A Computational Model for Colon Cancer Dynamics

Irina - Afrodita Roznovăţ

B.Sc.

A Dissertation submitted in part fulfilment of the

requirements for the award of

Doctor of Philosophy (Ph.D.)

to the

Dublin City University

Faculty of Engineering and Computing, School of Computing

Supervisors: Prof. Heather J. Ruskin,

Dr. Dimitri Perrin, Laboratory for Systems Biology, RIKEN Center for Developmental Biology, Kobe, Japan

September, 2014

Declaration

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

Signed: Irina - Afrodita Roznovăţ

ID No: 10117806

Date: 22.09.2014

Table of Content

List of Abbreviations

*Note: Extended summary of specialised terms appears in Glossary, (Appendix A).

5caC: 5-carboxylcytosine

5fC: 5-formylcytosine

5hmC: 5-hydroxymethylcytosine

5mC: methylated cytosine

AgentCrypt: Agent-based Model of Intestinal Crypt Dynamics

E-G Network Model: Epigenetic-Genetic Network Model

LogisticCrypt: Logistic Model on the Intestinal Crypt Structure

ABM: Agent-based Model

ASG: Age-Sensitive Gene

BN: Bayesian Network

CA: Cellular Automata

ChIP: Chromatin immunoprecipitation

CIESCI: Complexity of Interdependent Epigenetic Signals in Cancer Initiation

CIMP: CpG Island Methylator Phenotype

CIN: Chromosomal instability

CRC: Colorectal cancer

DDC: Density-Dependent Coefficient

Diff: Fully-differentiated cell

DNA: Deoxyribonucleic acid

DNAm: DNA methylation

DNMT: DNA methyltransferase

dNT: d-Network Threshold

DRC: Drug Response Curve

ENCODE: Encyclopaedia of DNA Elements

EWASs: Epigenome-wide association studies

FAP: Familial Adenomatous Polyps

GE: Gene expression

GEO: Gene Expression Omnibus

H-/H+ : Hypo-/ Hypermethylation

H1/H5, H2A, H2B, H3, H4: Families of histones

H3K27me3: Trimethylation of histone at lysine 27

HAT: Histone acetyltransferases

HDAC: Histone deacetylases

HDM: Histone demethylases

HM: Histone modifications

HMT: Histone methyltransferases

HNPCC: Hereditary nonpolyposis colorectal cancer

HPLC: High-performance liquid chromatography

ICGC: The International Cancer Genome Consortium

IHEC: The International Human Epigenome Consortium

IMET: Inhibitor Maximum Efficiency Time

MAP: MYH-Associated Plyposis

MCA: The methylation cycle number average

miRNA: Micro RNA

Mut: Mutation

NCBI: The National Center for Biotechnology Information

NGS: Next-Generation Sequencing

PCR: Polymerase chain reaction

PMR: Percentage of methylated reference

PMRA: Average percentage of methylated reference

Prog: Progenitor cell

RNA: Ribonucleic acid

SAM: Sensitive to an Age-related Methylation

SCI-SYM: The Centre for Scientific Computing and Complex Systems Modelling

SP: Signalling pathway

TCGA: The Cancer Genome Atlas

TSG: Tumour suppressor gene

List of Figures

List of Tables

Abstract

Cancer, a class of diseases, characterized by abnormal cell growth, has demonstrably high impact on human life: complex lifestyle changes, caused by malignancy, affect not only patients, but also family and friends. Cancer development has been linked to *genetic* and *epigenetic* abnormalities that affect the regulation of *key* genes that control cellular mechanisms. These alterations, which target *stem cells*, may be different or less immediately adverse from one person to another, as various risk factors are cumulative and variable in effect. However, a major issue in cancer research is the lack of precise information on tumour pathways; therefore the *delineation* of these and of the processes underlying disease proliferation is an important area of investigation.

Here, we present a hybrid computational model following a multi-scale approach, which has been developed for colorectal cancer dynamics by linking information from micro-molecular to cellular and tissue levels, (e.g. epigenetic events, stem cells, intestinal crypts). The current work aims to i) investigate genetic and epigenetic interdependencies leading to colon cancer initiation and progression; ii) to analyse influence of different risk factors at both molecular and cellular levels during cancer development; iii) to examine aberrant DNA methylation variation in malignant intestinal systems; iv) to evaluate the effect of inhibiting methylation modifications in abnormal colon crypts over time; and v) to assess the impact of deregulations in intestinal crypt dynamics with respect to tumour development. Given its crucial role in cancer development, *DNA methylation* is the main feature of the colorectal cancer model.

Computational modelling is performed at different levels. A network-based model,

namely *the Epigenetic-Genetic (E-G) Network Model*, has been developed to explore interdependencies between genetic and epigenetic events, recorded at different stages of colorectal cancer, (with a focus on gene relationships and tumour pathways). Micro-molecular modifications are studied in relation to *ageing* and *gender*, considered to be major risk factors in cancer development. Further, an agent-based model, *AgentCrypt*, with agents representing three cell types: *stem*, *progenitor* and *differentiated* cells in the *colon* (and with addition of a *Paneth* cell group in *small intestine*), has been developed to describe the dynamics of the intestinal crypt. Comparative analysis on methylation variation between colon and small intestine crypts during cancer initiation and under *carcinogen* influence is performed. A further extension concentrates on analysing the impact of potential *inhibitors* on methylation level in the intestinal crypt. Finally, a logistic model, *LogisticCrypt*, (focusing on cell division, differentiation and apoptosis rates and on cell competition related to the intestinal crypt space), has been built to provide information on the crypt structure at specific time points and to investigate time-intervals to occurrence of major crypt phenomena (such as 'crypt fission' and the 'bottleneck effect'), due to deregulations in intestinal cell number.

Keywords: *Epigenetic and Genetic events; gene relationship; tumour pathway; carcinogen; methylation inhibitor; stem cell dynamics; colon crypt; small intestine crypt; cross-comparative analysis; hybrid computational model; Bayesian network; agent-based; logistic model*.

To Elena, Claudiu, Mihai, Valentin and Vasile

Acknowledgments

First, I wish to thank God for giving me health during these years and strength to complete successfully my PhD.

I would like to acknowledge my supervisors, Prof. Heather J. Ruskin and Dr. Dimitri Perrin for helping me develop my scientific personality and for their continuous support and guidance they have provided during my PhD studies. I wish one day I could be as enthusiastic, energetic and diplomatic as they are. I would like also to acknowledge Dr. Ana Barat and Dr. Edel Hyland for our useful discussions on Epigenetics, and my undergraduate Professor, Dr. Liviu Ciortuz, who introduced me to Bioinformatics and Machine learning and encouraged me during the past several years. I am grateful for financial support from the CIESCI ERA Net Complexity Project, (EC/IRCSET) for two years and a final year top-up under the Embark Initiative from the Irish Research Council (IRC). I would like also to acknowledge Dr. Martin Crane, Dr. Anne Parle-McDermott and Dr. Huiru Zheng for valuable comments on my thesis and for making the examination process as enjoying as possible.

