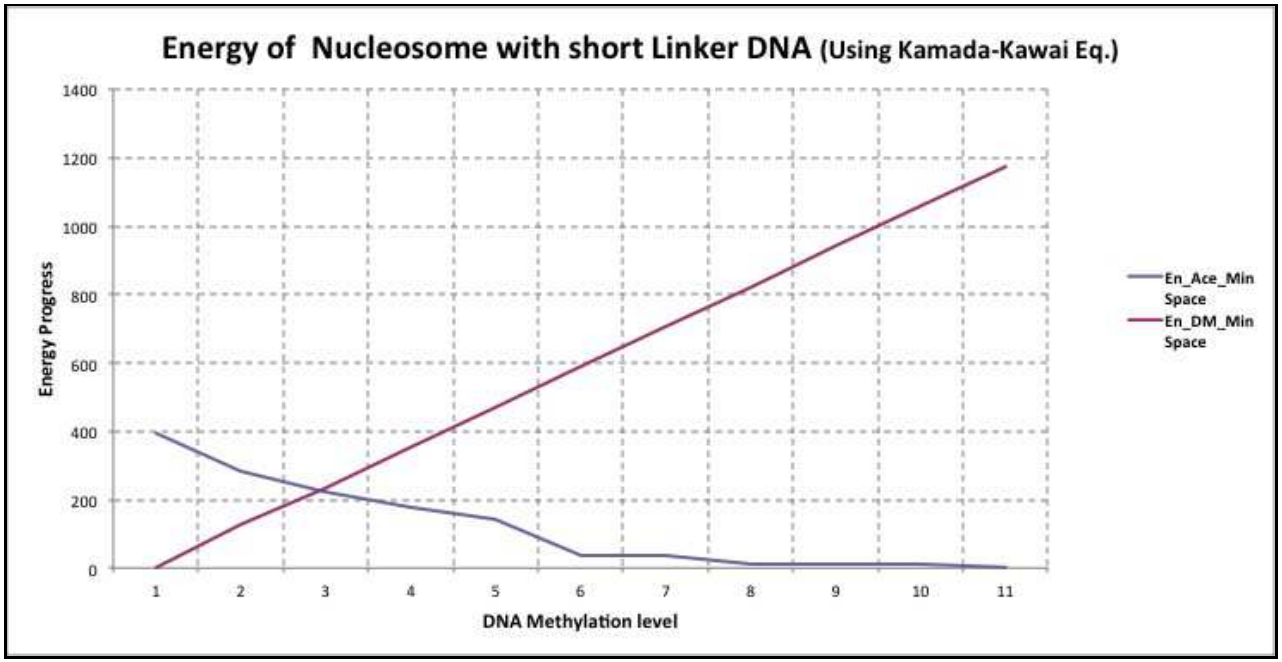


Figure 7.4: Nucleosome Energy (Less Linker DNA)

Change in Spring Potential Energy for different values of DNA methylation for least values of Linker DNA. Magenta color graph represents energy values calculated based on DNA methylation, extracted from micromodel outputs.

the corresponding changes in its epigenetic states was reported. This involved 3 steps – (i) apply EpiGMP model on a single gene and generate levels of gene expression, histone modifications and DNA methylation, (ii) calculate the spring energy for each nucleosome based on DM or average acetylation modifications and locus of nucleosomes in Euclidean 2-D plane, and (iii) use the energy values to decide on the movement of the nucleosome to either form a closed or open chromatin structure. The progress of energy values for different values of DM and linker DNA length was found to be robust and consistent. The framework to mimic the movement of nucleosomes to form an open chromatin or "Beads on a string" structure has been achieved, though corresponding results are not reported here. Simulation of a closed chromatin structure, which forms either a two-start zigzag or crossed linked is currently a work in progress.

The future work includes information on orientation of the nucleosome agents (angle of orientation) along the chromatin axis. Data obtained from experimental analysis reports that for an increase of 10 bp in linker length, there is a turn of 17 degrees in the nucleosome angle, [C. Wu et al., 2007]. This intrinsic property can be included in future simulations in order to choose one of the two zig-zag conformations for the chromatin fibre to be modelled. Previous work, [Raghavan, Ruskin, and Perrin, 2011] on frequency analysis of components in a human DNA sequence, can also be included in the simulation, to make the system more sensitive to DNA contents. There have been no previous attempts in modelling chromatin structure based on epigenetic signatures in the human genome. The simple agent-based model proposed here, although in its primary stages of development, helps to demonstrate the relation between epigenetic events and change in loci of nucleosome units along the chromatin axis. The final goal is to understand the underlying physical changes with regard to chromatin so as to detect anomalies during disease onset in the human genome.

Chapter 8

Conclusion & Future Directions

8.1 Summary

Chapter 2 elaborated on the key elements such as DNA methylation and histone modifications and their role in the field of Epigenetics. The multitude of such events in the layer of Human Epigenome presents a very complex scenario and a challenge for modelling. Discrete information obtained from previously attempted experimental trials has helped to establish a characteristic pattern of events inside the genome. Consequently it has helped to identify the need to investigate abnormal interactions among key elements that support cancer initiation. Part I explained the models developed to investigate influence of DNA sequence in Epigenetics and later quantify the interdependency between histone modifications and DNA methylation (EpiGMP micromodel or tool). Part II elaborates on the attempts to combine the DNA sequence pattern information and micromodel, followed by applying the micromodel to analyse (i) a network of colon cancer genes and, (ii) investigate chromatin remodelling/ structural dynamics (Agent-based modelling framework).

8.1.1 Main Findings and Developments

The major highlighting sections in the thesis are as follows,

- Chapter 3: The significance of DNA sequences and evolution of the nucleotide patterns in the large DNA sequences were described. The sequences are known to play a direct role in establishing differential DNA methylation across the genome.Two methods were used to spot frequency components in DNA sequence which include - the Fourier and Discrete Wavelet Transformations. This was followed by an elaborate discussion on the correlation between the DNA patterns elucidated through the methods and their impact in the Epigenetics.

- Chapter 4: This is the most important part of the thesis, which explained the development of a stochastic tool - EpiGMP (Epigenetic Molecular Prediction). This helped to model and develop a mathematical interdependency between DM and HM. EpiGMP can report expression levels and occurrence of histone modification for a gene. Based on Markov Chain Monte Carlo algorithm algorithms, EpiGMP can populate histone modifications for an initial input of DM values. This chapter concluded with results of specific histone modification levels and levels of gene expression.

- Chapter 5: This is a continuation of the EpiGMP work which explained how the tool was combined with pattern analysis from DNA sequences (Chapter 3.1). Here, a study was carried out to identify specific DNA sequence patterns throughout chromosome 21 in humans and analyse those regions for levels of histone modifications using the EpiGMP tool. Due to large data requirements (involving more than 400 gene blocks across chromosome 21), a parallel computing approach was adopted to expedite the time taken for the simulation process, (use of MPI and OpenMP routines in the EpiGMP framework).

- Chapter 6: This chapter elaborated about a real application of EpiGMP in combination with empirical data for carcinoma stage of colon cancer. Information on the conditional relation

among genes during the cancer stage was extracted from an in-house epigenetic database called *StatEpigen*. The use of such information from a database (simple and condition information) allowed the development of a network of conditionally affected genes. The network contained 85 nodes (or genes) and 121 edges (which represents occurrence of epigenetic or genetic events). The EpiGMP tool was applied to each gene or node in the network to report levels of genes expression based on histone modifications evolution. This was a crude yet encouraging step towards a model framework that could incorporate information across scales, (from molecules to genes), and investigate the effect at a system level.

- Chapter 7: Chromatin remodelling is the secondary effect caused by the interactions between histone proteins, DNA sequences and methylation levels. We proposed a simple agent-based framework to model the actual physical rearrangement of nucleosomes (made of DNA sequences and histone proteins) that results in a compact human genome. The EpiGMP tool was adopted to report temporal changes to a nucleosome agent. Furthermore, Kamada-Kawai algorithms were applied to utilize results from EpiGMP and decide consequent spatial movement of nucleosomes. Preliminary results such as the calculation of Energy values that decided physical arrangements of agents for different levels of DNA methylation were reported. The development of this framework is currently on going.

8.2 Future Work

The multi-layered Epigenetic computational model has helped set the stage for phenomenological modelling in Molecular Biology. Despite giving some ideas of the scope of what is involved the prototype has some clear limitations and these have emphasized some immediate and long-term improvements.

1. Addition of Data on Histone variants – As previously mentioned, since the field of epigenet-

ics is rapidly growing, more information about special types of histone proteins have been consistently reported for a while now. The contributions of these histone variants known to appear in special type of cell, (H2AX, CENP - special H3 type of histone, H3.1,2 and .3 etc) can be included in the nearest future.

2. Include information about DNA demethylation process in the model to study its consequence in the interactions among other epigenetic events.

3. In Chapter 6 complex signals across scales were reported but only for 1 stage of colon cancer. The same approach can be applied to other stages which include Polyps, Adenoma and Metastasis using data from StatEpigen.

4. A third layer to the work in Chapter 6 can be included, which would address a multi-system level perspective of progress of malignancies. This includes data of abnormal events in different systems of the human body such as Lungs, blood, breast, liver etc. This way, a system wide analysis of connectivity among genes in different cell types based on epigenetic and genetic events reported can be investigated.

5. Stem and differentiated cell types have a separate yet very crucial contribution to the progress of cancer in the body. The literature indicates that epigenetic profile of pluripotent stem cells are totally different from differentiated or progenitor cells and hence molecular aberrations in these cell types allow them to spread the abnormalities in genes to progenitor cells after the process of differentiation. This type of information on cell type can be included in the model to increase its capability to resemble the actual biological mechanisms, [Perrin, 2008; Perrin and Ruskin, 2010].

6. Chromatin Remodelling - Since this aspect of Epigenetics is in its primary stages many improvements to the work reported in Chapter 7 can be considered.

(a) The current graph theoretic approach can be modified to accommodate a larger chromatin structure, involving multi scaling approaches.

(b) A 3-Dimensional model of the Chromatin dynamics which closely resembles the real chromatin fibre can be developed in the near future.

7. Incorporation of Protein-DNA interactions, (Transcription and DNA methylation associated Factors), through Multi Agent Systems Modelling: In the real system, the actual process of transcription involves not only histone signatures and DNA methylation, but also use of other proteins that bind to the DNA and histones to cause a local perturbation in the structure and help induce a specific event. The presence and roles of such Transcription or DNA methylation associated factors or proteins can be modelled using "Multiple Agents" modelling techniques.

8.3 Conclusion & Final Remarks

The significance of epigenetic layer in the human genome has become highly recognized and with information being reported on a very regular basis. The model developed and reported in this thesis is a very significant advance towards investigating the complex signals in epigenetic layer of human genome. The theoretical and robust framework, (EpiGMP) reported and applied repeatedly here has successfully modelled the "driving" epigenetic molecular changes in the human genome. A bottom-up approach as adopted here has facilitated low-level information processing between different components so as to understand the evolution of phenotype or physical appearance of an organism at higher level especially during abnormal conditions. This is one of the few prototype attempts aimed at addressing a cascade of common events in the human epigenome in order to develop a mathematical relation between them. Researchers from both aspects, (biological laboratories and modelling using computer sciences), have to be aware that significant steps in unravelling

the workings of this Epigenetic layer in human genome has been achieved but is not yet complete. A comprehensive understanding of the complete picture in this field of research is a non-trivial task. Based on this awareness the final aim of this project is to identify bottleneck events in the field of epigenetics that aid in cancer progress, through these generic computational modelling attempts.

Source Code

The source code for the work described in Chapters 3 to 6 and Appendix A, can be accessed through `http://www.computing.dcu.ie/~kaghavan/work.html`. Details on the username and password required to download the source code can be obtained by contacting me through email. (`kaghavan[at]computing.dcu.ie` or `karthika.raghavan[at]gmail.com`).

Bibliography

A'Hearn, M. F., Ahern, F. J., and Zipoy, D. M. (1974). "Polarization Fourier Spectrometer for Astronomy". *Applied Optics* 13.5, pp. 1147–1157.

Ahlquist, T., Lind, G., Costa, V., Meling, G., Vatn, M., Hoff, G., Rognum, T., Skotheim, R., Thiis-Evensen, E., and Lothe, R. (2008). "Gene methylation profiles of normal mucosa, and benign and malignant colorectal tumors identify early onset markers". *Molecular Cancer* 7.1, p. 94.

Ajiro, K., Scoltock, A. B., Smith, L. K., Ashasima, M., and Cidlowski, J. A. (2010). "Reciprocal epigenetic modification of histone H2B occurs in chromatin during apoptosis in vitro and in vivo". *Cell Death and Differentiation* 17.6 (6), pp. 984–993.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002). *Molecular Cell Biology*. Fourth. New York: Garland Science.

Allis, C. D., Jenuwein, T., Reinberg, D., and Caparros, M. L. (2007). *Epigenetics*. Cold Spring Harbor Press.

Aparicio, O., Geisberg, J. V., Sekinger, E., Yang, A., Moqtaderi, Z., and Struhl, K. (2001). "Chromatin Immunoprecipitation for Determining the Association of Proteins with Specific Genomic Sequences In Vivo". In: *Current Protocols in Molecular Biology*. John Wiley & Sons Inc. Chap. 17.

Avramouli, A., Tsochas, S., Mandala, E., Katodritou, E., Ioannou, M., Ritis, K., and Speletas, M. (2009). "Methylation status of RASSF1A in patients with chronic myeloid leukemia". *Leukemia Research* 33.8, pp. 1130 –1132.

Bahar, A. Y., Taylor, P. J., Andrews, L., Proos, A., Burnett, L., Tucker, K., Friedlander, M., and Buckley, M. F. (2001). "The frequency of founder mutations in the BRCA1, BRCA2, and APC genes in Australian Ashkenazi Jews". *Cancer* 92.2, pp. 440–445.

Barat, A. and Ruskin, H. J. (2010). "A Manually Curated Novel Knowledge Management System for Genetic and Epigenetic Molecular Determinants of Colon Cancer". *The Open Colorectal Cancer Journal* 3, pp. 36–46.

Barber, C. M., Turner, F. B., Wang, Y., Hagstrom, K., Taverna, S. D., Mollah, S., Ueberheide, B., Meyer, B. J., Hunt, D. F., Cheung, P., and Allis, C. D. (2004). "The enhancement of histone H4 and H2A serine 1 phosphorylation during mitosis and S-phase is evolutionarily conserved". *Chromosoma* 112.7, pp. 360–371.

Baylin, S. B. and Ohm, J. E. (2006). "Epigenetic gene silencing in cancer – A mechanism for early oncogenic pathway addiction". *Nature Review Cancer* 6.2, pp. 107–116.

Bock, C., Walter, J., Paulsen, M., and Lengauer, T. (2007). "CpG Island Mapping by Epigenome Prediction". *PLoS Computational Biology* 3.6, e110.

