Epigenetic Modelling: DNA Methylation and working towards Model Parameterisation

Jyoti Porwal

A thesis submitted in partial fulfillment of the

requirements for the degree of

Master of Science (Research)

in the

School of Computing

Faculty of Engineering and Computing, School of Computing

Dublin City University,

Supervisor: Prof. Heather J. Ruskin,

Auxiliary Supervisor: Dr. J. Burns

June, 2011

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of master by research 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.

Signature:· · · · · · · · · · · · · · ·

ID: 57123365

Date: · · · · · · · · · · · · · · · · · ·

Contents

Abstract

The main focus of the research in this thesis is the investigation in DNA methylation mechanisms of epigenetics and the study of a specific database. As part of the latter work, the role of curation is described, and a new knowledge management system, PathEpigen[1] , is reported that is currently being developed for colon cancer in the Sci-Sym centre. The database deals with genetic and epigenetic interactions and contains considerable data on molecular events such as genetic and epigenetic events. The data curation includes biomedical and biological information. An efficient method was devised to extract biological information from the literature to process, manage and upgrade data. We present a Deterministic Finite Automata (DFA) model for the DNA methylation mechanism controlled by DNA methyltransferase (DNMT) enzymes. This thesis provides a brief introduction to epigenetics, a survey of ongoing research on computational epigenetics and a description of the DNA methylation database. Furthermore, it also gives an overview of DNA methylation and its importance in cancer. The DFA models three states of methylation frequency (normal, *de-novo* and hypermethylated) in the cell. It has been executed on input of random strings of size 100. Out of the strings considered, we found that 26%, 37% and 37% correspond to normal, *de-novo* (cancer initiation) and hypermethylated (cancer) states, respectively.

[1] The PathEpigen Knowledge Management System is continually being added to in terms of data content and additional features. It has also been renamed StatEpigen, to reflect its focus on the statistical information. It is currently found at http://statepigen.sci-sym.dcu.ie/index.php.

Acknowledgements

I would like to thank Prof. Heather J. Ruskin for her support, expertise and research insight. Her thoughtful advice helped me to maintain a sense of direction during the course of my work. She has been a pillar of strength and source of great inspiration. I am heartily thankful to my elder brothers Pavan Kumar and Bipin Kumar who motivated me to work in this project. I would like to thank the almighty who had been always there to help me to sustain the endurance during my present investigation by providing me such a golden opportunity and my parents for encouraging me to combat with the challenges coming in my life and showing faith in me. I want to express my gratitude to Dr. John Burns and Dr. Ana Barat, and to the whole research group at Syi-Sym center, DCU for the excellent atmosphere and for their stimulating discussions, their help with technical issues and general advice.

Jyoti Porwal

Date: June 2011

List of Figures

List of Tables

List of Abbreviations

Fr:	Frequency of tumour in event 2
CIMP:	CpG island methylator phenotype
MBD:	Methyl-CpG-binding domain protein
MBD 1:	Methyl-CpG-binding domain protein 1
MBD 2:	Methyl-CpG-binding domain protein 2
MBD 3:	Methyl-CpG-binding domain protein 3
MeCP-2:	Methyl CpG binding protein 2
NAM:	Normal adjacent mucosa
MSI:	Microsatellite instability
NTE1:	Number of tumour event1
Clinphat and CP:	Clinical pathology
DNMTs:	DNA methyltransferase
nb_samples:	Numbers of samples
LOH:	Loss of heterogeneity
HM:	Hypermethylation
Qnt:	Quantification
Phn:	Phenotype
Perc:	Percent
mut:	mutation
Ev:	Event
NL:	NULL
Y:	Yes
G:	gene
L:	Level
H:	High

Chapter 1

Introduction

1.1 Context

Epigenetics is a field in which profound investigations of DNA and chromatin modifications often have reversible and far-reaching impacts on inheritance. It appears in the literature as far back as the mid-19th century, although its conceptual origins date further back. Information management in the nucleus means that some of the genetic information is extraordinarily tightly packaged in the genome. Moreover, there is genetic information that must be on and active all of the time, such as housekeeping genes. Thus, epigenetics can resemble information management at home: things that you need all the time do not get stored away, but your old school records are kept packed in boxes in the attic. Computational methods for the analysis and interpretation of large epigenomic datasets can deepen our knowledge of epigenetics and disease and may even enable optimized therapies, thus establishing computational epigenetics as a relevant and exciting field at the intersection of epigenetic and bioinformatic research. Research in computational epigenetics consists of developing bioinformatic methods capable of solving epigenetic questions as well as theoretical modelling of DNA methylation.

1.2 Scope

In the present dissertation, we looked at elements needed to construct a simple microscopic model of epigenetic mechanisms, specifically those relating to DNA Methylation features. As initial objectives, this thesis both develops a deterministic model for analysing the activity of DNA methyltransferase enzymes on CpG islands (for tumour suppressor genes in cancer initiation) and curates data on colon cancer. Ideally, such a model would enable research into events at the molecular level that influence overall cancer growth. This thesis is divided into 5 chapters. Chapter (1) provides an overview of the content of the thesis. Chapter (2) gives a comprehensive discussion of epigenetic structure and the role of DNA methylation in both cell growth and cancer initiation. DNA methylation has a critical role in complex diseases such as cancer. The relationship between epigenetics and cancer is also discussed in this chapter. In addition, Chapter (2) provides an overview of the computational representations of epigenetic mechanisms and includes a discussion of the challenges and complexity involved in developing computational models for epigenetics. These models provide an interface between the biological concepts and the nature (or properties) of data. Furthermore, this chapter discusses the scope of computational epigenetics as well as epigenomic analyses and predictions. In Chapter (3), a model of DNA methylation is presented. Fundamentals of epigenetic micro-modelling using the DFA[1] model are

[1] In the theory of computation, a DFA is a deterministic finite state machine which accepts finite 'strings of symbols' over some alphabet. Set of strings form a DFA language.

detailed. Primary results from this prototype model are discussed. The importance of data curation in substantiating model forms is discussed in Chapter (4). Such data can be used to populate an in-house database, namely, PathEpigen[2], (Ruskin et al., 2008) and provides the basis for the parametrisation of more advanced model forms. Finally, conclusions are provided in Chapter (5). The main focus for our research is to lay the groundwork for data analysis and to develop a model for investigating complex epigenetic mechanisms such as DNA methylation.

[2] The PathEpigen Knowledge Management System is continually being added to in terms of data content and additional features. It has also been renamed StatEpigen, to reflect its focus on the statistical information. It is available at http://statepigen.sci-sym.dcu.ie/index.php.

Chapter 2

A Survey of Epigenetics

This chapter provides introductory background on genomes[1] and epigenomes[2]. Subsequent sections introduce epigenetics, examples of epigenetics and tools. In addition, the chapter summarises the DNA methylation developments in epigenetics and provides a global overview of the various DNA methylation processes and the applications of computational epigenetics.

2.1 Introduction

Any given biological cell contains the complete hereditary information for its species in the form of DNA known as the genome. The basic building blocks of DNA are the base pairs [A(Adenine)−T(Thiamine) and C(Cytosine)−G(Guanine)]. Long DNA chains (200 base pairs wrapped around a histone[3] octamer) are arranged in a structure known as the nucleosome. Multiple nucleosomes comprise a bundle that is known as

[1] The total DNA contents in a single chromosome in a organism.

[2] A subset of genes whose function is controlled by specific biochemical factors.

[3] It is small spheres that DNA wraps around. If the way that DNA is wrapped around the histones changes, gene expression can change as well.

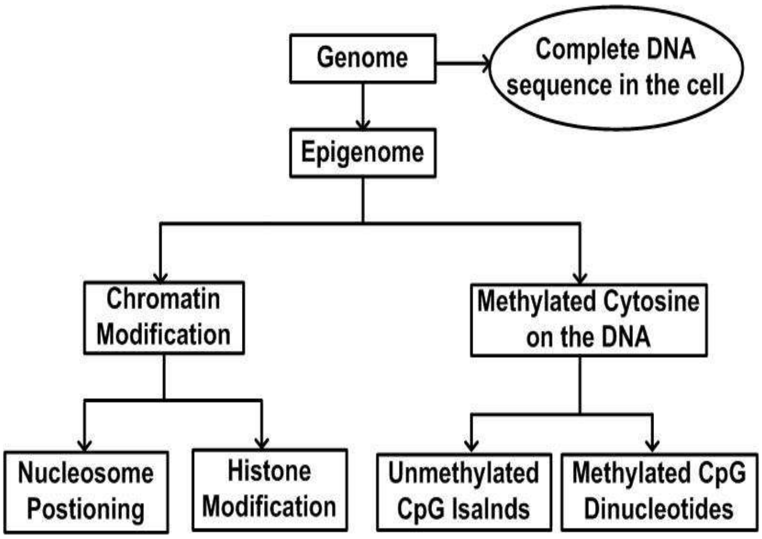


Figure 2.1: Possible states: Genome vs. Epigenome. Epigenome mainly considers two states: (i) Methylated states of the CpG sites on the DNA, and (ii) Chromatin modification. Methylated Cytosine is normally unmethylated at promoter and methylated in the rest of the genome. On the other hand, modification in chromatin get direct impacts on the nucleosome position and its protein (histone).

chromatin. Consequently, the sequential organisation of the DNA in a cell is as follows: DNA strand ⟶ [histone + DNA (200 base pair)] ⟶ nucleosome ⟶ chromatids[4] ⟶ chromosome[5]. Chemical modification of the DNA and histones that takes place without changing the entire sequence is known as 'epigenetics' (i.e., 'epi= over and 'genetics'). Generally, these changes occur on DNA (Cytosine and Thymine) nucleotides and the histone proteins. In the 'epigenome', we study the overall epigenetic state of a cell. Possible epigenetic states are illustrated in Figure (2.1).

[4] One of two identical strands into which a chromosome splits during cell division.

[5] A chromosome is an organised structure of DNA and protein, found in cells.

2.1.1 CpG Islands

CpG[6] dinucleotides occur randomly in the genome. Normally, CpG dinucleotides are found to be frequently methylated across the whole genome. In contrast, when CpG dinucleotides are found at high densities at the promoter[7] region of a gene, (mainly those genes that have a primary function in cell division), they are known as CpG islands (or CG) islands. A CpG island is defined as a sequence that has a G+C content greater than 60% and a CpG to GpC ratio of at least 0.6 or 200 GC base pairs, (Antequera and Bird, 1993). They are found at high densities at the 5' regulatory regions of genes and are normally unmethylated. Recent predictions have lowered the estimates of the number of CpG islands occurring across the whole genome to $\approx 27,000$ CG base pairs, (Antequera, 2003).

2.1.2 Epigenetics: History and Relevance

The biological term 'epigenetics' was introduced by, Waddington (1953). It refers to heritable changes of observable traits (phenotypes) in living beings. Epigenetic mechanisms control gene expression without any changes in the DNA sequence, (Bird, 2007). The study of epigenetics includes investigations of two major modifications of DNA and chromatin, (Feinberg, 2007):

- Post-replication DNA methylation (a covalent modification of Cytosine residue).
- Post-translational modification of histones such as histone acetylation and methylation respectively.

[6] "C" and "G" are connected by a phosphodiester bond on single stranded DNA.

[7] A promoter is a region of DNA that facilitates the transcription of a particular gene.

Functionally, these modifications of epigenetics act as a mark that regulates gene activity. A fundamental question in biological research is to address how a gene receives information when multicellular organisms evolve from a single stem cell, (Reik, 2007). Further, during development, there is a need to know which molecular mechanisms are involved in phenotypic inheritance, (Richards, 2006). Recently, epigenetics has attracted much attention in molecular biology and cancer research. In cancer research, the study of epigenetics opens up the possibility of new approaches, such as identifying gene methylation levels for the early diagnosis and treatment of primary cancers, (Jones and Baylin, 2007; Costa and Shaw, 2007). Typical biological experiments deal with enzymatic activity and analyse the up - or downregulation of gene activity, (Chen et al., 2003; Clark and Melki, 2002). To study gene regulation activity, it is important to investigate the role of factors that affect the degree of regulation and the phenotypic behaviour of epigenetic mechanisms, (Jiang et al., 2004). The parameter range of epigenetic patterns between organisms due to phenotypic variation, together with the time scale of cell stages, (Rando and Verstrepen, 2007), are very important for determining cancer initiation but are non-trivial to deduce. Further, these epigenetic changes can be used as a first step in the modelling of DNA methylation. The methylation profile in humans must satisfy certain criteria established by the epigenomic database. The following sections briefly introduce epigenetics tools and indicate a number of different modelling approaches taken to provide insights into this active research area.

2.1.3 Examples and Tools of Epigenetics

Epigenetic knowledge has many potential applications in medical science. One of the best examples of its relevance in humans is the transfer of information during stem cell differentiation. When a fertilised stem cell divides, a zygote is created, which transfers the information to other cell types, (Reik, 2007). This information transfer into cells is accomplished by epigenetic mechanisms, including DNA methylation and histone modifications. A second example of the importance of epigenetics is in human diseases, such as cancer, (Miyamoto and Ushijima, 2005) and epigenetic therapy (Brown and Strathdee, 2002). In some congenital diseases, such as Angelman and Prader Willi syndrome, epigenetics plays an important role, (Waterland and Jirtle, 2003) in terms of gene expression patterns. Another example concerning epigenetic modification is that some dietary supplements given to mammals cause epigenetic changes in gene expression that affect their phenotype and lead to cancer initiation, (Cooney et al., 2002). The Human Epigenome Project (HEP) has taken on the challenges and opportunities of high-resolution epigenomic analysis in multiple unrelated individuals. Two large-scale efforts, the Chromatin Immunoprecipitation (ChIP) and Bisulfite sequencing projects, have recently been completed and have analysed epigenetic markers (DNA methylation patterns) simultaneously in a single cell and at multiple stages during cell differentiation, (Barski et al., 2007; Mikkelsen et al., 2007). The growing resources (tools and techniques) available for the study of epigenetics offer new and revolutionary ways to study how environmental factors influence phenotype, (Rakyan et al., 2004; Eckhardt et al., 2006).

2.1.4 Epigenetics and Cancer

The misregulation of epigenetic mechanisms plays an important role in phenotype transmission and development, and it consequently affects uncontrolled cell division, such as cancer, (Grnbek et al., 2007; Martin and Zhang, 2007). Connections between cancer and epigenetic mechanism misregulation are found in many genes in cancer cells, including in the silencing of the MGMT (O6-methylguanine-DNA methyltransferase) gene. MGMT has a primary role in DNA repair mechanisms in cell division. When MGMT is silenced by methylation, DNA repair cannot occur during cell division. Therefore, this gene plays a decisive role in the development of cancer, (Soejima et al., 2005). Essentially, cancer is caused by alterations of two different classes of genes:

- Tumour suppressor genes[8] which inhibit cell growth and have normally unmethylated CpG islands in their promoter regions.
- Proto−oncogenes[9] which promote cell growth and have normally methylated CpG dinucleotides in the genome.

The various changes, including environmental changes that lead to gene misregulation, are involved in full cancer progression and in the determination of the final cancerous phenotype and are decided by the combined status of tumour suppressor genes and oncogenes. These epigenetic changes are due to covalent modifications of amino acid[10] residues in the histones around which the DNA is wrapped, together with changes in the methylation patterns of cytosine residues (C) in the context of CpG dinucleotides in the DNA, (Ro and Rannala, 2001). Furthermore, the heritable changes that occur due to DNA methylation are the basic markers of epigenetics, and they consist of different stages, such as hypermethylation[11] and hypomethylation[12] in cancer. Mi-

[8] A tumour suppressor gene (MGMT, MLH1, P53 etc), or anti-oncogene, is a gene that protects a cell from one step on the path to cancer. When this gene is mutated to cause a loss or reduction in its function, the cell can progress to cancer.

[9] A proto-oncogene (RAS, WNT, MYC, ERK gene etc.) is a normal gene that can become an oncogene due to mutations and increased expression.

[10] Amino acids are the structural units that make up histone proteins.

[11] An increase in the epigenetic methylation of cytosine residues in CpG island.

[12] A decrease methylation of cytosine residues in genomic DNA.

croRNA[13] inactivation by DNA methylation has also been studied in human cancer cells, (Lujambio et al., 2007).

2.1.5 Epigenetic Challenges in Cancer

Generally, mutations and chromosome abnormalities inhibit the function of tumour suppressor genes and initiate cancer. It is now clear that the inactivation of tumour suppressor genes due to abnormal epigenetic patterns, (Esteller, 2007; Portela and Esteller, 2010), can cause cancer. All of these results point to two specific challenges for the epigenetics of cancer:

1. To observe common patterns and functional relationships in the epigenetics of cancer cells in the genome.

2. To develop bioinformatic tools for improved cancer diagnosis, stage detection and therapy , (Esteller, 2007).

Furthermore, in terms of cancer prediction, epigenomics offers new opportunities for improving diagnostic technologies and therapeutic options.

2.2 Review of DNA Methylation

The DNA methylation mechanism controls gene activity. Faithful transfer of DNA methylation patterns between parents and daughter cells is essential for human development and for maintaining health, (Boyer et al., 2005). Dynamic alterations in DNA methylation patterns are thought to change gene expression, and initiate the

[13] MicroRNAs are short RNA molecules (22 nucleotides) found in all eukaryotic cells and it regulates post-transcriptional process by binding a complementary sequences on target mRNA transcripts and resulting gene silencing.

development of diseases such as cancer, (Stewart et al., 2005). However, methylation patterns pose difficulties in understanding cancer mechanisms. Predicting the methylation patterns in cancer is thus clearly a non-trivial step because cancer is a complex disease that is highly heterogeneous among patients. For the successful treatment and eradication of this disease, experts suggest that an individualised therapeutic strategy is needed that can identify both genetic and epigenetic defects, (Marte et al., 2008).

2.2.1 DNA Methylation in Normal Cells

DNA methylation occurs at the Cytosine bases of DNA after which the Cytosine dinucleotide is converted to 5-methylcytosine by the activity of DNA methyltransferase (DNMTs) enzymes, (Tajima and Suetake, 1998). During methylation, a methyl group is attached to the DNA. For example, in the Cytosine pyrimidine ring, the methyl group is attached to the number 5' carbon, which has the specific effect of reducing the expression of all genes in the sequence or fragment, as illustrated in Figure(2.2). This modification is transient and does not alter the original DNA sequence through inheritance and subsequent removal, (Clark et al., 1995). Such modifications characterise an epigenetic mechanism and are part of the epigenetic code (the DNA methylation pattern).

2.2.1.1 DNA Methylation and Gene Regulation

In recent decades, it has been found that DNA methylation plays a crucial role in gene repression by blocking promoters. The exact role of DNA methylation in gene expression is still unknown, but it has been hypothesised that DNA methylation impacts cell differentiation and embryonic development, (Suzuki and Bird, 2008). The main properties in DNA methylation include:

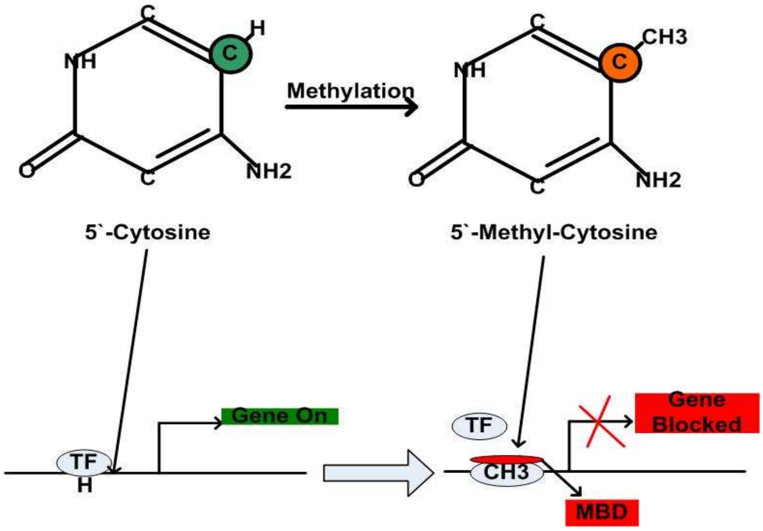


Figure 2.2: Methylation at Cytosine ring. Here C, carbon; H, hydrogen; NH, NH2, Amino group; CH3, methyl group; TF, transcription factor; O, oxygen, and 5' is the backbone direction (sugar base) for DNA, MBD is the DNA binding protein. The net effect of the methylation process is to turn off the formerly activated gene by attachment of methyl group and gene can not be expressed.

- Specific effects on reducing gene expression. In general, about 60-90% of all CpG sites are methylated, (Tucker, 2001). Unmethylated CpGs, which are grouped and are known as “CpG islands”, are present in the 5' regulatory regions of many genes.

- In cancer, CpG islands in a gene's promoter acquire abnormal hypermethylation (as presented in our model), resulting in heritable gene silencing.

- DNA methylation may impact the transcription of a gene in two ways: (i) through direct inhibition by transcription factors, and (ii), through indirect binding of MBD proteins; further details are provided in Section (2.2.2).

The most widely accepted hypothesis, however, is that DNA methylation facilitates the control of gene expression during development and cell differentiation, (Razin and Cedar, 1991). As described in recent research, (Herman and Baylin, 2003; Miller and Sweatt, 2007) it has been described that the long term memory storage in humans may be regulated by DNA methylation. However, in normal tissue, the methylation of particular subgroups of CpG island promoters can be detected. Promoter CpG islands undergo abnormal hypermethylation in diseases such as cancer, which may result in heritable transcriptional silencing, (Bird, 2002; Xue et al., 2009). Impacts on gene transcription due to DNA methylation may occur in two different ways:

1. The binding of transcriptional proteins to the gene may be physically impeded by DNA methylation itself, (Ordway et al., 2004).

2. More importantly, methyl-CpG-binding domain proteins (MBDs) may bind to methylated DNA, (Ng et al., 2000).

After this binding, MBDs recruit more proteins to the locus such as histone deacetylases and other chromatin remodelling proteins. These additional proteins may modify the histone, thereby forming “silent chromatin”, which is compact and inactive, (Ng et al., 2000; Hendrich and Bird, 1998). The link between DNA methylation and chromatin is inherently important in cancer; for example, the loss of methyl-CpG-binding protein 2 (MeCP2) has been implicated in Rett syndrome, (Jones et al., 1998). Similarly, the transcriptional silencing of hypermethylated genes in cancer is mediated by methyl-CpG binding domain protein2 (MDB2), (Singal and Ginder, 1999). DNA methylation modification patterns are established and maintained by DNA methyltransferases (DNMTs). There are a number of processes associated with DNA methylation, including imprinting and X-chromosome inactivation, (Yasukochia et al., 2010). Thus, DNA methylation is essential for the normal development of the cell. The following sections discuss the role of DNMTs in DNA methylation.

2.2.1.2 DNA Methyltransferases (DNMTs) in DNA Methylation Mechanisms

DNA methyltransferases are enzymes that catalyse DNA methylation, i.e., *de−novo* methylation and the maintenance of methylation, as shown in Figure(2.3). Their family encompasses the enzymes DNMT1, DNMT2, DNMT3A and DNMT3B. These enzymes are divided into two categories, (i) maintenance (DNMT1 and DNMT2) and (ii) de-novo (DNMT3A and DNMT3B) methyltransferases, (Doerfler et al., 2006). After replication, DNMT1 binds methyl groups to hemimethylated[14] DNA, as illustrated in Figure (2.3). DNMT3A and DNMT3B attach methyl groups to CpG dinucleotides of unmethylated DNA, as also shown in Figure (2.3). The formation of established methylation patterns in promoters and in the first exons[15] of human genomic DNA are executed by DNMT1, DNMT3A and DNMT3B. The patterns of DNA methylation are highly conserved in somatic cells[16] and are maintained during cell division by the methylation maintenance enzyme DNMT1, (Das and Singal, 2004; Momparler, 2003). These patterns of methylation are disrupted in major human diseases, such as imprinting disorders and cancer, (Jones and Liang, 2009), hence the importance of understanding how these patterns are established and maintained. It is generally believed that embryonic development and faithfully inherited methylation patterns are

[14] Double strand DNA where only one of the two strands is methylated. Important for regulating and protecting DNA.

[15] Active part of the DNA for transcription.

[16] Somatic cell (diploid) is any cells forming the body of an organism.

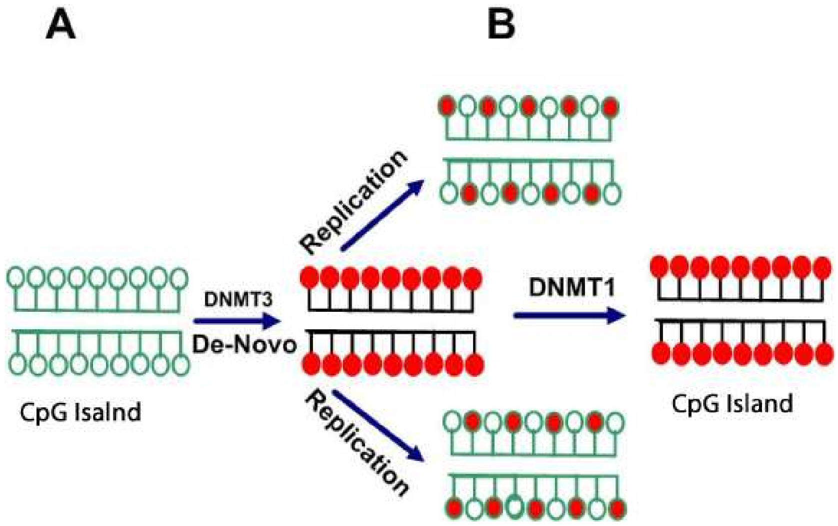


Figure 2.3: *De−novo* and maintenance methylation activity during replication. (A) Here, DNMT3 adds methyl group on unmethylated CpG sites, (B) After replication DNMT1 adds the methyl group to the hemimethylated CpG sites. Blocked circle and open circles methylated and unmethylated dinucleotides respectively.

established in somatic cells by a maintenance mechanism. As a component of the DNA replication complex, DNMT1 maintains DNA methylation by adding a methyl group to the 5'-position of the cytosine ring within the CpG dinucleotides of newly synthesised DNA strands. In genomic DNA, DNMT3A and DNMT3B establish new methylation patterns, (Brown and Strathdee, 2002; Das and Singal, 2004).

2.2.2 DNA Methylation and Cancer

Alterations in DNA Methylation, observed in cancer initiation during the past few decades, (Jiang et al., 2004; Kim et al., 2010), have attributed the role of methylation marker patterns to three things:

- Mutations at CpG dinucleotides involved in DNA hypomethylation and in the generation of oncogenic point mutations.