Time at DCU passed quickly, but I will always remember and treasure the great moments spent as colleagues and friends with people from around the World during these years. I thank Derek who did his best to find the most suitable accommodation when needed. Special thanks to Alina for encouragement given even before I started my PhD, to Emily, Irina and Pooyan for their positive attitudes and advices especially during the last several months and to Olaru Family for their warmth and kindness during my time in Dublin.

I will forever be grateful to my Family for their love and unconditional support. Warmest thanks to Valentin who was in the same time, patient and critical, calm and enthusiastic, ‘travelling’ together all the steps of this ‘trip’. I thank to You all.

Chapter 1

Introduction

1.1 Motivation

Colorectal cancer, (CRC), is the third most common cancer type world-wide, [Ferlay et al., 2014], accounting for around 8.6% and 13% of all new cancer cases in the US (2013), [SEER, 2013b], and UK (2010), [Cancer Research UK, 2014b], respectively. In addition, it has been reported to be the second leading cause of cancer-related deaths for the combined sexes in US (2013) and UK (2011), accounting for around 15% and 10% of all cancer deaths in these countries. In Ireland, CRC is the second most common cancer type, accounting for around 2,270 new cases diagnosed in 2009, [Irish Cancer Society, 2014], with a total of around 37,000 new cases identified during 1994-2011, [National Cancer Registry Ireland, 2014]. It is also the second most common cancer-related cause of death, [BowelScreen, 2014].

Epigenetic events, or the micro-molecular interactions that influence gene expression without altering the DNA sequence, have been detected (i) in the earliest stages of cancer initiation, (ii) in the ageing process and also (iii) in response to cellular stress. Over recent decades, diverse studies have shown that epigenetic modifications are *reversible*, [Yoo and Jones, 2006; Dworkin et al., 2009], and have revealed the potential of using this information in developing possible *treatments* for incipient cancer stages, [McCabe et al., 2012]. Therefore, understanding these phenomena, which are also considered *markers* for tumour

initiation, [Baylin and Jones, 2011], may increase the success of *cancer therapy* in affected patients. In addition, in the context of recent efforts on personalised medicine and targeted treatments, a novel research direction concentrates on the identification of intra-individual epigenetic variation linked to cancer predisposition and development, i.e. *epigenome-wide association studies, (EWASs)*. However, a major challenge of the EWASs is acquiring repetitive assays from the same individuals, particularly from internal tissues.

Given time and cost implications for laboratory experimentation and human epigenome studies, a range of computational models and tools has been developed over recent years, to help scientists and clinicians to better understand the impact of malignant molecular events on tumour pathways and to investigate new strategies that can be applied in cancer treatment, (e.g. [McCabe et al., 2012]). However, although extensive biocomputational work has been performed, to date, understanding of the interdependencies between genetic and epigenetic mechanisms leading to cancer initiation and progression is still limited. Therefore, given the current accessibility of genome-wide approaches, epigenetic-related experimental datasets require bioinformatics and computational approaches to assist in data analysis, hypothesis validation and results interpretation.

1.2 Scope and contribution

1.2.1 Research Objectives

This work aims to develop a simplified computational prototype of CRC dynamics, in order to explore aberrant mechanisms in CRC at different layers, i.e. *micromolecular*, *cellular* and *tissue* levels, and to identify critical events that can be used as potential targets for CRC therapy. A healthy intestinal crypt is characterised by specific relationships between cell division, differentiation and apoptosis rates in order to maintain *homeostatic*[1] control over time, [Eisenhoffer et al., 2012]. Consequently, interest centres on how modifications in these indicators can affect crypt dynamics and subsequent tumour development. Dis-

[1] homeostatic = characteristic related to the potential of a biological system, (e.g. an organism or cell) to conserve the internal stability, [Merriam-Webster, 2014].

ease phenotype has been associated with modifications of the DNA methylation (DNAm) patterns and it has been suggested that genome-wide analysis on DNAm variation can inform on tumour predisposition, [Rakyan et al., 2011]. Interdependencies exist between genetic and epigenetic events in malignant conditions and data on their occurrence and on *signalling pathways* in CRC have become increasingly available, (Chapter 3). Ageing and gender are considered major risk factors in tumour development, (Chapter 2). Abnormal micro-molecular modifications have been found to be accumulative in different systems over time, [Fraga et al., 2007], and gender differences have been reported for CRC development, [Ogino et al., 2006b]. In addition, external factors such as chemical radiation and environmental features are known to induce deregulation in biological system dynamics leading to tumour initiation and progression, (Section 2.6.2). Nevertheless, the reversibility property is of interest and several epigenetic drugs are already in use for blood cancer therapy, (Section 2.7). Thus, the main **Research Objective** of this Thesis is:

To build a model base for investigation of methylation level variation at cellular and tissue levels in human intestinal systems, with respect to i) genetic - epigenetic interdependencies, ii) patient characteristics, iii) potential methylation inhibitors and iv) carcinogen[2] influences during CRC initiation and progression, and to evaluate quantitatively the impact of deregulations on crypt dynamics with regard to tumour development.

The current work aims to establish plausibility of this information linkage for monitoring and prediction. Subsidiary research objectives include:

RO. 1. To quantify the impact of genetic and epigenetic interdependencies on the methylation level in CRC development and to estimate the influence of ageing and gender with respect to methylation level in abnormal colon cells;

RO. 2. To evaluate quantitatively methylation level variation in intestinal tissues under different external conditions during cancer initiation and progression;

[2] A carcinogen is a substance or an agent causing cancer.

RO. 3. To assess the effect of deregulation on intestinal crypt dynamics with regard to tumour development.

1.2.2 Contribution

Clearly, cancer is a complex process generated by abnormal genetic and epigenetic interdependent modifications in cellular mechanisms that exhibit different dynamics. Therefore, one formalism was never likely to be adequate and a joint network, agent-based and logistic approach has been implemented. The Colorectal Cancer Model contains three main computational components, namely:

I. **Epigenetic-Genetic (E-G) Network Model**, which explores i) the interdependencies between genetic and epigenetic events, recorded at different CRC stages, and ii) the way in which these are influenced by patient characteristics, (e.g. ageing, gender).

II. **AgentCrypt Model**, an agent-based model on crypt dynamics, which handles crypt *intra-* and *inter*-dependencies and performs a comparative analysis on DNAm level between the colon and small intestine tissues. Potential *carcinogens* are also considered and their impact on tumour pathways examined. Moreover, a set of potential methylation *inhibitors*, (defined by different prevention patterns of *de novo* methylation modifications over time), is investigated in abnormal intestinal crypts.

III. **LogisticCrypt**, a logistic model on crypt structure at specific time points, which focuses on cell division, differentiation and apoptosis rates and on cell competition related to crypt space.