Boyer, L. A., Plath, K., Zeitlinger, J., Brambrink, T., Medeiros., L. A., Lee, T. I., Levine, S. S., Wernig, M., Tajonar, A., Ray, M. K., Bell, G. W., Otte, A. P., Miguel Vidal, a. D. K. G., Young, R. A., and Jaenisch, R. (2006). "Polycomb complexes repress developmental regulators in murine embryonic stem cells". *Nature* 441.7091, pp. 349–353.

Brower, V. (2011). "Epigenetics: Unravelling the cancer code". *Nature* 471.7339, S12–S13.

Campbell, P. T., Newton, C. C., Dehal, A. N., Jacobs, E. J., Patel, A. V., and Gapstur, S. M. (2012). "Impact of Body Mass Index on Survival After Colorectal Cancer Diagnosis: The Cancer Prevention Study-II Nutrition Cohort". *Journal of Clinical Oncology* 30.1, pp. 42–52.

Carey, M. F., Peterson, C. L., and Smale, S. T. (2009). "Chromatin Immunoprecipitation (ChIP)". *Cold Spring Harbor Protocols* 2009.9, pdb.prot5279.

Cass, D., Hotchko, R., Barber, P., Jones, K., Gates, D., and Berglund, J. A. (2011). "The four Zn fingers of MBNL1 provide a flexible platform for recognition of its RNA binding elements". *BMC Molecular Biology* 12.1, p. 20.

Cedar, H. and Bergman, Y. (2009). "Linking DNA methylation and histone modification: Patterns and Paradigms". *Nature Review Genetics* 10.5, pp. 295–304.

Chahwan, R., Wontakal, S. N., and Roa, S. (2011). "The Multidimensional Nature of Epigenetic Information and Its Role in Disease". *Discovery Medicine* 11.58, pp. 233–243.

Chamberlain, S. J. and Lalandea, M. (2010). "Neurodevelopmental disorders involving genomic imprinting at human chromosome 15q11–q13". *Neurobiology of Disease* 39.1, pp. 13–20.

Chapman, B., Jost, G., and Pas, R. v. d. (2007). *Using OpenMP: Portable Shared Memory Parallel Programming (Scientific and Engineering Computation)*. The MIT Press.

Chi, P., Allis, C. D., and Wang, G. G. (2010). "Covalent histone modifications – miswritten, misinterpreted and mis-erased in human cancers". *Nature Reviews Cancer* 10.7, pp. 457–469.

Cho, H., Toyokawa, G., Daigo, Y., Hayami, S., Masuda, K., Ikawa, N., Yamane, Y., Maejima, K., Tsunoda, T., Field, H., et al. (2011). "The JmjC domain-containing histone demethylase KDM3A is a positive regulator of the G1/S transition in cancer cells via transcriptional regulation of the HOXA1 gene". *International Journal of Cancer* 131 (3), E179–89.

Choi, J. K. (2010). "Systems biology and Epigenetic gene regulation". *IET Systems Biology* 4 (5), pp. 289–95.

Clay, O., Schaffner, W., and Matsuo, K. (1995). "Periodicity of eight nucleotides in purine distribution around human genomic CpG dinucleotides". *Somatic Cell and Molecular Genetics* 21.2, pp. 91–98.

Collas, P. (2010). “The Current State of Chromatin Immunoprecipitation”. *Molecular Biotechnology* 45.1, pp. 87–100.

Collins, F. S., Patrinos, A., Jordan, E., Chakravarti, A., Gesteland, R., and Walters, L. (1998). “New Goals for the U.S. Human Genome Project: 1998–2003”. *Science* 282.5389, pp. 682–689.

Conlon, T., Ruskin, H. J., and Crane, M. (2009). “Seizure characterization using frequency-dependent multivariate dynamics”. *Computers in Biology and Medicine* 39.9, pp. 760–767.

Cowan, R. (1991). “Expected Frequencies of DNA Patterns Using Whittle’s Formula”. *Journal of Applied Probability* 28.4, pp. 886–892.

Cowley, D. E. and Atchley, W. R. (1992). “Quantitative Genetic Models for Development, Epigenetic Selection, and Phenotypic Evolution”. *Evolution* 46.2, pp. 495–518.

Cuddapah, S., Jothi, R., Schones, D. E., Roh, T., Cui, K., and Zhao, K. (2009). “Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains”. *Genome Research* 19.1, pp. 24–32.

Das, R., Dimitrova, N., Xuan, Z., Rollins, R. A., Haghighi, F., Edwards, J. R., Ju, J., Bestor, T. H., and Zhang, M. Q. (2006). “Computational prediction of methylation status in human genomic sequences”. *Proceedings of the National Academy of Sciences* 103.28, pp. 10713–10716.

Dijkstra, E. W. (1959). “A note on two problems in connection with graphs”. *Numerische Mathematik* 1, pp. 269–271.

Djokic, D. and Maidment, D. (1993). “Application of GIS network routines for water flow and transport”. *Journal of Water Resources Planning and Management* 119.2, pp. 229–245.

Doerfler, W. and Böhm, P. (2006). *DNA Methylation: Basics Mechanisms*. First. Springer.

Doerfler, W., Toth, M., Kochaneka, S., Achtena, S., Freisem-Rabiena, U., Behn-Krappaa, A., and Orenda, G. (1990). “Eukaryotic DNA Methylation – Facts and Problems”. *FEBS Letter* 286.2, pp. 329–333.

Doubleday, K. J. and Esunge, J. N. (2011). "Application of Markov Chains to Stock Trends". *Journal of Mathematics and Statistics* 7 (2), pp. 103–106.

Dubes, R. and Jain, A. (1989). "Random field models in image analysis". *Journal of applied statistics* 16.2, pp. 131–164.

Eckert, D., Biermann, K., Nettersheim, D., Gillis, A., Steger, K., Jack, H., Muller, A., Looijenga, L., and Schorle, H. (2008). "Expression of BLIMP1/PRMT5 and concurrent histone H2A/H4 arginine 3 dimethylation in fetal germ cells, CIS/IGCNU and germ cell tumors". *BMC Developmental Biology* 8, p. 106.

Ehrlich., M., Sanchez, C., Shao, C., Nishiyama, R., Kehrl, J., Kuick, R., Kubota, T., and Hanash, S. (2008). "ICF, an immunodeficiency syndrome: DNA methyltransferase 3B involvement, chromosome anomalies, and gene dysregulation". *Autoimmunity* 41.4, pp. 253–271.

Engeland, M. van, Roemen, G. M., Brink2, M., Pachen1, M. M., Weijenberg, M. P., Bruïne1, A. P. de, Arends, J.-W., Brandt2, P. A. van den, Goeij, A. F., and Herman, J. G. (May 2002). "KRAS mutations and RASSF1A promoter methylation in colorectal cancer". *Oncogene* 21.23, pp. 3792–3795.

Epps, J. (2009). "A Hybrid Technique for the Periodicity Characterization of Genomic Sequence Data". *EURASIP Journal on Bioinformatics and Systems Biology* 2009.

Esteller, M. (2008). "Epigenetics in Cancer". *New England Journal of Medicine* 358.11, pp. 1148–1159.

Esteller, M. (2007). "Cancer Epigenomics: DNA methylomes and histone-modification maps". *Nature Reviews Genetics* 8.4 (4), pp. 286–298.

Esteller, M., Risques, R.-A., Toyota, M., Capella, G., Moreno, V., Peinado, M. A., Baylin, S. B., and Herman, J. G. (2001). "Promoter Hypermethylation of the DNA Repair Gene O6-Methylguanine-DNA Methyltransferase Is Associated with the Presence of G:C to A:T Transition Mutations in p53 in Human Colorectal Tumorigenesis". *Cancer Research* 61.12, pp. 4689–4692.

Fernández, M. and Miranda-Saavedra, D. (2012). "Genome-wide enhancer prediction from epigenetic signatures using genetic algorithm-optimized support vector machines". *Nucleic Acids Research* 40.10, pp. 1–12.

Fullgrabe, J., Kavanagh, E., and Joseph, B. (2011). "Histone Onco – Modifications". *Oncogene* 30.31, pp. 3391–3403.

Gao, F. and Zhang, C.-T. (2006). "GC-Profile: a web-based tool for visualizing and analyzing the variation of GC content in genomic sequences". *Nucleic Acids Research* 34.2, pp. 686–691.

Gertz, J., Varley, K. E., Reddy, T. E., Bowling, K. M., and Pauli, F. (2011). "Analysis of DNA Methylation in a Three-Generation Family Reveals Widespread Genetic Influence on Epigenetic Regulation". *PLoS Genetics* 7.8, e1002228.

Gilbert, N. and Ramsahoye, B. (2005). "The relationship between chromatin structure and transcriptional activity in mammalian genomes". *Briefings in Functional Genomics & Proteomics* 4.2, pp. 129–142.

Glass, J. L., Fazzari, M. L., Ferguson-Smith, A. C., and Greally, J. M. (2004). "CG di-nucleotide periodicities recognized by the DNMT-3a-DNMT-3L complex are distinctive at retro-elements and imprinted domains". *Mammalian Genome* 20.9-10, pp. 633–643.

Glass, J. L., Thompson, R. F., Khulan, B., Figueroa, M. E., Olivier, E. N., Oakley, E. J., Zant, G. V., Bouhassira, E. E., Melnick, A., Golden, A., Fazzari, M. J., and Greally, J. M. (2007). "CG dinucleotide clustering is a species-specific property of the genome". *Nucleic Acid Research* 35.20, pp. 6798–6807.

Gonzalo, V., Lozano, J. J., Muñoz, J., Balaguer, F., Pellisé, M., Miguel, C. R. de, Andreu, M., Jover, R., Llor, X., Giráldez, M. D., Ocaña, T., Serradesanferm, A., Alonso-Espinaco, V., Jimeno, M., Cuatrecasas, M., Sendino, O., Castellví-Bel, S., and Castells, A. (Jan. 2010). "Aberrant Gene Promoter Methylation Associated with Sporadic Multiple Colorectal Cancer". *PLoS ONE* 5.1, e8777.

Goodman, J. W. (2005). *Introduction to Fourier Optics*. Third. Roberts and Company.

Goto, T., Mizukami, H., Shirahata, A., Sakata, M., Saito, M., Ishibashi, K., Kigawa, G., Nemoto, H., Sanada, Y., and Hibi, K. (2009). "Aberrant Methylation of the p16 Gene Is Frequently Detected in Advanced Colorectal Cancer". *Anticancer Research* 29.1, pp. 275–277.

Grama, A., Karypis, G., Kumar, V., and Gupta, A. (2003). *An Introduction to Parallel Computing*. 2nd ed. Addison Wesley.

Gropp, W. and Lusk, E. (2004). "Fault Tolerance in Message Passing Interface Programs". *International Journal of High Performance Computing Applications* 18.3, pp. 363–372.

Grunau, C., Renault, E., and Roizes, G. (2002). "DNA Methylation Database "MethDB": a user guide". *The Journal of Nutrition* 132.8, 2435S–2439S.

Guo, Y., Xiao, P., Lei, S., Deng, F., Xiao, G., Liu, Y., Chen, X., Li, L., Wu, S., Chen, Y., et al. (2008). "How is mRNA expression predictive for protein expression? A correlation study on human circulating monocytes". *Acta Biochimica et Biophysica Sinica* 40.5, pp. 426–436.

Hall, M., Shundrovsky, A., Bai, L., Fulbright, R., Lis, J., and Wang, M. (2009). "High-resolution dynamic mapping of histone-DNA interactions in a nucleosome". *Nature structural & molecular biology* 16.2, pp. 124–129.

Herzel, H., Weiss, O., and Trifonov, E. N. (1999). "10-11 bp periodicities in complete genomes reflect protein structure and DNA folding." *Bioinformatics* 15.3, pp. 187–193.

Hibi, K., Nakayama, H., Koike, M., Kasai, Y., Ito, K., Akiyama, S., and Nakao, A. (2002). "Colorectal Cancers with both p16 and p14 Methylation Show Invasive Characteristics". *Japanese Journal of Cancer Research* 93, pp. 883–887.

Hoffman, B. and Jones, S. (2009). "Genome-wide identification of DNA–protein interactions using Chromatin Immuno-Precipitation coupled with flow cell sequencing". *Journal of Endocrinology* 201.1, pp. 1–13.

Holliday, R. (2006). “Dual Inheritance”. In: *DNA Methylation: Basic Mechanisms*. Ed. by W. Doerfler and P. Bóhm. Vol. 301. Current Topics in Microbiology and Immunology. Springer Berlin Heidelberg, pp. 243–256.

Horne, H. N., Lee, P. S., Murphy, S. K., Alonso, M. A., Olson, J. A., and Marks, J. R. (2009). “Inactivation of the MAL Gene in Breast Cancer Is a Common Event That Predicts Benefit from Adjuvant Chemotherapy”. *Molecular Cancer Research* 7.2, pp. 199–209.

Hosid, S., Trifonov, E. N., and Bolshoy, A. (2004). “Sequence periodicity of Escherichia coli is concentrated in intergenic regions”. *BMC Molecular Biology* 5.1, p. 14.

Ito, T. (2007). “Role of Histone Modification in Chromatin Dynamics”. *Journal of Biochemistry* 141.5, pp. 609–614.

Jansen, A. and Verstrepen, K. J. (2011). “Nucleosome Positioning in *Saccharomyces Cerevisiae*”. *Microbiology and Molecular Biology Reviews* 75.2, pp. 301–320.

Jennings, N., Sycara, K., and Wooldridge, M. (1998). “A roadmap of agent research and development.” *Autonomous agents and multi-agents systems* 1.1, pp. 7–38.

Jenuwein, T. and Allis, C. D. (2001). “Translating the Histone Code”. *Science* 293.5532, pp. 1074–1080.

Jiang, C. and Pugh, B. F. (2003). “Nucleosome positioning and gene regulation: advances through genomics”. *Nature Review Genetics* 10 (3), pp. 161–172.

Jung, I. and Kim, D. (2009). “Regulatory Patterns of Histone Modifications to Control the DNA Methylation Status at CpG Islands”. *IBC* 1.4, pp. 1–7.

Kaiser, G. (1994). *A Friendly Guide to Wavelets*. Sixth. Birkhäuser.

Kaja, M., Kristian, R., and Janusz, M. B. (2011). “Databases and Bioinformatics Tools for the Study of DNA Repair”. *Molecular Biology International* 2011, pp. 1–9.

Kamada, T. and Kawai, S. (1989). “An algorithm for drawing general undirected graphs”. *Information processing letters* 31.1, pp. 7–15.

Kaneko, K. (2009). “Relationship among phenotypic plasticity, phenotypic fluctuations, robustness, and evolvability; Waddington’s legacy revisited under the spirit of Einstein.” *Journal of Biosciences* 34.4, pp. 529–542.