- The epigenetic silencing of tumour suppressor genes by DNA hypermethylation.

- The utilisation of experimental data in the study of DNA methylation patterns.

Aberrant patterns of DNA methylation may cause "incorrect" gene expression of certain genes, and in cancer, aberrant methylation, as well as both hypomethylation and hypermethylation, have been observed. DNA methylation is also influenced by histone modifications, (Singal and Ginder, 1999). The following sections briefly discuss the roles of DNA hypermethylation and hypomethylation in cancer development.

2.2.2.1 DNA Hypermethylation

DNA hypermethylation is the regional methylation of gene promoters in CpG islands that are normally unmethylated. This regional methylation of CpG islands is found in tumour suppressor genes in cancer. Some studies have found that regional hypermethylation in the promoter region of the RB[17] gene can reduce its expression by up to 8%, (Ohtani-Fujita et al., 1993). Consequently, cancer occurs due to aberrant methylation of the promoter regions of tumour suppressor genes (TSGs) that is linked to a loss of gene activity. These gene changes constitute a heritable state, not mediated by altered DNA sequence, that appears to be tightly linked to the formation of transcriptionally active chromatin[18]. All three methyltransferase enzymes (DNMT1, DNMT3A and DNMT3B) are overexpressed in human tumours, although only at moderate levels, (Rodenhiser and Mann, 2006). DNA methylation was the first epigenetic alteration to be observed in cancer cells, (Jiang et al., 2004). Hypermethylation of CpG islands in a tumour suppressor gene switches the gene off. Abnormal methylation in the cell

[17] RB stand for 'Retinoblastoma' and is involved in cell division maintenance.

[18] Active for gene expression.

can lead to the hypermethylated state and is therefore most likely involved in carcinogenesis, (Houshdaran et al., 2010). Two types of mechanisms are responsible for transcriptional repression via DNA methylation.

1. Methylation directly inhibits the binding of transcription factors at CpG dinucleotides present within the TF binding site (transcription factors), and these are sensitive to methylation, (Christensen et al., 2009). Figure (2.4) shows the direct impact of methylation by DNMT on a gene promoter in both a normal cell and a tumour cell.

2. Proteins specific for m5CpG dinucleotides bind to methylated DNA, which recruits m5CpG-binding (MeCP2) and MBD proteins, (Fatemi and Wade, 2006). Figure (2.5) illustrates transcriptional repression by MeCP-2 proteins.

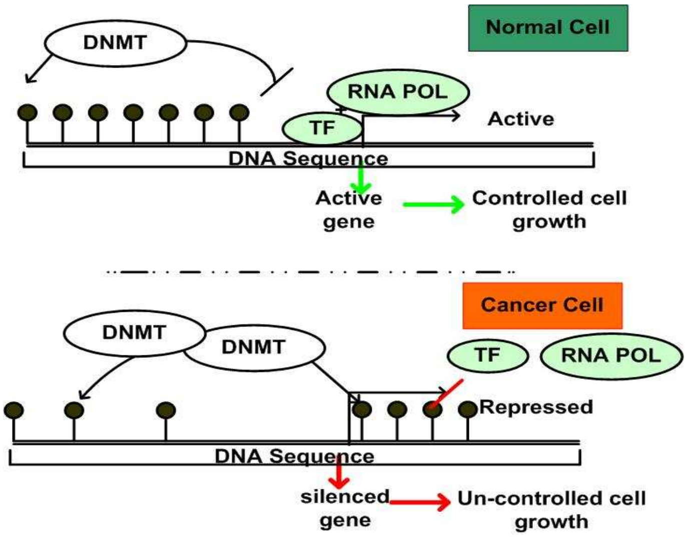


Figure 2.4: Direct Effects of DNA Methylation. Filled circle indicates methylated DNA.

MeCP1 and MeCP2 create steric obstacles by binding specifically to methylated DNA throughout the whole genome to disable the binding of TFs to promoter sequences.

The transcription of specific genes is repressed by MeCP1, and single methylated CpGs in DNA strands are bound by MeCP2, (Hendrich and Bird, 1998). The proteins in the MBD family are MBD1, MBD2, MBD3 and MBD4, and they are characterised by Kaiso complexes, which bind to methylated DNA, (Baylin and Herman, 2000; Das and Singal, 2004). The symmetrically methylated CpG dinucleotides are also bound by one of the MBD proteins, which inhibits gene expression by blocking TF interactions with the promoter, (Fujita et al., 2000). The MBD2 protein actively demethylates DNA *in vivo* and may bind to methylated DNA (Ng et al., 2000; Szyf et al., 2004). MBD3 is a component of the chromatin remodelling protein complex and, in association with MBD2, targets methylated DNA, (Ballestar et al., 2003). MBD4 is a Thymine and Uracil glycosylase involved in the repair of DNA mismatches formed during Cytosine and methylcytosine deamination, (Fujita et al., 2003; Hendrich et al., 1999).

2.2.2.2 DNA Hypomethylation

Hypomethylation is a process that reduces the methylation level of proto-oncogenes in the cell. Some studies have found that, in a large variety of hypomethylated tumour samples, the changes in the cell are not only correlated with altered methylation patterns but also with increased tumour progression, (Feinberg and Tycko, 2004). Consequently, particular DNA hypomethylations are linked to cancer initiation. Tumour cells also exhibit hypomethylation of CpG dinucleotides in various DNA regions that are responsible for increased gene expression, invasion and metastasis of cancer cells. Tumour cells display global DNA hypomethylation. Global hypomethylation leads to genomic instability and inappropriate activation of proto−oncogenes and transposable elements, (Feinberg, 2007). It appears that genomic DNA methylation levels, which

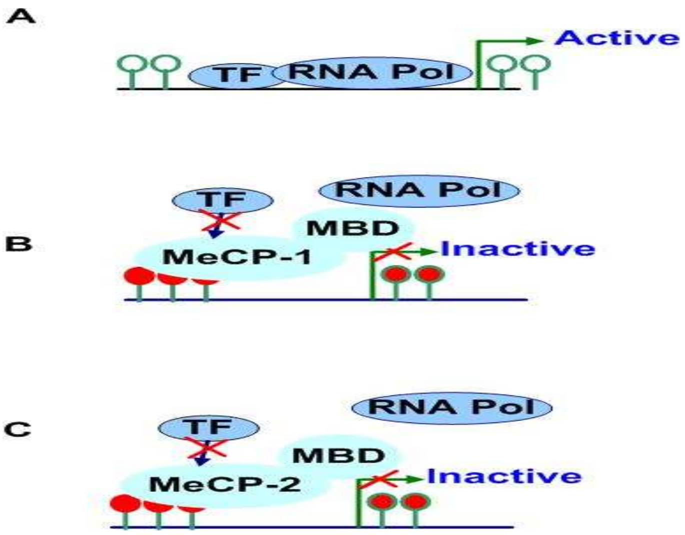


Figure 2.5: Transcriptional repression by MeCpG binding Proteins. (A) Active state of DNA for the transcription; (B) Inactive state of the DNA by binding of the MeCP-1 and MBD proteins in to methylated DNA; (C) Inactive state of DNA by binding of MeCP-2 and MBD proteins. Filled circles indicate methylated and unfilled circles unmethylated CpG dinucleotides.

are maintained by DNMT enzymes, are directly balanced within cells; the overexpression of DNMTs is linked to cancer in humans, (Feinberg and Tycko, 2004; Rodenhiser and Mann, 2006). In numerous cells, global hypomethylation has been observed that is responsible for the overexpression of proto-oncogenes, growth factors and genes involved in cancer cell proliferation, invasion and metastasis via their protein products, (Szyf et al., 2004). In some medical examples related to malignant cell metastasis and invasiveness, the most common protease is the PLAU (urokinase type plasminogen) enzyme, which is expressed in many cancers such as breast and prostate cancer, (Pakneshan et al., 2004, 2005). The precise roles of global DNA methylation (hypomethylation of CpG dinucleotides) in malignant cells remain unclear. Research indicates that one reason may be the complete or partial deficiency of numerous enzymes involved in methyl transport at the cellular level, (Steele et al., 2005).

2.2.3 DNA Methylation and Cancer Therapy

Currently, scientists are studying the connections between methylation abnormalities, gene expression and silencing, and various diseases such as cancer, (Robertson and Jones, 2000). The result of these studies will be invaluable for treating these diseases, as well as for understanding and preventing complications of treatment of cancer. For example, in cancer treatment, two types of DNMT inhibitors are used:

- The nucleoside analogue 5-Aza-2 deoxycytidine (5-Aza-CdR), which is incorporated into DNA and inactivates DNMTs, (Yoo and Jones, 2004).
- Antisense oligonucleotides that induce the degradation of DNMT1 mRNA and inhibit enzyme biosynthesis, (Yoo and Jones, 2004).

The reversibility of epigenetic changes is the key target in cancer therapy, (Cooney et al., 2002; Klisovic et al., 2008). DNMT inhibitors are drugs that change gene activity, and they include 5-Azacytidine (5-Aza-CR), 5-AzaCdR and 1-D ribofuranasyl-2(1H)-pyrimidinone (Zebularine), (Laird, 2005; Yoo and Jones, 2004). Furthermore, these enzymes induce the demethylation of the promoters of TSGs such as CDKN2A[19], RB1[20], MLH1[21] in cancer cells, (Yoo and Jones, 2004; Robertson and Jones, 2000). To correct silent gene expression patterns and revert cells back to more normal functions, DNMT inhibitors, such as 5-Aza-CR and 5-Aza-CdR, are used. It seems likely that

[19] Cyclin-dependent kinase inhibitor gene is multiple tumour suppressor, other name of this gene is p14; p19; p16,p16INK4A, p14ARF.

[20] RB1',retinoblastoma (RB) gene, involved to maintain cell divisions.

[21] It is DNA mismatch repair gene and protein is Mlh1, COCA2; FCC2; HNPCC; HNPCC2; MGC5172; hMLH1.

cancer diagnosis based on epigenetics, together with cancer therapies, will create possibilities in the future treatment.

2.3 Epigenomic Analysis

DNA methylation and histone modifications are studied in the epigenome. Here, we describe epigenomic analytical techniques that are essential to predicting epigenetic states. Epigenetic research utilises powerful experimental techniques to analyse DNA methylation. Some of these include: bisulphate modification associated with polymerase-chain-reaction and chromatin immunoprecipitation (ChIP), (Herman et al., 1996). Epigenomic techniques generally describe the epigenomes of human diseases by calculating the frequencies of different stages of DNA methylation, (Ballestar et al., 2003; Esteller, 2007).

2.3.1 Epigenomic Mapping Techniques

Three techniques that convert raw epigenetic data into a readable form are elucidated in Figure (2.6). These techniques are used to generate large amounts of epigenetic data, and they provide an efficient way to analyse epigenetic information in the genome. All three techniques are designed for genome-wide mapping and for acquiring epigenetic data in three stages:

1. Epigenetic information is converted to genetic information.

2. Then, genetic engineering techniques, such as tiling microarrays, are used to obtain DNA sequences.

3. Finally, computational algorithms are used to generalise the epigenetic informa-

tion from the DNA sequences.

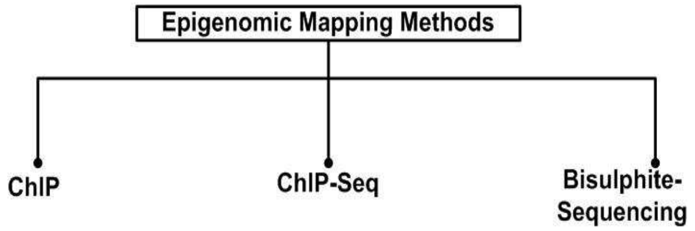


Figure 2.6: Methods of epigenomic mapping

The main goal of the techniques defined above is to map epigenetic information on a genome-wide scale. In some cases, researchers want to know whether the frequency of methylation in regions of the genome differs between different cell types or patient populations, (Baker, 2010). Brief details on the three epigenomic mapping techniques illustrated in Figure (2.6), are provided in the following two subsections.

2.3.1.1 ChIP and ChIP− seq Techniques

The objective of the ChIP (or ChIP−on−chip) technique is to use chromatin immunoprecipitation (ChIP) to find differences between normal and modified DNA. The main goal of ChIP is to locate protein binding sites that identify functional elements in the genome. Chromatin is extracted from DNA and protein and is fragmented into 500 base pair lengths. ChIP-seq is a variant of ChIP-on-chip that uses for the whole genome. The advantages of ChIP-seq technique is:

- Less data normalisation.
- That sequencing results are given in complete counts compared to hybridisation scores in ChIP−on−ChIP.
- That sequencing synthesis methods are highly cost-efficient.

2.3.1.2 Bisulphite sequencing

Another technique for the analysis of methylation is bisulphite sequencing. In bisulphite sequencing, bisulphite[22] converts methylated Cytosines into methylation-dependent SNPs[23]. The advantage of bisulphite sequencing is that DNA methylation patterns are detected in a single cell.

2.3.2 Epigenome Analysis Projects:

A number of projects have been established to try to manage large-scale epigenetic data and to develop tools for analysing the data. Six of the well-known epigenome mapping projects are illustrated in Figure (2.7), and a brief description of their objectives is provided below.

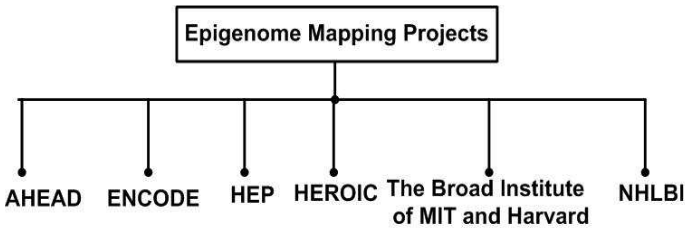

Epigenome Mapping Projects
AHEAD
ENCODE
HEP
HEROIC
The Broad Institute of MIT and Harvard
NHLBI

Figure 2.7: Schematic Diagram of Epigenome Mapping Projects

- The plan of the ‘AHEAD’ (Alliance for Human Epigenomics and Disease) project is to ‘genomicise’ epigenomics research and pave the way for breakthroughs in the prevention, diagnosis and treatment of human disease, (Jones, 2008). It coordinates with the human epigenome mapping project, (American Association for

[22] Bisulfite is an ion; (hydrogen sulfite of HSO3-). Salts containing the HSO3- ion are termed bisulfites also known as sulfite lyes.

[23] A single-nucleotide polymorphism (SNP) is a DNA sequence variation occurring when a single nucleotide A, T, C, or G in the genome. For example, two sequenced DNA fragments from different individuals, AAGCCTA to AAGCTTA, contain a difference in a single nucleotide. In this case we say that there are two alleles: C and T.

Cancer Research Human Epigenome Task Force and European Union, Network of Excellence, Scientific Advisory Board, 2008) and focuses on determining the epigenetic status of the cell.

- The 'ENCODE' (ENCyclopedia Of DNA Elements) project is an example of a close collaboration between experimental and computational biologists. The aim of this project is to functionally map the genome and identify key epigenetic questions, (Consortium, 2004).

- The 'HEP' (Human Epigenome project) focuses on high resolution epigenome analysis of multiple individuals. Current contributors to this project are 'The Wellcome Trust Sanger Institute' (UK), 'Epigenomics AG' (U.S.A., Bosten and Berlin) and the 'Centre National de Genotypage', (France), (Bradbury, 2003).

- 'HEROIC' (High-throughput Epigenetic Regulatory Organisation In Chromatin) is an integrated project of the 'European Commission. The main goals of this project are to understand the regulation of the whole human genome through analysis of ChIP−on ChIP− data and chromosome interactions. Time period of this project is 2005-11-01 - 2010-02-28.

- The Broad Institute of MIT and Harvard uses ChIP−seq data and maps genome-wide maps of mouse stem cells, to analyse the massive amounts of genome-related data that are being generated by scientists at the Broad Institute and around the world, (Mikkelsen et al., 2007). Currently, a wide range of critical biomedical projects[24] are in progress at this institute, such as genome mapping projects.

[24] Available at http://www.broadinstitute.org/science/projects/projects.

- The National Heart, Blood and Lung Institute of the National Institutes of Health (NHLBI) (USA) also uses ChIP−seq data to analyse the genome-wide chromatin status of human T-cells. The institute has three strategic branches structured to link, as indicated, successive stages of scientific discovery, (Barski et al., 2007). These are denoted as the following:
  - Form to function,
  - Function to causes
  - Cause to cures.
- Another project in this series is the "AACR Human Epigenome Task", (Jones and Martienssen, 2005). The main goals of this project are to finalise the number of epigenomes, to define histone modifications on the basis of epigenetic markers and to characterise pathological states such as cancer.

2.3.3 Epigenome Databases

The development of multiple epigenomic analytical techniques, including high-throughput assays, has resulted in datasets of increasing complexity and diversity. Here, we talk about the most important DNA methylation databases. DNA methylation data are useful for studying the covalent modification of the genome. A number of online epigenetics databases are available, and they contain information on DNA methylation frequencies for different tissues, cells and phenotypes, as illustrated in Table (2.1). All of these databases are available online and URL addresses are given in appendix (A). Important sources for methylation databases are MethDB, PubMeth, MeInfoText, methprimerDB and PathEpigen. These contain information on DNA methylation pat-

Database	**Number and Records**	**Country_Year**
MethDB	19,905 DNA methylation content data and 5,382 methylation patterns for 48 species, 1,511 individuals, 198 tissues and cell lines and 79 phenotypes	Germany_2006
MeInfoText	Gene methylation information across 205 human cancer types.	Taiwan_2006
PubMeth	5,000 records on methylated genes in various cancer types.	Belgium_2007
REBASE	22,000 DNA methyltransferases genes derived from GenBank.	New England_2005
MethPrimerDB	259 primer sets from human, mouse and rat for DNA methylation analysis.	Belgium_2003
The Histone Database	254 sequences from histone H1, 383 from histone H2, 311 from histone H2B, 1043 from histone H3 and 198 from histone H4, altogether representing at least 857 species.	USA_2007
ChromDB	9,341 chromatin-associated proteins, including RNAi-associated proteins, for a broad range of organisms.	AZ USA_2005
CREMOFAC	1725 redundant and 720 non-redundant chromatin-remodeling factor sequences in eukaryotes.	India_2009
KFEL*	DNA methylation data of human chromosomes 21, 22, male germ cells and DNA methylation profiles in monozygotic and dizygotic twins.	Canada_2003
MethyLogiX	DNA methylation data of human chromosomes 21 and 22, male germ cells.	Germany_2006
PathEpigen	5,500 genetic and epigenetic events.	Ireland_2009

Table 2.1: DNA methylation databases. *KFEL (The Krembil Family Epigenetics Laboratory) is a PAHO/WHO Collaborating Centre and fully affiliated with the University of Toronto.

terns across different species, tissues, cell types and phenotypes. The Pubmeth and PathEpigen databases contain information on the gene methylation profiles of specific cancer types. Our model frequencies are based on PubMeth (Belgium) and PathEpigen (Ireland). More details about these databases are presented in Chapter (3). Recently (Bock, 2008), the BiQ Analyzer software was developed to simplify methylation data analysis and to help visualise the resulting output. BiQ Analyzer facilitates the visualisation and quality control of DNA methylation data from bisulphite sequencingand presents a bioinformatic analysis of the National Methylome Project for the Chromosome 21 dataset, (Bock, 2008). Finally, large-scale data present clear epigenetic modifications in the genome, but it remains a significant challenge in computational epigenetics to find genomic and epigenomic activities, which are two prominent independent systems for identifying patterns related to inheritance.

2.4 Computational Epigenetics

Current research in computational epigenetics aims to analyse the available data related to cancer and to generate models to facilitate better treatment for this complex disease.

2.4.1 Evaluation

The role of computational epigenetics lies in the development and application of bioinformatic methods for solving epigenetic questions as well as in computational modelling and data analysis, (Bock, 2008; Perrin et al., 2008; Raghavan et al., 2010). The key things to epigenetics study are include individual or aggregated histone changes or the epigenomic (overall) picture, or how computation is going to help in this. One possible way is that it can handle the different dynamics of histone modification and DNA methylation changes they are very different wit the latter being more stable and the former showing cumulative effects over much shorter time periods. While complex to handle experimentally, both can be incorporated in a computational approach. In general, DNA sequences are invariant across tissues, but epigenetic marks show tissue-specific variations, (Jones and Liang, 2009). New paradigms, (such as epigenome mapping tools), have been designed to study complex diseases such as cancer, and computational methods have played an important role in interpreting epigenomic information. Examples (Bock et al., 2006) include comparing methylation patterns between cancer and normal tissues, (brain and blood), and predicting the preferential locations of epigenetic modifications.

2.4.2 Scope

Computational epigenetics combines elements of emerging fields such as, population genetics, evolutionary genetics, medical epigenetics, theoretical modelling, pattern recognition and computational biology. Modelling provides a new dimension in epigenetic research. For example, Genereux et al. (2005) and Siegmund et al. (2008) have developed a model for methylation mechanisms from gene expression profiles of a large heterogeneous cell population. Furthermore, Perrin and Ruskin (2010) analysed aberrant methylation pattens of the H. Pylori[25] infection for gastric cancer and Raghavan et al. (2010) used a stochastic approach to analysis of histone modification during transcription. The direct benefits of producing an accurate model for DNA methylation patterns and other features include a reduction in the number of experiments and the minimisation of the costs associated with them as well as a the establishment of a gen-

[25] Helicobacter pylori (H. pylori) is the bacteria and spread stomach inflammation such as ulcers.

eral framework for exploring a number of possible changes simultaneously. This can be achieved through computational investigations. Hence, based on simple initial parameter choices, dependent parameters of methylation mechanisms can be derived and their evolution montored, by simulating e.g. the dynamics of the methylation activity of methyltransferase expression in different cancer cells, (Fraga and Esteller, 2007; Jiang et al., 2004). Initial and derived values form the inputs to sensitivity analyses, which enable assessment of the robustness of different outcomes to initial parameter values. These parameters may be specific to a particular cancer type and vary for other cancer types, (Boyer et al., 2005). Thus, theoretical modelling can provide an in depth and quantitative understanding of epigenetic mechanisms. In medical epigenetics, such models can explore the role of epigenetic mechanisms in complex diseases, such as cancer, mental disorders and autoimmune diseases, (Brown and Strathdee, 2002). In addition, computational epigenetic modelling, in general, can lead to a better scientific understanding of the behaviour of many biological systems and can guide further experimentation. In recent decades, studies have shown e.g. that cooperative protein activity is required to maintain epigenetic states at the molecular level, (Cowley and Atchley, 1992; Sagal et al., 2001) and (Dodd et al., 2007; Reik, 2007). A key requirement is to understand methylation patterns in cancer progression and their affect on cell status, (Jones and Martienssen, 2005) and analysis of such patterns is an ongoing challenge, (Raghavan and Ruskin (2011)- private communication).

2.4.3 Challenges

Although dramatic progress has been made over the last few years, (Delcuve et al., 2009), DNA methylation for cancer detection remains, by and large, an 'academic exercise'. For example, there is an urgent need for an informative set of targets and for knowledge of how DNA methylation is altered to specify cancerous states. A number of genome-wide approaches have been developed, (Beck and Rakyan, 2008; Illingworth et al., 2008) to identify DNA methylation biomarkers for various types of cancer. However, the clinical validation of each candidate target with respect to both specificity and sensitivity is a slow and tedious process not to mention the difficulty of collecting a good cohort of clinical samples. Furthermore, effective recovery of DNA from bodily fluids (plasma DNA in particular) is essential for the successful use of DNA methylation profiling for early detection and disease outcome prediction. Finally, assay development for a faithful set of target genes tailored to clinical oncology demands extra resources and skills, which are simply lacking in the majority of research labs in both academic and clinical institutions. In this regard, the main challenges are:

1. Understanding the molecular mechanisms by which DNA methylation patterns are regulated because the DNMT1[26] and DNMT3[27] families of methyltransferases do not appear to have any sequence specificity beyond CpG dinucleotides, (Dodge et al., 2002).

2. Interpreting chromatin-based mechanisms: proposed to explain how DNA methyltransferases find their target sequences in the genome, (Bird, 2002).

Inevitably, a further major challenge for computational epigenetics is to generate appropriate data and to manage epigenomic databases, which can help build theoretical models. The main challenges here are:

[26] DNMT1 is the most abundant DNA methyltransferase in mammalian cells, and considered to be the key maintenance methyltransferase in mammals.

[27] DNMT3 is a family of DNA methyltransferases that could methylate hemimethylated and unmethylated CpG at the same rate.

1. The generation of experimental data, together with epigenomic data analysis , (Bock, 2008).

2. Epigenome mapping in different genomic and epigenomic regions for different individuals, (Jones, 2008).

3. Understanding the *dynamics* of epigenetic mechanisms to develop theoretical and empirical models, (Reik, 2007).

4. Reducing the cost of epigenomic mapping tools (software as well as wet-lab costs), so that different species can be analysed.

5. Data visualisation; for an example of web browsers available for the statistical analysis of epigenomic data, see Galaxy (http://bitbucket.org/galaxy).

2.4.4 Model Considerations

For modelling, processes are classified with respect to a set of criteria. Some considerations include:

2.4.4.1 Top-Down vs. Bottom-Up

For any system specified in abstract terms, there are two types of modelling approaches: (i) top-down and (ii) bottom-up. Top-down modelling is formal equation-based modelling that formulates and specifies an overview of the system. By contrast, bottom-up modelling is based on the individual elements of the system, which are first specified in great detail, in terms of their onherent characteristics and the ways in which they can interact with other elements. System elements are then linked together to form larger subsystems, which then in turn are linked, sometimes at many levels, until a complete top-level system is formed when, in theory, the system of equations should effectively aggregate or complete the local information to give a larger picture, (Szallasi et al., 2010).

2.4.4.2 *Ab-Initio* vs Template or Empirical Modelling

Ab-initio is also known as fundamental (or first principles) modelling, and it is based on what is known and what can be assumed about the components of a given system. People use this approach to make predictions about biological features. Example include (i) Molecular Dynamics, which tries to specify each component and the forces and interactions, which affect it, in order to determine electronic structure, (Tuckerman et al., 1996) (ii) Monte Carlo, used for many-body problems, which are too complex to allow for individual elements to be modelled, so a sampling technique is used to solve either the direct or indirect (substitute) problem from what is known /can be assumed about the 'particles' statistical behaviour, (Riley et al., 1995) (iii) Markov model variants, which exploit the Markov property to determine departure from random occupation of possible 'states' for system elements and many others. The Hidden Markov model, for example, which incorporates features of both *ab-initio* and template modelling, is widely-used to explore patterns in biological sequences, (Yoon, 2009; Krogh et al., 1994). The template or empirical model is typically developed from actual data, as the name suggests, and centres on the concepts of observation, dependency and agency, (Demin and Goryanin, 2008), with the objective in broad terms of finding close or distant matches.