The structure of the Colorectal Cancer Model and the relationships between model components are illustrated in Figure 1.1. Specifically, E-G Network incorporates on patient features and provides information on cell methylation level at specific time, which may be considered as input data for AgentCrypt. In addition, cell population dynamics are investigated in abnormal intestinal systems following both bottom-up and top-down approaches.

The CCM components described in the Thesis can be used in personalised medicine for CRC development. Specifically, they can be used for prediction (the E-G Net-

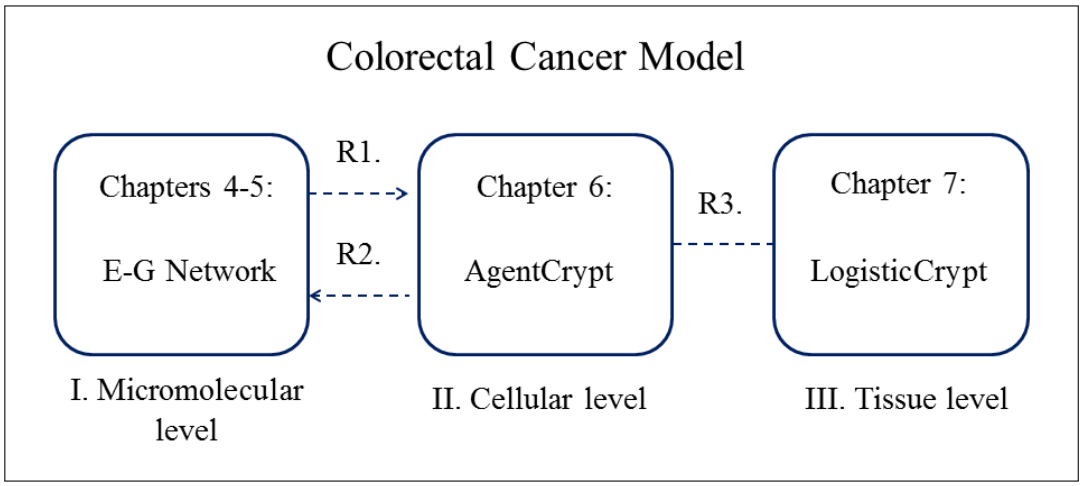


Figure 1.1: Structure of the Colorectal Cancer Model
The relationships between model components may refer to R1) average methylation level for gene network / intestinal cell; R2) patient features; R3) intestinal crypt dynamics, explored following bottom-up and top-down approaches.

work and the LogisticCrypt) and exploration of tumour dynamics, (all three). The AgentCrypt provides a framework for DNAm variation in comparative analysis of intestinal tissues and can be viewed as a proof of concept in EWASs. In addition, the AgentCrypt can be extended for investigation of intra- and inter-individual epigenetic variation in other cancer types, (e.g. liver, lung).

1.3 Thesis structure

Chapter 2 provides a brief introduction to Epigenetics and CRC mechanisms and considers three system layers, namely i) *micromolecular*, ii) *cellular* and iii) *tissue* levels. In Chapter 3, the basis for different computational models and databases developed for investigation of key-questions related to CRC dynamics, are reviewed. In addition, the importance and challenges of the multi-scale modelling approach are highlighted.

An initial version of the E-G Network model was presented in Roznovăţ and Ruskin [2013a]. This has been extended to include information on signalling pathway information and refinements of different sets of dynamics, (such as those involved in the methylation level update process), and is described in Chapter 4. Inclusion of ageing and gender influ-

ences in E-G Network was presented in Roznovăţ and Ruskin [2013c]. Methods proposed for cancer progression decision, details of implementation extensions and results obtained are described in Chapter 5. The AgentCrypt and the LogisticCrypt model components are presented in Chapters 6 and 7, respectively. (An earlier version of AgentCrypt and preliminary results on epigenetic inhibitor and carcinogen influences during CRC development were outlined in Roznovăţ and Ruskin [2013b].) Finally, Chapter 8 summarises the main findings and limitations of the overall Colorectal Cancer Model and proposes future research directions.

Appendix A provides a "Glossary" of key terms in the Biology of intestinal cancer systems, (which appear in text with the following format: **apoptosis˜**). These are gathered together for ease of reference in the text. Extended information related to CRC features, (e.g. key-genes, major types, age-group classification), is included in Appendix B. The lists of i) Epigenetics resources, ii) the several databases on cancer mechanisms and iii) BN-generator software, which were interrogated for the Colorectal Cancer Model development, with addition of iv) a gene group included in this thesis, (with information on gene symbol and names according to the GeneCards database, [Safran et al., 2010]), are tabulated in the Appendix C. Detailed information on the mathematical calculation of the relationships between cell division, differentiation and apoptosis rates within the intestinal crypt, (Chapter 7), is provided in Appendix D. Finally, Appendix E provides a list and summary abstracts of publications arising from this work.

Chapter 2

Biological background

2.1 Introduction

Colorectal cancer (CRC) has been a major area of focus of research for years due to its impact on human health. According to statistical studies, CRC is the third most common cancer type, after lung and breast cancers, and global numbers of new cases and deaths associated with it in 2012 have been estimated at around 1.3 million and 0.7 million, respectively, [Ferlay et al., 2014]. Over the last two decades, it has been shown that both *genetic* and *epigenetic* events can stimulate abnormal micromolecular modifications, leading to cancer development. A major characteristic of cancerous systems, making them difficult to understand and investigate, is cancer *diversity*; in effect, given no two individuals are the same, then no two cancers are the same, [Hutchinson, 2011]. However, in order to explore the aberrant phenomena involved in cancer initiation and progression, features of normal biological systems must be well-understood. To this end, an overview of mechanisms observed at *micro-molecular*, *cellular* and *tissue* levels in both normal and abnormal systems, (with a focus on the *intestinal* system and *CRC* development), is provided in this chapter. The first section briefly introduces principles of *Genetics* for reference in the rest of this Thesis. Secondly, the nature of *epigenetic events* and their involvement in cancer development are introduced, (Section 2.3), and are followed by a description of the intestinal crypt structure and dynamics, (Section 2.4). A summary of malignant system features is given in

Section 2.5, and information related to CRC initiation and progression, (e.g. CRC types, risk factors), is presented briefly in Section 2.6. Finally, an overview of cancer therapies involving epigenetic events is included, (Section 2.7).