Karlić, R., Chung, H., Lasserre, J., Vlahoviček, K., and Vingron, M. (2010). “Histone modification levels are predictive for gene expression”. *Proceedings of National Academy of Science* 107.7, pp. 2926–2931.

Kashtan, N., Itzkovitz, S., Milo, R., and Alon, U. (2002). *Mfinder tool guide*. Tech. rep. Department of Molecular Cell Biology, Computer Science, and Applied Mathematics, Weizman Institute of Science.

Kawamoto, Y., Tsuchihara, K., Yoshino, T., Ogasawara, N., Kojima, M., Takahashi, M., Ochiai, A., Bando, H., Fuse, N., Tahara, M., Doi, T., Esumi, H., Komatsu, Y., and Ohtsu, A. (2012). “KRAS mutations in primary tumours and post-FOLFOX metastatic lesions in cases of colorectal cancer”. *British Journal of Cancer*, pp. 1–5.

Klose, R., Kallin, E., and Zhang, Y. (2006). “JmjC-domain-containing proteins and histone demethylation”. *Nature Reviews Genetics* 7.9, pp. 715–727.

Koinuma, K., Kaneda, R., Toyota, M., Yamashita, Y., Takada, S., Choi, Y. L., Wada, T., Okada, M., Konishi, F., Nagai, H., and Mano, H. (2005). “Screening for genomic fragments that are methylated specifically in colorectal carcinoma with a methylated MLH1 promoter”. *Carcinogenesis* 26.12, pp. 2078–2085.

Kouzarides, T. (2007). “Chromatin Modifications and Their Function”. *Cell* 128.4, pp. 693–705.

Kuntz, K., Lansdorp-Vogelaar, I., Rutter, C., Knudsen, A., Ballegooijen, M. van, Savarino, J., Feuer, E., and Zauber, A. (2011). “A Systematic Comparison of Microsimulation Models of Colorectal Cancer The Role of Assumptions about Adenoma Progression”. *Medical Decision Making* 31.4, pp. 530–539.

Kurdistani, S. K. (2011). “Histone Modifications in Cancer Biology and Prognosis”. In: *Epigenetics and Disease*. Ed. by S. M. Gasser and E. Li. Vol. 67. Progress in Drug Research. Springer Basel, pp. 91–106.

Kwon, D., Vannucci, M., Song, J. J., Jeong, J., and Pfeiffer, R. M. (2008). “A Novel Wavelet-based Thresholding Method for the Pre-processing of Mass Spectrometry Data that Accounts for Heterogeneous Noise”. *Proteomics* 8.15, pp. 3019–3029.

Lercher, M. J., Urrutia, A. O., Pavlicek, A., and Hurst, L. D. (2003). “A unification of mosaic structures in the human genome”. *Human Molecular Genetics* 12 (19), pp. 2411–2415.

Lind, G., Ahlquist, T., Kolberg, M., Berg, M., Eknaes, M., Alonso, M., Kallioniemi, A., Meling, G., Skotheim, R., Rognum, T., Thiis-Evensen, E., and Lothe, R. (2008). “Hypermethylated MAL gene - a silent marker of early colon tumorigenesis”. *Journal of Translational Medicine* 6.1, p. 13.

Lind, G., Thorstensen, L., Lovig, T., Meling, G., Hamelin, R., Rognum, T., Esteller, M., and Lothe, R. (2004). “A CpG island hypermethylation profile of primary colorectal carcinomas and colon cancer cell lines”. *Molecular Cancer* 3.1, p. 28.

Li, R., Zhu, H., and Ruan, J. (2010). “De novo assembly of human genomes with massively parallel short read sequencing”. *Nucleic Acid Research* 20.2, pp. 265–272.

Loeb, L. A. (1994). “Microsatellite Instability: Marker of a Mutator Phenotype in Cancer”. *Cancer Research* 54.19, pp. 5059–5063.

Lossow, M. (2007). “A min-max version of Dijkstra’s algorithm with application to perturbed optimal control problems”. *PAMM* 7.1, pp. 4130027–4130028.

Lustig, B. and Behrens, J. (2003). “The WNT signaling pathway and its role in tumor development”. *Journal of Cancer Research and Clinical Oncology* 129 (4), pp. 199–221.

Madsen, B., Villesen, P., and Wiuf, C. (2008). “Short Tandem Repeats in Human Exons: A Target for Disease Mutations”. *BMC Genomics* 9.1, p. 410.

Malone, B. M., Tan, F., Bridges, S. M., and Peng, Z. (2011). “Comparison of four ChIP-Seq Analytical Algorithms Using Rice Endosperm H3K27 Trimethylation Profiling Data”. *PLoS ONE* 6.9, e25260.

Martinowich, K., Hattori, D., Wu, H., Fouse, S., He, F., Hu, Y., Fan, G., and Sun, Y. E. (2003). “DNA Methylation-Related Chromatin Remodeling in Activity-Dependent BDNF Gene Regulation”. *Science* 302.5646, pp. 890–893.

Martins, C. (2007). “Fish Cytogenetics”. In: Science Publisher Inc., Enfield. Chap. Chromosomes and repetitive DNAs: a contribution to the knowledge of fish genome. Pp. 421–453.

Matsumoto, M. and Nishimura, T. (1998). “Mersenne Twister: A 623-dimensionally equidistributed uniform pseudo-random number generator”. *ACM Transactions on Modeling and Computer Simulation* 8.1, pp. 3–30.

Ma, Z., Shah, R., Chang, M., and Benveniste, E. (2004). “Coordination of cell signaling, chromatin remodeling, histone modifications, and regulator recruitment in human matrix metalloproteinase 9 gene transcription”. *Molecular and cellular biology* 24.12, pp. 5496–5509.

Meng, C. F., Zhu, X. J., Peng, G., and Dai, D. (2009). “Promoter histone H3 lysine 9 di-methylation is associated with DNA methylation and aberrant expression of p16 in gastric cancer cells”. *Oncology Report* 22.5, pp. 1221–1227.

Merhavi, E., Cohen, Y., Avraham, B. C. R., Frenkel, S., Chowers, I., Pe’er, J., and Goldenberg-Cohen, N. (2007). “Promoter Methylation Status of Multiple Genes in Uveal Melanoma”. *Investigative Ophthalmology & Visual Science* 48.10, pp. 4403–4406.

Migliore, L., Migheli, F., Spisni, R., and Coppedé, F. (2011). “Genetics, Cytogenetics, and Epigenetics of Colorectal Cancer”. *Journal of Biomedicine and Biotechnology* 2011.

Miklós, I. (2003). “MCMC genome rearrangement”. *Bioinformatics* 19.2, pp. 130–137.

Miyamoto, S., Ito, K., Miyakubo, M., Suzuki, R., Yamamoto, T., Suzuki, K., and Yamanaka, H. (Mar. 2012). “Impact of pretreatment factors, biopsy Gleason grade volume indices and post-

treatment nadir PSA on overall survival in patients with metastatic prostate cancer treated with step-up hormonal therapy". *Prostate Cancer Prostatic Disease* 15.1, pp. 75–86.

Moco, S., Bino, R. J., Vorst, O., Verhoeven, H. A., Groot, J. de, Beek, T. A. van, Vervoort, J., and Vos, C. R. de (2006). "A Liquid Chromatography-Mass Spectrometry-Based Metabolome Database for Tomato". *Plant Physiology* 141 (4), pp. 1205–1218.

Morrison, N. (1994). *Introduction to Fourier Analysis*. Wiley-Interscience.

Murrell, A., Rakyan, V. K., and Beck, S. (2005). "From Genome to Epigenome". *Human Molecular Genetics* 14.1, pp. 3–10.

Myers, F. A., Chong, W., Evans, D. R., Thorne, A. W., and Crane-Robinson, C. (2003). "Acetylation of Histone H2B Mirrors that of H4 and H3 at the Chicken–Globin Locus but Not at Housekeeping Genes". *The Journal of Biological Chemistry* 278.38, pp. 36315–36322.

Nakayama, T. and Takami, Y. (2001). "Participation of histones and histone-modifying enzymes in cell functions through alterations in chromatin structure". *Journal of biochemistry* 129.4, pp. 491–499.

Nakazawa, T., Kondo, T., Ma, D., Niu, D., Mochizuki, K., Kawasaki, T., Yamane, T., Iino, H., Fujii, H., and Katoh, R. (2012). "Global histone modification of histone H3 in colorectal cancer and its precursor lesions". *Human Pathology* 43.6, pp. 834–842.

Nelson, P. and Perelson, A. (2002). "Mathematical analysis of delay differential equation models of HIV-1 infection". *Mathematical Biosciences* 179.1, pp. 73–94.

Netzer, P., Forster, C., Biral, R., Ruchti, C., Neuweiler, J., Stauffer, E., Schönegg, R., Maurer, C., Hüsler, J., Halter, F., and Schmassmann, A. (1998). "Risk factor assessment of endoscopically removed malignant colorectal polyps". *Gut* 43.5, pp. 669–674.

Newman, M. E. J., Watts, D. J., and Strogatz, S. H. (2002). "Random graph models of social networks". *Proceedings of the National Academy of Sciences of the United States of America* 99.1, pp. 2566–2572.

Nishida, H., Kondo, S., Matsumoto, T., Suzuki, Y., Yoshikawa, H., Taylor, T. D., and Sugiyama, J. (2012). "Characteristics of nucleosomes and linker DNA regions on the genome of the basidiomycete Mixia osmundae revealed by mono- and dinucleosome mapping". *Open Biology* 2.4.

Ozers, M. S., Warren, C. L., and Ansari, A. Z. (2009). "Determining DNA Sequence Specificity of Natural and Artificial Transcription Factors by Cognate Site Identifier Analysis". *Methods in Molecular Biology* 544.

Paar, V., Glunčić, M., Basar, I., Rosandić, M., Paar, P., and Cvitković, M. (2011). "Large tandem, higher order repeats and regularly dispersed repeat units contribute substantially to divergence between human and chimpanzee Y chromosomes". *Journal of molecular evolution* 72.1, pp. 34–55.

Pancione, M., Remo, A., and Colantuoni, V. (2012). "Genetic and Epigenetic Events Generate Multiple Pathways in Colorectal Cancer Progression". *Pathology Research International* 2012 (1), pp. 1–11.

Paponov, V. D., Gromov, P. S., and Rupasov, V. V. (1980). "Are the bonds between histone fraction and DNA of different strengths?" *Bulletin of Experimental Biology and Medicine* 90 (2). 10.1007/BF00844531, pp. 1058–1060.

Patra, S. K. and Szyf, M. (2008). "DNA methylation-mediated nucleosome dynamics and oncogenic *Ras* signaling". *FEBS Journal* 275.21, pp. 5217–5235.

Perrin, D. (Aug. 2008). "Multi-layered model of individual HIV infection progression and Multi-layered model of individual HIV infection progression and mechanisms of phenotypical expression". PhD thesis. Dublin City University: School of Computing.

Perrin, D. and Ruskin, H. J. (2010). "Cell type-dependent, infection-induced, aberrant DNA methylation in gastric cancer". *Journal of Theoretical Biology* 264 (2), pp. 570–577.

Qinqin, W., Weiqiang, Z., Jiajun, W., and Hong, Y. (2011). “Correlation between the flexibility and periodic dinucleotide patterns in yeast nucleosomal DNA sequences”. *Journal of Theoretical Biology* 284.1, pp. 92 –98.

Rabenseifner, R., Hager, G., and Jost, G. (2009). “Hybrid MPI/OpenMP Parallel Programming on Clusters of Multi-Core SMP Nodes”. In: *Proceedings of the 2009 17th Euromicro International Conference on Parallel, Distributed and Network-based Processing*. PDP ’09. IEEE Computer Society, pp. 427–436.

Raghavan, K., Roznovat, I., and Ruskin, H. J. (2011). “Complex Interdependant Epigenetic Signals in Cancer Initiation (CIESCI)”. *ASSYST/CSS* 22, pp. 5–6.

Raghavan, K., Ruskin, H. J., and Perrin, D. (2011). “Computational Analysis of Epigenetic Information in Human DNA Sequences”. In: *Proceedings of the International Conference on Bioscience, Biochemistry and Bioinformatics 2011*. Ed. by S. Baby. Vol. 5. International Proceedings of Chemical, Biological and Environmental Engineering, pp. 383–387.

Raghavan, K., Ruskin, H. J., Perrin, D., Burns, J., and Goasmat, F. (2010). “Computational Micromodel for Epigenetic Mechanisms”. *PLoS One* 5.11, e14031.

Rebhan, M., Chalifa-Caspi, V., Prilusky, J., and Lancet, D. (1998). “GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support.” *Bioinformatics* 14.8, pp. 656–664.

Reed, K., Poulin, M., Yan, L., and Parissenti, A. (2010). “Comparison of bisulfite sequencing PCR with pyrosequencing for measuring differences in DNA methylation”. *Analytical biochemistry* 397.1, pp. 96–106.

Riggs, A. D. and Xiong, Z. (2004). “Methylation and epigenetic fidelity”. *Proceedings of National Academy of Sciences* 101.1, pp. 4–5.

Routh, A., Sandin, S., and Rhodes, D. (2008). "Nucleosome repeat length and linker histone stoichiometry determine chromatin fiber structure". *Proceedings of the National Academy of Sciences* 105.26, pp. 8872–8877.

Rutter, C., Knudsen, A., and Pandharipande, P. (2011). "Computer Disease Simulation Models: Integrating Evidence for Health Policy". *Academic Radiology* 18.9, pp. 1077 –1086.

Rutter, C., Miglioretti, D., and Savarino, J. (2011). "Evaluating risk factor assumptions: a simulation-based approach". *BMC Medical Informatics and Decision Making* 7, pp. 11–55.

Rutter, C., Zaslavsky, A., and Feuer, E. (2011). "Dynamic Microsimulation Models for Health Outcomes A Review". *Medical Decision Making* 31.1, pp. 10–18.

Salz, J. and Weinstein, S. B. (1969). "Fourier Transform communication system". In: *Proceedings of the first ACM symposium on Problems in the optimization of data communications systems*. ACM, pp. 99–128.

Sawan, C. and Herceg, Z. (2010). "3 - Histone Modifications and Cancer". In: *Epigenetics and Cancer, Part A*. Ed. by Z. Herceg and T. Ushijima. Vol. 70. Advances in Genetics. Academic Press, pp. 57 –85.

Schones, D. E., Cui, K., Cuddapah, S., Roh, T.-Y., Barski, A., Wang, Z., Wei, G., and Zhao, K. (2008). "Dynamic Regulation of Nucleosome Positioning in the Human Genome". *Cell* 132.5, pp. 887–898.

Schreiber., F. and Schwöbbermeyer, H. (2005). "MAVisto: a tool for the exploration of network motifs." *Bioinformatics* 21, pp. 3572–3574.