2.4.4.3 Deterministic vs. Stochastic Modelling

A deterministic model is one in which every set of variable states is uniquely determined by model parameters and by sets of the previous of these variables, e.g., (A + B = C). Consequently, deterministic models perform the same way for a given set of initial conditions. Conversely, in a stochastic model, randomness is present, and variable states are not described by unique values but rather by probability distributions, (Wilkinson, 2006).

2.4.4.4 Parameterisation/Estimation

Parameterisation is the process of deciding and defining and/or deriving the parameters (or variables) necessary for the relevant specification of a model. Estimation is a process typically used to find the upper or lower bounds of a quantity that can not readily be computed precisely. Mostly, formal parameterisation is based on, e.g., known mechanisms or system values (in metamathematical terms), as well as on assumptions that can be made about the model's elements. With parameter values can be either assigned or estimated (based on actual data). The level at which these can be estimated from the data, i.e., whether at the system or unit level, can influence whether a top-down or bottom-up approach is chosen. This choice also influences the questions that can be addressed and how the information can be used for the system, (Wilkinson, 2006). The model focus, discussed in this thesis, together with the corresponding parameterisation work on PathEpigen, is deterministic and bottom-up: we used a Deterministic Finite Automata (DFA) machine for the simulation of methylation dynamics and this is described in Chapter (3).

2.5 Summary

The precise mechanisms underlying epigenetic patterns, such as DNA methylation, are still unknown, but information is accumulating. One important question in molecular genetics is: what patterns of DNA methylation are observed during normal and cancer cell divisions? Recent studies have shown that remarkable changes in DNA methylation in tumour suppresser genes promote the initiation of cancer development, (Teng et al., 2011). DNA methylation patterns in cell populations are diverse in cancers, but the last few decades have seen broad acceptance of the role of methylation. Some medical experimental examples have shown individually that a number of features, such as mutations at CpG dinucleotides, are involved in DNA hypomethylation and the generation of oncogenic point mutations, and these contribute to methylation change. A significant ongoing challenge is, clearly, to identify methylation patterns for all cancer stages. There is an urgent need for an informative set of targets and knowledge of how DNA methylation is altered to specify cancerous states, (Jones and Liang, 2009). Given that methylation patterns are highly heterogeneous among patients, computational modelling (which can exploit available experimental data and build on this through sensitivity analyses) is a good method for analysing and predicting the methylation patterns in different cancer cell types. Computational epigenetics thus uses bioinformatics knowledge and modelling methods and assumptions to complement experimental research. In recent research, on epigenomic datasets, computational methods are playing an increasingly important role in all areas of epigenetic research. Nevertheless, a detailed molecular description of epigenetic modifications and gene regulation does not yet exist, (Artyomov et al., 2010). We have outlined several possible modelling approaches, which are being employed to address this lack. Further, it is known that DNA methylation patterns in cancer tissues generally display more variation than those of normal tissues, so that computational modelling also have a role in the design and implementation new databases. As these resources develop, they are expected to greatly facilitate template model determination as well as parameterisation of *ab-initio* models in computational studies of epigenetics. Different aspects of epigenetic modifications at the cellular level can be explored using one or more modelling approaches and should lead to increased understanding of the complexity of molecular mechanisms, such as DNA methylation. A deterministic model for DNA methylation dynamics is discussed in the next chapter.

Chapter 3

A Deterministic Finite Automata (DFA) Model of DNA Methylation Dynamics

In this chapter we present a simple, prototype deterministic model of methylation dynamics for the promoter region of the MGMT (Methyl Guanine Methyl Transferase) gene. We briefly describe the epigenetic modelling methods, which are strongly empirically-based. A subsequent section describes the parameterisation for MGMT and methylation, gathered from the reported literature. The dynamics of DNA methylation with respect to the methyltransferase concentration patterns, is then discussed and we present an initial DFA model of DNA methylation and evaluate it based on random strings. These random string represents changing value of the DNA methyltransferase concentration.

3.1 Introduction

3.1.1 The Deterministic Finite Automata (DFA)

A Deterministic Finite Automata (DFA) is finite state machine accepting strings of symbols over some alphabet Σ. For each state, there is a transition arrow leading out to a next state for each symbol in the alphabet. The language L is simply a subset of $\Sigma^\star$ i.e, a set of strings over Σ. Languages accepted by DFA are known as regular language. Formally, a DFA is a five tuple system (Q, Σ, δ, q_0, F), consisting of

1. a finite set of states (Q)
2. a finite set of input symbols called the alphabet (Σ)
3. a transition function (δ : Q $\times$ Σ $\rightarrow$ Q)
4. a start state ($q_0 \in$ Q)
5. a set of accept states (F$\subseteq$ Q)

The DFA is a simple 'machine' driver and has been used extensively in providing logic to vending machines for example. We describe a DFA by its transition function δ using a table of states and input symbols with an arrow for a start state and a star (asterisk) for an end state (for example, see Table [(3.7 and 3.8)]). In this chapter we model the DNA methylation dynamics using the DFA.

3.2 Epigenetic Micro-modelling

Epigenetic modelling is concerned with developing a framework for answering epigenetic questions. In particular, at the "micro" level, we seek to formulate expressions that represent the relationships among key entities and that describe how these work and are influenced by internal and external conditions. For example, a basic question is how different measurable factors affect the way in which epigenetic mechanisms are involved in cancer initiation. Modelling can help to explain how high-level changes occur and how these affect low-level changes, (Dodd et al., 2007). Thus, just some of the challenges include modelling the interactions between phenotypic and environmental factors and modelling the transfer of heritable traits in different cells by known epigenetic mechanisms. The advantage of a computational approach to epigenetics is that this may be used in conjunction with quantitative data to test possible solutions. Here, we present an overview of epigenetic models:

- A study done by Cowley and Atchley (1992) presents a multivariate quantitative genetic model of epigenetic and maternal effects[1]. The model explains how genetic correlations can arise through epigenetic effects.

- Haslberger et al. (2006) presented the interaction between genetic and epigenetic mechanisms in cancer development and observed the interaction between the genome and the environment. This biomedical concept includes environmental health aspects regarding epigenetic and genetic modifications.

- Das et al. (2006) wrote a program (called HDFINDER) that predicts the methylation landscape in human brain DNA for all 22 autosomes[2].

- A useful approach for DNA methylation analysis was provided by Houseman

[1] This is a specific effect not just an effect. In terms of heritability the phenotype of an organism is determined not only by the environment it experiences and its genotype, but also by the environment and phenotype of its mother.

[2] An autosome is a chromosome that is not a sex chromosome i.e. an equal number of copies of the chromosome in males and females.

et al. (2008). It is a "beta mixture model"[3] that uses a "recursive-partitioning algorithm"[4] for normal and ageing tissues.

- Artyomov et al. (2010) developed a computational model of epigenetic and genetic regulatory networks for pluripotent cells[5]. They determined how cellular identity is maintained and transformed during cell division. This study provides some experimental predictions from the model for understanding the maintenance and transformation of cellular identity.

- Raghavan et al. (2010) used a stochastic approach to DNA methylation and transcription with a contribution of histone modifications.

- Perrin and Ruskin (2010) presents a model on aberrant DNA methylation for the H. Pylori[6] infection in gastric crypt[7] and investigate the sensitivity of different genes and cell types for the gastric cancer[8].

- Recently, Ruskin and Perrin (2011) submitted a special report (private communication) on "Multi-scale modelling of epigenetic mechanisms" in Transactions in Biomedical Engineering (IEEE).

The following section provides an overview of models of epigenetic mechanisms, such as DNA methylation and chromatin remodelling.

[3] This approach to solve a variety of problems related to correlations of gene-expression levels.

[4] A algorithm for addressing the problem of decision tree construction.

[5] A stem cell that can give rise to more than one differentiated cell type.

[6] Helicobacter pylori (H. pylori) is the bacteria responsible for most ulcers and many cases of stomach inflammation (chronic gastritis).

[7] These are gland cells in the stomach.

[8] Gastric cancer (or stomach cancer), develops in any part of the stomach and may spread throughout the stomach and to other organs such as the esophagus, lungs, lymph nodes, and the liver.

3.2.1 Contextual Chromatin Modelling

Global studies of the epigenome have shown that certain histone modification variants and post-translational modifications[9] are correlated with DNA methylation. For example, trimethylation of H3 at lysine[10] 4 is associated with gene expression activity. Therefore, it is impacted by DNA methylation, (Rice and Allis, 2001). Mono-, di- and tri- methylation of this histone inhibits gene expression, and this activity of the gene impacts its nucleosomal conformation, (Takeshima et al., 2006; Okuwaki and Verreault, 2004). It has been suggested that specific modifications, such as DNA and histone methylation, change the nucleosomal conformation; therefore, gene expression can be affected during replication, (Workman and Kingston, 1998). Subsequently, as described above, statements regarding methylation maintenance activity may be more complex than existing models. The available database suggests that some modifications are necessary for the existing models of DNA methylation activity. The first epigenetic tool for quantifying chromatin states was provided by Bio-Rad Laboratories [http://www.biorad.com/], the "EpiQ chromatin analysis kit. This kit provides quantitative information about chromatin accessibility for gene expression.

3.2.2 DNA Methylation Modelling

As mentioned earlier in section (2.2.1.1), DNA methylation is an important epigenetic modification that impacts gene expression. The aim of DNA methylation studies is to understand how methylated and unmethylated cytosine residues are preserved through cell division in different biological processes. In this context, the fundamental questions

[9] Modification of the proteins after translation process.

[10] An amino acid which directly affected by histone modification, therefore, modified lysine impacts on gene expression.

of DNA methylation research are:

- How are methylation patterns maintained during replication at the cellular level, (Jones and Liang, 2009)?

- How are patterns quantified with respect to *de-novo* and maintenance methylation activities, (Jair et al., 2006) and (Xie et al., 2011)?

Consequently, it is important to understand the regulation of methylation reactions in regard to the enzymatic activity of DNMTs. Several studies have provided information on the maintenance methylation mechanism involving DNMT activity during replication, (Ordway and Curran, 2002), but they do not fit with the experimental data. The original hypothesis from experimental evidence has been supported as to the existence of both *de-novo* and maintenance DNMTs in the methylation process, (Herman et al., 1996; Costello and Plass, 2001). Unfortunately, investigating DNA methylation is challenging due to its transient nature. Computational methods for DNA methylation are used to identify specific sets of methylated genes, and computational modelling provides insights into the role of DNA methylation pattern variations among different tissues, (Yi and Goodisman, 2009). Computational methods such as hidden Markov model profiling and the Multiple Motif Scanning program can be used to analyse known methyltransferases and to predict new ones , (Petrossian and Clarke, 2009). The following section presents the pros and cons of established models of DNA methylation.

3.3 Established Models of DNA Methylation Dynamics

3.3.1 A Population-Epigenetic Model for DNA Methylation Patterns:

Genereux et al. (2005) used a population epigenetic approach to understand methylation dynamics during DNA replication for the FMR1 gene. This model calculates site-specific rates of *de-novo* and maintenance methylation in cell populations. The pros and cons of this model are as follows:

- The merits of this useful mathematical model are that it shows how methylated and unmethylated CpG sites in the FMR1 gene are transmitted and how methylation dynamics are maintained during DNA replication.
- The downside of this model is that it does not explain DNMT activity during DNA replication.

3.3.2 DNA methylation modelling for cancer cell populations:

Siegmund et al. (2008) used a Bayesian computation method and observed DNA methylation patterns for cancer under different evolutionary factors. The pros and cons of this model are as follows:

- The benefits of the cancer cell population model is that Siegmund et al. (2008) concludes that methylation patterns of human cancer are consistent with replication errors and the growth of certain simple cancers, and it provides a statistical analysis for the increased size of methylated tumours.

- The downside is that it does not explore the methylation dynamics in the cell population.

3.3.3 How DNA methylation patterns are maintained:

Jones and Liang (2009) present a model for maintaining DNA methylation patterns involving methyltransferases (DNMT1 and DNMT3A and DNMT3B). The model focuses on how DNMT1, DNMT3A and DNMT3B maintain methylation activity during DNA replication. The pros and cons of this model are as follows:

- Advantage: it addresses DNA methylation maintenance by the activity of DNMTs during cell division.

- Disadvantage: it needs to address the quantification of DNMT levels during cell division. It is important to check the error rate of methylation levels for gene expression notification.

3.4 A New Baseline DFA Model of Methylation Dynamics

As mentioned earlier, it is important to understand the quantitative influence of DNMTs on genes in different cancer cell types. In this context, we propose an early stage model to elucidate how methylation activity is regulated by DNMTs for different genes. The aim of our model is to derive DNA methylation patterns from the activities of methyltransferases. The DFA model accepts random strings of DNMTs and predicts the methylation patterns of genes. In the simplest terms, such a model can be used in conjunction with actual data to check the methylation status. In extended form, other pathways can be included to examine hyper-, hypo- and *de-novo* methylated forms. An advantage is that the "benchmarking" of the DFA model can be used to predict the cancer stage from quantitative DNMT data without the need for experimentation.

3.4.1 Comparison of the Established Model with the DFA Model

As described in sections (3.2.2 and 3.3), it is important to emphasise that DNMT1 and DNMT3 quantification reflect *de-novo* and maintenance methylation levels. In this context, our DFA model addresses the quantification of DNMT activity for different methylation levels of a particular gene. We utilise a methylation dynamics language that represents the up- and downregulation of DNMT expression. Goals of using this method include finding the quantitative values of DNMTs and understanding the methylation patterns of different genes. Subsequently, our DFA model also provides a potential approach for quantifying methylation activities with respect to *de-novo* and maintenance of methylation research, as mentioned in section (3.2.2). It is hardly the whole answer, but rather describes a way of going about it i.e. an approach. In two useful studies (Jones and Liang, 2009; Jair et al., 2006) on DNMT activity, a DFA model is used also to address the problem of DNMT quantification. Our DFA model is used to quantitate DNMTs at different methylation stages. It facilitates the analysis of methylation levels, as previously described, (Jair et al., 2006).

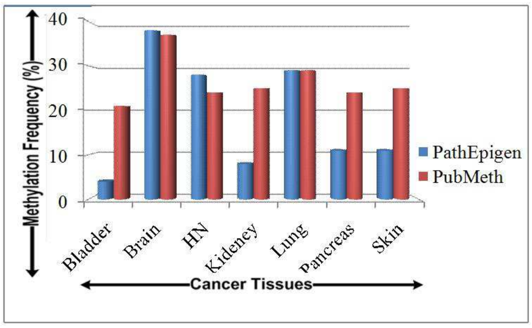


Figure 3.1: MGMT Methylated Frequency Comparison. HN indicates Head and Neck cancer.

3.5 Data Sets

3.5.1 Data on Observed Frequencies for MGMT Gene

MGMT, (Methyl Guanine Methyl Transferase), is a tumour suppressor gene. It is found on chromosome 10 in humans and acts as DNA repair protein during cell division. The CpG island in the MGMT promoter region is $\approx$ 780 bp in length, including 97 CpG dinucleotides, (Soejima et al., 2005). Normally, the MGMT promoter region is unmethylated in all tissues, (Esteller et al., 1999; Ongenaert et al., 2007). Silencing of this gene plays a critical role in cancer initiation, (Soejima et al., 2005). The MGMT methylation status of the tissues are varied: some of them show heavily methylated CpG islands and others show hypomethylation, regardless of gene silencing. In this work, we analysed some specific database such as PathEpigen and Pubmeth for MGMT CpG hypermethylated frequency variations in different tissue and found that an average hypermethylated frequency for the MGMT varies in between $4-38$ (PathEpigen) and $21-37$ (PubMeth), shown in Table (3.1). These are wide ranges. According to the PathEpigen and Pubmeth database average CpG hypermethylation frequencies for

(a) PathEpigen Database

Tissue	Frequency(%)
Bladder	4
Brain	38
kidney	8
Lung	29
Pancreas	11
Skin	11
Head and Neck	28

(b) PubMeth Database

Tissue	Frequency (%)
Bladder	21
Brain	37
Kidney	25
Lung	29
Pancreas	24
Skin	25
Head and Neck	24

Table 3.1: MGMT CpG hypermethylation frequency variation for different databases

MGMT promoter region are $\approx 4 - 38$. Figure (3.1) illustrates comparison between cancer tissues of PubMeth and PathEpigen database. In Figure (3.1), X-axis denotes methylated frequencies (%) for MGMT gene and Y-axis represents different cancer tissues. Tables [(3.2) and (3.3)] represents frequency of MGMT promoter regions for the colorectal and lung cancer respectively. Table (3.2) depicts an average methylation range in between $0 - 12$ for the normal tissues, with again a broad range of MGMT % for cancer initiation, i.e. $12 - 46$. Additionally, a similar assessment for the lung cancer, gives % MGMT of 0 for the normal tissues and between $14 - 36$ for the cancer

Frequency(%) of MGMT		
Cancer type	Normal tissue	Cancer
Colorectal adenocarcinoma	0	32
Colorectal carcinoma	0-10	3-44
Colorectal adenoma	0-10	8-50
No subtype specified	0-12	12-46

Table 3.2: MGMT frequency(%) variation for colorectal cancer from Pubmeth database

Frequency(%)		
Cancer type	Normal tissue	Cancer
No subtype specified	0	2-39
†SCL	0	20
§NSCL-squamous cell carcinoma	0	14-36

Table 3.3: MGMT frequency(%) variation for lung cancer from Pubmeth database. Table (3.3) §Non-small cell lung cancer; †Small cell lung cancer.

tissues, illustrated in Table (3.3). According to these data, we estimate an average frequency range for normal $(0 - 10)$, *de-novo* $(10 - 30)$ and hypermethylation (≥ 30) of the MGMT promoter region.

3.5.2 Dynamics of DNA Methylation Mechanisms

Since DNA methylation is an important mechanism for regulation of gene activity. It is important to know how methylation patterns are established on the CpG sites at a gene promoter. Methylation patterns are initiated and maintained, as noted in Chapter (2), mainly by combined activity of two enzymes: DNMT1 (D1) and DNMT3B (D2). Table (3.4) represents the functions of DNMT1 and DNMT3B in DNA methylation mechanisms. With respect to cancer initiation, observed expressions data of DNMT1 and DNMT3B from the literature are given in Table (3.5). Here, a fold is a concentration

Enzyme	Function	Reference
DNMT1	Maintain DNA methylation and low *de-novo* activity.	(Doerfler et al., 2006)
DNMT3B	*De-novo* methylation activity and very low maintenance of DNA metylation.	(Doerfler et al., 2006)

Table 3.4: Function of DNMTs in the DNA methylation mechanisms

Expression level	DNMT1	DNMT3B	References
Overexpressed	>2 fold*	>7.5 fold	(Robertson et al., 1999)
Average	4 and 10-40 fold	3.1 fold	(Robertson and Jones, 2000)
Hemimethylation	7-20 fold	3-7 fold	(Hesketh, 2009)
Hemimethylation	5-30 fold	>7.5 fold	(Jair et al., 2006)

Table 3.5: DNMT1 and DNMT3B expression. *2 fold means that an increase level of enzyme to two times the original amount

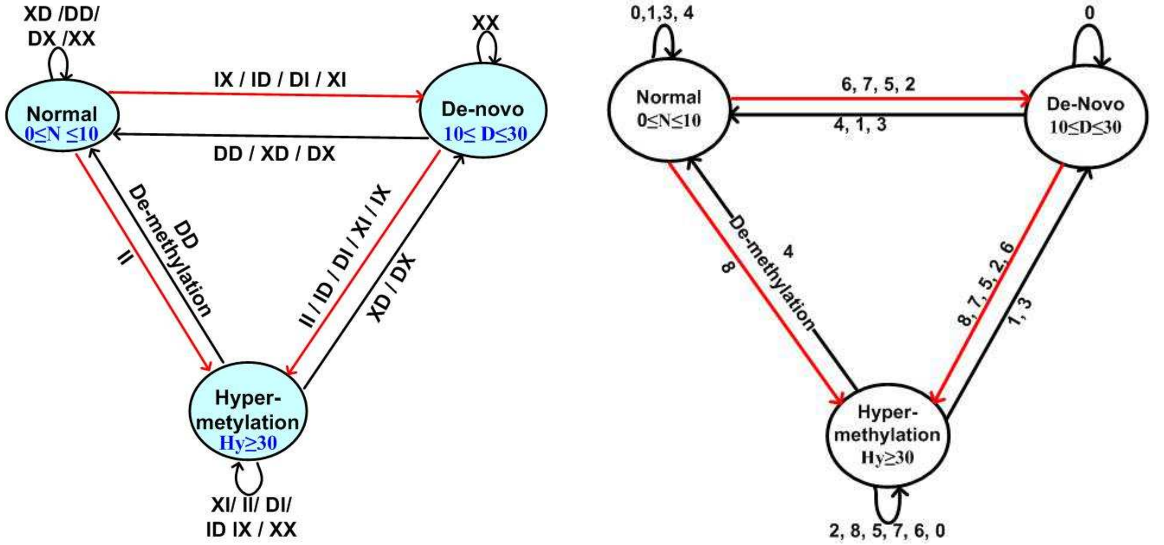
of the DNMT1 and DNMT3B from normal to other stages of methylation. Detailed information for the DNA methyltransferase is available at http://www.uniprot.org/. These DNA methyltransferase enzymes provide maintenance (DNMT1) and *de-novo* (DNMT3B) methylation patterns in a cell. Two possible conditions for the methylation state instigated by the activity of methyltransferase in a cell, (Laird, 1996) are given below:

1. Steady-state; when DNMT3B concentration level > DNMT1 concentration and a limiting level of DNMT1 operates.

2. Higher- state; when DNMT1 concentration level > DNMT3B and a limiting level of DNMT3B operates.

3.6 The *DFA* Model

In this section, we describe a new approach for analysing methylation dynamics. Figure (3.2(a)) shows possible methylation patterns by the activity of the enzymes for CpG island methylation. The methylation dynamics has three states for the methylation frequency as illustrated in the Figure (3.2(a)): (i) normal, (ii) *de-novo* and (iii) hypermethylated. The arrow in Figure (3.2(a)) indicates the change in two states after



(a) *DFA* State of DNA methyltransferase for double bit input: Strings of even length, {(IX/ID/DI/XI),(II/ID/DI/XI/IX)and(II)} over red colour arrows imply that increased or decreased concentration of D1 and D2 for the normal to *de-novo*, *de-novo* to hypermethylation and normal to hypermethylation transition states. Conversely, strings, {(DD/XD/DX), (XD/DX) and (DD)} over the black arrows show that D1 and D2 increased or decreased concentration for the *de-novo* to normal, hypermethylation to *de-novo* and hypermethylation to normal transition states respectively.

(b) *DFA* State of DNA methyltransferase for single bit input: Strings of alphabet, {(6,7,5,2),(8,7,5,2)and(8)} over red colour arrows imply that increased or decreased concentration of D1 and D2 for the normal to *de-novo*, *de-novo* to hypermethylation and normal to hypermethylation transition states. Conversely, strings, {(4,1,3), (1,3) and (4)} over the black arrows show that D1 and D2 increased or decreased concentration for the *de-novo* to normal, hypermethylation to *de-novo* and hypermethylation to normal transition states respectively.

Figure 3.2: State diagram of CpG Methylation dynamics at a promoter region for MGMT. Red colour arrows indicate transition from normal to cancer state.

receiving input of a string of even length over the alphabet $\Sigma = \{I, D, X\}$; where I, D and X represent increased, decreased and zero concentrations of DNMT1 and DNMT3B enzymes respectively and N, D, and Hy indicates normal, *de-novo* methylation and hypermethylation stages of a CpG island for gene promoter. For example, if we take even length input string IXIDDIXI, DFA_1 is in normal state (q_n), as illustrated in Figure (3.3). In the first step after consuming 'two bit' input IX, DFA_1 moves to state q_{dn} and the string left is IDDIXI. In the next step it consumes ID and moves to state q_{dn}. Continuing in this way then, after consuming the entire string IXIDDIXI, DFA_1 stops in q_{dn} (final *de-novo* state). We can simplify both the $DFA's$ (Figure (3.4)) by identifying the set {XX,XD,XI,DX,DD,DI,IX,ID,II} with the set {0,1,2,3,4,5,6,7,8} as shown in Table (3.6). After rewriting the $DFA's$ using the alphabet from the set {0,1,2,3,4,5,6,7,8} we get single bit input DFA, illustrated in (3.2(b)). The DFA accepts numeric strings which is verified from the Table (3.6). One can ask the following two questions from the dynamic as shown in Figure (3.2(a)).

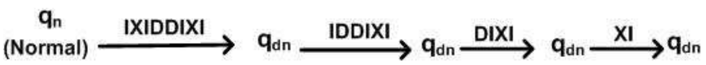


Figure 3.3: An example of transition state for the two bit input. q_n and q_{dn} represents normal, *de-novo* states.

1. If the methylation frequency is normal then what concentration patterns of D1 and D2 are necessary to retain the methylation level in a normal state? Here D1 and D2 are DNMT1 and DNMT3B respectively.

2. If the methylation frequency is normal then what are the concentration patterns of D1 and D2 required to convert it into a hypermethylated frequency?