2.2 A brief introduction to Genetics

The *cell* is the basic structural and functional unit of the biological systems. Its functions are encoded by *genes*, which are segments of DNA. *DNA*, (or *deoxyribonucleic acid*) is well-known to the public due to recent research breakthroughs, (as well as popular TV culture), and is the macromolecule that contains genetic information, (inherited from both parents). The DNA helix form has been regularly illustrated and consists of two strands of *nucleotides* formed from a nucleobase, a sugar and a phosphate chemical compound. The primary *nucleobases* are adenine (A), cytosine (C), guanine (G) and thymine (T), which form A-T and C-G base pairs. *RNA*, (*ribonucleic acid*), is a macromolecule, involved in gene coding, regulation and expression. It has a single strand form with A, C, G and uracil (U), which replaces the thymine nucleobase from DNA. The *proteins* are large molecules performing the functions encoded by genes, (e.g. cell division, differentiation). This transcription & translation process is usually summarised by the *Central Dogma* of molecular biology, [Crick, 1970]:

$$\text{DNA} \rightarrow \text{RNA} \rightarrow \text{protein (from chains of amino acids).}$$

Histones are proteins that package the DNA sequence in *nucleosome*, (i.e. basic units), which are grouped and form *chromatin*. Chromatin be found in two forms in cells, namely *heterochromatin*, (highly condensed chromatin, characterised by low levels of gene transcription), and *euchromatin*, (less compacted, characterised by high levels of gene transcription). *Chromosomes* are pairs of organized DNA packages in cells, where deletions or duplications of parts of chromosomes are possible, resulting in a state called *chromosomal instability*.

Two types of genes have been identified in the human genome[1], namely the *non-coding RNA* genes, (briefly presented in subsection 2.3.3), and *protein-coding genes*. Gene expression is dictated by its *promoter*[2] and regulatory elements, located typically at distant regions from the gene body.

Mutations, (i.e. abnormal modifications of the DNA sequence), can occur with high rates during the cell cycle, primarily during division, (when DNA is replicated and genetic information is transmitted from mother to daughter cells). However, there are several mechanisms that protect cells from mutation. The DNA repair mechanism, dictated by the *DNA repair genes*[3], is active constantly in normal systems to identify such aberrant changes and correct them. If mutations can not be corrected, two major actions can be induced in normal systems, namely i) programmed cell death, (i.e. *apoptosis*), or ii) *senescence*, i.e. when a cell is not considered to undergo further divisions and dies eventually. However, in abnormal systems, a mutated cell can escape from these mechanisms and continue division, (often with faster dynamics), leading to tumour development. This occurs due to deregulation of the expression patterns of three major gene types, namely ***tumour suppressor genes˜****, (TGS)*, ***proto-oncogenes˜*** and DNA repair genes. A TSG is a gene that protects a cell from progressing increasingly along the path to cancer and its inactivation leads to disease predilection because if correct expression of the protein is inhibited, the apoptosis of abnormal cells cannot be dictated, [Desper et al., 1999]. Initially, a normal cell contains proto-oncogenes, which can become ***oncogenes˜*** in tumour cells due to mutation or epigenetic activation, [Vogelstein and Kinzler, 2004]. Oncogenes can induce aberrant cell proliferation. Aberrant modifications of the DNA repair genes are associated with failure of correcting the DNA sequence errors that arise before cell division.

[1]The human genome, i.e. the complete set of DNA information found in a human body, has 23 pairs of chromosomes and approximative 20,000 protein-coding genes.

[2]Promoter is a segment of the DNA sequence, usually located just before a gene. In the transcription process, it acts as a binding site for RNA polymerase (an enzyme that synthesises RNA from DNA template) and transcription factors, (proteins binding to specific DNA sequences). Generally, a gene contains a single promoter; however, recently, genes with two promoters have been also identified.

[3]Example: MLH1 gene, with its aberrant expression leading to genomic instability and cell proliferation, Table B.1.

2.3 Epigenetic Mechanisms in Normal and Malignant Systems

Although cells share essentially the same genetic information within a body, various cell types exist and these show highly diversified gene expression patterns reflecting different functionalities. Over recent decades, biological research has focused on studying *epigenetic* phenomena, i.e heritable changes in chromatin structure that do not include modifications in DNA sequence, but affect gene expression, [Allis et al., 2007]. Conversely, the 'all-genetic' paradigm, which considers that phenotype is influenced only by genotype and environment, has been abandoned in favour of this detailed description, which includes epigenetic mechanisms, (i.e. an additional layer of gene regulation, as illustrated in Figure 2.1).

Major epigenetic mechanisms that have been identified as assisting gene regulation within a body are i) DNA methylation, ii) histone modification, and more recently, iii) small non-coding RNAs. These mechanisms form part of the **epigenome**~[4], which has rapid dynamics compared to the genome, and can be influenced by cell environment, [Alegría-Torres et al., 2011]. Epigenetic events have been detected in the earliest stages of different malignancies and act as **biomarkers**~[5] for cancer initiation, [Baylin and Jones, 2011]. Additionally, their unique *reversibility* property, (not shared by genetic mutations, [Yoo and Jones, 2006; Dworkin et al., 2009]), is already exploited in some therapy regimes, (Section 2.7). Therefore, understanding of epigenetic alterations can give improved insight on the way in which aberrant modifications characterize cancer initiation and progression.

2.3.1 DNA methylation

DNA methylation (DNAm) is a molecular process that involves the addition of a methyl group to a cytosine ring, [Bestor, 2000; Bird, 2002]. In addition, adenine methylation has been reported recently in prokaryotes[6]; however, this modification has been observed less in

[4] Epigenome includes the totality of epigenetic information in an organism.

[5] In Genetics/Epigenetics, a biomarker can refer to an event, (or a modification), associated with risk to disease development. Biomarkers can be used for identification of overall tumour occurrence, but also of specific tumour stages and subtypes, and can help in tumour prognosis, informing on tumour response, and in treatment selection.

[6] A prokaryote is a simplistic organism, composed from a single cell that has no distinct nucleus, [Merriam-Webster, 2014].