Schrem, H., Klempnauer, J., and Borlak, J. (2002). "Liver-Enriched Transcription Factors in Liver Function and Development. Part I: The Hepatocyte Nuclear Factor Network and Liver-Specific Gene Expression". *Pharmacological Reviews* 54.1, pp. 129–158.

Schulz, F., Wagner, D., and Weihe, K. (2000). "Dijkstra's algorithm on-line: an empirical case study from public railroad transport". *Journal of Experimental Algorithmics (JEA)* 5, p. 12.

Segal, E. and Widom, J. (2009). “What controls Nucleosome positions?” *Trends in Genetics* 25.8, pp. 335–343.

Segal Eranand Fondufe-Mittendorf, Y., Chen, L., Thastrom, A., Field, Y., Moore, I. K., Wang, P. Z., and Widom, J. (2006). “A genomic code for nucleosome positioning”. *Nature* 442.0028-0836 (7104), pp. 772–778.

Sell, S. (2004). “Stem cell origin of cancer and differentiation therapy”. *Critical reviews in oncology/hematology* 51.1, pp. 1–28.

Shakya, K., Ruskin, H. J., and O’Connell, M. J. (2012). “The Landscape for Epigenetic/Epigenomic Biomedical Resource”. *Epigenetics* 7 (9), pp. 982–986.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). “Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks”. *Genome Research* 13.11, pp. 2498–2504.

Shiraz, O., Galehdari, H., Yavarian, M., Geramizadeh, B., Kafilzadeh, F., and Janfeshan, S. (2012). “Comparative Study of Direct Bisulfite Sequencing PCR and Methylation Specific PCR to Detect Methylation Pattern of DNA”. *Middle-East Journal of Scientific Research* 11.4, pp. 445–449.

Silverman, B. and Linskera, R. (1986). “A measure of DNA periodicity.” *Journal of Theoretical Biology* 118.7, pp. 295–300.

Smits, R., Kartheuser, A., Jagmohan-Changur, S., Leblanc, V., Breukel, C., Vries, A. de, Kranen, H. van, Krieken, J. van, Williamson, S., and Edelmann, W. (1997). “Loss of APC and the entire chromosome 18 but absence of mutations at the Ras and TP53 genes in intestinal tumors from Apc1638Ns, a mouse model for APC-driven carcinogenesis.” *Carcinogenesis* 18.2, pp. 321–327.

Strachan, T. and Read, A. P. (1999). *Human Molecular Genetics*. 2nd ed. New York: Wiley-Liss.

Strahl, B. and Allis, C. (2000). “The language of covalent histone modifications”. *Nature* 403.6765, pp. 41–45.

Su, A. I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K. A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., Cooke, M. P., Walker, J. R., and Hogenesch, J. B. (2004). “A gene atlas of the mouse and human protein-encoding transcriptomes”. *Proceedings of the National Academy of Sciences* 101.16, pp. 6062–6067.

Su, J., Zhang, Y., Lv, J., Liu, H., Tang, X., Wang, F., Qi, Y., Feng, Y., and Li, X. (2010). “CpG_MI: a novel approach for identifying functional CpG islands in mammalian genomes”. *Nucleic Acids Research* 38.1, e6.

Sundararajan, N., Mao, D., Chan, S., Koo, T.-W., Su, X., Sun, L., Zhang, J., Sung, K.-b., Yamakawa, M., Gafken, P. R., Randolph, T., McLerran, D., Feng, Z., Berlin, A. A., and Roth, M. B. (2006). “Ultrasensitive Detection and Characterization of Posttranslational Modifications Using Surface-Enhanced Raman Spectroscopy”. *Analytical Chemistry* 78.11, pp. 3543–3550.

Sun, J. M., Chen, H. Y., Espino, P. S., and Davie, J. R. (2007). “Phosphorylated serine 28 of histone H3 is associated with destabilized nucleosomes in transcribed chromatin”. *Nucleic Acids Research* 35.19, pp. 6640–6647.

Suzuki, H. I., Yamagata, K., Sugimoto, K., Iwamoto, T., Kato, S., and Miyazono, K. (2009). “Modulation of microRNA processing by p53”. *Nature* 490 (7254), pp. 529–533.

Szerlong, H. J. and Hansen, J. C. (2011). “Nucleosome distribution and linker DNA: connecting nuclear function to dynamic chromatin structure”. *Biochemistry and Cell Biology* 89.1, pp. 24–34.

Takai, D. and Jones, P. A. (2002). “Comprehensive analysis of CpG islands in human chromosomes 21 and 22”. *Proceedings of National Academy of Science* 99.6, pp. 3740–3745.

Tanaka, Y., Yamashita, R., Suzuki, Y., and Nakai, K. (2010). "Effects of Alu elements on global nucleosome positioning in the human genome". *BMC Genomics* 11.1, p. 309.

Taplick, J. (1998). "Histone H4 acetylation during interleukin-2 stimulation of mouse T cells". *FEBS Letters* 436.3, pp. 349–352.

Tiwari, S., Ramachandran, S., Bhattacharya, A., Bhattacharya, S., and Ramaswamy, R. (1997). "Prediction of probable genes by Fourier analysis of genomic sequences". *Computer applications in the biosciences : CABIOS* 13.3, pp. 263–270.

Tolstorukov, M. Y., Kharchenko, P. V., Goldman, J. A., Kingston, R. E., and Park, P. J. (2009). "Comparative analysis of H2A. Z nucleosome organization in the human and yeast genomes". *Genome research* 19.6, pp. 967–977.

Tsonis, A. A., Kumar, P., Elsner, J. B., and Tsonis, P. A. (1996). "Wavelet analysis of DNA sequences". *Physical Review E* 53.2, pp. 1828–1834.

Turner, B. M. (2001). *Chromatin and Gene Regulation – Mechanisms in Epigenetics*. 2nd Edition. BlackWell Science Ltd.

Tu, Z. and Zhu, S. (2002). "Image segmentation by data-driven Markov chain Monte Carlo". *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 24.5, pp. 657–673.

Ushijima, T., Watanabe, N., Okochi, E., Kaneda, A., Sugimura, T., and Miyamoto, K. (2003). "Fidelity of the methylation pattern, its variation in the genome". *Genome Research* 13.5, pp. 868–874.

Vogel, S. de, Weijenberg, M. P., Herman, J. G., Wouters, K. A. D., Goeij, A. F. P. M. de, Brandt, P. A. van den, Bruine, A. P. de, and Engeland, M. van (2009). "MGMT and MLH1 promoter methylation versus APC, KRAS and BRAF gene mutations in colorectal cancer: indications for distinct pathways and sequence of events". *Annals of Oncology* 20.7, pp. 1216–1222.

Vogelstein, B. and Kinzler, K. W. (2002). *The Genetic Basis of Human Cancer*. 2nd. McGraw-Hill, Medical Publication Division.

Volpe, T., Schramke, V., Hamilton, G., White, S., Teng, G., Martienssen, R., and Allshire, R. (2003). "RNA interference is required for normal centromere function infission yeast". *Chromosome Research* 11.2, pp. 137–146.

Voss, R. F. (1992). "Evolution of Long-range Fractal Correlations and 1/f noise in DNA base sequences". *Physical Review Letters* 68.25, pp. 3805–3808.

Waddington, C. H. (1942). "The Epigenotype". *Endeavour* 1, pp. 18–20.

Weisenberger, D., Siegmund, K., Campan, M., Young, J., Long, T., Faasse, M., Kang, G., Widschwendter, M., Weener, D., Buchanan, D., et al. (2006). "CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer". *Nature genetics* 38.7, pp. 787–793.

Wernicke, S. and Rasche, F. (2006). "FANMOD: a tool for fast network motif detection". *Bioinformatics* 22.9, pp. 1152–1153.

Whittle, P. (1955). "Some Distribution and Moment Formulae for the Markov Chain". *Journal of the Royal Statistical Society Series B (Methodological)* 17.2, pp. 235–242.

Wijk, B. C. M. van, Stam, C. J., and Daffertshofer, A. (Oct. 2010). "Comparing Brain Networks of Different Size and Connectivity Density Using Graph Theory". *PLoS ONE* 5.10, e13701.

Williams, K., Christensen, J., and Helin, K. (2011). "DNA methylation: TET proteins – guardians of CpG islands?" *EMBO reports* 13 (1), pp. 28–35.

Woodcock, C., Grigoryev, S., Horowitz, R., and Whitaker, N. (1993). "A chromatin folding model that incorporates linker variability generates fibers resembling the native structures". *Proceedings of the National Academy of Sciences* 90.19, pp. 9021–9025.

Wu, C., Bassett, A., and Travers, A. (2007). "A variable topology for the 30-nm chromatin fibre". *EMBO reports* 8.12, pp. 1129–1134.

Wu, H. and Zhang, Y. (2011). "Mechanisms and functions of Tet protein-mediated 5-methylcytosine oxidation". *Genes & Development* 25.23, pp. 2436–2452.

Wyrick, J. J. and Parra, M. A. (2008). “The role of histone H2A and H2B post-translational modifications in transcription: A genomic perspective”. *Biochimica et Biophysica Acta* 1789.1, pp. 37–44.

Xing, M., Cohen, Y., Mambo, E., Tallini, G., Udelsman, R., Ladenson, P. W., and Sidransky, D. (2004). “Early Occurrence of RASSF1A Hypermethylation and Its Mutual Exclusion with BRAF Mutation in Thyroid Tumorigenesis”. *Cancer Research* 64.5, pp. 1664–1668.

Yasuhara, J. C., DeCrease, C. H., and Wakimoto, B. T. (2005). “Evolution of heterochromatic genes of Drosophila”. *Proceedings of National Academy of Science* 102.31, pp. 10958–10963.

Yin, H. and Lin, H. (2007). “An epigenetic activation role of Piwi and a Piwi associated piRNA in Drosophila melanogaster”. *Nature* 450.7167, pp. 304–308.

Yoo, C. B. and Jones, P. A. (2006). “Epigenetic therapy of cancer: past, present and future”. *Nature Reviews Drug Discovery* 5.1, pp. 37–50.

Yu, H., Zhu, S., Zhou, B., Xue, H., and Han, J. (2008). “Inferring causal relationships among different histone modifications and gene expression”. *Genome Research* 18.8, pp. 1314–1324.

Zhang, K. and Dent, S. Y. R. (2005). “Histone modifying enzymes and cancer: going beyond histones”. *Journal of cellular biochemistry* 96.6, pp. 1137–1148.

Zhang, L., Eugeni, E. E., Parthun, M. R., and Freitas, M. A. (2003). “Identification of novel histone post-translational modifications by peptide mass fingerprinting”. *Chromosoma* 112.2, pp. 77–86.

Zhang, L., Su, X., Liu, S., Knapp, A. R., Parthun, M. R., Marcucci, G., and Freitas, M. A. (2007). “Histone H4 N-Terminal Acetylation in Kasumi-1 Cells Treated with Depsipeptide Determined by Acetic Acid. Urea Polyacrylamide Gel Electrophoresis, Amino Acid Coded Mass Tagging, and Mass Spectrometry”. *Journal of Proteome Research* 6.q, pp. 81–88.

Zhao, J., Yang, X. W., Li, J. P., and Tang, Y. Y. (2001). “DNA sequences classification based on Wavelet Package analyses”. In: *WAA ’01: Proceedings of the Second International Conference on Wavelet Analysis and Its Applications*. Springer-Verlag.

Zheng, C. and Hayes, J. J. (2003). “Structure and Interactions of Core Histone Tail Domains”. *Bioplymers* 68.4, pp. 539–546.

Zou, H., Harrington, J., Shire, A., Rego, R., Wang, L., Campbell, M., Oberg, A., and Ahlquist, D. (2007). “Highly methylated genes in colorectal neoplasia: implications for screening”. *Cancer Epidemiology Biomarkers & Prevention* 16.12, pp. 2686–2696.

Part III

Appendix

Appendix A

Comparison of specific dinucleotides associated with Human Epigenome

A.1 Introduction

In this chapter, analysis of specific dinucleotides – (i) involved in structural condensation of large DNA sequences and chromatin compaction such as TT and AA and, (ii) Transcription associated dinucleotide such as GC is carried out again using Fourier and Wavelet Analysis techniques. For the first time, a correlation analysis is performed between CG dinucleotides, (reported previously in Chapter 3) and the above mentioned dinucleotides in order to compare their relative distribution levels across the Human Genome.

A.1.1 Relevance of GC dinucleotides

Analysis of the GC contents, (CG and GC dinucleotides) in all eukaryotic organisms has been the main focus in functional genomics and epigenomics. Although, GC contents are generally under-represented in the human genome, the GC-rich regions have a high density of genes and

CpG islands, [Lercher et al., 2003]. Following the sequencing of the human genome, segments of chromosomes ($>$ 300 Kilobases) were categorized into different *Isochores*[1], based on their GC contents, (L1, L2 - regions with low GC density; H1, H2 and H3 - regions with high GC density). While the relevance of CG dinucleotides in genes and CpG islands have been dealt in detail in the literature, [Bock et al., 2007; Raghavan, Ruskin, and Perrin, 2011; Takai et al., 2002], the focus on GC dinucleotides have been quite less in comparison. This dinucleotide type has been reported in association with RNA splicing and recognition site for "Muscleblind-Like" (MBNL1) protein in humans, [Cass et al., 2011].

A.1.2 Relevance of AA and TT dinucleotides

AA and TT dinucleotides are found most abundantly in the Human genome but comparatively less in frequency in GC rich regions. Several experimental and computational analyses have revealed the contribution of AA/TT dinucleotides in chromatin structure dynamics, [Hosid et al., 2004; Nishida et al., 2012; Y. Segal E. F.-M. et al., 2006; Tanaka et al., 2010], especially the contribution of 10bp periodicity in aligning DNA sequence to histone proteins. Although this periodicity is high among yeast and other lower organisms, it is quite weakly present in mammals, especially humans. Two persistent challenges with regard to these 2 dinucleotides are investigation of the factors responsible for "damping" of 10 bp in humans and nature of other significant patterns. These reasons suggest that analyses of periodicities of AA/TT and GC in comparison to the most analysed CG dinucleotides in human sequences are very essential.

A.2 Methods

The following subsections revise and explain in detail, the utility of the transformation techniques and definition of Covariance of Wavelet coefficients as a powerful method to compare the distribu-

[1] A large segment of the genome where the a high degree of uniform distribution of GC contents is present.

tion of different dinucleotides in the human sequences.

A.2.1 Fourier and Discrete Wavelet Transformation

The same set of methods, involving equations (3.1) and (3.2) from chapter 3 is used here to analyse AA/TT and GC. Additionally, the covariance for the amplitudes of the wave function is calculated to understand the difference between all datasets. Unlike the input used in Chapter 3, a larger dataset containing 20 genes from each of the 22 chromosome in Human genome, (a total of 440 genes and introns each) were selected for Fourier analysis, based on their association with a diseases reported. The coding regions, (gene) and the non coding (Introns) were extracted using UCSD genome browser, (refer Table D in Appendix D for complete list of genes). The same contiguous sequences from Table 3.1 in Chapter 3 were chosen for Comparison of wavelet analysis (for dinucleotides CG, GC and AA/TT), and covariance of wavelet coefficients.