D1	D2	Meaning	Two bit Input	Single bit Input
X	X	Both D1 and D2 are zero	XX	0
X	D	D1 is zero and D2 is decreased	XD	1
X	I	D1 is zero and D2 is increased	XI	2
D	X	D1 is decreased and D2 is zero	DX	3
D	D	Both D1 and D2 are decreased	DD	4
D	I	D1 is decreased and D2 is increased	DI	5
I	X	D1 is increased and D2 is zero	IX	6
I	D	D1 is increased and D2 is decreased	ID	7
I	I	Both D1 and D2 are increased	II	8

Table 3.6: Variations of D1 and D2 enzymes Concentrations. Two and single bit DFA Input

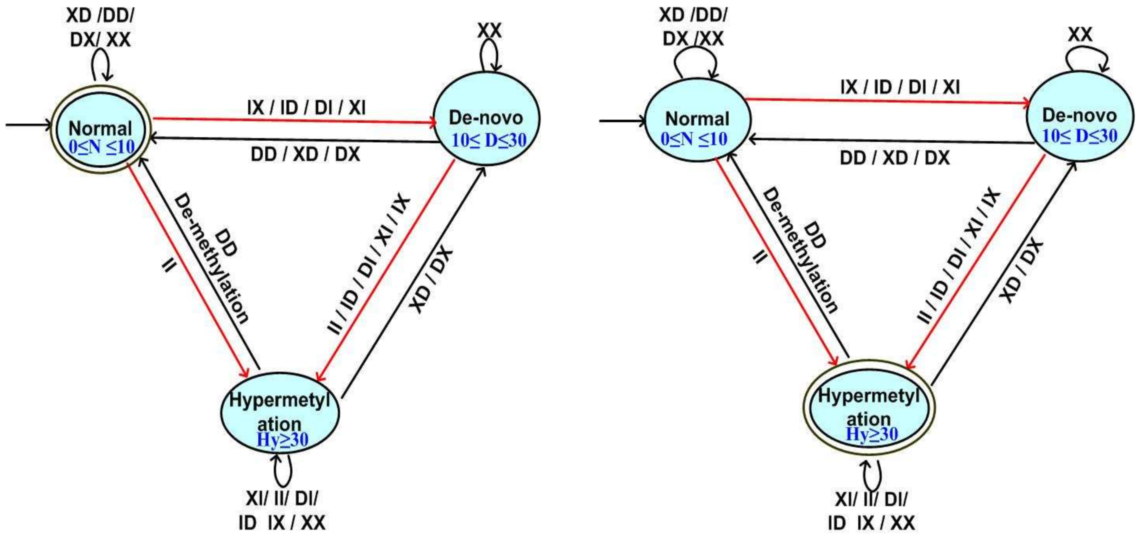
Motivated by these two questions we consider two DFA models as described in Figure (3.4). In this example, current DFA is present two automata state as shown in Figure, (3.4(a) and 3.4(b)).

DFA_1 When start and end state is normal, illustrated in Figure (3.4(a)).

DFA_2 When start is normal and end state is hypermethylated, illustrated in Figure (3.4(b)) .

3.6.1 The $DFA's$ Transition States

The deterministic finite automata (DFA) machine accepts regular languages (Sipser, 2008). Our DFA model reads single bits of the input strings at a given time. Here, we are considering only the following two DFA cases: (i) When start and end states are the normal state (denoted as q_n), illustrated in Figures (3.5(a)), and (ii) when start state is normal and end sate is hypermethylated (denoted as q_{hy}), illustrated in Figure



(a) DFA for normal state (DFA_1) (b) DFA for cancer state (DFA_2)

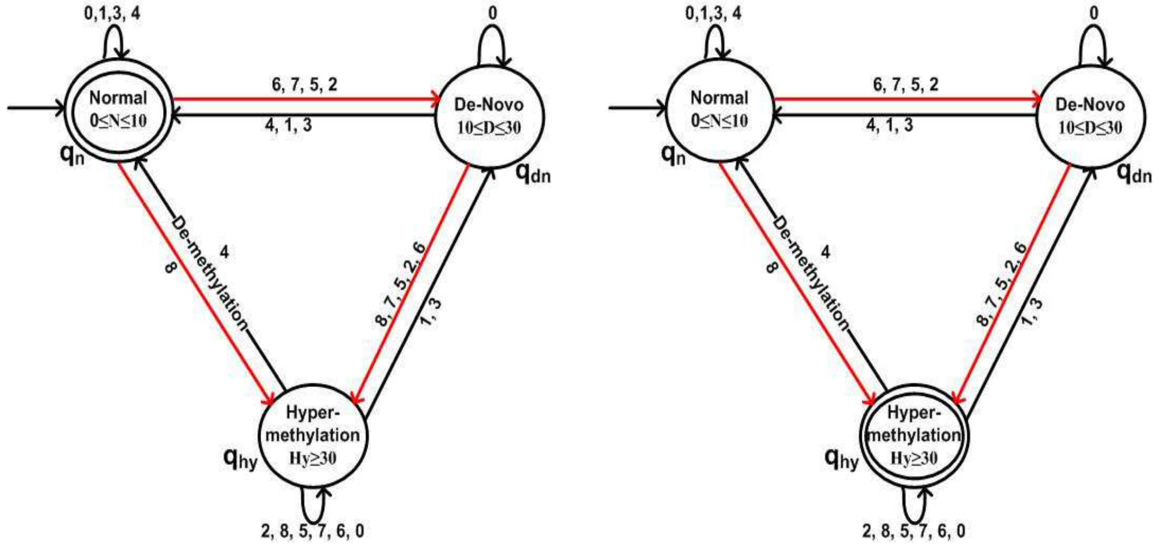
Figure 3.4: Double bit DFA of case 1 and 2

(3.5(b)). Figure (3.5(a)) and (3.5(b)) represents single bit DFA for the normal and hypermethylated states respectively.

Transition states for the DFA_1 and DFA_2 are illustrated in Tables (3.7) and (3.8) respectively. Next example (Figure 3.6) shows how DFA_1 accepts the string 258014.

δ	0	1	2	3	4	5	6	7	8
$\longrightarrow q_n{}^*$	q_n	q_n	q_{dn}	q_n	q_n	q_{dn}	q_{dn}	q_{dn}	q_{hy}
q_{dn}	q_{dn}	q_n	q_{hy}	q_n	q_n	q_{hy}	q_{hy}	q_{hy}	q_{hy}
q_{hy}	q_{hy}	q_{dn}	q_{hy}	q_{dn}	q_n	q_{hy}	q_{hy}	q_{hy}	q_{hy}

Table 3.7: Transition function for DFA_1. q_n, q_{dn} and q_{hy} implies normal, *de-novo*, and hypermethylation sates and ($\longrightarrow q_n{}^*$) denotes start and end state for DFA_1. δ is the transition function.



(a) DFA for normal state (DFA_1)

(b) DFA for cancer state (DFA_2)

Figure 3.5: Single bit DFA of case 1 and 2. q_{dn} is denotes as *de-novo* state.

δ	0	1	2	3	4	5	6	7	8
$\longrightarrow q_n$	q_n	q_n	q_{dn}	q_n	q_n	q_{dn}	q_{dn}	q_{dn}	q_{hy}
q_{dn}	q_{dn}	q_n	q_{hy}	q_n	q_n	q_{hy}	q_{hy}	q_{hy}	q_{hy}
${q_{hy}}^*$	q_{hy}	q_{dn}	q_{hy}	q_{dn}	q_n	q_{hy}	q_{hy}	q_{hy}	q_{hy}

Table 3.8: Transition function for DFA_2. Here, ($\longrightarrow q_n$) and (${q_{hy}}^*$) start and end state for the DFA_2.

Example:

Suppose DFA_1 is in normal sate (q_n) and the input string is 258014 (see Figure (3.6)). In step 1, after consuming 2, DFA_1 moves to state q_{dn} and the string left is 58014. In the next step it consumes 5 and moves to state q_{hy} and the string left is 8014. Continuing in this way after consuming the entire string 258014, DFA_1 stops in q_n (final normal state).

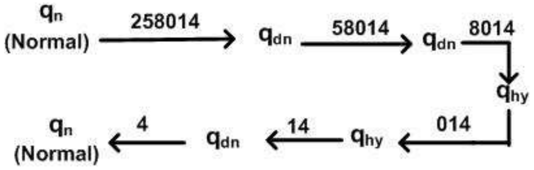


Figure 3.6: An example of transition states of proposition 1.

3.7 Results

3.7.1 Language of the DFA Models

We present some languages[11] accepted by DFA_1 in the following proposition.

Proposition 1 Let $\Sigma = \{0, 1, 2, 3, 4, 5, 6, 7, 8\ \}$ be the alphabet of symbols. If Σ^* denote the set of string of all possible size then some of the strings accepted by DFA_1 are from the set

$$L = L_1 \cup L_2 \cup L_3 \cup L_4$$

where:

$$\begin{aligned} L_1 &= \{x \in \Sigma^* \mid x_1 \in \{0,1,3,4\}\text{and } x_k = 4\} \\ L_2 &= \{x \in \Sigma^* \mid x_1 \in \{2,5,6,7\}\text{and } x_k = 4\} \\ L_3 &= \{x \in \Sigma^* \mid x_1 \in \{2,5,6,7,8\}, x_{k+1} \in \{0,1,3,4\} \text{ and } x_k = 4\} \\ L_4 &= \{x \in \Sigma^* \mid x_1 = 8 \text{ and } x_k = 4\}. \end{aligned}$$

[11] The set L provides only some of strings accepted by DFA_1. There are might be more strings that are accepted by DFA_1.

Hence, some of the strings accepted DFA_1 are either set of L_1 or L_2 or L_3 or L_4. It would be an interesting future task to find other strings accepted by DFA_1.

3.8 Discussion and Analysis

We used the Pubmeth and PathEpigen databases to access methylation frequencies for the MGMT gene in different tissues, as illustrated in Table (3.1). One reason the MGMT gene was chosen was because its silencing plays a prime role in the initiation of all cancers (Soejima et al., 2005). MGMT silencing disturbs DNA repair mechanisms, and errors occurring during DNA replication are preserved. We analysed MGMT frequency variations and found that it is $0 - 10\%$ methylated in normal tissues and $10 - 30\%$ methylated in cancer tissues, as depicted in Tables (3.3 and 3.2). We note two features of methylation dynamics based on DNMT activity:

1. Altered concentrations of D1 and D2 in normal to hypermethylated states.

2. Altered concentrations of D1 and D2 in hypermethylated to de-methylated states.

We examined 3 states (normal, *de-novo*[12] and hypermethylated) within the MGMT promoter region, as explained in Figure (3.2(a)). The distribution of the '3' methylation states is denoted as methylation levels from the variations in the activities of the DNMTs. The 'zero' value indicates an unmethylated promoter. Given the significance of alternate methylation states, we applied the DFA method to analyse each methylated state. The DFA estimates two conditions for each methylation states: (i) steady state and (ii) higher state, as described in section (3.5.2). We used an even length 'two bit' alphabet after the simplification of 'one bit' input [Table (3.6)] to simulate the

[12] *De-novo* is an intermediate state between normal and hypermethylation.

Methylation State	Cancer state	Cancer Prediction (%)
Normal	No cancer	26
de-novo	Cancer initiation	37
Hypermethylated	Cancer	37

Table 3.9: Cancer prediction based on the results of Table 3.10

DFA model. We found 9 possible transitions states for each methylation state. We found possible strings as results for the methylation dynamics analysis, given in the proposition in section (3.7.1). In the early stage of the model, we analysed some examples of each DNA methylation state (normal, *de-novo* and hypermethylated), as shown in Table (3.10). Table (3.10) gives DFA_1 and DFA_2 outputs for 30 random strings of length 100 generated by the software[13]. Table (3.9) presents cancer predictions based on the analysis of random strings as given in Table (3.10). Here, we represent normal state *de-novo* methylation and cancer initiation via hypermethylation. Based on the random data from Table (3.10), we obtained predictions. We analysed 30 random strings of length 100 and found that out of the 30 strings, 26%, 37% and 37% represented normal (no cancer), *de-novo* (cancer initiation) and hypermethylation (cancer) states.

Once we have real data of methylation frequency, we can check the methylation states of the gene and cell populations. The DFA model developed by us can be used for predicting cancer by obtaining methylation frequency corresponding to DNA methyltransferase (DNMTs) up- down regulation activity.

[13] Available at, for random string (http://www.psychicscience.org/random.aspx) and DFA simulator (http://home.arcor.de/kai.w1986/dfasimulator/).

Random String	DFA_1	DFA_1 **Output**	DFA_2	DFA_2 **Output**
7578327821753237363012444443760764118817045755887287728067368855055243728163241753765330686541168520	HM	Reject	HM	Accept
6641458674714267058815431638248061852821001630063128435528560528812234630140545261068573080868836341	Normal	Accept	Normal	Reject
0613521763672268248447040575535582261003656500041573200770884174758288153786782610858858526145770721	DN	Reject	DN	Reject
0334271605505862018582250250824288214027655454801265528735147403354547225100727125555240845882657083	DN	Reject	DN	Reject
6134233437177203770837562858825864831710608661472818712017260367871144058084360735732005878748468671	DN	Reject	DN	Reject
6483463307755131707444154186312275108583055043144554526307428840828577288488612847264256067155554716	DN	Reject	DN	Reject
5058125427861200025682267414557045683330515216545410460713878507552343615563345381224531113067450030	Normal	Accept	Normal	Reject.
6411055652865161284650501605446278685833317656818108237586038728622551455258350226812782546455884440	Normal	Accept	Normal	Reject
8325182865506168423148660855407015606607202751256631173802823833736126608383115568211031235628004137	DN	Reject	DN	Reject
5427175522757131212421878666846456063247250525408282780743356383830333130133434473628318356262850288	HM	Reject	HM	Accept
3672120121273300484708763055172804704463537360523083810373537622155047302676260386665876614238070670	HM	Reject	HM	Accept
1067070552880200538724732877825364831144467172287424140871264387148271020830407256726774688136002280	HM	Reject	HM	Accept
5671422635086578183584501471864136687515181455738342330648686014121260644418802300440504268312705454	Normal	Accept	Normal	Reject
0380813507154512662251443468837647567738068576131608183750265625735386182021106871337543271731215056	HM	Reject	HM	Accept
4160725884257654636548114112830662402303676053728637734613454603038825167271628114587835161868673070	HM	Reject	HM	Accept
1235707022573552325746355410806534331783466553766233818814683375420884100337626856086274543550278502	HM	Reject	HM	Accept
1213687353624360531072376275114630552587625255484180542482222026555858457312356671456507551754621168	HM	Reject	HM	Accept
0500788301103067505635180554738341557642107644361517505276408542635512208037566164564156074877168281	DN	Reject	DN	Reject
2382057572031543321660154730165610333634462704386376864088586362665557657761634627848735544718716243	Normal	Accept	Normal	Reject
2012525570804530085465348024684603002282235572770248521760772065141147008368461003565228202754738450	DN	Reject	DN	Reject
4188563572300517088166253211830276635175851301856562777748035630000037112358512510253463784282754037	DN	Reject	DN	Reject
7043654465443864217366801508003876460130110331655757653005257485528161360576208010845268334206247444	Normal	Accept	Normal	Reject
1686808387758386641847430848716206557213637670222331053671176604733064880703068832664710470340404021	Normal	Accept	Normal	Reject
3880562642350707737013255143765651342063782706105672414254802354502066558170352341706138202564151600	DN	Reject	DN	Reject
8324576565875150180246445201171840121145803134637414112538684106277684843834718675646788754023447341	Normal	Accept	Normal	Reject
7174000181830611473304877047636010238055332374004142571370752232137388648748868124307023212757651120	DN	Reject	DN	Reject
1751033740686004220034410674253835355214625011502200883107623204325238301331288454514281021737382557	HM	Reject	HM	Accept
5027624043350063467155473886431738271505277402658502424862500317821327565478681207282127865542180246	DN	Reject	DN	Reject
0408676441143660253702480886184236420261164717532750374067248368824752208010653400513148807614820500	HM	Reject	HM	Accept
3442555722675364326480711743177663666383254466713081004530772755867673168103581554082381457314730685	HM	Reject	HM	Accept

Table 3.10: Data for random strings. Here, DN and HM denotes *de-novo*, and hypermethylation state respectively.

Chapter 4

Data Curation and Annotation for PathEpigen

An overview of the “bioinformatics” or investigation towards parametrisation of future models of epigenetic mechanisms is provided in this chapter. The first section gives an introduction of the PathEpigen database. The data curation process is subsequently discussed, “single and “double relation” are defined and examples are provided. In concluding on the work of this chapter we highlight some of the key requirements for biomedical resources in terms of querying and extraction of data for empirically-based model.

4.1 Introduction

The Sci-Sym laboratory in DCU, has been involved in investigate biomedical resource strands, and latterly in designing and developing, for its various investigations, a resource for epigenetics. In particular, work on PathEpigen, aims to integrate genetic and epigenetic molecular determinants of (colon) cancer (in the first instance). This knowledge management platform consist of a database and a user interface and its primary development has been by Dr. Barat and co-workers, (Ruskin et al., 2008). The database correlates molecular events to various colon cancer phenotypes, with the focus on early-stage examples. Currently, the resource contains 6,445 records on colon cancer. These represent correlations between epigenetic and other molecular events such as gene expression and mutations. The resource is expected to facilitate future modelling and quantitative research in the cancer field, especially in the area of early diagnosis and risk assessment. The objective of PathEpigen is to integrate and comprehend the involvement of genetic and epigenetic events in various phenotypes of colon cancer. The resource stores and integrates colon cancer molecular data in a specified computational format. It also collects specific clinico-pathological factors, different phenotypes and statistical information about these such as the relative frequency of occurrence of, or incidence of, a molecular event, given the Phenotype of the sample: i.e. P(Event | Phenotype). The PathEpigen database is implemented in MySQL, following a relational schema generated according to an expert assessment of the data structures encountered in the specialist literature. Efforts are being made to support extension and refinement of the resource, and to incorporate additional details, as new data are published. To provide high quality identification and storage of pathological molecular determinants, the resource brings together, for common analysis, data obtained from different experimental platforms, (epigenetic, genetic, mutation and other). Therefore, the database is designed to support storage for a variety of distinct types of data, to allow for dependent events such as "single" and "double relations", and allows for expansiion and growth of available data, with future provision built to be as flexible as possible in order to accommodate new data types. This is achieved through PathEpigen's own internal classification system, designed to facilitate both the curation and querying processes. As the project has expanded, it has generated additional ramifications that need dedicated research and development inputs. In the following subsections, the emphasis on colon cancer is given, and my contributions to project tasks are described, providing a further dimension to the investigation of requirement for epigenetic modelling.

4.2 Colon Cancer

Colon cancer, (also known as colorectal cancer), is a malignancy in which there is uncontrolled cell growth on the inner side or end of the colon. It arises from the epithelial cells[1] that line the colon[2] of the gastrointestinal tract[3]. It is exacerbated by genetic and epigenetic changes in the Wnt signalling pathway, (Segditsas and Tomlinson, 2006). Both genetic and epigenetic processes are involved in colon cancer progression. Some points concerning the progression of colon cancer involving the analysis of its molecular genesis are as follows:

- Genomic instability, (Lengauer et al., 1997), which is due to genetic and epigenetic alterations of housekeeping gene such as APC, MGMT and MLH1.
- Heritability[4] character, (Kinzler and Vogelstein, 1996), which is due to genetic defects in germ line cells[5].

[1] It is a membranous tissue covering internal organs and other internal surfaces of the body.

[2] The colon (large intestine) is the last part of the digestive system and extracts water and salt from solid wastes.

[3] The human gastrointestinal tract refers to the stomach and intestine.

[4] Heredity is the passing of traits to offspring.

[5] The germline of a mature or developing individual are those cells that have genetic material that

Subsequently, the progression of colon cancer is activated and de-activated by oncogenes and tumour suppressor genes, respectively. Consistent with the facts described above, colon cancer is initiated by alterations in the Wnt signalling system and progresses through subsequent sequential events.

4.2.1 Symptoms and Risk Factors

Colon cancer progression is marked by no specific symptoms. Colon cancer development depends on the location of the tumour in the colon, and it spreads elsewhere in the body to become metastatic or constitutional (affecting the whole body). Here, we report some specific factors that facilitate colon cancer invasion, such as alcohol use, age, polyps[6], and environmental factors. A study by Cho et al. (2004) has found that if a person takes in more than 30 grams of alcohol per day, they are at higher risk for colon cancer. Ageing is also a risk factor. Most patients are in their $60s$ or $70s$, and patients under 50 are uncommon unless a family history of early colon cancer is present, (Lengauer et al., 1997). Polyps of the colon, particularly adenomatous polyps, (Segditsas and Tomlinson, 2006) present an additional risk. The removal of colon polyps, identified by colonoscopy reduces the subsequent risk. Industrial countries are at a higher risk compared to less developed countries that traditionally have high-fibre/low-fat diets. Studies of migrant populations have revealed a role for environmental factors, particularly diet, in the aetiology of colorectal cancers, (Cho et al., 2004).

may be passed to a child.

[6] A polyp is a fleshy growth, occurring on the lining of the colon.

4.2.2 Specific Affected Genes

Important genes are frequently affected in colon cancer, including APC, MGMT and MLH1.

4.2.2.1 APC:

Adenomatous polyposis coli (APC) is a tumour suppressor gene that regulates the transcription of critical cell proliferation factors by interacting with transcription factors. Loss of APC function initiates a sequence of changes at both the molecular and histological levels, (Fodde, 2002). Mutations in the APC gene alter transcription and initiate familial adenomatous polyposis (FAP)[7]and sporadic colon cancer. They block the Wnt pathway, and therefore, Wnt is turned off. Subsequently, either chromosomal instability (CIN) or microsatellite instability (MSI), occurs, which appear early in tumour progression, (Segditsas and Tomlinson, 2006). High APC mutation rates have been reported, with APC mutated in up to 70% of all early stage colon cancers, (Miyaki et al., 1994). APC hypermethylation initiates transcriptional silencing independently of somatic mutations. Hypermethylation of promoters occurs at an early stage in tumourigenesis, (Lee et al., 2009). APC is associated with FAP and somatic mutations in colon cancer. Hypermethylation in the APC promoter region alters gene activation mechanisms. One study (Esteller et al., 2001) suggested that ≈ 34% of promoter APC methylation is found in colon cancer.

[7] An inherited condition in which numerous polyps form mainly in the epithelium of the large intestine.

4.2.2.2 MGMT:

MGMT gene encodes a protein known as methylated-DNA-protein-cysteine methyltransferase, (Lee et al., 2009). It is a DNA repair protein that is involved in defending cells from O6-methylguanine (O6-MeG)[8] in DNA. Inactivation and promoter hypermethylation of the MGMT gene frequently occurs in colon cancer, (Shen et al., 2005). MGMT silencing is associated with a G (Guanine) to A (Adenine) point mutation in the KRAS[9] gene during colon cancer progression, (Esteller et al., 2000). Thus, MGMT promoter methylation is a marker for colon cancer. One study has shown (Shima et al., 2010) MGMT promoter hypermethylation in 325 tumours (38%) and the loss of MGMT expression in 37% of colon cancers. Silencing of MGMT has been shown to be associated with and to precede the appearance of G-to-A point mutations in the KRAS gene during colorectal tumourigenesis.

4.2.2.3 MLH1:

MLH1 (MutL Homolog 1) is a mismatch repair gene (MMR) that provides a signal for DNA repair proteins. It fixes errors that occur during DNA replication, (Jacob and Praz, 2002). In HNPCC (Hereditary Non Polyposis Colorectal Cancer), MLH1 predisposition causes colon cancer more than other cancers, such as stomach or blood cancer. Genetically, mutated MLH1 leads to DNA mismatch repair mechanisms and causes the accumulation of mutations in the genome, thereby initiating microsatellite instability (MSI) and promoting carcinogenesis, (Bronner et al., 1994). When MLH1 becomes inactive, errors are left during cell division and genomic instability increases.

[8] 6-O-Methylguanine is a derivative of the nucleobase guanine in which a methyl group is attached to the oxygen atom. It base-pairs to thymine rather than cytidine, causing a G:C to T:A mutation in DNA.

[9] A oncogene which is activate in the cancer genome.

Subsequently, cells divide continuously and errors accumulate in the DNA. Finally, the cells become inactive and form tumours in the colon. As described above, DNA mismatch repair deficiency impacts MLH1 promoter methylation, (Perucho, 2000). MLH1 hypermethylation contributes to DNA mismatch repair mechanisms and leads to genomic instability, (Leung et al., 1999).

4.2.3 Genetic changes in Colon Cancer

Colon cancer is a multi-step process that is accompanied by genetic alterations to the normal colonic epithelium, which leads to adenomatous polyps,[10] (Lea et al., 2009). Colon cancer steps are driven by mutations in different classes of genes, such as tumour suppressor genes, proto-oncogenes and mismatch repair genes.

1. The key tumour-suppressor gene (TSG) changes in colon cancer are loss of heterogenicity at chromosome-5 (in the APC gene), chromosome-18, (in the DCC tumour-suppressor gene), and chromosome-17 (p53). For example, a mutation in the APC gene at the chromosome 5q-marker is an early event in colon cancer progression. It causes inherited familial adenomatous polyposis syndrome,[11] (Gryfe et al., 1997).

2. Activating mutations in proto-oncogenes such as KRAS are also an early event in colon cancer.

3. Mismatch repair genes such as MGMT and MLH1 play a role in the predisposition of colon epithelial cells to cancer. If these gene are mutated, errors will start to

[10] An adenoma is a type of polyp which is pre-malignant.

[11] It is collection of hundreds to thousands of colonic adenomatous polyps. When these polyps are left untreated, initiates colon cancer.

accumulate at rates of hundreds or thousands per cycle.

4.2.4 Epigenetic Emphasis on Colon Cancer

Epigenetics is the inheritance of basic gene expression information in the absence of base sequence changes. Detailed studies of epigenetic mechanisms are described in chapter (2). Epigenetic modifications are also characteristic of colon cancer and include DNA methylation, histone acetylation and gene silencing mediated by small non-coding RNA (microRNA), (Wong et al., 2006). The promoter sites of tumour suppressor genes have been found to be methylated in colon cancer. Some tumour suppressor genes such as APC, MGMT and MLH1 have been found to be hypermethylated in colon cancer. As mentioned earlier, DNA hypermethylation and hypomethylation are related to gene silencing and genomic instability, respectively. Activation of proto-oncogenes such as KRAS has also been detected in colon cancer. The range of aberrant methylation of the APC, MLH1, and MGMT genes is 20%, 15% and 50%, respectively, for colon cancer, (Lind et al., 2004). A study (Ramreza et al., 2008) found that 47% of genes are nonmethylated, and 41% are methylated in 82 colon cancer samples. The MGMT, APC and MLH1 genes are most frequently methylated in colon cancer. The promoters of these genes are methylated in the early stages of colon cancer progression. In addition, hypermethylation of these gene correlates with age, (Christensen et al., 2009). Promoter methylation of the tumour suppressor gene SFRP causes activation of the Wnt singling pathway in colon cancer, (Suzuki et al., 2004).Therefore, aberrant methylation patterns are very common and the analysis of these changes is important in colon cancer development.