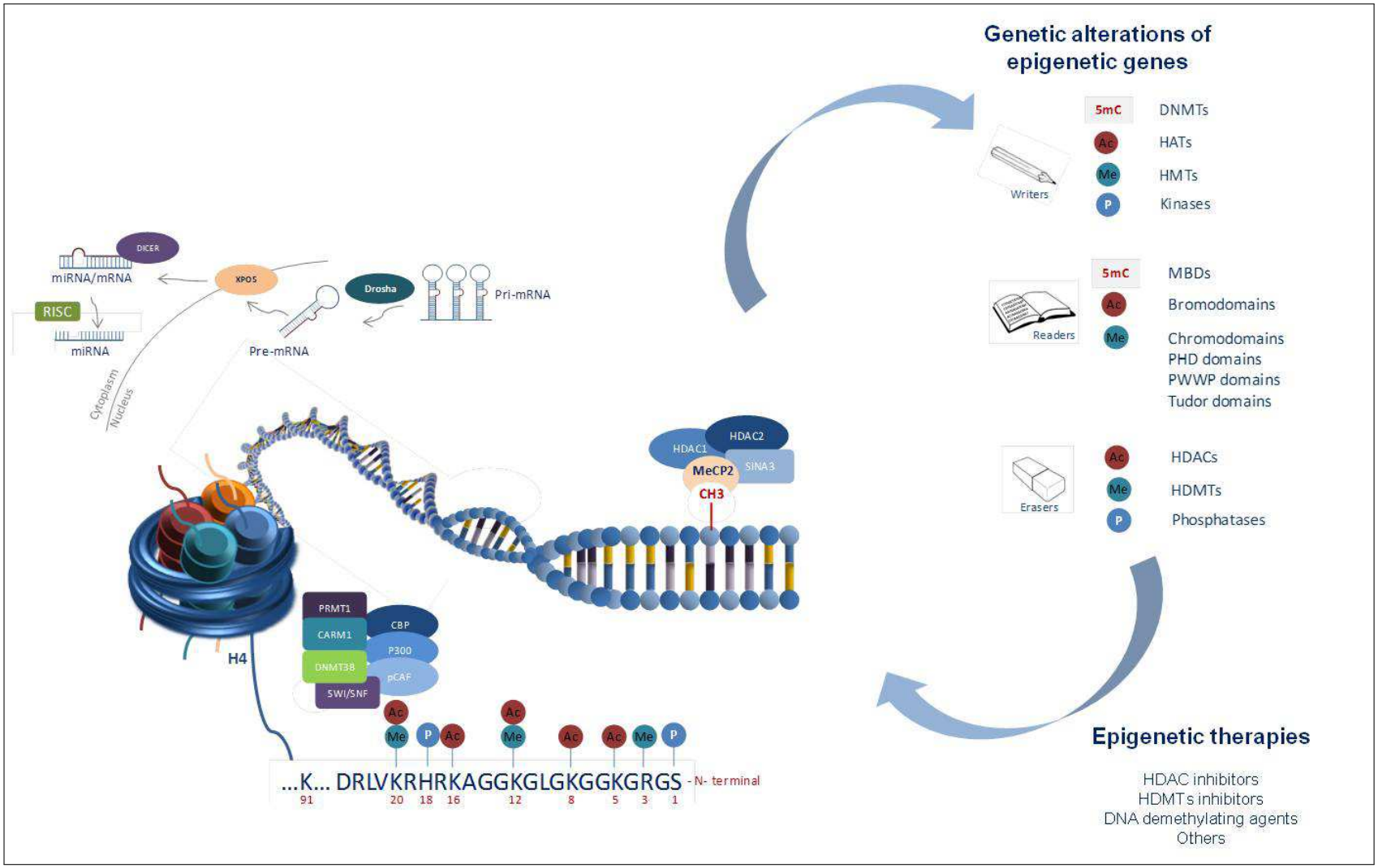


Figure 2.1: Epigenetic events - Simplified representation

Epigenetic events - the heritable changes in chromatin structure that affect gene expression, but not the DNA sequence, [Allis et al., 2007] - have been identified in the earliest stages of different cancers, [Baylin and Jones, 2011]. As *markers* for cancer initiation, epigenetic events are already exploited in some cancer treatment strategies, due to the *reversibility* property, (Section 2.7). Histone modifiers include writers/eraser, i.e. that add/remove a mark, (e.g. acetylse/deacetylase), and readers - that 'interpret' a marker, (e.g. Bromodomain), [Tarakhovsky, 2010]. Reprinted by permission from Berdasco and Esteller [2013]. "Springer and Springer-Verlag/ Human Genetics, vol. 132, 2013, pp. 359 - 383, Genetic syndromes caused by mutations in epigenetic genes, Berdasco, Mara and Esteller, Manel, Fig. 1, (Copyright 2013, Springer-Verlag Berlin Heidelberg) is given to the publication in which the material was originally published, by adding; with kind permission from Springer Science and Business Media"

mammals, ([Aguilar and Craighead, 2013] and references therein). The DNAm process is driven by the *DNA methyltransferase (DNMT)* family of *enzymes*[7], (reviewed in [Carey et al., 2011]). Specifically, *de novo* methylation is catalysed by DNMT3A, DNMT3B and DNMT3L, while DNMT1 is involved in maintaining DNAm patterns during DNA replication. Recently, other cytosine patterns, (i.e. 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC), [8]), have been also identified and potentially associated with the intermediate states during the DNA demethylation process, [Kriaucionis and Heintz, 2009; Ito et al., 2011; Pfaffeneder et al., 2011].

In cancer development, aberrant DNAm patterns have been found in two distinct forms: *hypo-(H-)* and *hyper-(H+)* methylation. DNA hypomethylation, an epigenetic event specific to cancer initiation, is associated with proto-oncogene activation, [Bjornsson et al., 2004; Allis et al., 2007], and chromosomal instability, [Herman and Baylin, 2003]. Both of these can promote cell proliferation and lead to an increase in cell predisposition to mutations. Of particular interest in cancer research are those 'CpG islands' from the promoter region of different TSGs. A *CpG island* is a genomic region of at least 200 base pairs, containing a high frequency of CG dinucleotides, (the CG percentage $> 50\%$, i.e. cytosine and guanine comprise more than half of the total nucleotides present), and with the Observed-to-Expected (OE) CpG ratio[9] $> 60\%$, [Gardiner-Garden and Frommer, 1987]. Normally unmethylated, CpG islands become targets for DNA hypermethylation during pathogenesis, commonly leading to unscheduled gene silencing, [Herman and Baylin, 2003; Bjornsson et al., 2004; Allis et al., 2007; Meissner et al., 2008].

[7] Enzymes refer to biological molecules that act as catalysts in biochemical reactions.

[8] 5hmC, 5fC and 5caC are DNA methylation variants resulted from sequential processes of methylation cytosine (5mC) oxidation, (regulated by the Tet (ten eleven translocation) protein group), as illustrated by the following relationships:

$$\text{5mC} \xrightarrow{oxidation} \text{5hmC} \xrightarrow{oxidation} \text{5fC} \xrightarrow{oxidation} \text{5caC, (dictated by Tet proteins).}$$

[9] The OE CpG ratio is calculated based on formula proposed in Gardiner-Garden and Frommer [1987]:

$$OE = \frac{CG\ number}{C\ number \times G\ number} \times Sequence\ length. \tag{2.1}$$

2.3.2 Histone modifications

Similar to DNAm, modification of histones influences changes in chromatin structure, [Jenuwein and Allis, 2001]. There are five known families of histones H1/H5, H2A, H2B, H3 and H4, grouped into two categories: the *core* group, (H2A, H2B, H3, H4) and the *linker* histones, (H1 and H5), [Ito, 2007]. In addition, several variations of these histones have been identified and denoted collectively as histone variants, (e.g. H2A.X, H3.3), (reviewed in [Sarma and Reinberg, 2005]). Histone modification (HM) may be cumulative and may take a number of forms, including *methylation*, *acetylation*, *phosphorylation*, *ubiquitylation*, and *sumoylation*, [Kouzarides, 2007; Suganuma and Workman, 2008]. The histone modification nomenclature[10] is usually by histone name, (e.g. H1, H3, H4), followed by the amino acid abbreviation and position in the protein, (e.g. K9 for Lysine at position 9), [Turner, 2005]. The type and number of modifications, (e.g. Me3 for trimethylation), can be specified also, [Turner, 2005]. For example, H3K27me3 refers to trimethylation of histone at lysine 27, [Cedar and Bergman, 2009].