A.2.2 Covariance of Wavelet Coefficients

In order to spot the characteristic patterns of each dinucleotide in genes and introns, Fourier transformation is applied to each sequence type. Following Fourier, MODWT is applied on undifferentiated and contiguous sequences, to check the occurrence of such highly frequent patterns. This step helps to identify where high frequency components of different dinucleotides occur and hence predict functional significance based on their loci. The presence of multi-components (dinucleotides) in sequences can be further clarified by calculating the covariance of wave coefficients for different inputs at all scales. This covariance analysis is generally performed on many functions to check if the frequency components vary at "different time-horizons", [Conlon et al., 2009]. Hence an overall investigation of how CG dinucleotides vary when compared to other epigenetically significant dinucleotides such as GC, AA/TT, is briefly described in these subsequent sections.

$$\upsilon_{fg}(\tau_j) = \frac{1}{M_j} \sum_{t=L_j-1}^{N-1} \tilde{D}_{j,t}^{f(t)} \tilde{D}_{j,t}^{g(t)} \tag{A.1}$$

$$M_j = N - L_j + 1 \tag{A.2}$$

where,

$\upsilon_{\mathrm{fg}}(\tau_{\mathrm{j}})$ = covariance at scale "j".

$\tilde{D}_{j,t}^{f(t)}$= wavelet coefficient matrix for function f(t).

$\tilde{D}_{j,t}^{g(t)}$ = wavelet coefficient matrix for function g(t).

N=maximum scale possible, (scale = log of length of input data).

A.3 Results

The details of the dataset used for Fourier is explained in Table D. Inputs for Wavelet Transformation and covariance analysis are given in Table 3.1. The Fourier and Wavelet maps of CG dinucleotides are given in Chapter 3 (Figures 3.1, 3.2, 3.3 and 3.4), which is used to compare with other dinucleotides here.

A.3.1 Fourier Analysis

GC dinucleotides

The Fourier analysis of GC dinucleotides is shown in Figure A.3.1. This base pair has not been explicitly analysed before unlike CG dinucleotides. The most highlighting part of the frequency map is that the graph corresponding to Introns, (blue), contains the tallest peak for a period of 6 bp. Several analyses before have insisted that periodicities, which are multiples of 2 correspond to the repeat regions in the human genome. These include 2,4,6 and 8 bp periodicities.

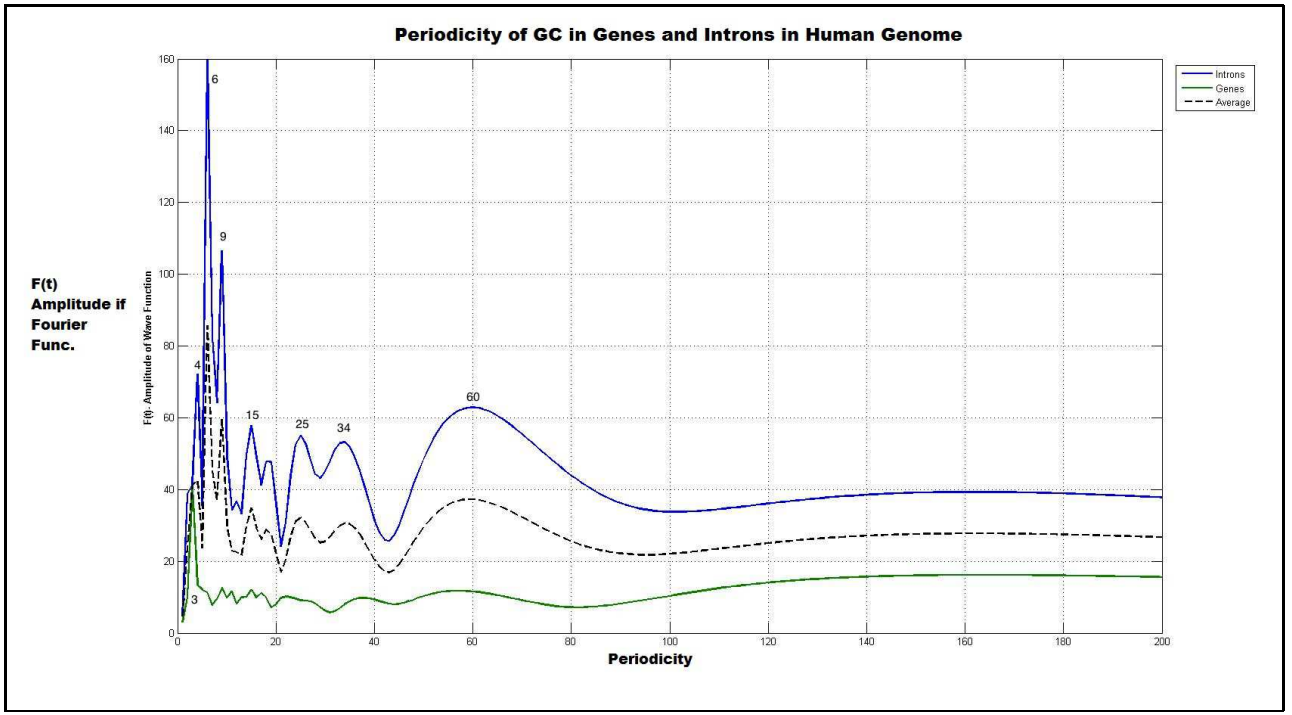


Figure A.1: GC Fourier Map

Average of Amplitude of frequency of GC dinucleotides in all 440 genes, (green colour graph) and intron regions (blue colour graph) in the Human Genome. Dash lines represent the average of amplitude values in Genes and Introns for unto a periodicity of 200 bp.

S. No.	Dinucleotide	Intron - Characteristic Periodicity	Genes - Characteristic Periodicity
1.	CG	2,4,8,10-11,16,20,25,36,55	3
1.	GC	4,6,9,15,25,34,60	3
2.	AA	2,11,16,20,25,34,53	3
3.	TT	2,11,16,20,25,34,53	3

Table A.1: Dinucleotide Frequency Table
Characteristic Periodicities of CG, GC and AA/TT dinucleotides in Genes and Intron regions in the Human Genome

Another supporting evidence from analysis of Short dispersed Tandem Repeats, (STR), in human genome conducted recently, [Madsen et al., 2008] insist that the human genome is heavily made up of such tandem repeats with length of 3 or 6 or 9 bp sequences. Hence it is quite possible that this GC dinucleotide is present in most STR sequences with length of 6bp that are found in many non-coding regions in the human genome. The same explanation can be given to the presence of a peak with a period of 9bp as well. In fact the amplitude level for the peak at periodicity of 6bp can be used as a characteristic marker for describing GC distribution in human genome, when compared with other organisms. GC and CG dinucleotides, share a peak at 4bp and 25 bp in intron regions apart from the peak formed by the characteristic 3bp spacing associated with the graph that represents the genes.

AA and TT dinucleotides

Figures A.3.1 and A.3.1 describe the average of amplitudes over all 440 introns and genes across the human genome. The two figures show the distributions of AA and TT up to a maximum periodicity of 200 bp. This pair of dinucleotides have been the focus of attention due to their active effect in secondary chromosomal compaction in lower organisms such as yeast, [Tanaka et al., 2010]. The structural properties of DNA molecules play a major role in - chromatin condensation, transcription

regulation and control of other cellular activities which include Cell division and differentiation, [Qinqin et al., 2011]. Analysis of AA and TT dinucleotides gained immense attention after the significant role of their 10 bp periodicity in nucleosome positioning and chromatin remodelling was discovered in Yeast, [Jansen et al., 2011; Y. Segal E. F.-M. et al., 2006]. A further analysis by [Tolstorukov et al., 2009], revealed that this specific periodicity is present but not explicit in human genome, [Tolstorukov et al., 2009]. In Figures A.3.1 and A.3.1, the graph that represents Introns (blue), have an explicit similarity due the appearance of peaks at the same period. While a small peak at the 10bp period is obvious, unreported and new peaks appears at of 20, 25, 34 and 53bp periodicity.

A previous in depth analysis of CG dinucleotides by [Raghavan, Ruskin, and Perrin, 2011], discussed a characteristic rise at periodicity levels of 20 and 25. Irrespective of the fact that the distribution of GC contents is complementary to AA/TT across the genome, these two sets share a common increase at 20 and 25 bp. This faint but obvious marker behaviour of dinucleotides in human genome was previously postulated by [R. Li et al., 2010], about the presence of several 25-mers across the human genome, [R. Li et al., 2010]. The other bigger peaks in these two figures include a very obvious peak at 34 and 53 periodicity. The 34-bp periodicity is also a commonality between CG dinucleotides and this pair. The significance for this common peaks at 20 and 34 is not well understood, but the frequency of 25bp has been postulated to be linked with interference RNA or *iRNA* mechanisms in human genome, [Raghavan, Ruskin, and Perrin, 2011].

The only possible explanation for the big peak at 34bp periodicity is the presence of a significant quantity of satellite (repeat) sequences. Analysis of mouse and fish genome revealed the presence of mini satellites of size 34, [Martins, 2007; Paar et al., 2011]. A strong possibility is that humans could have inherited or evolved to form sequences containing lots of AA/TT dinucleotides separated by 34 bp. The same explanation is applicable to the peak at a period of 53 for AA and TT.

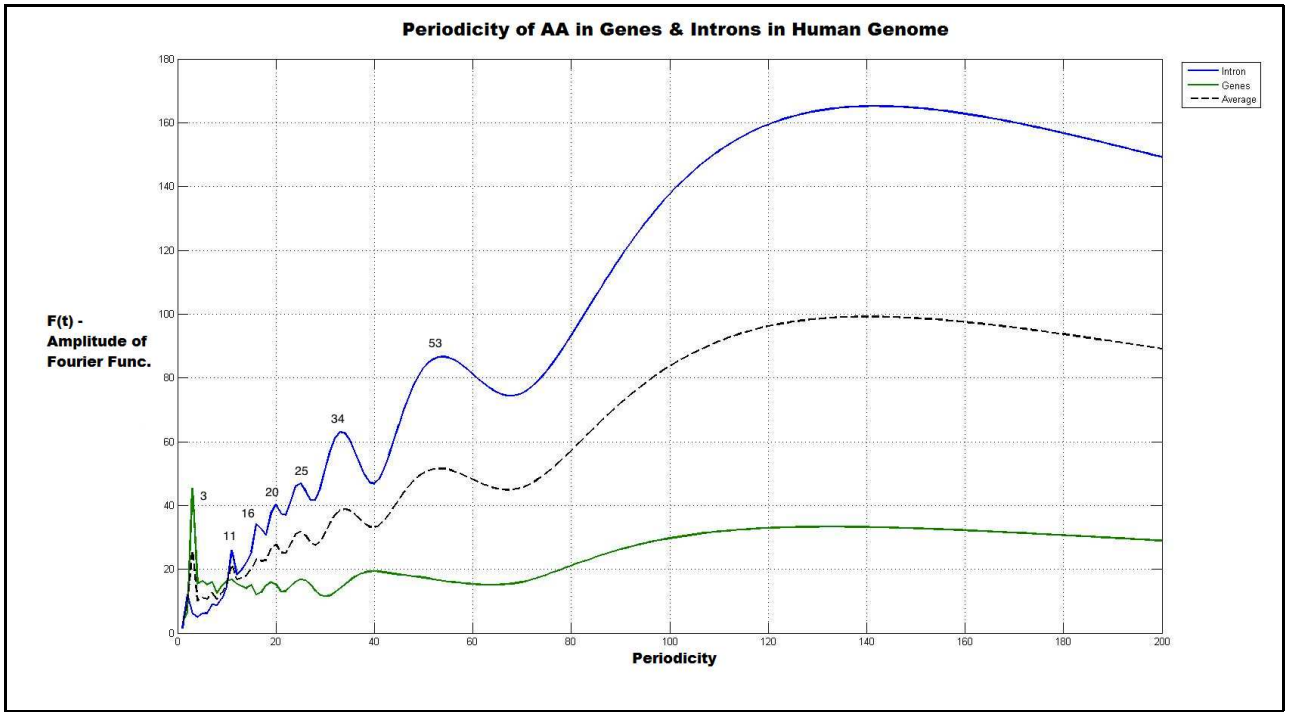


Figure A.2: AA Fourier Map

Average of Amplitude of frequency of TT dinucleotides in all 440 genes, (green colour graph) and intron regions, (blue colour graph) in the Human Genome. Dash lines represent the average of amplitude values in Genes and Introns for unto a periodicity of 200 bp.

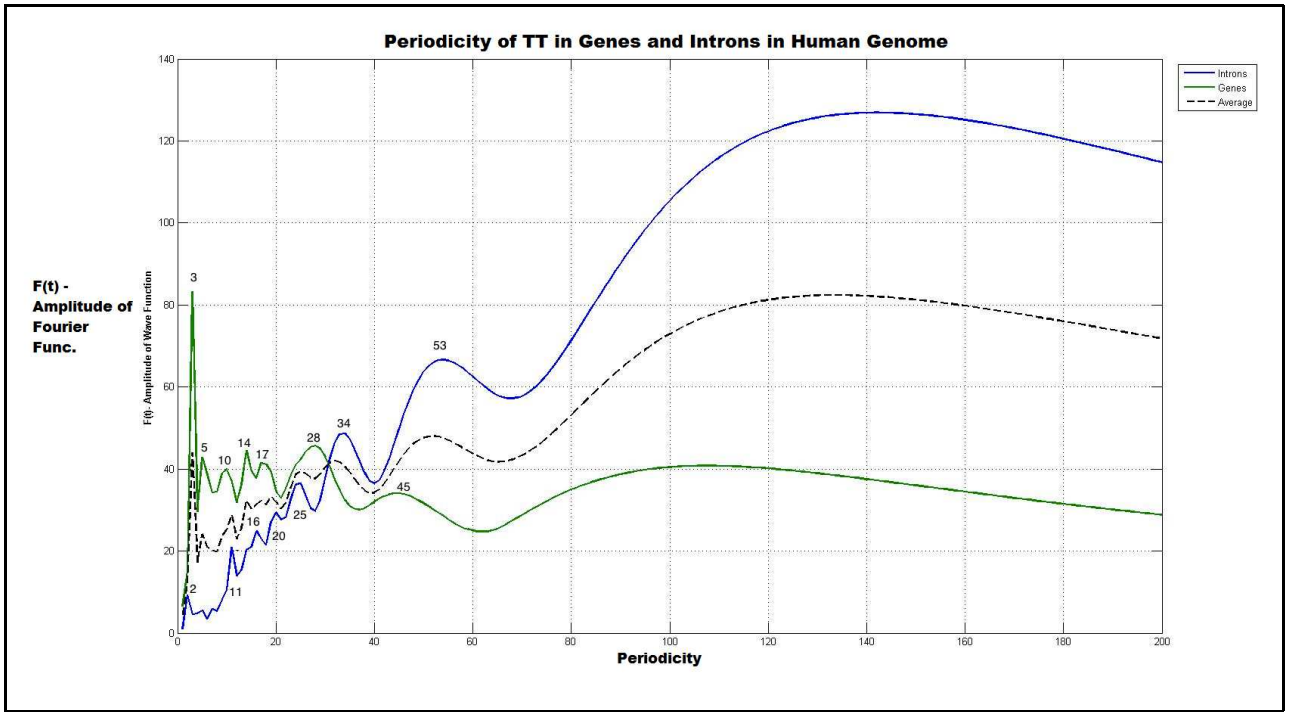


Figure A.3: TT Fourier Map

Average of Amplitude of frequency of AA dinucleotides in all 440 genes, (green colour graph) and intron, (blue colour graph) regions in the Human Genome. Dash lines represent the average of amplitude values in Genes and Introns for unto a periodicity of 200 bp.