4.3 Data Curation

Data curation has been critical to the development of biology from Linnaeus to UniProt, while the careful collection and organisation of data has been the source from which new hypotheses and understanding have emerged, (Thornton, 2009). Data curation involves extracting useful data from active and on-going research resources using bioinformatics and biological knowledge. The aim of data curation is to improve data discovery and retrieval, while preserving its quality, improving its value, and providing data re-use, (such as examining it in conjunction with other data or as inputs for modelling). Data curation involves important features, such as: authentication, archiving, management and an emphasis on representation of ideas clearly, in order to support its verification and validation. As biomedical and biological information continues to grow at staggering rates, efficient methods need to be devised to extract biological information from the literature, then to process and manage it in order to achieve worth while upgrading. Without such efforts, there is little prospect for successful information synthesis. In this context, *manual curation* remains essential, but should be limited to those parts that can not be done automatically which, unfortunately, tends to be extensive particularly for new fields of study. This intensive, manual curation of the scientific literature is currently crucial to population of the PathEpigen platform, (as opposed to text-mining), due to its acknowledged high accuracy, (Winnenburg et al., 2008). Text mining is a high-throughput computational technique that scales easily, but is capable of making large errors, due to the complexity of natural language. The information sought for PathEpigen is thus too complex to be subject to automatic curation at the current time of writing. Our records link together a number of concepts such as genotype and phenotype, found in different parts of the published literature, under different formats and from different assays; such data are not readily amenable to automatisation. The last decade of experience gained at the European Bioinformatics Institute [http://www.ebi.ac.uk/] and other bioinformatics centres, such as EMBL, (Heidelberg; http://www.embl.de/index.php) has shown that the best people to develop and design biomedical resources are those who initially dedicate time to data curation, because they have a fuller understanding of the structure and information content of the data. Some challenges for the data curation task include:

1. Development of ontology-driven GUIs which play an important role in many semantically working environments, such as knowledge management systems.

2. Improvement in presentation of data to users, e.g. in showing statistical relations between molecular events e.g., the probability of the gene being expressed and its dependence on the strength of its epigenetic regulation mechanisms.

3. Improvement in data submission systems.

One aim while working on data curation for PathEpigen, was to participate in the design of its data submission system which is currently being developed in Sci-Sym. While PathEpigen was in its initial stages of development, MySQL code was used in order to insert new data. This approach is fairly rudimentary, and the reason for adopting it was ease of use together with initial pilot resources. In particular, early work involved a relatively small number of records in a limited number of tables, which provide the main data-types in the database. However, as the resource has continued to grow, the initial method of curation has proved both cumbersome and time-consuming and a data submission interface is partly now designed.

4.3.1 The PathEpigen Data Model

PathEpigen is a manually curated, web-accessible knowledge management system, integrating comprehensive information initially, on molecular events for the genetics and epigenetics of colon cancer, and their relations to various phenotypes, (Barat and Ruskin, 2010). The database includes tables, containing basic information on genotypes and phenotypes of colon cancer. These tables are the building blocks, which can be used to represent interesting correlations between the events. There are two ways by which PathEpigen represents correlations between genotype and phenotype:

1. The frequency of finding a molecular event in the context of a given phenotype; defined as *"single relation"*.
2. The frequency of finding a molecular event in the context of another molecular event and a given phenotype defined as a *"double relation"*.

The relation-entity association scheme of the database is illustrated in Figure (4.1). "Single relations" record a molecular event, a phenotype and the frequency of that event. In probabilistic terms, a "single relation" can be expressed as:

$$Event\_Freq = Probability(Event|\ Phenotype) \tag{4.1}$$

Equation (4.1) thus represents the correlation between a molecular event and it's associated phenotype. For example, if MGMT promoter methylation is observed in 38% of 23 tubular adenoma[12] samples, the main fields of a "single relation" record from the database can be shown as for Table (4.1): From the point of view of conditional

[12] A benign epithelial tumour of glandular origin.

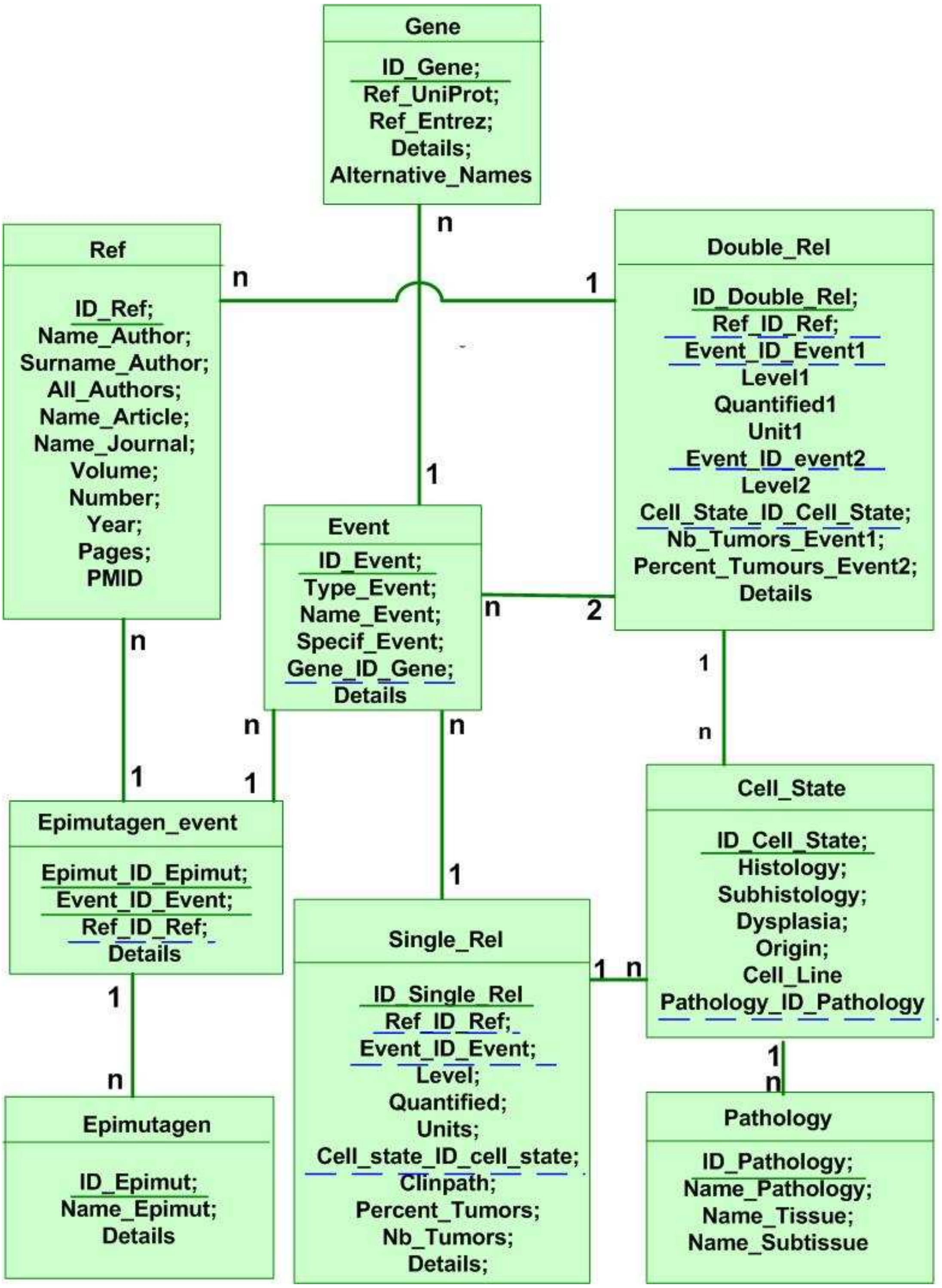


Figure 4.1: Entity Relationship for PathEpigen Database

Ref	Gene	Event	Quanti	Units	Level	Phen	Clinpat	Fr	nb_samp
PMID	MGMT	HM	NULL	NULL	YES	adenoma	NULL	0.38	23

Table 4.1: Inserting example for the “Single relation”

Ref	G1	Ev1	EID1	L1	Qnt	Unit	NTE1	G2	Ev2	EID2	L2	Fr	Phe	CP
PMID	APC	mut	128	Y	NL	NL	47	APC	M	128	Y	0.980	10	NL

Table 4.2: Inserting example for the “Double relation”

probabilities, this can be written as:

$$P(MGMT\, promoter\, methylation = YES|Phenotype = colon\, tubular\, adenoma) \tag{4.2}$$

where MGMT promoter methylation and phenotype are variables which have taken the values “YES” and “adenoma” respectively. The level (or status) of the molecular event is “YES”, because the promoter methylation is present, (without any information on the methylation intensity in this particular case). Certain references in the literature will in addition give information such as the fact that the promoter is highly methylated, or totally methylated. In those cases, the variable level in Table (4.1) takes the values, “HIGH” and “TOTAL” respectively. A “Double Relation” is represented by Equation (4.2). The records give two events, their qualitative or quantitative data, and their summary statistics; the number of samples in which the first event is present, in the case of the given phenotype, and the incidence of the second event in this number of samples. For example, assuming that 47 colon carcinoma[13] samples, with mutations in the APC gene have been verified for APC promoter methylation, and 98% samples were found to be methylated, the corresponding “double relation” is given by Table (4.2), Equation (4.3) following: In this table it can be written as P(APC promoter

[13] Any malignant tumour derived from epithelial tissue.

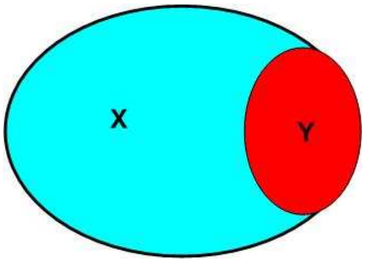


Figure 4.2: Venn diagram of “Double Relationship”. Where ‘X’ and ‘Y’ are probability of event_1 and event_2 respectively.

methylation = YES — APC mutation = YES and Phenotype = colon carcinoma). The variable phenotype can take other values such as colon adenoma, colon polyps, normal colon and others.

$$Event\_2\_Freq = Probability(Event\_1|Event\_2 \; and \; Phenotype) \tag{4.3}$$

Equation (4.3) represents a “double relation” from the statistical point of view, that is: the frequency of event_2 can be defined by the probability of occurrence on event_2 in samples characterised by event_1 and the given phenotype. Consequently, equation (4.3) can also be depicted as the Venn diagram in Figure(4.2).

4.4 Methods and Techniques

4.4.1 Data curation for PathEpigen

In order to extract biological information from bulk sources, expert analysis, comparison and annotation is needed. Further, an extensive knowledge of the role of genetic and epigenetic modifications in cancer development is required for data curation for PathEpigen. Background work experience, (involving detailed study of the epigenetic molecular determinants of cancer), has provided this and this expertise has been further improved upon for the current curation work, which has involved considerable data fitting to the PathEpigen data model. Tasks included were:

- Familiarisation with the data model of the database.
- Assessment and minimisation of the error rate in manual data curation.

A strategy was devised for accomplishment of these tasks:

A. Based on specific keywords, related to molecular determinants of colon cancer, relevant articles have been selected from the PubMed[14] resource.

B. In this context new sources of data such as gene expression data from new microarray experiments, paired with methylation data, have been explored.

C. Text-mining facilitated the extraction of accurate, comprehensive information from literature for manual curation.

D. Further, articles have been selected by reading their abstracts.

E. Data related to epigenetic, genetic and phenotypic information have been extracted, by examining the full manuscripts.

One of the main objectives of this work is to find correlations and associations between new genetic and epigenetic events; such events are found in the colon cancer phenotype and molecular events are already known as cancer hallmarks. Therefore, much attention has been paid to the simultaneous reporting or outcome between molecular events. The following thus formed the focus of the curation work:

[14] A literature resource for the biomedical and life sciences journal literature. Available at www.ncbi.nlm.nih.gov/pubmed.

1. Clear distinction between different colon cancer phenotypes in order to provide insight on pathology initiation and dynamics.
2. Extraction of quantitative data for the resource.
3. Provision of non-redundant data i.e. high-quality manually annotated data curated in relation to various types of cancers
4. Correlations between molecular events.
5. Analysis of molecular events, according to clinical pathological factors.
6. Selection of information on environmental factors, and their interaction with epigenetic and genetic events.

4.4.2 Details of Curation Process

The curation process of individual scientific reference contained the following detailed steps:

a. The phenotype characterising the samples involved in the specific study, was isolated and compared against the PathEpigen classification of phenotypes. If the given phenotype existed already in the classification, its ID was kept for further annotation of the "single" and "double relations" found in the reference. Otherwise, a new phenotype was added to the classification and allocated an identification number.

b. All molecular events found in the samples were isolated. Again, a molecular event was checked against the database and introduced if not already present.

c. "Single relations" were curated. The molecular event and the phenotype were recorded, together with the total number of samples analysed and the frequency of the molecular event found in these samples.

d. "Double relations" were dealt with similarly to "c" above.

PathEpigen also supports annotation of relations with clinico-pathological factors. For instance, if MGMT methylation frequency was observed in 73% of 15 proximal colon cancer samples then the "single relation" record has the form as illustrated in Table (4.3). PathEpigen also differentiates between primary tumours and cell lines; the reason for this is that, as cell lines are passed on through many generations, they are likely to acquire new modifications, absent from the primary tumours that the original cell lines were collected from. Thus, each cell line is recorded as a distinct phenotype. In the case of cell line phenotype, a "single relation" has the form shown in Table (4.4). In this table a "single relation" contains a molecular event methylation, statistical information about it, such as number of samples analysed, incidence of the event, i.e. 'YES' in the Level column, and any qualitative or quantitative information available on the event. The phenotype is CaCo2 cell lines.

Ref	Gene	Event	Quanti	Units	Level	Phen	CP	Perc	No._samp
PMID	MGMT	methylation	NULL	NULL	YES	C.C.*	proximal	73	15

Table 4.3: Inserting "Single relation" example for the clinical pathological. C.C.* represents colon cancer.

Ref	Gene	Event	Quanti	Units	Level	Phen	Clincphat	Perc	nb_samp
PMID	NDRG2	M+	NULL	NULL	YES	CaCo2	NULL	1	1

Table 4.4: Inserting "Single relation" example for the quantified unit

4.4.3 Algorithm

To support the curation process and data integration, a program which converts the initial text annotation format to an SQL format has been developed.

```
Open the input file ;
Check name of the file (If file not correct, show message” File could not open”);
if (fi! = NULL) then
    while (!feof(fi)) do
        ch = fgetc(fi)        // get character from file ;
        if ($ch == -1$) then break;    /* ch == -1 indicates end of the file */
        Put the file pointer p* at 1st position after reading and each line file pointer
        increase by one ;
        Initialize new_line to empty string ;
        Read the whole line from the input file ;
        p = line        // p* points to the line;
        while ($*$p! = 0) do
            if (strcmp(new_line,” ”)) then        /* If new line is empty */
                Put ”(”at the beginning of the line;
            else
                else put ”,”);
                Copy line in to pointer p*;
            endif
            while (($*p$! = ′new line′) && ($*p$ ! = ′tab′)) do
                If there is more letter in this line;
                Advance of the p pointer;
            endw
            change the ’ t’ or ’ n’ to ’ 0’, and then;
            Advance p* to the next word;
            strcat(new_line, q)        // Copy new line in to q* ;
        endw
        Put the ”)” at end of the line;
        Put the ”;” after bracket;
        Write the line in to out put file;
        Advances the pointer to the next line in the output file;
        Show message that “Job Done”;
    endw
endif
```

Algorithm 1: Algorithm of the developed Code

The complete code has been implemented using C++ language. This code helps to direct SQL query input to the database. A sequential view of the code is provided in Algorithm (1).

4.5 Results

At present, the PathEpigen resource contains 6,445 records. Of these, the curation efforts described here have added some 647 new records, including 430 “single relations”, describing early colon cancer phenotypes and given molecular events in colon cancer populations, as well as 212 “double relations”, all issuing from the manual curation of 50 scientific articles. Data were submitted for verification, quality control and validation. As part of this work, 24 new PathEpigen entries were included, which describe molecular events for 21 genes and 3 phenotypes, such as Gastric CRC, carcinoma with undifferentiated sub-histology, and colon cancer, originating from Crohn’s disease. An example of “double relations” for the new phenotype of Gastric CRC is illustrated in Table (4.5). In Table (4.5) statically possible outcomes for events MSI and CIMP are as follows: All samples for event_1 (MSI)= 16. Possible outcome frequencies of event_2 (CIMP) are 0.250 and 0.375 i.e. these event_2 frequencies may impact on the cell phenotype, represented in Figure (4.3) (A) and (B) respectively. Significant data, with regard to genetic and epigenetic correlation have been extracted from articles such as (Miyakura et al., 2001). In this article MLH1 gene promoter methylation and

Ref	G1	Ev1	EID1	L1	Qnt	Units	NTE1	G2	Ev2	EID2	L2	Fr	Phen	CP
262	NL	MSI	115	H	NL	NL	16	NL	CIMP	116	H	0.250	713	NL
262	NL	MSI	115	L	NL	NL	16	NL	CIMP	116	H	0.375	713	NL

Table 4.5: “Double relation” of the article (Kim et al., 2005). ‘713’ ID number represents Gastric cancer phenotype.

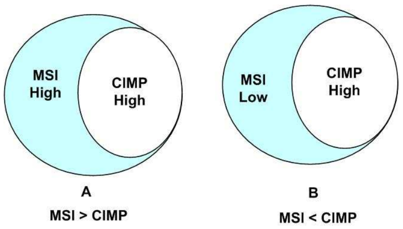


Figure 4.3: Diagrammatic representation of “Double Relationship” for the table (4.5).

expression status has been analysed in the context of colon cancer samples. The main reason for selection of this article as one of those for curation is because MLH1 function is involved in DNA mismatch repair and DNA mutations have been associated with colon cancer. Mammals without MLH1 are infertile and spermatocytes exhibit high levels of prematurely separated chromosomes and cell cycle arrest. MLH1 and the DNA repair mechanism, (of which MLH1 is a member), are directly linked to ageing. Thus, inclusion of data from this article contributes towards refining the probabilities of epigenetic events in the colon cancer population. For instance, “single relations” from this reference can be represented as given in Table (4.6). Clinico-pathological factors here split the population in two with regards to MLH1 methylation. Table (4.7) shows a reciprocal relation between MLH1 methylation with colon cancer phenotype in proximal as opposed to distal locations. This correlation shows the difference in occurrence frequencies of tumour characteristics, MLH1 methylation according to two

Ref	Gene	Event	Quanti	Unit	Level	Phen	CP	Fr	No_samp
259	MLH1	Methylation	NULL	NULL	YES	C.C.	proximal	0.521	46
259	MLH1	Methylation	NULL	NULL	YES	C.C.	distal	0.381	42

Table 4.6: “Single relation” of the article (Miyakura et al., 2001). C.C. represents colon cancer.

different clinicopathological factors: proximal[15] and distal[16] location of the tumours. This example illustrates that in some cases clinico-pathological factors are associated with different routes or pathways in oncogenesis.

Another example from this article is related to high MSI tumours. MSI tumours are related to hypermethylation phenotype and have been found preferentially in proximal colon tumours. This is illustrated in Table (4.7). The article, (Miyakura et al., 2001) concludes that MLH1 promoter methylation appears 'upstream', as an early event of carcinogenesis. This impact on the MLH1 expression of normal adjacent mucosa is illustrated in Table (4.8). In addition to these summarised elements, further information from articles with respect to environmental factors has been included: for example, dietary factors show hallmark changes in colon cancer (Pufulete et al., 2003).

Ref	Gene	Event	Quant	Units	Level	Phn	Clinphat	Fr	nb_samples
259	NULL	MSI	NULL	NULL	HIGH	colon cancer	proximal	1	88

Table 4.7: "Single relation_2" from the article (Miyakura et al., 2001)

Ref	Gene	Event	Quant	Units	Level	Phn	Clinphat	Fr	nb_samples
259	MLH1	exp	NULL	NULL	YES	NAM	NULL	0.965	85

Table 4.8: "Single relation_3" from the article (Miyakura et al., 2001)

Further, an example from the article Royce et al. (2009) of "double relations" for the sporadic colon cancer phenotype is illustrated in Table (4.9). In Table (4.9) statically possible outcomes for genes SAMD4 and TGFBR2 are as follows: All samples for gene 1 (SAMD4) = 79 and 24. Possible outcome frequencies of the gene 2 (TGFBR2) are 0.228 and 0.125, i.e. theses TGFBR2 frequencies may impact on the cell phenotype.

[15] Toward the beginning, the nearer of two (or more) items, e.g. proximal portion of the intestine at right side.

[16] Anatomically located far from a point of reference, such as an origin or a point of attachment

Ref	G1	Ev1	L1	Qnt	Units	NTE1	G2	Ev2	L2	Fr	Phen	CP
PMID	SAMD4	exp	Y	NL	NL	79	TGFBR2	mut	Y	0.228	10	NL
PMID	SAMD4	exp	N	NL	NL	24	TGFBR2	mut	Y	0.125	10	NL

Table 4.9: "Double relation" of the article (Royce et al., 2009).

259	NULL	MSI	NULL	NULL	HIGH	colon cancer	proximal	1	88

Table 4.10: Format of C++ developed code

Format:

The developed C++ code converts the data in table format to My SQL format. To illustrate, consider data in the format below, which converts to

Example:

My SQL format as follows:
INSERT INTO single_rel(REF_ID_REF, EVENT_ID_EVENT, Quantified, Units, Level, PHENOTYPES, Clinpath, frequency_tumours, Nb_tumours) VALUES
(256, 115, NULL, NULL, 'HIGH', 10, 'proximal', 1, 88);

4.6 Conclusions

PathEpigen is a peer-reviewed knowledge base, initially for colon cancer epigenetics and pathways, and functions as a resource for both initial analysis and data-mining. This curated database is integrated with phenotypical and genotypical interaction networks which can be accessed computationally. All the data are extracted from non-curated resources, (which include information on colon cancer). Other benefits of PathEpigen include the provision of direct access to data required to facilitate development of models at the molecular level and the combination of highly dispersed information, for its current particular focus on simultaneous occurrence of annotated epigenetic and genetic information for colon cancer phenotypes. To support the curation process and data integration, C++ code, which converts text annotation format to SQL format, has been developed, as illustrated in Algorithm (1), and has been successfully used to convert the curated data from tabular to SQL format, speeding up the process of data entry to the system. The advantage of this approach is that the tabular annotation format contains additional data fields to guide the curation and to minimise error, while the SQL format addresses the information by its IDs in the relational scheme. Finally, the knowledge accumulated, while studying the molecular determinants of colon cancer and the corresponding genotype-phenotype relations, has provided experience enabling further contributions to the design of the querying interface of the resource.

Chapter 5

Conclusions

Great progress has been made in the exploration of epigenetic mechanisms in both normal and cancerous tissues. Epigenetic research mainly focuses on cancer, but it has provided new insights into other kinds of diseases, such as neurological, (Javierre et al., 2010) and autoimmune diseases, (Urdinguio et al., 2009). A challenging task in epigenetic research is to determine how various epigenetic entities interact and what mechanisms convey sequence specificity to the enzymes involved. Gene silencing is a critical precursor in cancer cell stimulation because it changes the dynamic interplay between *de-novo* methylation and demethylation of CpG islands. Gene expression activity depends on DNA methylation. Precise mechanisms of DNA methylation are still unknown. Some medical examples also make clear that oncogenic point mutations at CpG dinucleotides are involved in DNA hypomethylation. It seems clear that DNA methylation is directly or indirectly involved in cancer development. One of the most important questions in molecular genetics is concerned with the patterns of DNA methylation during cell division. An established model for DNA methylation explores the methylation patterns in the cell population, and the current data on DNA methylation inheritance suggests that methylation patterns are propagated, (Jones and Liang, 2009). Further questions involve the interactions of DNMTs with nucleosomes and the quantification of methylation at the initiation of replication and also immediately after nucleosome wrapping. Moreover, the placement of methylation patterns and the regulation of their specificity remains uncharacterised. In addition, unanswered questions, such as what are the roles and functions of micro-RNAs in humans in the context of epigenetics, are highlighted in Chapter (2).

As described in Chapter (2), the main goals of computational epigenetic research are to find and analyse the epigenetic states in the genome and to begin to understand epigenetic inheritance. In this regard, it is important to know the methylation patterns in both normal and cancerous tissues. Long-term studies and bioinformatic analysis will be used to predict the variation in DNA methylation mapping in the human population with large scale epigenomic data. The ability to confidently detect aberrant methylation in diseased patients will depend on the range of DNA methylation variation in healthy individuals. It will be challenging to see whether comparative epigenomics can improve our ability to identify functionally significant sites in the human genome, as is the case for comparative genomics. International epigenomic analysis projects such as the ENCODE Project and the AHEAD project are creating whole-genome, high-resolution maps of epigenetic modifications, such as human DNA methylation patterns, histone modifications and nucleosome positioning, in healthy and diseased tissues. Current epigenomic projects deal with methylation mechanisms in the context of proofreading, the repair of epigenetic information that occurs during DNA replication and how DNA methylation patterns are maintained in normal and cancer cells, (American Association for Cancer Research Human Epigenome Task Force and European Union, Network of Excellence, Scientific Advisory Board, 2008). A web address has been provided for the methylation database in Appendix (A).

Computational epigenetics projects are being developed by the Sci-Sym centre in DCU, and a number of aspects besides the assessment of DNA methylation profiles and knowledge management are being explored. In the present dissertation, we have studied epigenetic mechanisms in the context of DNA methylation dynamics by the activity of DNMTs, and we addressed unanswered questions of methylation quantification in cancer initiation as described above. The DFA model focuses, Chapter [3] on methylated frequency variations due to corresponding DNMT alterations implemented for possible low level changes, and it provides complete DFA terms for the normal and hypermethylated states. An advantage of this model is its analysis of more random data and its prediction of methylation states on the basis of DNMT activity. This study can be extended to find maintenance methylation patterns during replication. From the current results, it is clear that methylation activity varies for DNMTs, and these variations may alter methylation patterns, which affect cancer stages. In Chapter(4), we have shown data curation work and extraction for a new Knowledge Management System, PathEpigen. The database deals with genetic and epigenetic interactions and contains considerable data on corresponding molecular events. To support the curation process and data integration, which converts the text annotation format to SQL format, C++ code has been developed. Screen shots of the actual interface of the PathEpigen database, a summary of 647 new records and information on the 24 new entries in table format that summarise the genes involved are given in Appendix (B).

Bibliography

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. Mol. Cancer, 7:94.

American Association for Cancer Research Human Epigenome Task Force and European Union, Network of Excellence, Scientific Advisory Board (2008). Moving ahead with an international human epigenome project. Nature, 454(7205):711–715.

Antequera, F. (2003). Structure, function and evolution of CpG island promoters. Cellular and Molecular Life Sciences, 60:1647–1658.