Two of the most studied histone modifications are histone methylation and acetylation, where addition/removal of specific chemical compounds are driven by different protein families, including the histone methyltransferase (HMT)/ demethylases (HDM), and acetyltransferases (HAT)/ deacetylases (HDAC) respectively. Histone acetylation is characterised by rapid changes in the acetyl state at specific lysine part and histone methylation is defined as addition of maximum three methyl groups to lysine and arginine **residues**~[11].

The effect of HM depends on both the histone and the modification types. For example, H3K27me3 and H4K20me2 have been associated with transcriptionally inactive genome regions, (i.e. gene repression), while both acetylation and methylation on H3K4 have been linked to transcriptionally active chromatin, (i.e. gene activation), [Kouzarides, 2007]). Moreover, the trimethylation and acetylation of H3K9 have been associated with gene silencing and activation, respectively, [Koch et al., 2007].

[10] The Brno nomenclature is a standard system used for histones and histone modification notation.

[11] A residue is a small remainder from a substance, resulted following a chemical/ physical process applied to the considered substance.

2.3.3 Small non-coding RNAs

Small non-coding RNAs (sncRNAs), (including the three major categories: *small interfering RNA (siRNA)*, *micro RNA (miRNA)* and *piwi-interacting RNA (piRNA)*), are involved in different developmental phases and are considered to have a major role in protecting cells against various external influences such as viral infection, [Carthew and Sontheimer, 2009; Ghildiyal and Zamore, 2009; Mattick et al., 2009]. The sncRNAs are also implicated in directing the patterns of other epigenetic events, such as DNA methylation, [Carthew and Sontheimer, 2009; Ghildiyal and Zamore, 2009; Mattick et al., 2009].

MicroRNAs, (or miRNAs), are small sequences of RNA, ($\sim$ 19 - 22 nucleotides), which are involved in the cell cycle, (e.g. division, differentiation, apoptosis). They can act as TSGs or oncogenes and can be abnormally *down-* or *up-regulated* during disease development, [Negrini et al., 2009], as a consequence also of aberrant DNA methylation patterns, [Yang et al., 2009]. Over recent years, it has been shown that miRNAs can be considered *biomarkers* for malignant tumour stages, (including for CRC, [Schetter et al., 2011]).

2.3.4 Interplay between DNA methylation and Histone modification

It has been recently established that DNA methylation and histone modifications can influence one another, [Vaissière et al., 2008; Cedar and Bergman, 2009; Ikegami et al., 2009]. In Cedar and Bergman [2009], authors reported that histone acetylation depend on unmethylated DNA, leading to an open chromatin structure, while non-acetylated histones are associated with DNA methylation, determining more compact chromatin. It was also observed that DNA methylation level can be increased by histone methylation, [Cedar and Bergman, 2009]. Moreover, DNAm was found to inhibit some histone modifications, such as H3K4 methylation, [Cedar and Bergman, 2009].

In addition, different dynamics have been reported for DNAm and HM. For instance, DNAm is considered to be highly stable in nature, while HM occurrence can follow a much faster dynamic, (for example, histone deacetylase is a rapid mechanism).

2.4 The Intestinal Crypt

The *intestinal crypt* is considered to be the structural and functional unit of the intestinal tissues, (e.g. colon, small intestine). The small intestine and colon crypts present a number of both similarities and differences with respect to their structure, (e.g. cell location, number and cycle duration). The intestinal crypt cell population consists of *stem*, *progenitor* and *differentiated cells* in the colon, with the addition of the *Paneth cell* group in the small intestine.

2.4.1 Crypt Structure and Dynamics

Normal stem cells are undifferentiated cells that can divide in an unlimited way, in order to produce new stem cells, or differentiate into specialized cells, [Reya and Clevers, 2005]. Additionally, they have the property of conserving epigenetic patterns during cell division and transmitting these from generation to generation, [Probst et al., 2009]. A stem cell can divide either *asymmetrically*, (when a stem cell and a progenitor cell are born), or *symmetrically*, when two identical stem cells are produced, after which it is replaced by the daughter stem cell, [Humphries and Wright, 2008]. The lifecycle duration for a stem cell is linked to the time period for cell division, which has been determined to be approximately seven days for the human colon, [Potten et al., 2003; Frank, 2007], and five days for the small intestine, [Potten et al., 2003], respectively. The stem cell compartment is found at the base of the *colonic* crypt, [Bock, 2012], or proximate to the crypt base in the *small intestine*, [Clevers and Bevins, 2013]. Progenitors, which are located in the crypt to the left and right of the stem cell group, are more specialized than stem cells and undergo a limited number of divisions, while differentiated cells, (i.e. *enterocytes*, *goblet* and *endocrine* cells), occupy the upper part of the colon crypt and represent fully-specialized cells, [Khalek et al., 2010; Vaiopoulos et al., 2012].

In normal conditions, i.e. in healthy phenotypes, a major difference between the structures of the colon and small intestine crypts is given by the presence of *Paneth cells* at the base of the small intestine crypt, [Sancho et al., 2003], and their absence from the colon

crypt, [Nicolas et al., 2007], as illustrated in Figure 2.2. Paneth cells are differentiated cells that protect the stem cell group within the small intestine crypt against bacterial infections, [Umar, 2010]. Although Paneth cells can be observed sometimes in the *ascending*[12] colon, their presence in this tissue has been associated with a certain disease stage, [Humphries and Wright, 2008]. Unlike progenitors and differentiated cells that migrate to the top of the intestinal crypt, Paneth cells move closer to crypt base, [Kim et al., 2005]. The Paneth cell cycle has been determined to be equal to around 20-30 days, [Sancho et al., 2003; Clevers and Bevins, 2013].

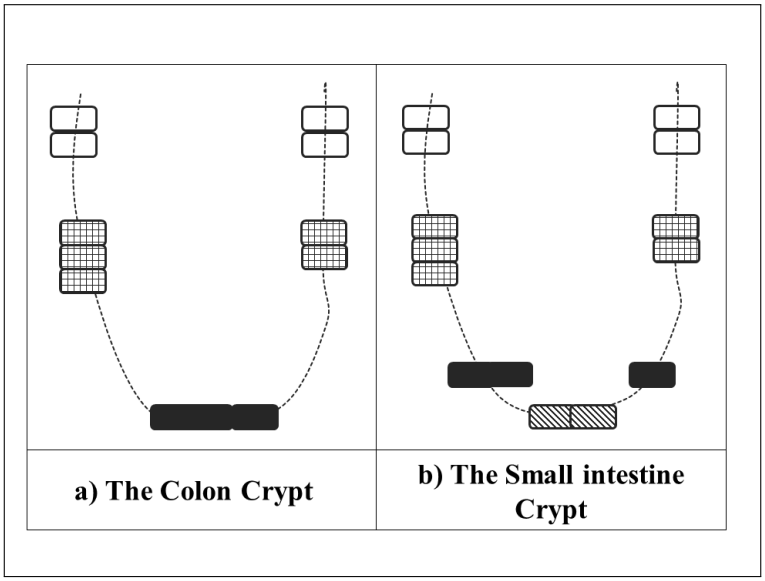


Figure 2.2: Simplified structure of a) colon and b) small intestine crypts - Comparison
Cells are illustrated by rectangles with different shading patterns: i) stem cell (dark-shaded rectangle) at the base/ close to the base of colon and small intestine crypts, respectively; ii) progenitor cells, (grid pattern shading), to left and right of stem cell group; iii) differentiated cells, (unshaded rectangle), at the top of the crypts; iv)Paneth cells, (diagonal-hashed shading), occur only at the base of the small intestine crypt. Image adapted from McDonald et al. [2006].