Another unreported fact is that when CG periodicity map was plotted for a maximum period of 200 bp, (results not shown here), peaks at periods 36 and 55 which were exactly 2 nucleotide positions away from the appearance of peaks for AA and TT, were apparent. This underlying observation explains as to why that human genome has accumulated a lot of random repeats, which contain specific (but well spaced) dinucleotides.

A.3.2 Comparison of different Dinucleotides using Wavelet Transformation

Figure A.4 depicts the wavelet coefficients for different dinucleotides (considered as different functions here), for a wavelet scale factor of 11 for contiguous sequence from chromosome 21, (Contig sequence id: NT_113952.1, refer Table 3.1 in Chapter 3). This scale factor was most suitable for further analysis and comparison with Fourier[2] The main focus of this analysis was to identify the location of specific frequencies of specific dinucleotides along the length of the DNA sequences. It is quite clear that the dataset representing the information on AA/TT dinucleotides have higher values of coefficients compared to CG and GC. Another interesting factor is that AA and TT dinucleotides have the same high frequency components, (Table A.1 and Figures A.3.1 and A.3.1) but they are not present in the same location. This is quite obvious in Figure A.4 where the graphs associated with AA and TT have a complementary coefficients values between the intervals 28,000-30,000bp and 140,000-145,000bp.

There is a steep increase and peak formation in the AA wavelet coefficients values, unlike TT, which has a dip in the same loci but succeeds by steep peak present a few 1000 nucleotides away. The same behaviour is obvious in other contiguous sequences. It must be noted that both GC and CG associated wavelet coefficient values are very small in amplitude probably due the under representation of these dinucleotides in the Human genome when compared to AA/TT. The GC

[2] Although we applied MODWT for all scales, for analysing all dinucleotides, we do not share the results in order to simplify the presentation here.

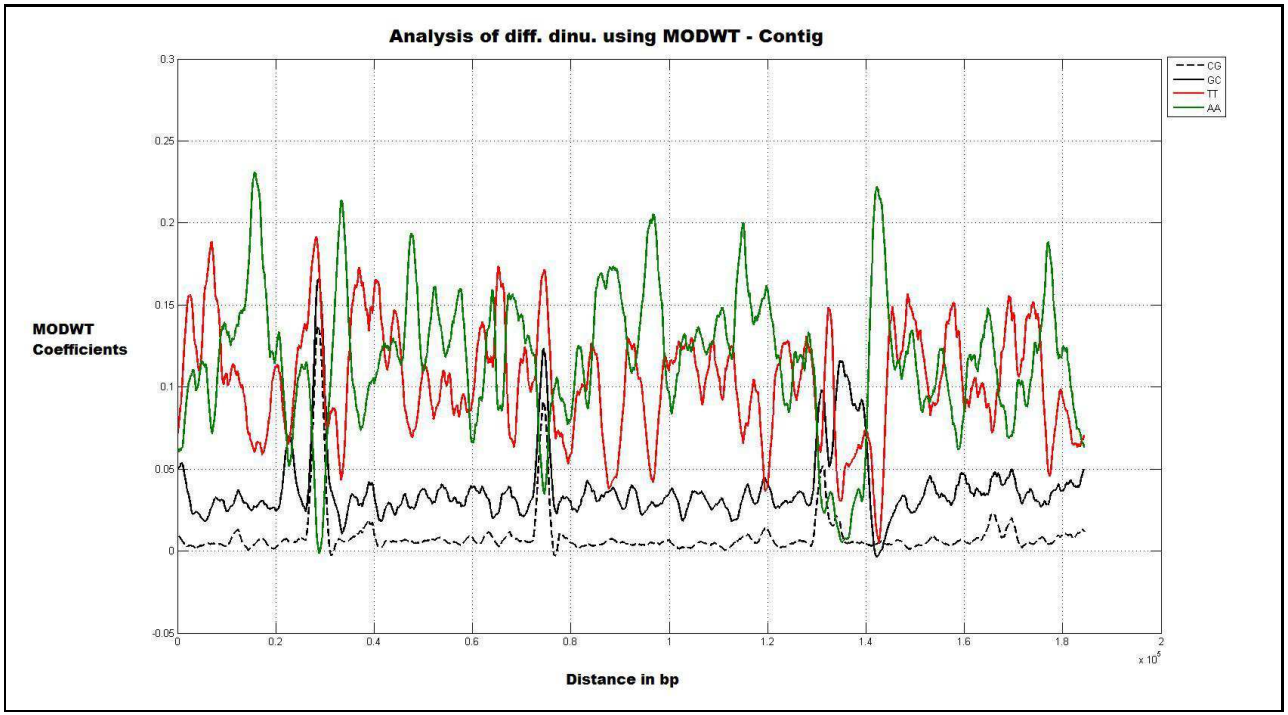


Figure A.4: Dinucleotides Wavelet Comparison

Comparison of Wavelet Coefficients at scale 12 for CG(dash line), GC(black), AA(green) and TT(red). The dataset used was Contiguous sequence with id: NT_113952.1 from Human Genome.

values shadows the distribution of CG dinucleotide associated wavelet coefficients. This confirms the fact that peaks for both datasets; contain genes or CpG islands where they have the same distributions, (between the intervals 28,000-30,000bp, 70,000-80,000bp and 130,000-141,000bp). Among graphs obtained for all the 6 contiguous sequences, (results not shown here), the GC and CG follow a similar trend. AA and TT associated values do not follow a well-defined order of patterns in comparison to GC contents, but do appear complementary to each other.

A.3.3 Covariance Analysis

As a confirmation of our recordings from Fourier and MODWT, of CG, TT/AA and GC dinucleotides, the covariance analysis of the wavelet coefficients for the same contiguous sequence in chromosome 21 is performed. In Figure A.3.3 the progress of covariance of CG with GC, TT/AA were obtained for all scales possible (maximum scale (j) = 18). When CG and GC was compared, both showed maximum variation at lower decomposition levels or scales, as the datasets do not share a lot of common frequency patterns. This covariance decreased for higher scales, because, from our wavelet analysis, we found that the CG and GC distributions are located very near to one another probably following the same frequency of appearance.

An opposite behaviour was noted for CG when compared with AA and TT dinucleotides. Fourier analysis indicates that these dinucleotides share more than 2 common frequency patterns (periodicity of 11, 20 and 25bp). But wavelet analysis indicates that the features shared among CG and AA/TT are not present in the same loci of DNA sequences. This reflects in the covariance of CG-TT and CG-AA at higher scales. A recent analysis reported that human genome is prone to constant mutational changes thus allowing a high conversion of “G” to “T” and “C” to “A”, [Esteller et al., 2001]. This can be stated as a strong reason to support that these 3 dinucleotides share many common features but are not present at the same locations. The similiar behaviour of covariance between AA and TT dinucleotides can be predicted. This conclusion is based on the fact that AA

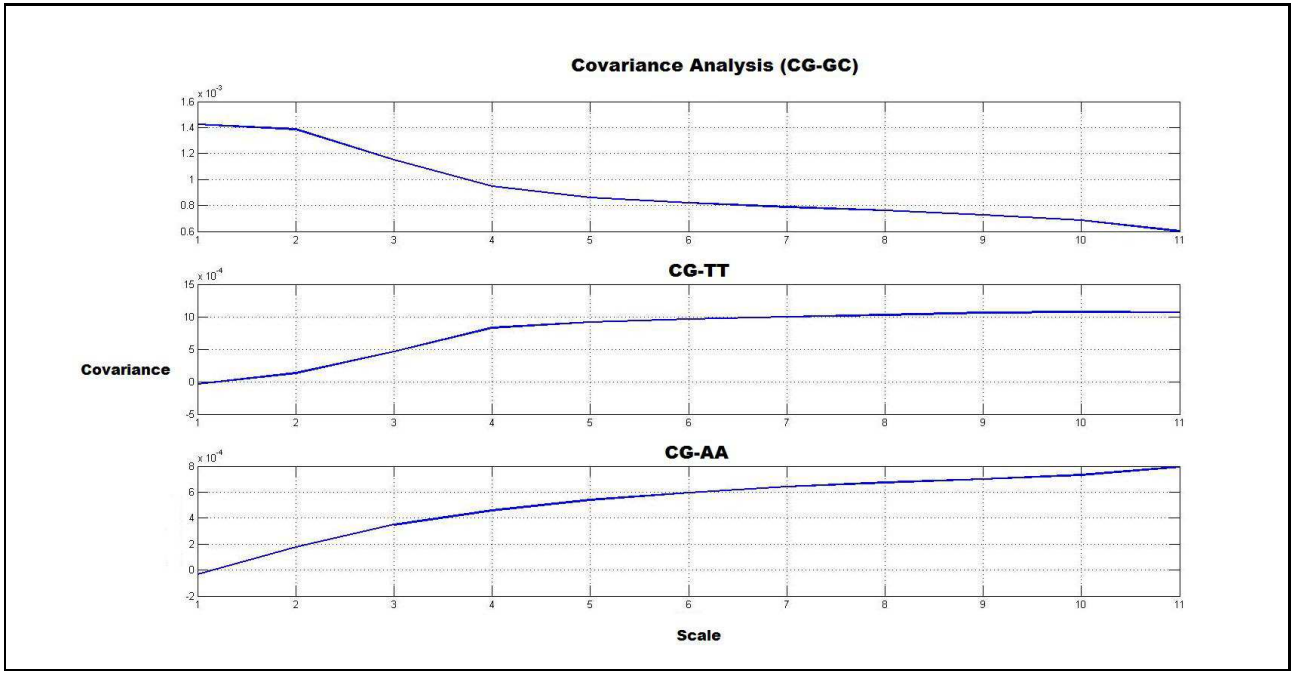


Figure A.5: Wavelet Covariance Analysis

Covariance of Wavelet Coefficients among CG, GC, AA and TT dinucleotides for scale or decomposition level of 11 for Contiguous Sequence 1 - id: NT_113952.1

and TT share the set of distributions or patterns but are not necessarily located at the same place as indicated in Figure A.4.

A.4 Summary

In this chapter an in-depth comparative analysis of the global distribution of different dinucleotides were performed using Fourier and Wavelet transformation techniques. Calculation of covariance of wavelet coefficients of different dinucleotides to understand the difference in their distribution was further explained in detail. New and distinct patterns from the frequency maps of Fourier analysis were reported. Most of these were consistently associated with the non-coding short and long tandem repeat sequences in the genome. Apart from identifying the commonalties among the different dinucleotides, (using Fourier), Discrete Wavelet Transformation helped to locate these common frequencies along DNA sequences. The covariance analysis indicated that different dinucleotides with similiar periodicities were not necessarily co-located along the length of DNA sequences.

The Human genome has evolved over several years thus accumulating a lot of information in the form of complex gene sequences and repeat patterns. Identifying and interpreting the multitude of patterns is not a simple task. The methods described here are the most efficient in terms of extracting useful information for further comprehension. While the use of CG, AA/TT and GC dinucleotides have been discussed in brief; the use of other contributing dinucleotides or small sequences that repeatedly occur in the genome can be performed. This type of analyse in combination with properties of DNA molecules will help to understand the structural condensation of the compact human genome and also events such as DNA replication and Transcription in the genome that depend on local changes in the DNA structure.

Appendix B

Publications and Conference Papers

Book Chapters

1. Raghavan, K., and Ruskin, H.J.,*Modeling DNA Methylation Dynamics*, in Dr. T. Tatarinova and Dr. O. Kerton, DNA Methylation - From Genomics to Technology, (ISBN-9789535103202, 1st Ed., 3-28), InTech Publications, 2012.

Conference Publications

2. Raghavan, K. and Ruskin, H.J., *Computational Epigenetic Micromodel - Framework for Parallel Implementation and Information Flow*. Proceedings of 8th International Conference on Complex Systems 2011 (ICCS) , Boston, USA, June 2011.

Abstract: Following the sequencing of the human genome, attempts to characterise the human epigenetic profile and determine gene expression have focused on the roles played by key elements, such as dynamic histone modifications, and stable DNA methylation. In this paper, we elaborate on the development of a novel micro model, which aims to predict molecular events leading to quantification of gene expression. This bottom-up approach facilitates

low-level information processing and enables an overview of the information exchange between scales, whereby gene expression controls the phenotype or physical appearance of an organism at higher level. The focus here is testing and refinement of the model for a large dataset, in order to assess epigenetic change influence at genome level for abnormal or disease conditions. Evolution of several independent gene blocks, based on molecular interactions across the human chromosome 21, is analysed and impact discussed. Parallel implementation of the micro-model, (using MPI) to handle the large data load, is also described.

3. Raghavan, K., Ruskin H.J. and Perrin, D., *Computational Analysis of Epigenetic Information in Human DNA Sequences*. International Proceedings of Chemical, Biological and Environmental Engineering,Vol. (5), May 2011.

Abstract: Over the last few years, investigations of human epigenetic profiles have identified key elements of change to be Histone Modifications, stable and heritable DNA methylation and Chromatin remodeling. These factors determine gene expression levels and characterise conditions leading to disease. In order to extract information embedded in long DNA sequences, data mining and pattern recognition tools are widely used, but efforts have been limited to date with respect to analyzing epigenetic changes, and their role as catalysts in disease onset. Useful insight, however, can be gained by investigation of associated dinucleotide distributions. The focus of this paper is to explore specific dinucleotides frequencies across defined regions within the human genome, and to identify new patterns between epigenetic mechanisms and DNA content. Signal processing methods, including Fourier and Wavelet Transformations, are employed and principal results are reported.

Journal Publications

4. Raghavan, K., Ruskin, H.J., Perrin, D., , Burns, J., and Goasmat, F., *Computational Micro Model for Epigenetic Mechanisms*, PLoS One,Vol. 5(11): e14031., November 2010.