Antequera, F. and Bird, A. (1993). Number of CpG islands and genes in human and mouse. Genetics, 90:11995–11999.

Artyomov, M. N., Meissner, A., and Chakraborty, A. K. (2010). A model for genetic and epigenetic regulatory networks identifies rare pathways for transcription factor induced pluripotency. PLoS Comput Biol, 6(5):e1000785+.

Baker, M. (2010). Epigenome: mapping in motion. Nature Methods, 7(3):181–186.

Ballestar, E., Paz, M. F., Valle, L., Wei, S., Fraga, M. F., Espada, J., Cigudosa, J. C., Huang, T. H., and Esteller, M. (2003). Methyl-CpG binding proteins identify novel sites of epigenetic inactivation in human cancer. EMBO J., 22:6335–6345.

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, 3:36–46.

Barski, A., Cuddapah, S., Cui, K., Roh, T. Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell, 129:823–837.

Baylin, S. B. and Herman, J. G. (2000). DNA hypermethylation in tumorigenesis: epigenetics joins genetics. Trends Genet., 16:168–174.

Beck, S. and Rakyan, V. K. (2008). The methylome: approaches for global DNA methylation profiling. Trends Genet., 24:231–237.

Bird, A. (2002). DNA methylation patterns and epigenetic memory. Genes Dev., 16:6–21.

Bird, A. (2007). Perceptions of epigenetics. Nature, 447:396–398.

Bock, C. (2008). Computational Epigenetics Bioinformatic methods for epigenome prediction, DNA methylation mapping and cancer epigenetics. PhD thesis, Max Plank Institute, Germany.

Bock, C., Paulsen, M., Tierling, S., Mikeska, T., Lengauer, T., and Walter, J. (2006). CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats and predicted DNA structure. PLoS Genetics, 2:e26.

Boyer, L. A., Lee, T. I., Cole, M. F., Johnstone, S. E., Levine, S. S., Zucker, J. P., Guenther, M. G., Kumar, R. M., Murray, H. L., Jenner, R. G., Gifford, D. K., Melton, D. A., Jaenisch, R., and Young, R. A. (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. Cell, 122:947–956.

Bradbury, J. (2003). Human Epigenome Project Up and Running. PLoS Biol., 1(3):316–319.

Bronner, C. E., Baker, S. M., Morrison, P. T., Warren, G., Smith, L. G., Lescoe, M. K., Kane, M., Earabino, C., Lipford, J., and Lindblom, A. (1994). Mutation in the DNA mismatch repair gene homologue hMLH1 is associated with hereditary non-polyposis colon cancer. Nature, 368(6468):258–261.

Brown, R. and Strathdee, G. (2002). Epigenomics and epigenetic therapy of cancer. Trends Mol. Med., 8:S43–48.

Chen, T., Ueda, Y., Dodge, J. E., Wang, Z., and Li, E. (2003). Establishment and maintenance of genomic methylation patterns in mouse embryonic stem cells by DNMT3A and DNMTB. Mol. Cell. Biol., 23:5594–5605.

Cho, E., Smith-Warner, S., Ritz, J., van den Brandt, P., Colditz, G., Folsom, A., Freudenheim, J., Giovannucci, E., Goldbohm, R., Graham, S., Holmberg, L., Kim, D., Malila, N., Miller, A., Pietinen, P., Rohan, T., Sellers, T., Speizer, F., Willett, W., Wolk, A., and Hunter, D. (2004). Alcohol intake and colorectal cancer: a pooled analysis of 8 cohort studies. Ann Intern Med., 140(8):I55.

Christensen, B. C., Houseman, E. A., Marsit, C. J., Zheng, S., Wrensch, M. R., Wiemels, J. L., Nelson, H. H., Karagas, M. R., Padbury, J. F., Bueno, R., Sugarbaker, D. J., Yeh, R.-F. F., Wiencke, J. K., and Kelsey, K. T. (2009). Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context. PLoS genetics, 5(8):e1000602+.

Clark, S. J., Harrison, J., and Frommer, M. (1995). CpNpG methylation in mammalian cells. Nat. Genet., 10:20–27.

Clark, S. J. and Melki, J. (2002). DNA methylation and gene silencing in cancer: which is the guilty party? Oncogene, 21:5380–5387.

Consortium, T. E. P. (2004). The encode (encyclopedia of dna elements) project. Science, 306(5696):636–640.

Cooney, C. A., Dave, A. A., and Wolff, G. L. (2002). Maternal methyl supplements in mice affect epigenetic variation and DNA methylation of offspring. J. Nutr., 132:2393S–2400S.

Costa, S. and Shaw, P. (2007). 'Open minded' cells: how cells can change fate. Trends Cell Biol., 17:101–106.

Costello, J. and Plass, C. (2001). Methylation matters. J. Med. Genet., 38:285–303.

Cowley, D. and Atchley, W. (1992). Quantitative Genetic Models for Development Epigenetic Selection and Phenotypic Evolution. Evolution, 46(2):495–518.

Das, P. M. and Singal, R. (2004). DNA methylation and cancer. J Clin. Oncol., 22(22):4632–4642.

Das, R., Dimitrova, N., Xuan, Z., Rollins, R., Haghighi, F., Edwards, J., Ju, J., Bestor, T., and Zhang, M. (2006). Computational prediction of methylation status in human genomic sequences. PNAS, 103(28):10713–10716.

Delcuve, G. P., Rastegar, M., and Davie, J. R. (2009). Epigenetic control. J. Cell. Physiol., 219:243–250.

Demin, O. and Goryanin, I. (2008). Kinetic Modelling in Systems Biology. Chapman and Hall/CRC Mathematical and Computational Biology.

Deng, G., Kakar, S., Tanaka, H., Matsuzaki, K., Miura, S., Sleisenger, M., and Kim, Y. (2008). Proximal and distal colorectal cancers show distinct gene-specific methylation profiles and clinical and molecular characteristics. Eur. J. Cancer, 15(12):1290–301.

Derks, S., Postma, C., Moerkerk, P. T., van den Bosch, S. M., Carvalho, B., Hermsen, M. A., Giaretti, W., Herman, J. G., Weijenberg, M. P., Bruïne, A. P., Meijer, G. A., and van Engeland, M. (2006). Promoter methylation precedes chromosomal alterations in colorectal cancer development. Cell Oncol., 28(5-6):247–257.

Dodd, I. B., Micheelsen, M. A., Sneppen, K., and Thon, G. (2007). Theoretical analysis of epigenetic cell memory by nucleosome modification. Cell, 129:813–822.

Dodge, J. E., Ramsahoye, B. H., Wo, Z., Okano, M., and Li, E. (2002). De novo methylation of MMLV provirus in embryonic stem cells: CpGversus non-cpg methylation. Gene, 289(1):41–48.

Doerfler, W., Bohm, P., Sneppen, K., and Thon, G. (2006). DNA Methylation Basic Mechanism. Springer, Verleg Berlin Heidenelberg.

Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V. K., Attwood, J., Burger, M., Burton, J., Cox, T. V., Davies, R., Down, T. A., Haefliger, C., Horton, R., Howe, K., Jackson, D. K., Kunde, J., Koenig, C., Liddle, J., Niblett, D., Otto, T., Pettett, R., Seemann, S., Thompson, C., West, T., Rogers, J., Olek, A., Berlin, K., and Beck, S. (2006). DNA methylation profiling of human chromosomes 6, 20 and 22. Nat. Genet., 38:1378–1385.

Esteller, M. (2007). Cancer epigenomics: DNA methylomes and histone-modification maps. Nat. Rev. Genet., 8:286–298.

Esteller, M., Corn, P. G., Baylin, S. B., and Herman, J. G. (2001). A gene hypermethylation profile of human cancer. Cancer research, 61(8):3225–3229.

Esteller, M., Hamilton, S., Burger, P., Baylin, S., and Herman, J. (1999). Inactivation of the DNA Repair Gene O6-Methylguanine-DNA Methyltransferase by Promoter Hypermethylation is a Common Event in Primary Human Neoplasia. Cancer Research, 59:793–797.

Esteller, M., Toyota, M., Sanchez-Cespedes, M., Capella, G., Peinado, M. A., Watkins, N. D., Issa, J.-P. J., Sidransky, D., Baylin, S. B., and Herman, J. G. (2000). Inactivation of the DNA Repair Gene O6-Methylguanine-DNA Methyltransferase by Promoter Hypermethylation Is Associated with G to A Mutations in K-ras in Colorectal Tumorigenesis. Cancer Research, 60(9):2368–2371.

Fatemi, M. and Wade, P. A. (2006). MBD family proteins: reading the epigenetic code. J Cell Sci, 119(15):3033–3037.

Feinberg, A. P. (2007). Phenotypic plasticity and the epigenetics of human disease. Nature, 447:433–440.

Feinberg, A. P. and Tycko, B. (2004). The history of cancer epigenetics. Nat. Rev. Cancer, 4:143–153.

Fodde, R. (2002). The APC gene in colorectal cancer. European Journal of Cancer, 38(7):867–871.

Fraga, M. F. and Esteller, M. (2007). Epigenetics and aging: the targets and the marks. Trends Genet., 23:413–418.

Fujita, N., Shimotake, N., Ohki, I., Chiba, T., Saya, H., Shirakawa, M., and Nakao, M. (2000). Mechanism of transcriptional regulation by methyl-CpG binding protein MBD1. Mol. Cell. Biol., 20:5107–5118.

Fujita, N., Watanabe, S., Ichimura, T., Tsuruzoe, S., Shinkai, Y., Tachibana, M., Chiba, T., and Nakao, M. (2003). Methyl-CpG binding domain 1 (MBD1) interacts with the Suv39h1-HP1 heterochromatic complex for DNA methylation-based transcriptional repression. J. Biol. Chem., 278:24132–24138.

Genereux, D. P., Miner, B. E., Bergstrom, C., and Laird, C. D. (2005). A population epigenetic model to infer site specific methylation rates from double stranded dna methylation patterns. PNAS, 102(16):5802–5807.

Grnbek, K., Hother, C., and Jones, P. (2007). Epigenetic changes in cancer. Journal Compilation, 115:1039–1059.

Gryfe, R., Swallow, C., Bapat, B., Redston, M., Gallinger, S., and Couture, J. (1997). Molecular biology of colorectal cancer. Curr. Probl. Cancer, 21(5):233–300.

Haslberger, A., Varga, F., and Karlic, H. (2006). Recursive causality in evolution: A model for epigenetic mechanisms in cancer development. Medical Hypotheses, 67(6):1448 – 1454.

Hellebrekers, D., Lentjes, M., van den Bosch, S., Melotte, V., Wouters, K., Daenen, K., Smits, K., Akiyama, Y., Yuasa, Y., Sanduleanu, S., Khalid-de Bakker, C., Jonkers, D., Weijenberg, M., Louwagie, J., van Criekinge, W., Carvalho, B., Meijer, G., Baylin, S., Herman, J., de Brune, A., and van Engeland, M. (2009). GATA4 and GATA5 are potential tumour suppressors and biomarkers in colorectal cancer. Clin. Cancer Res., 15(12):3990–7.

Hendrich, B. and Bird, A. (1998). Identification and characterization of a family of mammalian methyl-CpG binding proteins. Mol. Cell. Biol., 18:6538–6547.

Hendrich, B., Hardeland, U., Ng, H. H., Jiricny, J., and Bird, A. (1999). The thymine glycosylase MBD4 can bind to the product of deamination at methylated CpG sites. Nature, 401:301–304.

Herman, J. G. and Baylin, S. B. (2003). Gene silencing in cancer in association with promoter hypermethylation. N. Engl. J. Med., 349:2042–2054.

Herman, J. G., Graff, J. R., Myhnen, S., Nelkin, B. D., and Baylin, S. B. (1996). Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands. Proc. Natl. Acad. Sci. U.S.A., 93:9821–9826.

Hesketh, C. (2009). Importance of DNA methylation. A Scrap of Notebook Paper, Website. http://www.epidna.com/content.php?contentid=28.

Hibi, K., Kodera, Y., Ito, K., Akiyama, S., and Nakao, A. (2005). Aberrant methylation of HLTF, SOCS-1, and CDH13 genes is shown in colorectal cancers without lymph node metastasis. Dis. Colon Rectum, 48(6):1282–6.

Houseman, E., Christensen, B., Yeh, R., Marsit, C., Karagas, M., Nelson, H., Wiemels, J., Zheng, S., Wiencke, J., and Kelsey, K. (2008). Model-based clustering of DNA methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions. BMC bioinformatics, 9:365.

Houshdaran, S., Hawley, S., Palmer, C., Campan, M., and Olsen, M. (2010). DNA Methylation Profiles of Ovarian Epithelial Carcinoma Tumors and Cell Lines. PLoS, 5(2):e9359.

Howard, J., Frolov, A., Tzeng, C., Stewart, A., Midzak, A., Majmundar, A., Godwin, A., Heslin, M., Bellacosa, A., and JP., A. (2009). Epigenetic downregulation of the DNA repair gene MED1/MBD4 in colorectal and ovarian cancer. Cancer Biol. Ther., 8(1):94–100.

Illingworth, R., Kerr, A., Desousa, D., Jrgensen, H., Ellis, P., Stalker, J., Jackson, D., Clee, C., Plumb, R., Rogers, J., Humphray, S., Cox, T., Langford, C., and Bird, A. (2008). A novel CpG island set identifies tissue-specific methylation at developmental gene loci. PLoS Biol., 6(1):e22.

Jacob, S. and Praz, F. (2002). DNA mismatch repair defects: role in colorectal carcinogenesis. Biochimie, 84(1):27–47.

Jair, K., Bachman, K., Suzuki, H., Ting, A., Rhee, I., Yen, R., Baylin, S., and Schuebel, K. (2006). De novo CpG Island Methylation in Human Cancer Cells. Cancer Res., 66(2):628 – 692.

Javierre, B. M., Fernandez, A. F., Richter, J., Al-Shahrour, F., Martin-Subero, J. I., Rodriguez-Ubreva, J., Berdasco, M., Fraga, M. F., O'Hanlon, T. P., Rider, L. G., Jacinto, F. V., Lopez-Longo, F. J., Dopazo, J., Forn, M., Peinado, M. A., Carreño, L., Sawalha, A. H., Harley, J. B., Siebert, R., Esteller, M., Miller, F. W., and Ballestar, E. (2010). Changes in the pattern of DNA methylation associate with twin discordance in systemic lupus erythematosus. Genome Research, 20(2):170–179.

Jiang, Y. H., Bressler, J., and Beaudet, A. L. (2004). Epigenetics and human disease. Annu Rev Genomics Hum Genet, 5:479–510.

Jones, P. A. (2008). Moving AHEAD with an international human epigenome project. Nature, 454:711–715.

Jones, P. A. and Baylin, S. B. (2007). The epigenomics of cancer. Cell, 128:683–692.

Jones, P. A. and Liang, G. (2009). Rethinking how DNA methylation patterns are maintained. Nature Reviews Genetics, doi:10.1038/nrg2651:1–7.

Jones, P. A. and Martienssen, R. (2005). A blueprint for a Human Epigenome Project: the AACR Human Epigenome Workshop. Cancer Res., 65:11241–11246.

Jones, P. L., Veenstra, G. J., Wade, P. A., Vermaak, D., Kass, S. U., Landsberger, N., Strouboulis, J., and Wolffe, A. P. (1998). Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. Nat. Genet., 19:187–191.

Kanai, Y., Ushijima, S., Kondo, Y., Nakanishi, Y., and Hirohashi, S. (2001). DNA methyltransferase expression and DNA methylation of CPG islands and peri-centromeric satellite regions in human colorectal and stomach cancers. Int. J. Cancer, 91(2):205–12.

Kim, H. C., Kim, J. C., Roh, S. A., Yu, C. S., Yook, J. H., Oh, S. T., Kim, B. S., Park, K. C., and Chang, R. (2005). Aberrant CpG island methylation in early-onset sporadic gastric carcinoma. J. Cancer Res. Clin. Oncol., 131:733–740.

Kim, M., Lee, J., and Sidransky, D. (2010). DNA methylation markers in colorectal cancer. Cancer and Metastasis Reviews, 29(1):181–206.

Kinzler, K. W. and Vogelstein, B. (1996). Lessons from hereditary colorectal cancer. Cell, 87(2):159–170.

Klisovic, R. B., Stock, W., Cataland, S., Klisovic, M. I., Liu, L., Blum, W., Green, M., Odenike, O., Godley, L., Burgt, J., Laar, E. V., Cullen, M., Macleod, A., Besterman, J., Reid, G., Byrd, J., and G., M. (2008). A phase I biological study of MG98, an oligodeoxynucleotide antisense to dnamethyltransferase 1, in patients with high-riskmyelodysplasia and acutemyeloid leukemia. Clinical Cancer Research, 14(8):2444–2448.

Krogh, A., Brown, M., Mian, I. S., Sjölander, K., and Haussler, D. (1994). Hidden markov models in computational biology. applications to protein modeling. Journal of molecular biology, 235(5):1501–1531.

Laird, P. (1996). The role of DNA methylation in cancer genetics and epigenetic. Annu. Rev. Genet., 30:441–64.

Laird, P. W. (2005). Cancer epigenetics. Human Mol. Genet., 1:65–76.

Lea, I. A., Jackson, M. A., and Dunnick, J. K. (2009). Genetic pathways to colorectal cancer. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis, 670(1-2):96–98.

Lee, B. B. B., Lee, E. J. J., Jung, E. H. H., Chun, H.-K. K., Chang, D. K. K., Song, S. Y. Y., Park, J., and Kim, D.-H. H. (2009). Aberrant methylation of APC, MGMT, RASSF2A, and Wif-1 genes in plasma as a biomarker for early detection of colorectal cancer. Clinical cancer research, 15(19):6185–6191.

Lengauer, C., Kinzler, K. W., and Vogelstein, B. (1997). Genetic instability in colorectal cancers. Nature, 386(6625):623–627.

Leung, S., Yuen, S., Chung, L., Chu, K., Chan, A., and Ho, J. (1999). hMLH1 promoter methylation and lack of hMLH1 expression in sporadic gastric carcinomas with high-frequency microsatellite instability. Cancer Res., 59(1):159–64.

Lind, G. E., Thorstensen, L., Løvig, T., Meling, G. I., Hamelin, R., Rognum, T. O., Esteller, M., and Lothe, R. A. (2004). A CpG island hypermethylation profile of primary colorectal carcinomas and colon cancer cell lines. Molecular cancer, 3:28.

Lujambio, A., Ropero, S., Ballestar, E., Fraga, M. F., Cerrato, C., Setin, F., Casado, S., Suarez-Gauthier, A., Sanchez-Cespedes, M., Git, A., Gitt, A., Spiteri, I., Das, P. P., Caldas, C., Miska, E., and Esteller, M. (2007). Genetic unmasking of an epigenetically silenced microRNA in human cancer cells. Cancer Res., 67:1424–1429.

Marte, B., Eccleston, A., and Nath, D. (2008). Molecular cancer diagnostics. Nature, 452(7187):547–589.

Martin, C. and Zhang, Y. (2007). Mechanisms of epigenetic inheritance. Current Opinion Cell Biology, 19(3):266–72.

Mikkelsen, T., Ku, M., Jaffe, D. B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T. K., Koche, R. P., Lee, L., Mendenhall, E., O'Donovan, A., Presser, A., Russ, C., Xie, X., Meissner, A., Wernig, M., Jaenisch, R. N. C., Lander, E. S., and Bernstein, B. E. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cell. Nature, 448:553–560.

Miller, C. A. and Sweatt, J. D. (2007). Covalent modification of DNA regulates memory formation. Neuron, 53:857–869.

Miyaki, M., Konishi, M., Kikuchi-Yanoshita, R., Enomoto, M., Igari, T., Tanaka, K., Muraoka, M., Takahashi, H., Amada, Y., Fukayama, M., Maeda, Y., Iwama, T., Mishima, Y., Mori, T., and Koike, M. (1994). Characteristics of Somatic Mutation of the Adenomatous Polyposis Coli Gene in Colorectal Tumors. Cancer Res, 54(11):3011–3020.

Miyakura, Y., Sugano, K., Konishi, F., Ichikawa, A., Maekawa, M., Shitoh, K., Igarashi, S., Kotake, K., Koyama, Y., and Nagai, H. (2001). Extensive methylation of hMLH1 promoter region predominates in proximal colon cancer with microsatellite instability. Gastroenterology, 121:1300–1309.

Miyamoto, K. and Ushijima, T. (2005). Diagnostic and therapeutic applications of epigenetics. Jpn. J. Clin. Oncol., 35:293–301.

Momparler, R. (2003). Cancer epigenetics. Oncogene, 22:6479–6483.

Ng, H. H., Jeppesen, P., and Bird, A. (2000). Active repression of methylated genes by the chromosomal protein MBD1. Mol. Cell. Biol., 20:1394–1406.

Nosho, K., Shima, K., Kure, S., Irahara, N., Baba, Y., Chen, L., Kirkner, G., Fuchs, C., and Ogino, S. (2009). JC virus T-antigen in colorectal cancer is associated with p53 expression and chromosomal instability, independent of CpG island methylator phenotype. Neoplasia, 11(1):1290–301.

Ohtani-Fujita, N., Fujita, T., Aoike, A., Osifchin, N., Robbins, P., and Sakai, T. (1993). CpG methylation inactivates the promoter activity of the human retinoblastoma tumor-suppressor gene. Oncogene, 8(4):1063–7.

Okuwaki, M. and Verreault, A. (2004). Maintenance DNA methylation of nucleosome core particles. The journal of biological chemistry, 279(4):2904–2912.

Ongenaert, M., Neste, L., Meyer, T., Menschaert, G., Bekaert, B., and Criekinge, W. (2007). Pubmeth: a cancer methylation database combining text-mining and expert annotation. Nucleic Acids Research, 36:1–5.

Ordway, J. and Curran, T. (2002). Methylation matters: modeling a manageable genome. Cell Growth Differ, 13(4):149–62.

Ordway, J., Williams, K., and Curran, T. (2004). Transcription repression in oncogenic transformation: common targets of epigenetic repression in cells transformed by FOS, RAS or DNMT1. Oncogene, 23:3737–48.

Pakneshan, P., Szyf, M., and Rabbani, S. A. (2005). Methylation and inhibition of expression of uPA by the RAS oncogene: divergence of growth control and invasion in breast cancer cells. Carcinogenesis, 26:557–564.

Pakneshan, P., Ttu, B., and Rabbani, S. A. (2004). Demethylation of urokinase promoter as a prognostic marker in patients with breast carcinoma. Clin. Cancer Res., 10:3035–3041.

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

Perrin, D., Ruskin, H. J., Crane, M., and Walshe, R. (2008). Epigenetic modelling. ERCIM News, 72:46.

Perucho, M. (2000). Genetic and Epigenetic Modification of MLH1. Am J Pathol., 157(3):1052–1053.

Petrossian, T. and Clarke, S. (2009). Bioinformatic Identification of Novel Methyltransferases. Epigenomics, 1(1):163–175.

Piepoli, A., Cotugno, R., Merla, G., Gentile, A., Augello, B., Quitadamo, M., Merla, A., Panza, A., Carella, M. Maglietta, R., D'Addabbo, A., Ancona, N., Fusilli, S., Perri, F., and Andriulli, A. (2009). Promoter methylation correlates with reduced NDRG2 expression in advanced colon tumour. BMC Med. Genomics, 3(2):11.

Portela, A. and Esteller, M. (2010). Epigenetic modifications and human disease. Nat Biotech, 28(10):1057–1068.

Pufulete, M., Emery, P. W., and Sanders, T. A. (2003). Folate, DNA methylation and colo-rectal cancer. Proc. Nutr. Soc., 62:437–445.

Raghavan, K. and Ruskin, H. J. (2011). Computational epigenetic micromodel for cancer prediction. In Private communication, ECCS.

Raghavan, K., Ruskin, H. J., Perrin, D., Goasmat, F., and Burns, J. (2010). Computational Micromodel for Epigenetic Mechanisms. PLoS Comput Biol, 5(11):e14031.

Rakyan, V. K., Hildmann, T., Novik, K. L., Lewin, J., Tost, J., Cox, A. V., Andrews, T. D., Howe, K. L., Otto, T., Olek, A., Fischer, J., Gut, I. G., Berlin, K., and Beck, S. (2004). DNA methylation profiling of the human major histocompatibility complex: a pilot study for the human epigenome project. PLoS Biol., 2:e405.

Ramreza, N., Bandresa, E.and Navarrob, A., Ponsb, A., Jansab, S., Morenoc, I., Martnez-Rodenas, F., Zaratea, R., Bitartea, N., Monzob, M., and Garca-Foncillasa, J. (2008). Epigenetic events in normal colonic mucosa surrounding colorectal cancer lesions. European Journal of Cancer, 44:2689–2695.

Rando, O. J. and Verstrepen, K. J. (2007). Timescales of genetic and epigenetic inheritance. Cell, 128:655–668.

Razin, A. and Cedar, H. (1991). DNA methylation and gene expression. Microbiological, 55(2):451–458.

Reik, W. (2007). Stability and flexibility of epigenetic gene regulation in mammalian development. Nature, 447:425–432.

Rice, J. and Allis, C. (2001). Histone methylation versus histone acetylation: new insights into epigenetic regulation. Current opinion in cell biology, 13:263–273.

Richards, E. J. (2006). Inherited epigenetic variation–revisiting soft inheritance. Nat. Rev. Genet., 7:395–401.

Riley, M. R., Buettner, H. M., Muzzio, F. J., and Reyes, S. C. (1995). Monte carlo simulation of diffusion and reaction in two-dimensional cell structures. Biophysical journal, 68(5):1716–1726.

Ro, S. and Rannala, B. (2001). Methylation patterns and mathematical models reveal dynamics of stem cell turnover in the human colon. Proc Natl Acad Sci U S A., 98(19):10519–10521.

Roberts, R. J., Vincze, T., Posfai, J., and Macelis, D. (2005). REBASE–restriction enzymes and DNA methyltransferases. Nucleic Acids Res., 33:D230–2.

Robertson, K. D. and Jones, P. A. (2000). DNA methylation: past, present and future directions. Carcinogenesis, 21:461–467.

Robertson, K. D., Uzvolgy, E., Liang, G., Talmadge, C., Sumegi, J., A., G. F., and Jones, P. A. (1999). The human DNA methyltransfearse(DNMTs) 1, 3a and 3b: coordinate mRNA expression in normal tissues and overexpression in tumors. Nuclic Acid Research, 27(11):2291–2298.

Rodenhiser, D. and Mann, M. (2006). Epigenetics and human disease: translating basic biology into clinical applications. CMAJ, 174:341–348.