In addition, differences between the small intestine and colon crypts have been reported with regard to cell numbers. While the colon crypt contains around 16 - 19 stem cells, [Brittan and Wright, 2004; Khalek et al., 2010], within a total of 2000 cells, [Nicolas et al., 2007; Khalek et al., 2010; Vaiopoulos et al., 2012], the small intestine crypt is shorter, with

[12]The colon has four major subdivisions: the ascending, the transverse, the descending, and the sigmoid colon, ([Fritsch and Kühnel, 2008], page 202).

1-6 stem cells within a total of around 250 cells, [Sancho et al., 2003]. Abnormal cell population growth, (as a consequence of deregulation in cell cycle control mechanisms), has been associated with tumour pathways, [Humphries and Wright, 2008; Jin et al., 2009], and aberrant augmentation in the Paneth cell group has been related to Crohn's disease, (a major type of **inflammatory bowel disease~**[13] linked to increased risk for intestinal cancer development, [Canavan et al., 2006; Clevers and Bevins, 2013]).

2.4.2 Intra- and inter-crypt influences

Intercellular communication can occur between adjacent cells through 'gap junctions', (regulated by the *connexin* family). The *connexin gene* is considered to have a tumour suppressor property and downregulation of its expression, caused by aberrant epigenetics events, has been linked to deregulation in cell division, differentiation and apoptosis in several human cancer types, (reviewed in Vinken et al. [2009]).

Inter-crypt *interactions* have been reported also in the colon epithelium[14], where an aberrant crypt can influence its neighbours' **homeostasis~**, (i.e. the balance or stable state that characterises the internal environment of a normal biological system), by inducing deregulation in the local environment, [Humphries and Wright, 2008].

2.5 Malignant Systems: a brief overview

Cancer is caused by the deregulation of key genes that control cellular mechanisms, such as *division*, *differentiation*, *apoptosis*, and *movement*. Cellular processes are directly regulated by TSG and proto-oncogenes, (denoted as 'gatekeepers'). The activity of a TSG can be reduced in malignant cells by genetic mutations, insertions and deletions or epigenetic silencing, a condition that causes failure of apoptosis within abnormal cells, [Vogelstein and Kinzler, 2004]. In addition, aberrant oncogene activation can increase the formation of

[13] Inflammatory Bowel Diseases refer to disorders of the Gastrointestinal tract that cause intestine inflammation. Crohn's disease and ulcerative colitis are two major types of inflammatory bowel diseases, [Canavan et al., 2006; Clevers and Bevins, 2013; Merriam-Webster, 2014].

[14] Epithelium is one of the four major animal tissue types, (together with connective, muscle and nervous), and is characterised by high cell division rate and small intercellular gaps. The epithelium ensures functions of e.g. tissue protection, hormone secretion, nutrient absorption, sensation detection, [Merriam-Webster, 2014].

those proteins that dictate excessive cell growth, while alterations of DNA repair or *caretaker* genes allow a rapid accumulation of modifications in other genes that control the cell cycle, [Michor et al., 2004]. Finally, abnormal modifications of another gene category, i.e. *landscaper* genes, lead to increased cell susceptibility to malignant transformation, [Michor et al., 2004]. A simplified representation of interdependencies between genetic and epigenetic mechanisms and the influence of several risk factors in biological systems leading to cancer development is illustrated in Figure 2.3.

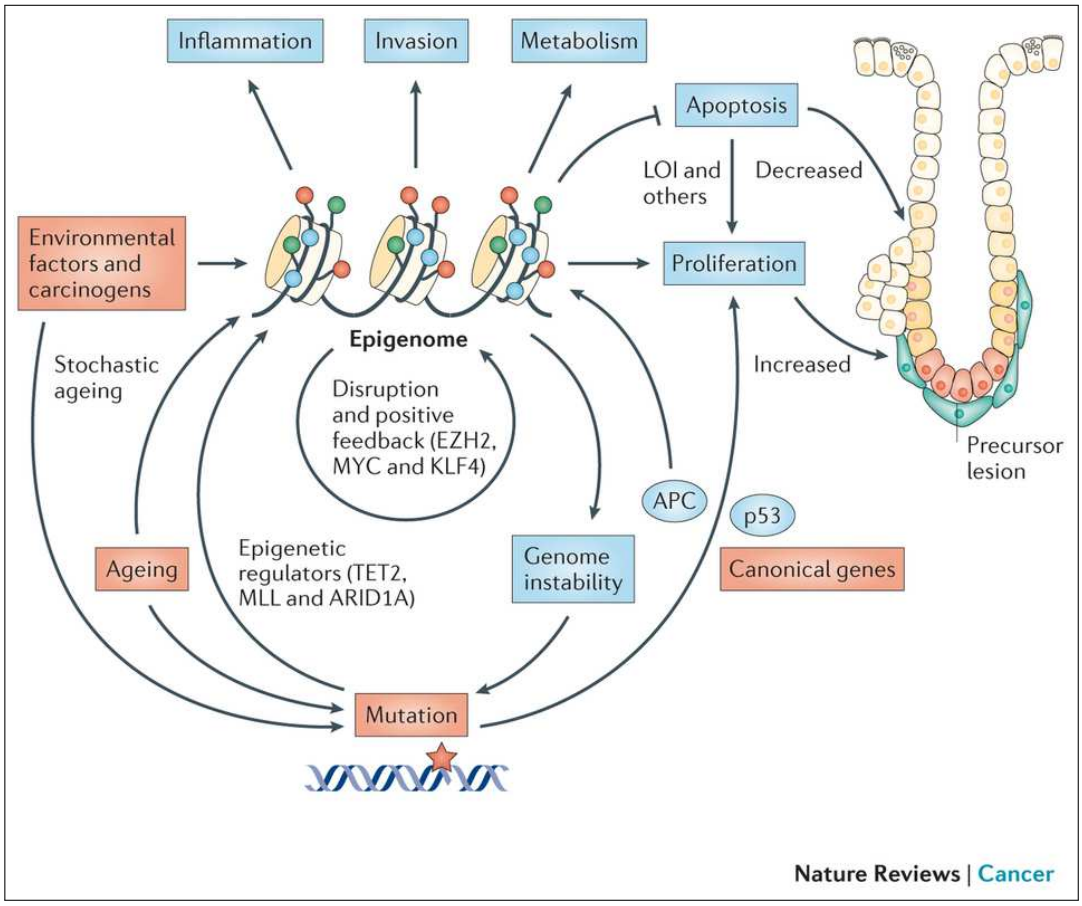


Figure 2.3: Malignant system features - Simplified representation
Aberrant genetic and epigenetic modifications of different genes deregulate the cell cycle causing abnormal cell proliferation and lead to cancer development. Tumour pathways are influenced also by different risk factors, such as ageing, gender, lifestyle and environmental features, (discussed in subsection 2.6.2). Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Cancer, ([Timp and Feinberg, 2013]), copyright (2013).