Abstract: Characterization of the epigenetic profile of humans since the initial breakthrough on the human genome project has strongly established the key role of histone modifications and DNA methylation. These dynamic elements interact to determine the normal level of expression or methylation status of the constituent genes in the genome. Recently, considerable evidence has been put forward to demonstrate that environmental stress implicitly alters epigenetic patterns causing imbalance that can lead to cancer initiation. This chain of consequences has motivated attempts to computationally model the influence of histone modification and DNA methylation in gene expression and investigate their intrinsic interdependency. In this paper, we explore the relation between DNA methylation and transcription and characterize in detail the histone modifications for specific DNA methylation levels using a stochastic approach.

Appendix C

Glossary

1. **Antibody hybridization** – This techniques which is used to identify methylated DNA sequences using specific anti-5-methyl-Cytosine (mC) Antibodies that bind to the methylated short DNA sequences in genomic arrays.

2. **Autocorrelation** – Autocorrelation of patterns is an extension for periodicity, i.e. appearance of a pattern after a lag or distance of "k" base pairs.

3. **Bisulphite Sequencing** – A technique where unmethylated cytosine residues are converted to uracil when allowed to react with sulphite compounds, and sequence discrimination (of methylated and unmethylated) is achieved by designing the short nucleotide sequences to overlap and bind with potential sites of DNA methylation (mainly CpG dinucleotides).

4. **Chromatin Immuno-Precipitation (ChIP)** – The isolation, using specific antibodies, of chromatin fragments that are bound by a particular nuclear factor or associated with a particular histone-modification signature. The immunoprecipated DNA can subsequently be analysed with specific PCR primers, [Esteller, 2007].

5. **ChIP-Seq.** – Experiments conducted to check for protein-DNA interactions combining chromatin immuno precipitation and massively parallel DNA sequencing techniques or microar-

ray (chip) experiments, [Collas, 2010].

6. **ChIP-on-Chip** – A combination of chromatin immunoprecipitation with hybridization to genomic microarrays that is used to identify DNA sequences bound to a particular nuclear factor or with a specific histone- modification profile, [Esteller, 2007].

7. **CG dinucleotides** – a dinucleotide is made of 2 base pairs, in this case C is followed by G.

8. **CpG Islands** - These are specific types of sequences present near the promoter of a Gene. CpG islands are immensely focussed in relation to hypermethylation event during cancer initiation. A sequence can be defined as CpG islands if, (i) The total length of the sequence is more than 200bp, (ii) The total amount o GC contents is $> 50\%$, (iii) an observed/expected CG dinculeotide ratio which must be $>60\%$. (The observed/expected CG ratio is calculated by (Num of CG/(Num of C + Num of G) * Total length of Sequence).

9. **Cloning (Allele Specific)** – Method of duplicating Alleles or gene forms into more than one copy.

10. **Hemi-Methylated** – DNA sequences which have one of its double strands methylated.

11. **High Performance Liquid Chromatography (HPLC)** – is a chromatographic technique used to separate a mixture of compounds in analytical chemistry and biochemistry with the purpose of identifying, quantifying and purifying the individual components of the mixture.

12. **Genomic/Parental Imprinting** – An epigenetic process which allows the Human genome to express only one copy (maternal or paternal origin) to be expressed by methylating the imprinted gene copy. Sometimes both alleles or copies of the same genes are expressed, but most cases allow only one to be expressed. Diseases associated with imprinting disorder are Beckwith-Wiedemann syndrome, Silver-Russell syndrome, Angelman syndrome and Prader-Willi syndrome.

13. **Kinetochore** – A protein found in chromatids which play a crucial role during cell division by acting as a bolster to all the spindle fibres that form during the metaphase of mitosis.

14. **Manual Sequencing** – A simple and direct way of reading through the linear order of nucleotides in a DNA sequence by allowing 4 copies of the sequence to pass through 4 channels that helps to detect only one type of nucleotide at a time.

15. **Mass Spectrometry** – Mass spectrometry (MS) is an analytical technique that measures the mass-to-charge ratio of charged particles. It is used for determining masses of particles, for determining the elemental composition of a sample or molecule, and for elucidating the chemical structures of molecules, such as peptides and other chemical compounds.

16. **Methyl Specific Digital Karyotyping (MSDK)** – A modified version of digital karyotyping (technique used to detect gene copy number variations throughout the genome at high resolution) which is used for the identification of novel differentially methylated sites, based on separating methylated and unmethylated sites.

17. **Microarray Technology** – A microarray is a chip with several minute sockets containing DNA fragments which are allowed to be exposed and hybridized with complementary external DNA sequences to ask to check gene functions after treating the chip further with some tests. e.g. affymetrix microarray chip technology.

18. **Microsatellite Stable, (MSS)** – Human genome consist of simple sequence repeats, (SSR) formed by dinucleotides or short oligonucleotides, (e.g. CACACACA). The contents and length of these are common and normal in normal human cells. The National Caner Institute defined 5 marker SSR which are affected during Colon cancer. Hence scientists check for the presence of these SSR within coding regions of tumor suppressor genes using various detection techniques. If the alleles of these genes do not contain the marker SSR, the cells

are reported to have a condition called Microsatellite Stability or MSS. If the alleles or genes DO contain these marker SSR, the condition is named as **Microsatellite Instability** or **MSI**.

19. **mRNA** - An intermediate product whose function is to encode the information from DNA molecules, which if transferred to the site of translation help to assemble amino acids are assembled to form long chains of protein

20. **miRNA** – Micro RNA or miRNA are a type of RNA Molecules whose length is approximately 22bp whose main function is to silence mRNA sequences. These small sequences, are known to be most commonly utilized in gene regulations.

21. **Non-Coding Regions** – The non coding regions referred here in this analysis are the segments in-between exons/coding regions and are removed during translation or protein production phase.

22. **Polymerase Chain Reaction (PCR)** – A molecular biology technique that is used to amplify or produce large (usually of the order thousands or millions) copies of a single or small number of DNA segments.

23. **PyroSequencing** – A DNA sequencing method in which light is emitted as a result of an enzymatic reaction, each time a nucleotide is incorporated into the growing DNA chain. As applied to methylation detection, methylation-dependent DNA sequence variation, which is achieved by sodium bisulphite treatment, is treated as a kind of single nucleotide polymorphism (SNP) of the C-T type, and is subjected to conventional SNP typing, [Esteller, 2007].

24. **RAS-Pathway** – The human cells contain a group of genes called “Ras” involved in transmission of signals from the environment. The ras-pathway involves activation of these genes to produce corresponding proteins which in turn activate genes responsible for cell growth differentiation and survival. The ras genes can undergo mutation to attain a permanent ex-

pressive state leading to cancer initiation. Hence its a pathway of immense interests to cancer biologists, e.g. Mutation of K-Ras genes is an indicator of Colon Cancer, [Kawamoto et al., 2012].

25. **Restriction Landmark Genomic Scanning (RLGS)** – A methylation detection and analysis technique which involves DNA radioactive labelling at methylation-specific cleavage sites (cleavage sites are where restriction enzymes cut the DNA polymer) and size-fractionated in one dimension. The digestion or trimmed products are further digested with two more restriction enzymes and then are separated by a technique called two-dimensional electrophoresis, yielding a number of scattered hot spots of DNA methylation. The resulting graph gives plain areas and those with spots which represent methylated and unmethylated regions from genome respectively.

26. **Sensitive Methyle Light (MSP)** – A DNA methylation detection techniques, which is usually combined with bi-sulphite treatment and polymerase chain reaction to detect either unmethylated or totally methylation sequences. The MethylLight relies on fluorescence based detection of required targets from a batch of samples.

27. **Time Series Analysis methods** – Set of statistics based methods and transformation techniques applied to study appearance of specific components in the spatially-varying data such as in DNA sequences, e.g. Fourier Transformation and Wavelet Transformation Techniques.

28. **Tumorigenesis** – The process of initiating or instigating production of tumors.

29. **WNT Pathway** – A intermediate pathway that signals many proteins, and consequently allow expression of genes responsible of cell growth, differentiation and survival. This family of proteins are highly conserved across organisms and have recently been linked to cancer initiations and progress.

Appendix D

Datasets and Auxiliary Information

This chapter provides auxiliary images and tables associated with the work reported in Chapters 3 to 7. The first is a table of 121 conditional events, extracted from StatEpigen database. This dataset presented in the table was analysed in the two-part model proposed in Chapter 6. The second is a list of images containing information about types of patterns formed by a subgraph made of 4 nodes. These patterns were detected by mfinder tool in the network that was developed by using data from the StatEpigen database. Thirdly a list containing 440 genes is provided and these genes were used to understand the distribution of different dinucleotides in Appendix A. Lastly, a list of useful websites or url are provided.

Table of Dataset from StatEpigen

S No	Gene 1	Gene 2	Event 1	Event 2	Frequency
1	MLH1	DAPK1	HypMet	HypMet	0
2	PYCARD	TP53	HypMet	Mut	0
3	CDKN2AP14	DAPK1	HypMet	HypMet	0.047
4	CDKN2AP16	DAPK1	HypMet	HypMet	0.066
5	TIMP3	DAPK1	HypMet	HypMet	0.1
6	TIMP3	APC	HypMet	HypMet	0.125
7	CDKN2AP14	APC	HypMet	HypMet	0.14
8	MGMT	BRAF	HypMet	Mut	0.14
9	MLH1	JCVT	HypMet	Exp	0.15
10	RASSF1A	CDKN2AP16	HypMet	HypMet	0.154
11	DAPK1	APC	HypMet	HypMet	0.188
12	MAL	HOXA9	HypMet	HypMet	0.2
13	MGMT	APC	HypMet	Mut	0.21
14	MGMT	TIMP3	HypMet	HypMet	0.217
15	RASSF2	KRAS	HypMet	Mut	0.238
16	SOCS1	JCVT	HypMet	Exp	0.24
17	CDKN2AP16	APC	HypMet	HypMet	0.244
18	CDKN2AP14	JCVT	HypMet	Exp	0.25
19	RUNX3	JCVT	HypMet	Exp	0.25
20	MLH1	APC	HypMet	HypMet	0.25
21	IGF2	JCVT	HypMet	Exp	0.26
22	CACNA1G	JCVT	HypMet	Exp	0.27
23	APBA3	JCVT	HypMet	Exp	0.29
24	WRN	JCVT	HypMet	Exp	0.3
25	CDKN2AP14	CDKN2AP16	HypMet	HypMet	0.303
26	CRABP1	JCVT	HypMet	Exp	0.31
27	NEUROG1	JCVT	HypMet	Exp	0.31
28	SFRP1	SFRP4	HypMet	HypMet	0.31
29	MGMT	TP53	HypMet	Mut	0.31
30	CRABP1	HIN1	HypMet	HypMet	0.32
31	TIMP3	CDKN2AP14	HypMet	HypMet	0.325
32	MGMT	HLTF	HypMet	HypMet	0.326
33	APBA1	JCVT	HypMet	Exp	0.33
34	HIC1	JCVT	HypMet	Exp	0.33
35	HIC1	BRAF	HypMet	Mut	0.33
36	RASSF2	BRAF	HypMet	Mut	0.33

Table 1 – *Continued from previous page*

S No	Gene 1	Gene 2	Event 1	Event 2	Frequency
37	SFRP2	SFRP4	HypMet	HypMet	0.333
38	MGMT	KRAS	HypMet	Mut	0.34
39	CHFR	JCVT	HypMet	Exp	0.35
40	MGMT	JCVT	HypMet	Exp	0.35
41	APBA3	BRAF	HypMet	Mut	0.35
42	CDKN2AP16	JCVT	HypMet	Exp	0.36
43	IGFBP3	JCVT	HypMet	Exp	0.37
44	CDKN2AP16	RUNX3	HypMet	HypMet	0.382
45	CDKN2AP16	HLTF	HypMet	HypMet	0.4
46	ID4	BRAF	HypMet	Mut	0.41
47	CDKN2AP16	CDKN2AP14	HypMet	HypMet	0.422
48	CRABP1	CDKN2AP16	HypMet	HypMet	0.44
49	CRABP1	MLH1	HypMet	HypMet	0.44
50	MAL	MGMT	HypMet	HypMet	0.475
51	CRABP1	MGMT	HypMet	HypMet	0.48
52	CRABP1	NR3C1	HypMet	HypMet	0.48
53	APBA1	BRAF	HypMet	Mut	0.48
54	APC	JCVT	HypMet	Exp	0.481
55	MLH1	MGMT	HypMet	HypMet	0.49
56	CDKN2AP16	TIMP3	HypMet	HypMet	0.49
57	RUNX3	HIN1	HypMet	HypMet	0.5
58	ROBO1	SLIT2	HypMet	HypMet	0.5
59	TIMP3	JCVT	HypMet	Exp	0.517
60	CDKN2AP16	BRAF	HypMet	Mut	0.52
61	APC	MCC	HypMet	HypMet	0.524
62	MGMT	CDKN2AP16	HypMet	HypMet	0.543
63	MLH1	CDKN2AP14	HypMet	HypMet	0.562
64	RUNX3	MLH1	HypMet	HypMet	0.563
65	SFRP1	SFRP5	HypMet	HypMet	0.58
66	RARB	JCVT	HypMet	Exp	0.594
67	SFRP2	SFRP5	HypMet	HypMet	0.61
68	RASSF1A	SLIT2	HypMet	HypMet	0.615
69	UNC5C	DCC	HypMet	HypMet	0.62
70	MLH1	ST3GAL1	HypMet	HypMet	0.625
71	CRABP1	RUNX3	HypMet	HypMet	0.64

Table 1 – *Continued from previous page*

S No	Gene 1	Gene 2	Event 1	Event 2	Frequency
72	MLH1	CHFR	HypMet	HypMet	0.68
73	RUNX3	NR3C1	HypMet	HypMet	0.68
74	APC	TP53	HypMet	Mut	0.688
75	MLH1	TIMP3	HypMet	HypMet	0.69
76	MLH1	CDKN2AP16	HypMet	HypMet	0.75
77	MLH1	FBXL7	HypMet	HypMet	0.75
78	MLH1	LYPD1	HypMet	HypMet	0.75
79	DCC	UNC5C	HypMet	HypMet	0.75
80	BMP3	EYA2	HypMet	HypMet	0.756
81	MLH1	HLTF	HypMet	HypMet	0.769
82	CDKN2AP16	MCC	HypMet	HypMet	0.794
83	CCNA1	CDKN2B	HypMet	Exp	0.8
84	MAL	ADAMTS1	HypMet	HypMet	0.8
85	RUNX3	TGFBR2	HypMet	Mut	0.82
86	MLH1	BRAF	HypMet	Mut	0.83
87	MLH1	DPYSL3	HypMet	HypMet	0.833
88	MLH1	PTPRO	HypMet	HypMet	0.833
89	CDKN2AP16	SLIT2	HypMet	HypMet	0.833
90	BMP3	ALX4	HypMet	HypMet	0.87
91	MLH1	ADAMTS19	HypMet	HypMet	0.875
92	MLH1	NRG2	HypMet	HypMet	0.875
93	MLH1	SLC30A10	HypMet	HypMet	0.875
94	MLH1	SLC30A3	HypMet	HypMet	0.875
95	BMP3	VIM	HypMet	HypMet	0.9
96	SFRP1	SFRP2	HypMet	HypMet	0.91
97	SFRP2	SFRP1	HypMet	HypMet	0.96
98	DAPK1	FAS	HypMet	Exp	1
99	DAPK1	FASLG	HypMet	Exp	1
100	PTEN	JCVT	HypMet	Exp	1
101	MLH1	BMP3	HypMet	HypMet	1
102	MLH1	C13ORF21	HypMet	HypMet	1
103	MLH1	CBS	HypMet	HypMet	1
104	MLH1	CLGN	HypMet	HypMet	1
105	RUNX3	CRABP1	HypMet	HypMet	1
106	MLH1	EVL	HypMet	HypMet	1