Royce, S., Alsop, K., Haydon, A., Mead, L., Smith, L., Tesoriero, A., Giles, G., Jenkins, M., Hopper, J., and Southey, M. (2009). The role of SMAD4 in early onset colorectal cancer. Colorectal Dis., 12(3):213–9.

Ruskin, H. J., Barat, A., and Sarbu, L. (2008). Novel Database for Genetic and Epigenetic Mechanisms in Colon Cancer. ERCIM News, 73.

Ruskin, H. J. and Perrin, D. (2011). Multi-scale modelling of epigenetic mechansisms. In Private communication, Transactions in Biomedical Engineering Special issue, IEEE.

Sagal, E., Tasker, B., Gasch, A., and Friedman, N. (2001). Rich probabilistic model for gene expression. Bioinformatics, 7:S247–S252.

Segditsas, S., Sieber, O., Rowan, A., Setien, F., Neale, K., Phillips, R., Ward, R., Esteller, M., and Tomlinson, I. (2008). Promoter hypermethylation leads to decreased APCmRNA expression in familial polyposis and sporadic colorectal tumours, but does not substitute for truncating mutations. Exp. Mol. Pathol., 85(3):201–6.

Segditsas, S. and Tomlinson (2006). Colorectal cancer and genetic alterations in the Wnt pathway. Oncogene, 25(57):7531–7537.

Seshimo, I., Yamamoto, H., Mishima, H., Kurata, A., Suzuki, R., Ezumi, K., Takemasa, I., Ikeda, M., Fukushima, T., and Tsujinaka, T. (2006). Expression and mutation of smad4 in poorly differentiated carcinoma and signet-ring cell carcinoma of the colorectum. Journal of experimental and clinical cancer research, 25(3):435–444.

Shen, L., Kondo, Y., Rosner, G. L., Xiao, L., Hernandez, N. S., Vilaythong, J., Houlihan, S. P., Krouse, R. S., Prasad, A. R., Einspahr, J. G., Buckmeier, J., Alberts, D. S., Hamilton, S. R., and Issa, J.-P. J. (2005). MGMT Promoter Methylation and Field Defect in Sporadic Colorectal Cancer. Journal of the National Cancer Institute, 97(18):1330–1338.

Shima, K., Morikawa, T., Baba, Y., Nosho, K., Suzuki, M., Yamauchi, M., Hayashi, M., Giovannucci, E., Fuchs, C., and Ogino, S. (2010). Promoter methylation, loss of expression and prognosis in 855 colorectal cancers. Cancer Causes and Control, 22(2):301–9.

Siegmund, K. D., Marjoram, P., and Shibata, D. (2008). Modeling DNA methylation in a population of cancer cells. Statistical applications in genetics and molecular biology, 7(1).

Singal, R. and Ginder, G. (1999). DNA methylation. Blood, 93:4059–4070.

Sipser, M. (2008). Introduction to the Theory of Computation. PWS publishing company, Boston MA02116.

Soejima, H., Zhao, W., and Mukai, T. (2005). Epigenetic silencing of the MGMT gene in cancer. Biochem. Cell Biol., 83:429–437.

Steele, W., Allegrucci, C., Singh, R., Lucas, E., Priddle, H., Denning, C., Sinclair, K., and Young, L. (2005). Human embryonic stem cell methyl cycle enzyme expression: modelling epigenetic programming in assisted reproduction? Reprod. Biomed. Online, 10:755–766.

Stewart, D., Donehower, R., Eisenhauer, E., Wainman, N., Shah, A., Bonfils, C.and MacLeod, A., Besterman, J., and Reid, G. (2005). A phase i pharmacokinetic and pharmacodynamic study of the DNA methyltransferase 1 inhibitor MG98 administered twice weekly. Ann. Oncol., 14:766–774.

Suzuki, H., Watkins, N. D., Jair, K.-W., Schuebel, K. E., Markowitz, S. D., Dong, Pretlow, T. P., Yang, B., Akiyama, Y., van Engeland, M., Toyota, M., Tokino, T., Hinoda, Y., Imai, K., Herman, J. G., and Baylin, S. B. (2004). Epigenetic inactivation of SFRP genes allows constitutive WNT signaling in colorectal cancer. Nat Genet, 36(4):417–422.

Suzuki, M. M. and Bird, A. (2008). DNA methylation landscapes: provocative insights from epigenomics. Nat. Rev. Genet., 9:465–476.

Szallasi, Z., Stelling, J., and Periwal, V. (2010). System Modeling in Cellular Biology. The MIT Press.

Szyf, M., Pakneshan, P., and Rabbani, S. (2004). DNA methylation and breast cancer. Biochem. Pharmacol., 68:1187–1197.

Tajima, S. and Suetake, I. (1998). Regulation and function of DNA methylation in vertebrates. J. Biochem., 123:993–999.

Takeshima, H., Suetake, I., Shimahara, H., Ura, K., Tate, T., and Tajima, S. (2006). Distinct DNA methylation activity of Dnmt3a and Dnmt3b towards naked and nucleosomal DNA. J. Biochem., 139:503–515.

Teng, I., Hou, P., Lee, K., Chu, P., Yeh, K., Jin, V., Tseng, M., Tsai, S., Chang, Y. D., Wu, C., Sun, H., Tsai, K., Jeng, L., Nephew, K., Huang, T., Hsiao, S., and Leu, Y. (2011). Targeted methylation of two tumor suppressor genes is sufficient to transform mesenchymal stem cells into cancer stem/initiating cells. Cancer Res., Epub ahead of print:402–415.

Thornton, J. (2009). Data curation in biology - past, present and future. http://precedings.nature.com/documents/3225/version/1.

Tian, X., Zhang, Y., Sun, D., Zhao, S., Xiong, H., and Fang, J. (2009). Epigenetic silencing of LRRC3B in colorectal cancer. Scand J. Gastroenterol, 44(1):79–84.

Tucker, K. L. (2001). Methylated cytosine and the brain: a new base for neuroscience. Neuron, 30:649–652.

Tuckerman, M. E., Ungar, P. J., von Rosenvinge, T., and Klein, M. L. (1996). Ab initio molecular dynamics simulations. The Journal of Physical Chemistry, 100(31):12878–12887.

Urdinguio, R. G., Sanchez-Mut, J. V., and Esteller, M. (2009). Epigenetic mechanisms in neurological diseases: genes, syndromes, and therapies. Lancet neurology, 8(11):1056–1072.

Waddington, C. (1953). Genetic assimilation of an acquired character. Evolution Int. J. Org. Evolution, 7:118–126.

Waterland, R. A. and Jirtle, R. L. (2003). Transposable elements: targets for early nutritional effects on epigenetic gene regulation. Mol. Cell. Biol., 23:5293–5300.

Wilkinson, D. J. (2006). Stochastic Modelling for Systems Biology (Chapman & Hall/CRC Mathematical & Computational Biology). Chapman and Hall/CRC, 1 edition.

Winnenburg, R., Wchter, T., Plake, C., Doms, A., and Schroeder, M. (2008). Facts from text: can text mining help to scale-up high-quality manual curation of gene products with ontologies? Brief Bioinformatics, 9:466–478.

Wong, J. J. L. J., Hawkins, N. J. J., and Ward, R. L. L. (2006). Colorectal cancer - a model for epigenetic tumorigenesis. Gut, 56(1):140–148.

Workman, J. and Kingston, R. (1998). Alterations of nucleosome structure as a mechanisms of transcrptional regulation. Annu. Rev. Biochemistry, 67:545–79.

Xie, H., Wang, M., de Andrade, A., Bonaldo, M., Galat, V., Arndt, K., Rajaram, V., Goldman, S., Tomita, T., and Soares, M. B. (2011). Genome-wide quantitative assessment of variation in DNA methylation patterns. Nucleic Acids Research.

Xue, C., Yu, J., Liu, H., and Chen, J. (2009). Epigenetics and cancer. Journal of Biomedical Engineering, 26:1133–1136.

Yasukochia, Y., Maruyamab, O., Mahajana, M., Paddenc, C., Euskirchend, G., Schulze, V., Hirakawaf, H., Kuharag, S., Pana, X.H.and Newburgerc, P., Snyderd, S., and Weissmana, S. (2010). Xchromosome-wide analyses of genomic DNA methylation states and gene expression in male and female neutrophils. PNAS, 107(8):3704–3709.

Yi, S. and Goodisman, M. (2009). Computational approaches for understanding the evolution of DNA methylation in animals. Epigenetics, 13(4):551–556.

Yoo, C. and Jones, P. (2004). DNA methyltransferase inhibitors in cancer therapy. Am. Assoc. Cancer Res. Educ. Book, 2:333–337.

Yoon, B. (2009). Hidden markov models and their applications in biological sequence analysis. Current Genomics, 10(6):402–415.

Yu, J., Tao, Q., Cheng, Y., Lee, K., Ng, S., Cheung, K., Tian, L., Rha, S., Neumann, U., Rcken, C., Ebert, M., Chan, F., and Sung, J. (2009). Promoter methylation of the wnt/beta-catenin signaling antagonist Dkk-3 is associated with poor survival in gastric cancer. Cancer, 115(1):49–60.

Appendix A

Web address for Computational Epigenetics Sources

A.1 Web addresses

Web addresses of the Table (2.1) for the Chapter (2):

1. Chip: http://www.chiponchip.org/.
2. NIH: http://www.nhlbi.nih.gov.
3. The broad institute of MIT and Harvard: http://www.broadinstitute.org.
4. AACR Human Epigenome Task: http://www.aacr.org/default.aspx.
5. KFEL: http://www.scienceblog.com/cms/company/krembil-family-epigenetics-laboratory.
6. MeInfoText: mit.lifescience.ntu.edu.tw.
7. PubMeth: http://www.pubmeth.org.
8. MethPrimer: medgen.ugent.be/methprimerdb.
9. The Histone Database: http://genome.nhgri.nih.gov/histones.

10. ChromDB:http://www.chromdb.org/.
11. CREMOFAC ‘CREMOFAC’ has included 720 non-redundant chromatin-remodeling factor sequences and it is available at http://www.jncasr.ac.in/cremofac/.
12. MethyLogiX: MethyLogiX DNA methylation database’ has male germ cells and late-onset Alzheimer’s disease. http://www.methylogix.com/genetics/database.shtml.htm.
13. PathEpigen: http://statepigen.sci-sym.dcu.ie/.
14. REBAS [New England Biolabs] (Roberts et al., 2005): http://rebase.neb.com/rebase.
15. ENCODE: http://www.genome.gov/10005107.
16. HEROIC: http://www.heroic-ip.eu.
17. EpiGRAPH: http://epigraph.mpi-inf.mpg.de/WebGRAPH/.
18. BiQ Analyzer software: http://biq-analyzer.bioinf.mpi-sb.mpg.de/.
19. Human Epigenome project: http://www.epigenome.org.
20. Epigenie: http://www.epigenie.com/Home.html.

Appendix B

A Summary of New Database Records

B.1 Objective

The main aim of the PathEpigen database is to become a biomedical resource for genetic and epigenetic molecular events of various early-stage colon cancer phenotypes. It is expected to facilitate future modelling and quantitative research in the cancer field in the area of early diagnosis. One aim for the ongoing data curation for PathEpigen is to provide more statistically sound relationships between genetic and epigenetic events, e.g., the probabilities of gene expression and methylation states for a given phenotype. Also, data submission is improved in the context of inserting new data. The resource provides two ways of analysing molecular event data:

1. *"Single relation"*, i.e, Event_Freq = Probability (Event — Phenotype).
2. *"Double relation"*, i.e., Event_2_Freq = Probability (Event_1 — Event_2 and Phenotype).

B.2 New Records for the PathEpigen

As described above in chapter (4), it contains 6,445 records on colon cancer, and my contribution adds 647 records from the manual curation of 50 scientific reports, which includes 430 and 212 “single” and “double” relations, respectively. It includes 24 new entries and describes genetic and epigenetic events for 21 genes [as illustrated in Table (B.1)] and 3 phenotypes, including gastric CRC carcinoma subhistology (undifferentiated) and CRC Crohn’s disease. The aim of choosing these studies was to analyse new molecular events for colon cancer. Here, we summarise 24 new entries out of 647 in the database, as illustrated in Table (B.1 and B.2). These gene are influenced by both genetic and epigenetic events and we describe some specific events of “ single and “double relation” type as illustrated in Table (B.2) and (B.3) respectively. It is interesting to see the “single” and “double” relations of new gene entries, such as DKK3,

Gene	Event	References
LRRC3B	Expression	Tian et al. (2009)
LRRC3B	Methylation	Tian et al. (2009)
NR3C1	Methylation	Ahlquist et al. (2008)
DKK3	Expression	Yu et al. (2009)
DKK3	Promoter methylation	Yu et al. (2009)
MED1	Expression	Howard et al. (2009)
TGFBR2	Frmashift mutation	Royce et al. (2009)
MINT2	Methylation	Kanai et al. (2001)
APC	Mono-allelic mutation	Segditsas et al. (2008)
SMAD4	Expression	Royce et al. (2009)
SOCS-I	Methylation	Hibi et al. (2005)
HOXA9	Methylation	Ahlquist et al. (2008)
SCGB3AI	Methylation	Ahlquist et al. (2008)
GATA4	Methylation	Hellebrekers et al. (2009)
GATA5	methylation	Hellebrekers et al. (2009)
HLTF	Methylation	Hibi et al. (2005)
RUNX3	Methylation	Ahlquist et al. (2008)
PIK3CG	Expression	Nosho et al. (2009)
SST	Methylation	Deng et al. (2008)
COX-2	Expression	Nosho et al. (2009)
HPGD	Methylation	Piepoli et al. (2009)

Table B.1: New entries in Genetic and Epigenetic Events

Gene	Event	Quantified	Unit	NbSmp	Level	Freq	Phe	CnecalPath	cell line
NDRG2	Methylation	NULL	NULL	8	YES	0.500	CC	NULL	NULL
NDRG2	Expression	ABNORMAL	1.6(up-reg)	1	YES	NULL	NULL	NULL	HCT116
PIK3CA	Mut	NULL	NULL	666	YES	0.153	CC	NULL	NULL
SMAD4	Expression	NULL	NULL	109	NO	0.248	CC	NULL	NULL
MED1	Expression	NULL	NULL	39	YES	1	CC	NULL	$31-40$
MED1	Expression	NULL	NULL	39	YES	1	CC	NULL	$41-50$
DKK-3	Expression	NULL	NULL	9	NO	0.667	CC	NULL	NULL
DKK-3	Methylation	NULL	NULL	128	YES	0.523	CC	NULL	NULL
DKK-3	Methylation	NULL	NULL	64	YES	0.516	CC	Male	NULL
DKK-3	Methylation	NULL	NULL	64	YES	0.484	CC	Female	NULL
LRRC3B	Methylation	NULL	NULL	31	YES	0.774	CC	Female	NULL
LRRC3B	Expression	NULL	NULL	31	LOW	0.548	CC	Female	NULL
APC	Mutation	NULL	NULL	63	YES	0.301	CC(Sporadic)	NULL	NULL
APC	MAM*	NULL	NULL	112	YES	0.366	CC(Sporadic)	NULL	NULL
MINT2	Methylation	NULL	NULL	34	YES	0.176	CC	NULL	NULL
GATA4	Methylation	NULL	NULL	90	YES	0.700	CC	NULL	NULL
GATA4	Methylation	NULL	NULL	47	YES	0.659	CC	Distal	NULL
GATA4	Methylation	NULL	NULL	41	YES	0.780	CC	NULL	NULL
GATA5	Methylation	NULL	NULL	77	YES	0.792	CC	NULL	NULL
GATA5	Methylation	NULL	NULL	36	YES	0.861	CC	Proximal	NULL
HLTF	Methylation	NULL	NULL	96	YES	0.521	CC	NULL	NULL
RUNX3	Methylation	NULL	NULL	25	YES	0.640	CC	NULL	NULL
SOCS-1	Methylation	NULL	NULL	61	YES	0.082	CC	NULL	NULL
SCGB3AI	Methylation	NULL	NULL	49	YES	0.184	CC	NULL	NULL
NR3CI	Methylation	NULL	NULL	63	YES	0.032	CC	NULL	NULL
SST	Methylation	NULL	NULL	74	YES	0.910	CC(Sporadic)	NULL	NULL

Table B.2: "Single relation" for the new entries in Genetic and Epigenetic Events. C.C = Colon Cancer; Sporadic colon cancer occurs in people who have no family history of the disease. MAM*:Mono allelic Mutation

GATA4 and GATA5.

Gene1	Event1	Level	Quantified	Unit	NbTu	Gne2	Event2	Level2	Freq	Phe	Cnecal
SAMD4	Expression	YES	NULL	NULL	79	NULL	TGFBR2	YES	0.228	CC	NULL
NDRG2	Methylation	YES	NULL	NULL	8	NULL	MSI	LOW	0.125	CC	NULL
PIK3CG	Mutation	YES	NULL	NULL	102	JVCT	Exp	YES	0.340	CC	NULL
SCGB3AI	Methylation	YES	NULL	NULL	9	NULL	MSI	YES	0.889	CC	NULL
HOXA9	Methylation	YES	NULL	NULL	12	NULL	MSI	YES	0.583	CC	NULL
HOXA9	Methylation	YES	NULL	NULL	12	NULL	MSI	NO	0.417	CC	NULL
COX-2	Expression	YES	NULL	NULL	645	JCVT	Exp	YES	0.370	CC	NULL
RUNX3	Methylation	YES	NULL	NULL	16s	NULL	MSI	YES	1	CC	NULL

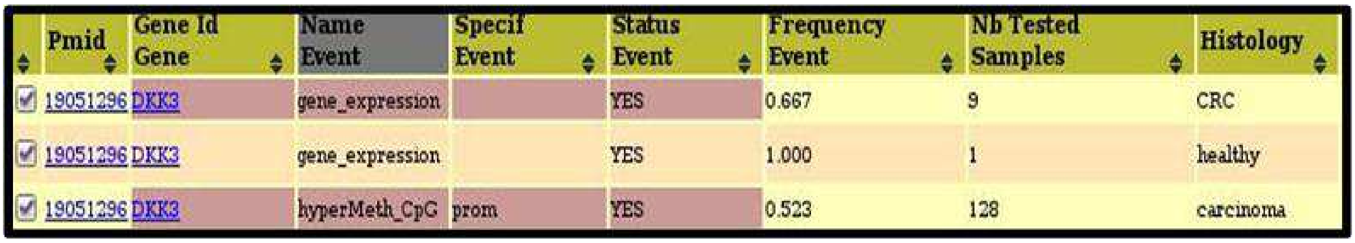
Table B.3: "Double relation" for new entries in Genetic and Epigenetic Events. CC = Colon Cancer

B.3 Screen-shots of the new entries

B.3.1 Screen-Shot for the DKK3 gene

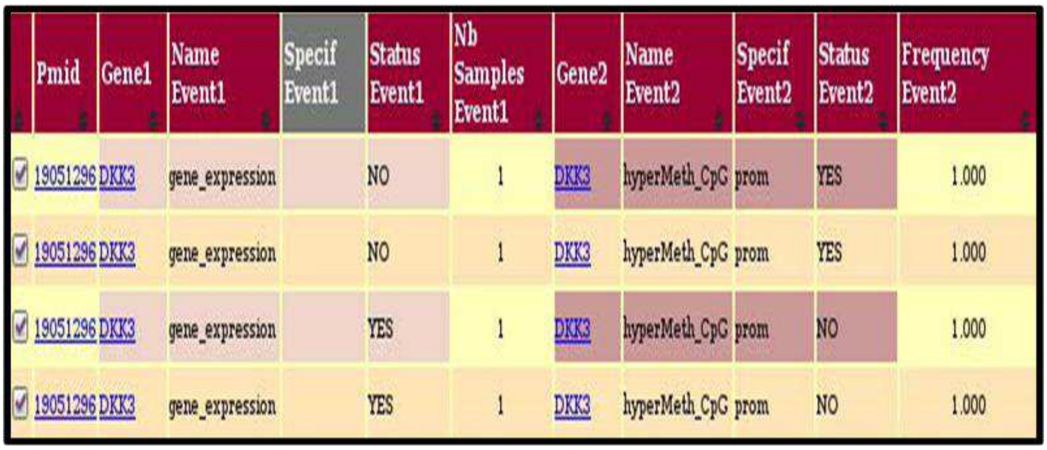
The DKK3 (Dickkopf homolog 3) gene is a member of the Dickkopf family and plays an important role in embryonic development. It inhibits Wnt-regulated processes such as antero-posterior axial patterning, limb development, somitogenesis and eye formation.

Generally, DKK3 is expressed at 0.667 frequency in 9 tissue samples and is hypermethylated at its promoter at 0.523 frequency in 128 tissue samples, as illustrated in



	Pmid	Gene Id Gene	Name Event	Specif Event	Status Event	Frequency Event	Nb Tested Samples	Histology
☑	19051296	DKK3	gene_expression		YES	0.667	9	CRC
☑	19051296	DKK3	gene_expression		YES	1.000	1	healthy
☑	19051296	DKK3	hyperMeth_CpG	prom	YES	0.523	128	carcinoma

Figure B.1: "Single relation"screen-shot for the DKK3 gene



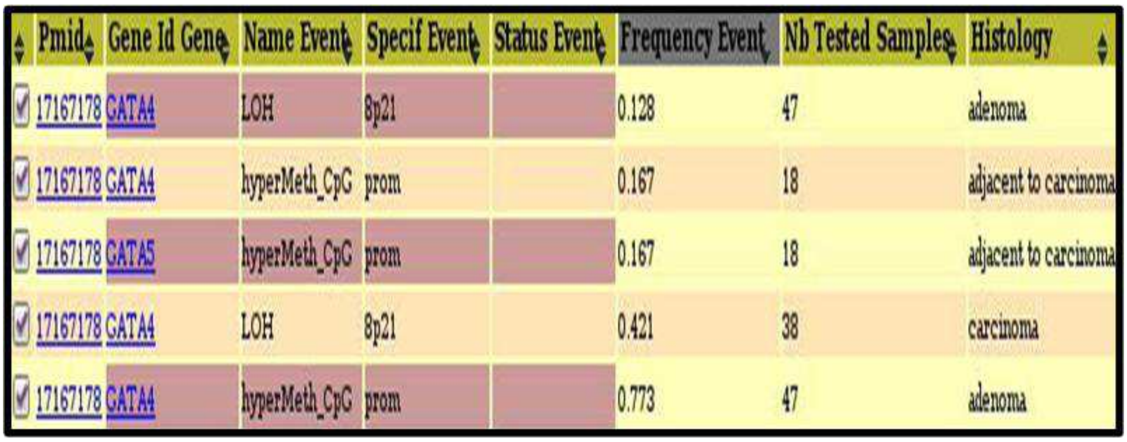
	Pmid	Gene1	Name Event1	Specif Event1	Status Event1	Nb Samples Event1	Gene2	Name Event2	Specif Event2	Status Event2	Frequency Event2
☑	19051296	DKK3	gene_expression		NO	1	DKK3	hyperMeth_CpG	prom	YES	1.000
☑	19051296	DKK3	gene_expression		NO	1	DKK3	hyperMeth_CpG	prom	YES	1.000
☑	19051296	DKK3	gene_expression		YES	1	DKK3	hyperMeth_CpG	prom	NO	1.000
☑	19051296	DKK3	gene_expression		YES	1	DKK3	hyperMeth_CpG	prom	NO	1.000

Figure B.2: "Double relation" screen-shot for the DKK3 gene

Figure (B.1). When DKK3 is 100% methylated, it is not expressed, as illustrated in Figure (B.2).

B.3.2 Screen-Shot for the GATA4 and GATA5 gene

The GATA4 gene encodes a member of the GATA family of zinc-finger transcription factors. GATA5 encodes a protein known to bind to hepatocyte nuclear factor-1alpha (HNF-1alpha), and this interaction is essential for cooperative activation of the intestinal lactase-phlorizin hydrolase promoter. Both genes (GATA4 and GATA5) are transcriptional activators. Their hypermethylation was found in 18 tissue samples at 0.167 frequency, and LOH occurs at 8p21 chromosome at 0.421 frequency in 38 tissues. GATA5 is maximally hypermethylated at its promoter with a frequency of 0.773 in 47 tissue samples, (Derks et al., 2006), as illustrated in (B.3).



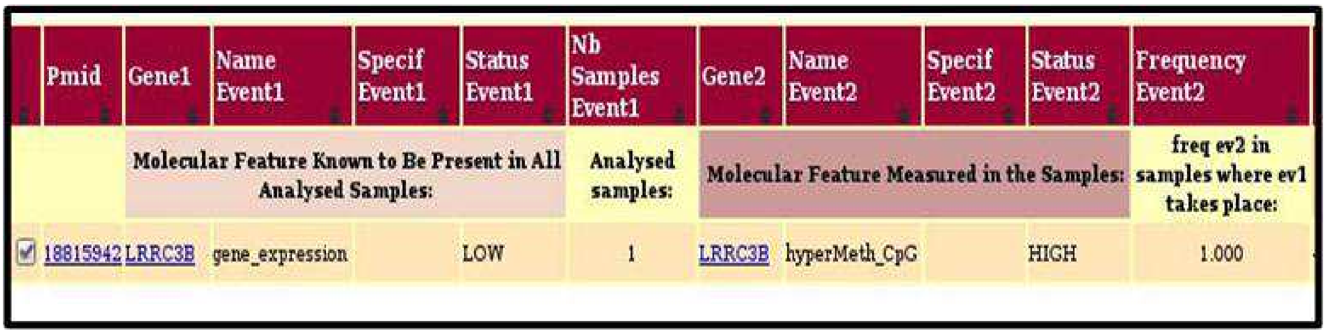
	Pmid	Gene Id Gene	Name Event	Specif Event	Status Event	Frequency Event	Nb Tested Samples	Histology
☑	17167178	GATA4	LOH	8p21		0.128	47	adenoma
☑	17167178	GATA4	hyperMeth_CpG	prom		0.167	18	adjacent to carcinoma
☑	17167178	GATA5	hyperMeth_CpG	prom		0.167	18	adjacent to carcinoma
☑	17167178	GATA4	LOH	8p21		0.421	38	carcinoma
☑	17167178	GATA4	hyperMeth_CpG	prom		0.773	47	adenoma

Figure B.3: "Single relation" screen-shots of the GATA4 and GATA5 genes

B.3.3 Screen-Shot for the LRRC3B gene

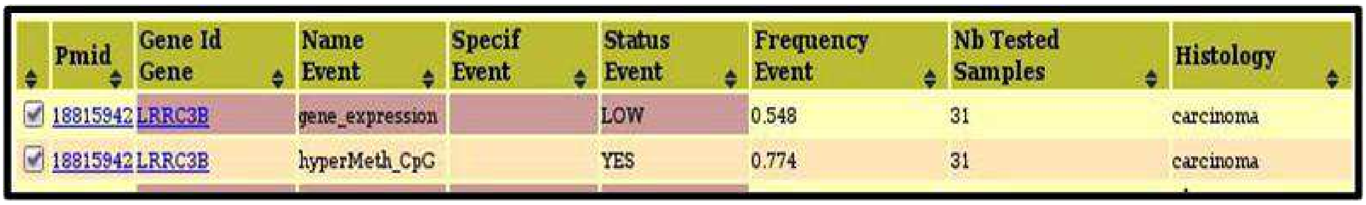
LRRC3B (Leucine Rich Repeat Containing 3B) is a tumour suppressor gene playing an important role in the pathogenesis of colorectal cancer (CRC), (Tian et al., 2009).