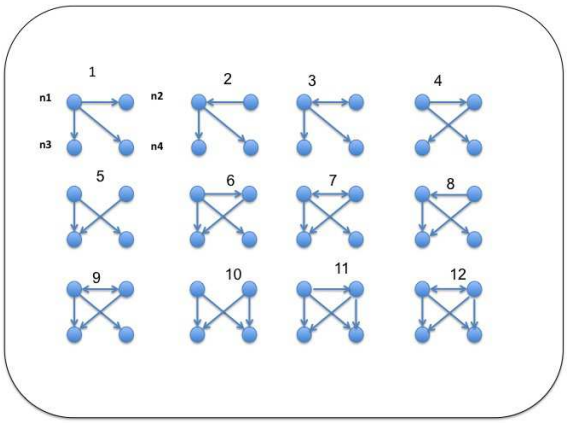
Table 1 – *Continued from previous page*

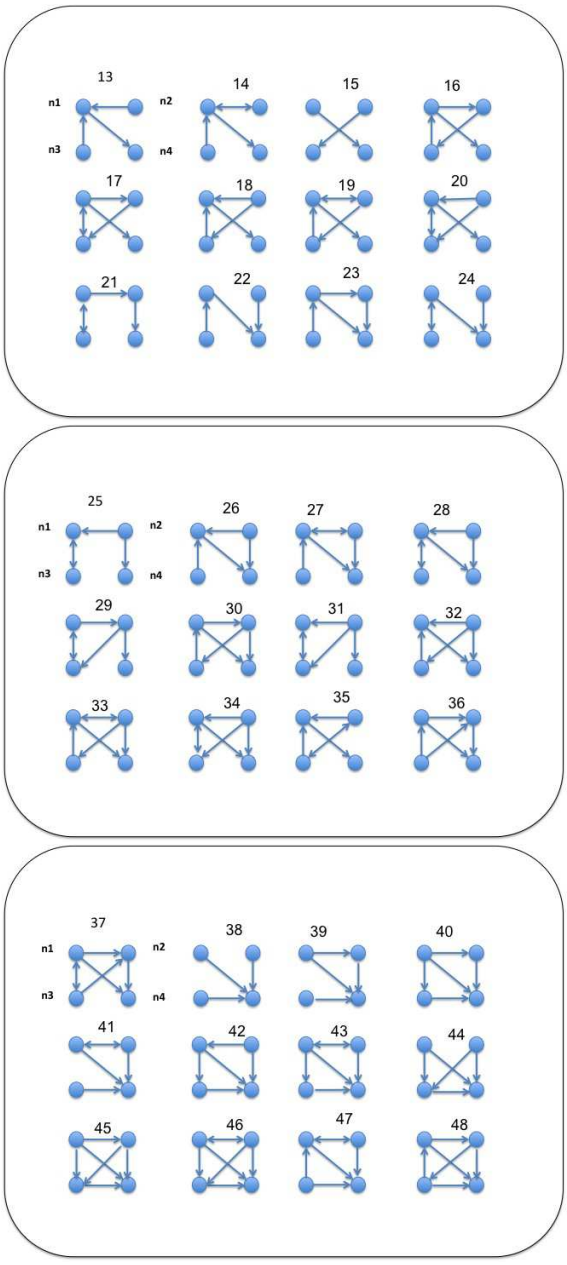
S No	Gene 1	Gene 2	Event 1	Event 2	Frequency
107	MLH1	FLJ37464	HypMet	HypMet	1
108	MLH1	FLJ41549	HypMet	HypMet	1
109	MLH1	GDF7	HypMet	HypMet	1
110	MLH1	IMAGE5728979	HypMet	HypMet	1
111	MLH1	KCNK13	HypMet	HypMet	1
112	MLH1	KIT	HypMet	HypMet	1
113	MLH1	LOC283887	HypMet	HypMet	1
114	MLH1	LRRC4	HypMet	HypMet	1
115	MLH1	MED12L	HypMet	HypMet	1
116	MLH1	NELL2	HypMet	HypMet	1
117	MLH1	NPHS2	HypMet	HypMet	1
118	MLH1	PAPLN	HypMet	HypMet	1
119	MLH1	PLEKHC1	HypMet	HypMet	1
120	MLH1	SOX7	HypMet	HypMet	1
121	MLH1	TCF7L1	HypMet	HypMet	1

Figure D.1: StatEpigen Dataset

The Conditional events obtained from StatEpigen Database; Mut = Mutation, HypMet = HyperMethylation, Exp = Gene Expression. Note: Conditional events with a zero frequency of occurring have also been included from StatEpigen.

Patterns detected in Motifs of size 4




13
n1
n2
n3
n4
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48

D6

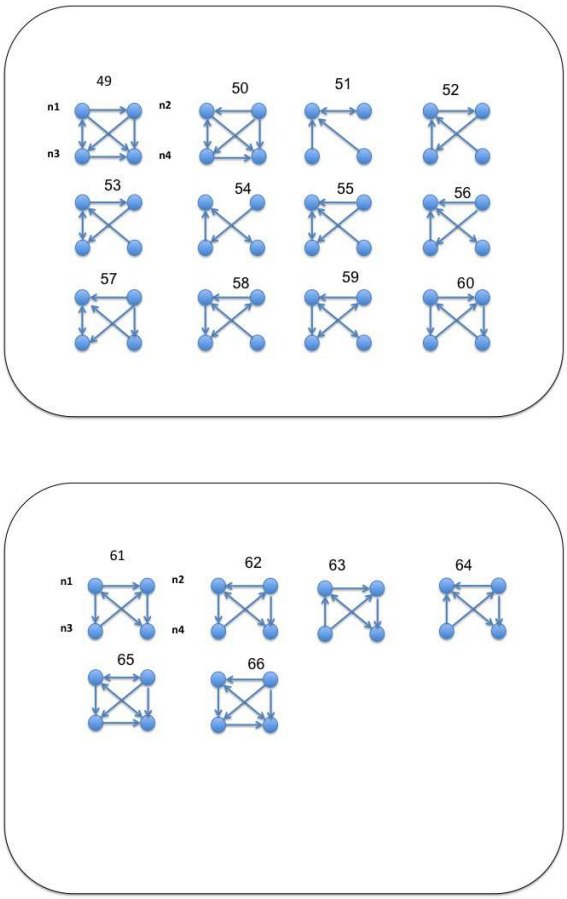


Figure D.2: Patterns in Motifs of size 4

As described in Appendix A, a global analysis of dinucleotide frequencies in many DNA sequences across the human genome was carried using time series analysis methods. The names of genes used in this analysis are given below. From each chromosome, DNA sequences of 20 genes associated with disease was extracted using UCSD genome browser.

1. Chromosome 1:

GPR153, ISG15, NOL9, RERE, PRAMEF6, PRAMEF15, C1orf201, PAFAH2, CATSPER4, SFN, PTCH2, FMO5, OR10T2, C1orf192, GLRX2, OR4F16, LRP8, AGRN, TMEM88B, CDC2L1

2. Chromosome 2:

ETM2, TMEM18, TSSC1, TTC15, ALLC, KIDINS220, RRM2, RPL6P4, NTSR2, NBAS, RPLP1P5, MSGN1, LAPTM4A, GPR113, CGREF1, UCN, GCKR, BRE, MYADML, CRIM1

3. Chromosome 3:

HPC5, SCA29, RPL23AP38, CRBN, MRPS35P1, GRM7, OVCAS1, OXTR,RAD18, LHFPL4, CAMK1, JAGN1, TMEM111, CYCSP11, VHL, ATP2B2, UBM2, SYN2, MKRN2, TMEM40

4. Chromosome 4:

ZNF595, ZNF718, MYL5, GAK, TMEM175, FGFRL1, SPON2, CTBP1, MAEA, SLBP, WHSC1, RNF4, GRK4, RGS12, DOK7, ADRA2C, EVC2, CRMP1, FGFBP1, GPR125

5. Chromosome 5:

SDHA, AHRR, TPPP, TRIP13, SLC12A7, LOC728613, LOC442131, CCT5, ROPN1L, ANKH, ZNF622, BASP1, CDH12, CDH9, BDA1B, MTMR12, C1QTNF3, PRLR, IL7R, PRKAA1

6. Chromosome 6:

IRF4, WRNIP1, RPP40, CAGE1, SSR1, BMP6, TMEM14B, HIVEP1, KAAG1, SCGN, ZNF165, ZNF193, HLA-G, PPP1R11, HLA-B, LST1, LY6G5B, HSPA1L, CYP21A2, TAPBP

7. Chromosome 7:

   PDGFA, GPER, INTS1, EIF3B, GNA12, SDK1, RBAK, CYTH3, ZNF12, COL28A1, HDAC9, IL6, GPNMB, GHRHR, SEPT(07), GPR141, RAMP3, EGFR, CCT6A, ZNF273

8. Chromosome 8:

   FBXO25, MYOM2, ANGPT2, SPAG11B, RP1L1, BLK, GATA4, TUSC3, FGF20, MTMR7, MTUS1, FGF17, TNFRSF10B, PNMA2, PTK2B, GULOP, GPR124, STAR, FGFR1, TCEA1

9. Chromosome 9:

   CDKN2A, PAX5, PTCH1, TAL2, NUP214, COL5A1, GLE1, DBH, SURF1, NOTCH1, ABCA1, TOPORS, VLDLR, TEK, TRPM6, FKTN, LMX1B, AK1, LHX3, INVS

10. Chromosome 10:

    NODAL, PHYH, GATA3, FLJ14813, MSMB, NCOA4, CCDC6, ZNF365, NEUROG3, SFTPA2, RPS24, LIPA, ABCC2, TLX1, RBP4, RBP4, MXI1, TCF7L2, PAX2, KIAA1279

11. Chromosome 11:

    HRAS, NUP98, CDKN1C, TPP1, SMPD1, PTH, TSG101, KCNJ11, LMO2, CAT, MAPK8IP1, RAPSN, CD82, CCND1, NUMA1, NDUFS8, ROM1, CLLS1, CALM, PPP2R1B

12. Chromosome 12:

SUOX, LRP6, OLR1, CDKN1B, AICDA, ETV6, KCNA5, TPI1, CACNA2D4, YEATS4, ACVR1B, KRT8, KRT1, KRT4 KRT5 KRT75, KIF5A, CDK4, MDM2, IGF1

13. Chromosome 13:

    FLT3, RNF6, SPG20, FREM2, RFXAP, RB1, ATP7B LIG4, SLITRK1, ING1, PCCA ZIC2, SLC10A2, IRS2, COL4A1, GRK1, ERCC5, FOXO1A, PHF11, IPF1

14. Chromosome 14:

    GOLGA5 SPG3A, DAD1, NRL, PABPN1, PCK2, TCRA, ANG, CEBPE, COCH, MYH7, PSMA6, FOXG1B, FBXO33, FANCM, GCH1, SIX6, MTHFD, POMT2, MLH3

15. Chromosome 15:

    BCL8, RPS17, NDN, UBE3A, GABRB3, OCA2, SNRPN, NOLA3, SLC12A1 PLCB2, BUB1B, RAD51A, RAB27A, FBN1, MPI, AQP9, HCN4, IGF1R, RECQL3, SEPS1

16. Chromosome 16:

    SCNN1B, MHC2TA, MEFV, ERCC4, ERCC4, PMM2, GLIS2, ABAT, LMF1, CREBBP, MYH11, HSD3B7, PALB2, IL21R, SLC5A2, PHKB, FTO, MMP2, CBFB, AMLCR2

17. Chromosome 17:

    SCOD1, PRPF8, MPDU1, TP53, ELAC2, PMP22, FLCN, RAI1, FOXN1, RNF135, JJAZ1, THRA, SGCA, KRT12, NAGLU, PHB, RARA, NME1, CSH1, AXIN2

18. Chromosome 18:

TTR, RBBP8, DSG1, MALT1, TCF4, FVT1, BCL2, TNFRSF11A, CYB5A, CTDP1, MYO5B, MADH4, LPIN2, NDUFV2, MC2R, DYM, DSG2 SETBP1 MADH4, FVT1

19. Chromosome 19:

    CRTC1, NDUFS7, PSORS6, SH3GL1, LMNB2, TCF3, GAMT, ELA2, CFD, LPSA, LYL1, IL12RB1, HPCQTL19, BCL3, TGFB1, ITPKC, PRPF31, TSEN34, HPC15, STK13

20. Chromosome 20:

    PRNP, SLC4A11, PANK2, PROKR2, RSPO4, JAG1, BMND7, C20orf7, VSX1, MMP9, SRC, GRD2, AURKA, KCNQ2, COL9A3, VAPB, STX16, SOX18, PTPN1, THBD

21. Chromosome 21:

    PRSS7, IFNGR2, KCNE1, MRAP, IFNAR2, SOD1, CRFB4, KCNE2, ITGB2, CBS, FTCD, PFKL, RUNX1, COL6A1, COL6A2, PCNT2, CSTB, LIPI, TMPRSS3, APP

22. Chromosome 22:

    CECR, SMARCB1, GGT2, TCN2, CRYBA4, SNAP29, MIF, MYH9, TBX1, BCR, TMPRSS6, RAC2, SOX10, UPK3A, ECGF1, ALG12, TRMU, LGALS2, TIMP3, FBXO7

List of useful websites or URL

1. Gene cards: `http://genecards.org`
2. MODWT Package: `http://www.atmos.washington.edu/~wmtsa/`
3. NCBI database: `www.ncbi.nlm.nih.gov`

4. StatEpigen: `http://statepigen.sci-sym.dcu.ie`

5. UCSD Genome Browser: `http://genome.ucsc.edu/`

6. UniGene: `http://www.ncbi.nlm.nih.gov/unigene`