Low expression of LRRC3B occurs when this gene is 100% methylated at its promoter region, as illustrated in Figure (B.4). Consequently, LRRC3B is expressed at the low frequency of 0.548 and is hypermethylated at its promoter with 0.774 frequency in 31 tissue samples, as depicted in Figure (B.5).



	Pmid	Gene1	Name Event1	Specif Event1	Status Event1	Nb Samples Event1	Gene2	Name Event2	Specif Event2	Status Event2	Frequency Event2
		Molecular Feature Known to Be Present in All Analysed Samples:				Analysed samples:	Molecular Feature Measured in the Samples:				freq ev2 in samples where ev1 takes place:
☑	18815942	LRRC3B	gene_expression		LOW	1	LRRC3B	hyperMeth_CpG		HIGH	1.000

Figure B.4: "Double relation" screen-shot of the LRRC3B gene

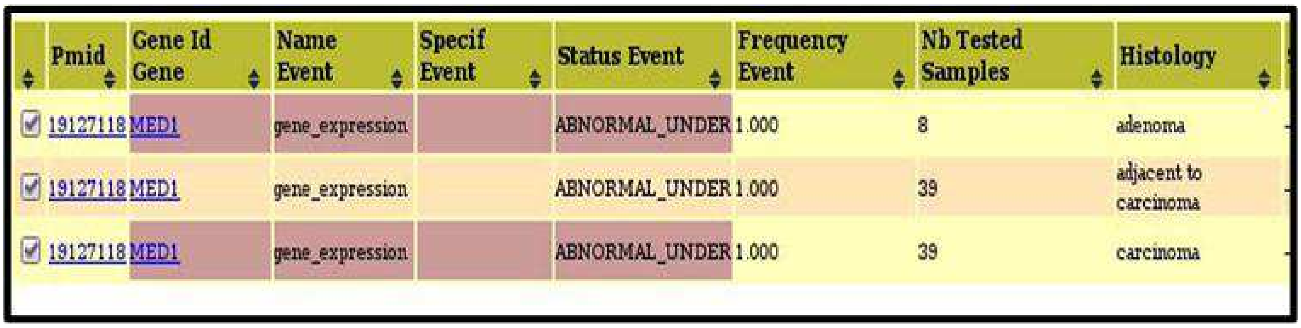


	Pmid	Gene Id Gene	Name Event	Specif Event	Status Event	Frequency Event	Nb Tested Samples	Histology
☑	18815942	LRRC3B	gene_expression		LOW	0.548	31	carcinoma
☑	18815942	LRRC3B	hyperMeth_CpG		YES	0.774	31	carcinoma

Figure B.5: "Single relation" screen-shot of the LRRC3B gene

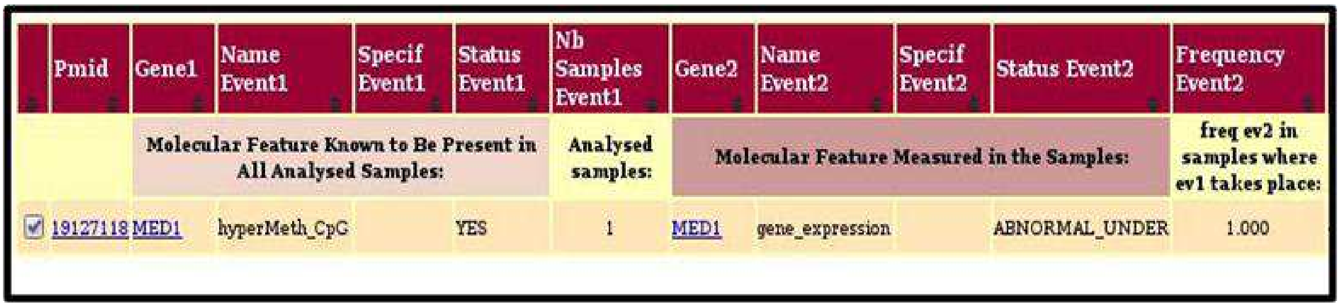
B.3.4 Screen-Shot for the MED1 gene

The MED1 gene encodes a protein mediator of RNA polymerase II transcription subunit 1. It plays a main role in transcriptional activation and recognises transcriptional enhancer sites in DNA. Abnormal expression of MED1 appears in adenoma adjacent carcinoma. When CpG sites of MED1 are methylated, it shows abnormal gene expression, as depicted in Figures (B.6 and B.7).



	Pmid	Gene Id Gene	Name Event	Specif Event	Status Event	Frequency Event	Nb Tested Samples	Histology
☑	19127118	MED1	gene_expression		ABNORMAL_UNDER	1.000	8	adenoma
☑	19127118	MED1	gene_expression		ABNORMAL_UNDER	1.000	39	adjacent to carcinoma
☑	19127118	MED1	gene_expression		ABNORMAL_UNDER	1.000	39	carcinoma

Figure B.6: "Single relation" screen-shot of MED1



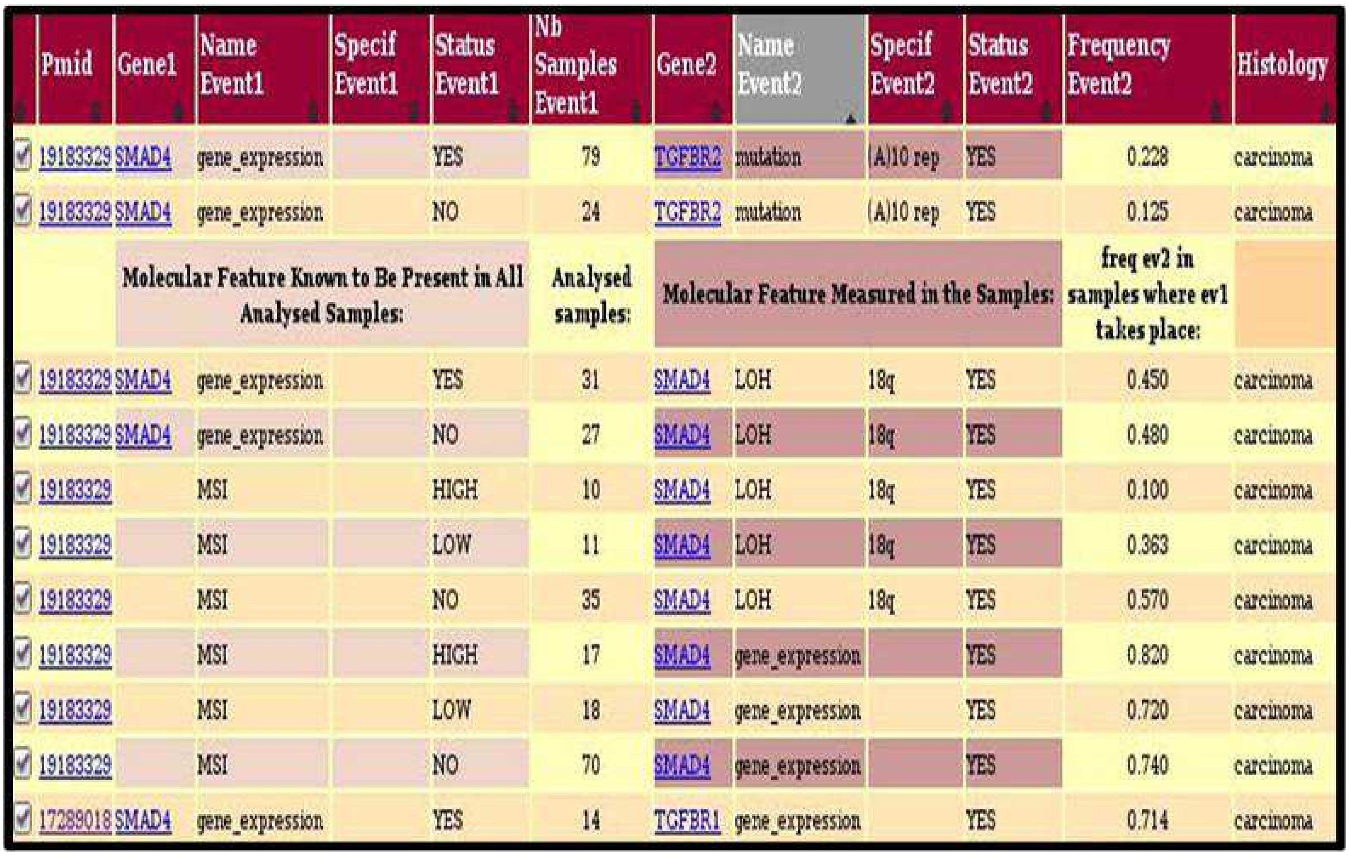
	Pmid	Gene1	Name Event1	Specif Event1	Status Event1	Nb Samples Event1	Gene2	Name Event2	Specif Event2	Status Event2	Frequency Event2
		Molecular Feature Known to Be Present in All Analysed Samples:				Analysed samples:	Molecular Feature Measured in the Samples:				freq ev2 in samples where ev1 takes place:
☑	19127118	MED1	hyperMeth_CpG		YES	1	MED1	gene_expression		ABNORMAL_UNDER	1.000

Figure B.7: "Double relation" screen-shot of MED1

B.3.5 Screen-Shot for the SMAD4 gene

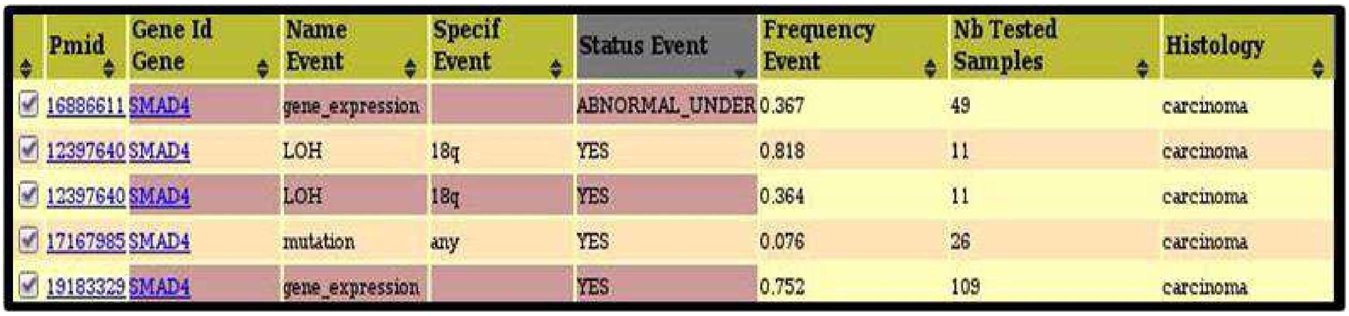
The SMAD4 gene is a member of the SMAD family of signal transduction proteins, (Seshimo et al., 2006). The proteins are activated by transmembrane serine-threonine receptor kinases in response to TGF-beta signalling. Seventy-nine tissue samples showed mutations in both SMAD4 and TGFBR2 ('YES') with a frequency of 0.228. Additionally, SMAD4 and TGFBR2 were expressed at a frequency of 0.714 in 14 samples.

In contrast, SMAD4 expression is 'YES' when the second event is 'LOH' in 31 tissue samples with 0.450 frequency, as shown in (B.8). In addition, 'MSI' highly corresponding to 'LOH' events was observed in 10 tissue samples with a frequency of 0.100,



	Pmid	Gene1	Name Event1	Specif Event1	Status Event1	Nb Samples Event1	Gene2	Name Event2	Specif Event2	Status Event2	Frequency Event2	Histology
☑	19183329	SMAD4	gene_expression		YES	79	TGFBR2	mutation	(A)10 rep	YES	0.228	carcinoma
☑	19183329	SMAD4	gene_expression		NO	24	TGFBR2	mutation	(A)10 rep	YES	0.125	carcinoma
		Molecular Feature Known to Be Present in All Analysed Samples:				Analysed samples:	Molecular Feature Measured in the Samples:				freq ev2 in samples where ev1 takes place:	
☑	19183329	SMAD4	gene_expression		YES	31	SMAD4	LOH	18q	YES	0.450	carcinoma
☑	19183329	SMAD4	gene_expression		NO	27	SMAD4	LOH	18q	YES	0.480	carcinoma
☑	19183329		MSI		HIGH	10	SMAD4	LOH	18q	YES	0.100	carcinoma
☑	19183329		MSI		LOW	11	SMAD4	LOH	18q	YES	0.363	carcinoma
☑	19183329		MSI		NO	35	SMAD4	LOH	18q	YES	0.570	carcinoma
☑	19183329		MSI		HIGH	17	SMAD4	gene_expression		YES	0.820	carcinoma
☑	19183329		MSI		LOW	18	SMAD4	gene_expression		YES	0.720	carcinoma
☑	19183329		MSI		NO	70	SMAD4	gene_expression		YES	0.740	carcinoma
☑	17289018	SMAD4	gene_expression		YES	14	TGFBR1	gene_expression		YES	0.714	carcinoma

Figure B.8: "Double relation" screen-shot of SMAD4



	Pmid	Gene Id Gene	Name Event	Specif Event	Status Event	Frequency Event	Nb Tested Samples	Histology
☑	16886611	SMAD4	gene_expression		ABNORMAL_UNDER	0.367	49	carcinoma
☑	12397640	SMAD4	LOH	18q	YES	0.818	11	carcinoma
☑	12397640	SMAD4	LOH	18q	YES	0.364	11	carcinoma
☑	17167985	SMAD4	mutation	any	YES	0.076	26	carcinoma
☑	19183329	SMAD4	gene_expression		YES	0.752	109	carcinoma

Figure B.9: "Single relation" screen-shot of SMAD4

and'NO' 'MSI' was observed in 35 samples. Furthermore, MSI was 'HIGH', 'LOW' and 'NO' in 17 (at frequency 0.820), 18 (at frequency 0.720), and 70 (at frequency

0.740)tissue samples, respectively, corresponding to SMAD4 expression. However, 24 tissue samples without SMAD4 gene expression but with TGFBR2 mutated show a frequency of 0.125, as illustrated in Figure (B.8). Furthermore, SMAD4 'expression', 'LOH' and 'mutation' with the frequency 0.367, 0.818 and 0.076 in 49, 11, 26 samples respectively. Furthermore, 109 samples show SMAD expression with a frequency of 0.752, as illustrated in Figure (B.9).

Appendix C

Glossary

The Cell: A cell is the basic unit of life. There are millions of different types of cells such as human blood cell and bacteria cells etc.

Stem Cell: A biological cell found in all multicellular organisms, that can divide through mitosis and differentiate into diverse specialized cell types and can self renew to produce more stem cells.

DNA and RNA: These are two different nucleic acids found in the cells of every living organism. RNA is single-stranded while DNA is a double-stranded helix. DNA contains the genetic information of an organism, and this information dictates how the body cells would construct new proteins according to the genetic code of the organism.

Nucleotide: These are the main building blocks for DNA and RNA. Two nucleotides on opposite complementary DNA or RNA strands that are connected via hydrogen bonds are called a base pair (bp). These are Adenine (A), which forms a base pair with Thymine (T and Guanine (G), which forms a base pair with Cytosine (C). In RNA, thymine is replaced by uracil (U).

Chromosome:DNA is organized into structures called chromosomes, which are duplicated during cell division. These chromosomes would then release the genetic codes that will be transcribed and carried by the RNA.

Nucleosome position: Each of our billions of cells contains about two metres or six feet of nuclear DNA. All of this DNA has to be packed into a nucleus that is about 10 microns or one hundredth of a millimetre across. In order to fit all this DNA into this tiny space most of the DNA strand is wrapped into tiny loops called nucleosomes. This arrangement of the DNA affects on physical structure of the cell and the organisms, also readout of the DNA sequence and moderate the expression of genes.

Promoter: A region of DNA that facilitates the transcription of a particular gene.

Transcription: A process of creating a complementary RNA copy of a sequence of DNA. During transcription, a DNA sequence is read by an enzyme (RNA polymerase), which produces a complementary, antiparallel RNA strand.

Translation: In translation, RNA (messenger RNA) produced by transcription is decoded by the ribosome[1] to produce a specific amino acid chain, or polypeptide, that will later fold into an active protein.

[1] A cell component where translation takes place.

Post Translational Modification (PTM): It is the chemical modification of a protein after its translation and includes acetylation, methylation and alkylation etc.

Histone Modification: Histone modification affects chromosome function through at least two distinct mechanisms. (i) alter the electrostatic charge of the histone resulting in a structural change in histones or their binding to DNA, (ii) binding sites for protein recognition modules, such as the bromodomains or chromodomains, that recognize acetylated lysines or methylated lysine, respectively.

Human Genome Project (HGP): HGP (1990-2003) was a 13-year project coordinated by the U.S. Department of Energy and the National Institutes of Health and completed in 2003. The Goal of this project were to identify ≈ 20,000 - 25,000 gene and determine the sequences of the 3 billion chemical base pairs, improve tools for data analysis in human DNA.

The Epigenome: A component part of the covalent structure of DNA, methylated cytosines located in the dinucleotide sequence CG and a noncovalent module.

Human Epigenome Project(HEP): A multinational science project and coordinated by "The Wellcome Trust Sanger Institute UK", " Epigenomics AG Germany/USA" and "The Centre National de Gnotypage France". Goal of this project is identify, catalog, and interpret genome-wide DNA methylation patterns of all human genes in all major tissues, (Bradbury, 2003)

Methyl Binding Protein (MBD): The MBD family, has been characterized at both the biochemical and genetic levels and members are MeCP2 (for specific sequences), MBD2 (bind with somewhat relaxed specificity to methylated DNA), MBD3 and MBD4, (Fatemi and Wade, 2006).

Chromatin Modification: A special modification on chromosome which impacts histone and nucleosome positions.

CpG island methylator phenotype (CIMP): CIMP refers to the concordant methylation of a group of genes in cancer.

Transcription Factor (TF): A TF is a protein that binds to specific DNA sequences and controls the flow transcription. It performs by promoting, blocking the recruitment of RNA polymerase[2]. to specific genes.

Evolutionary Genetics: A broad field of studies that attempts to account for evolution in terms of changes in gene and genotype frequencies within populations and the processes that convert the variation with populations into more or less permanent variation between species.

Microsatellite Instability (MSI): It is repeated sequences of DNA highly variable from person to person, each individual has microsatellites of a set length. These repeated sequences are common, and normal. The most common microsatellite in humans is a dinucleotide repeat of Cytosine and Adenine.

Chromosomal Instability: Irregularity in chromosome structure during cell division which occurs in cancer.

[2] The enzyme that performs the transcription of genetic information from DNA to RNA.

Wnt signaling pathway: A network of proteins which plays role in embryogenesis and cancer and their interactions with receptors on target cells.

PraderWilli syndrome (PWS): is a rare genetic disorder in which seven genes on chromosome 15 (q 11-13) are deleted or unexpressed (chromosome 15q partial deletion) on the paternal chromosome.

DNA Repair: DNA repair refers to a collection of processes by which a cell identifies and corrects damage to the DNA molecules that encode its genome.

RNA polymerase (RNApol): is an enzyme that produces RNA. In cells, RNAP is needed for constructing RNA chains from DNA genes as templates, a process called transcription. RNA polymerase enzymes are essential to life and are found in all organisms and many viruses.

Exon: An exon is a nucleic acid sequence that is represented in the mature form of an RNA molecule.

TGFBR2 Gene: This gene encodes a member of the Ser/Thr protein kinase family and the TGFB receptor subfamily. The encoded protein is a transmembrane protein that has a protein kinase domain, forming a heterodimeric complex with another receptor protein, and its function is legend binding, forming a receptor complex consisting of two type II and two type I transmembrane serine/threonine kinases.

Transforming growth factor beta (TGF) signaling pathway: The pathway which is involved regulation of the developing embryo including cell growth, cell differentiation, apoptosis, cellular homeostasis and other cellular functions.

Histone methylation: Histone methylation is the modification of certain amino acids in a histone protein by the addition of one, two, or three methyl groups. It turns the allowing transcription factors and other proteins to access the DNA or by encompassing their tails around the DNA, thus, restricting access to the DNA.

Cell Line: Specific cells that can grow indefinitely given the appropriate medium and conditions.

MLH1: MLH1 gene is located on the short arm of chromosome 3 at position 21.3. It is also known as COCA2, FCC2, hMLH1 and HNPCC2.

HNPCC: Hereditary Non Polyposis Colorectal Cancer (HNPCC) is an autosomal dominant genetic condition which has a high risk of colon cancer.

House Keeping Gene: A housekeeping gene is typically a constitutive gene that is required for the maintenance of basic cellular function, found in all cells of an organism.

PathEpigen: The PathEpigen Knowledge Management System is continually being added in terms of data content and additional features. It has now been renamed StatEpigen (first referring to PathEpigen), to reflect its focus on the statistical information and currently available at http://statepigen.sci-sym.dcu.ie/. However, as it was known as PathEpigen during the period within which the work reported here was performed, we have retained the original name in this thesis.

Appendix D

Publication

D.1 ICG 2008

One poster publication was presented during this project.

Porwal, J. and Ruskin, H.J. and Perrin, D. and Roche, D.and Burns, J.(2008). Microscopic Model of Epigenetic Mechanisms (poster). XX International Congress of Genetics, Berlin, Germany.

Abstract

"An initial microscopic model of epigenetic mechanisms is proposed, linking gene expression changes to cancer initiation. Emergent properties of complex tumours are influenced by aberrant modification of DNA methylation through the process of cell division. The objective here is to map the input for changes."

Microscopic Model of Epigenetic Mechanisms

J. Porwal, H.J. Ruskin, D. Perrin, D. Roche and J. Burns

DCU

email: {jporwal}@computing.dcu.ie

Abstract

An initial microscopic model of epigenetic mechanisms is proposed, linking gene expression changes to cancer initiation. Emergent properties of complex tumours are influenced by aberrant modification of DNA methylation through the process of cell division. The objective here is to map the input for changes.

Introduction

Epigenetic mechanisms influence cell phenotype through heritable regulation of gene expression.[1]

The building blocks:

- *Chromatin = Cluster of DNA + Proteins(histones).*

Designed to package DNA into smaller volume in the cell and to control gene expression.

- *Nucleosome = Fundamental unit of chromatin.*

Each nucleosome unit consist of 147 base pairs of DNA and four base pairs of histone proteins(Fig 1).

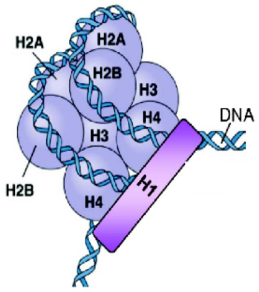


Figure 1: Nucleosome, fundamental unit of chromatin(adapted from C.Brenner,PhD thesis,Universite Libre de Bruxelles, 2005)

- *Epigenetic mechanisms = DNA methylation, Histone acetylation and Histone methylation.*

- Alterations in gene expression arise during development and cell proliferation and persist through cell divisions.
- DNA methylation is a post-replicative process of cytosine residues. CpG sequences are methylated during cell division.(C=cytosine, G=guanine, p=phosphate)
- Epigenetic changes result in DNA methylation errors; (Hypermethylation or Hypomethylation).

Fig 2 shows hypermethylation on CpG islands (>=200 base pairs of CpG dinucleotide) on promoter (regulatory region on the DNA).

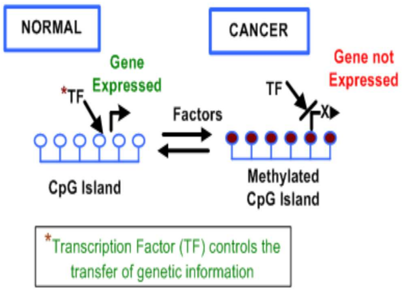


Figure 2: Change in DNA hypermethylation

Change of epigenetic factors directly impacts cell cycle in stem cell

- Environmental factors which alter epigenetic mechanisms include: smoking, radiation, nickel etc.[2][3]

Objective

Development of a microscopic representation of epigenetic mechanisms, providing hierarchy of changes and early detection of cancer initiation.

The Model

- An object-oriented model of epigenetic mechanisms focusing on individual nucleosomes and actual DNA sequences.

Fig 3 shows basic structure of the model. Each unit represents a class.

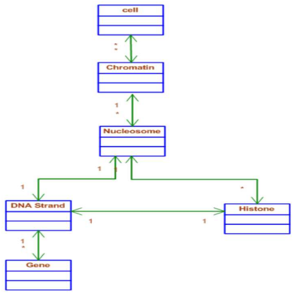


Figure 3: Structure of Model

- Model Layers:

$$P = G + E + EpiG \qquad (1)$$

Where P = phenotype (An organism behaviour); G = Genetic factors; E = Environmental factors; EpiG = Epigenetics(changes in gene expression stable between cell divisions)[4]

$$EpiG = E_g + G_{env} \qquad (2)$$

where E_g is the presence of an enzyme group with certain characteristics, G_{env} is a set of genes affected by environmental factors.

- Each environmental entity is represented by specific agents.Types of agents are inherited from a common class. Interactions are implemented by various C++ classes.

Conclusion

It is possible for low level changes to be modelled such that high level phenomena can be attributed to epigenetic events.

References

1. Bock C., Lengaur T., “Computational epigenetics”, Bioinformatics, 24:1-10, 2008.
2. Feinberg A.P., Tycko B., “The history of cancer epigenetics”, Nature Reviews, 4:143-153, 2004.
3. Herceg Z., “Epigenetics and cancer: towards an evalution of the impact of environmental and dietary factors”, Mutagenesis, 22:91-103, 2007.
4. Perrin D., Ruskin H.J., Crane M., Walshe R., “Epigenetic Modelling”, ERCIM NEWS, 72:46, 2008.