Signalling pathways Of particular interest in cancer research are those genes involved in signalling pathways. By definition, *a signalling pathway (SP)* is a chain of molecules that work together to control cell functions. An aberrant event recorded for the first molecule of a SP will induce abnormal activation on the second one. This process is repeated for all molecules of the SP and results, after activation of the last molecule in the chain, in loss of cell function. Studies of signalling pathways can help in the design of targeted therapeutic strategies that enable slowing down or halting of the malignant cellular modifications. Some of the well-known analyses of signalling pathways in cancer research include WNT/β-catenin, [Clevers, 2006], TGF-β, [Derynck et al., 2001], MAPK, Eph/Ephrin and Notch, [De Matteis et al., 2013]. Mutations in these pathways have a high impact on the development of various cancer types, (such as colon, breast, ovary, lung cancer, leukaemia), in both hereditary and **sporadic**~[15] forms, [Derynck et al., 2001; Clevers, 2006]. A comprehensive database of curated signalling pathways involved in human cancer development is the *KEGG PATHWAY* database, [Kanehisa et al., 2014].

Stem Cell hypothesis In recent years, the Cancer Stem Cell hypothesis, which postulates that the initiation and progression of cancer are associated with accumulation of aberrant genetic and epigenetic modifications within stem or stem-like cells, has become increasingly accepted as a plausible tumorigenesis hypothesis, [Boman and Wicha, 2008; Clevers, 2011; Papailiou et al., 2011]. Conversely, diminishing scientific support has been given to the traditional hypothesis of cancer development, which considers that every cell within the malignant population has the ability to promote disease, [Khalek et al., 2010; Molina-Pea and lvarez, 2012; Vaiopoulos et al., 2012]. Stem-like cells are abnormal cells that are changed during disease development and present similar characteristics to a normal stem cell, in terms of proliferation capacity and transgenerational epigenetic information transmission, [Johannes et al., 2009].

[15] Cancer sporadic refers to cases when individuals develop cancer, without having inherited DNA modifications that can increase malignancy risk.

Cancer Stages Cancer staging denotes the severity and degree of tumour development and extension, and provides a guideline for the medical profession in terms of identifying suitable treatment required. Thus, *Overall Stage Grouping* is a system used to describe cancer progression and classifies tumours/ disease status as I, II, III, and IV (as well as 0), where the most advanced cancer stage is denoted by IV. Stage 0 is referred to also as 'carcinoma in situ'. Defined as a cell having high potential to become abnormal, *carcinoma in situ* corresponds e.g. to *ductal carcinoma in situ* of breast cancer, to *Bowen's disease* in skin melanomas, or to *adenoma*[16] phenotypes in CRC. In addition, the designations *carcinoma* and *invasive carcinoma* correspond to Stages I, II, III, while *metastasis* refers to the Stage IV tumour phase, where cancer has spread throughout the body or to another organ. Thus, a tumour can progress from carcinoma in situ, to invasive carcinoma and, potentially, to metastasis.

2.6 Colorectal Cancer (CRC)

CRC development is a complex multi-step process associated with deregulations of both genetic and epigenetic events targeting stem cells in colon and rectal tissues. A brief description of several CRC features is included in this section. Information on CRC initiation and progression phases is provided, together with a summary of several *key*-genes identified in CRC and an overview of both sporadic and hereditary CRC forms. Further, 'crypt fission' and the 'bottleneck effect', two major intestinal crypt phenomena, for which an increase in occurrence rate has been associated with CRC development risk, are also described. Finally, the impact of several risk factors for CRC initiation and progression is discussed.

Benign to Malignant in CRC development Initially, cell growth rate increases in several crypts leading to **polyp˜** formation at epithelium's surface. Polyps are considered to be **be-**

[16]Colon adenoma = phenotype characterised by benign modifications in colonic epithelium, which can become malignant if not removed, leading to colon cancer.

nign˜[17], (i.e. Stage 0), but one type, (namely the *colon adenomatous polyps*), may increase their size and become cancerous over time, if not removed. Once presenting **malignant˜** characteristics, adenoma is referred as *adenocarcinoma*. As the disease progresses, cancer cells accumulate aberrant changes that facilitate cell *proliferation* within polyps and eventually *migration*. At Stage I, more of the inner colonic epithelium is involved, and in Stage II, the tumour affects nearby tissues, but not the **lymph˜** nodes; tumour extension to these is thus specific to Stage III. Finally, CRC can invade other organs, such as the liver and lungs[18], leading to **metastasis˜**. In addition, **recurrent˜** CRC includes those cases when malignant modifications reappear after treatment, affecting parts of the colon, rectum or other organs.

Key genes in CRC development Information on the characteristics of several key-genes identified in CRC development is provided in Table B.1, in Appendix B, together with principal references. The list includes TP53[19], (a TSG with crucial role in cell cycle control and whose mutation/deletion was detected in more than 50% of human diseases [Knudson, 2001]), APC[20], (abnormally changed in the earliest CRC stages [Suehiro et al., 2008]), RASSF1A[21], KRAS[22], BRAF[23] and MGMT[24], (highly mutated and hypermethylated in CRC, [Grady and Markowitz, 2002; Suehiro et al., 2008; Dworkin et al., 2009; Ahmed et al., 2013]), MCC[25] and MLH1[26], (in which alterations were associated with sporadic and hereditary forms of CRC, respectively, [Niv, 2007; Fukuyama et al., 2008]).

Sporadic and Hereditary forms of the CRC CRC presents in both *sporadic* and *hereditary* forms, [Cunningham et al., 2010; Al-Sohaily et al., 2012; De Matteis et al., 2013]. Most cases are sporadic, ($\approx$ 85% of total cases, [Half et al., 2009]), with major groups associated

[17] While benign refers to non-cancerous modifications/ features, malign indicates cancer presence.
[18] The liver and lungs are the most-affected organs by colorectal metastasis
[19] TP53: Tumour Protein p53, [Safran et al., 2010]
[20] APC: Adenomatous Polyposis Coli, [Safran et al., 2010]
[21] RASSF1A: Ras-associated domain family member 1, [Safran et al., 2010]
[22] KRAS: v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog, [Safran et al., 2010]
[23] BRAF: v-raf murine sarcoma viral oncogene homolog B
[24] MGMT: O^6-methylguanine-DNA methyltransferase, [Safran et al., 2010]
[25] MCC: Mutated In Colorectal Cancers, [Safran et al., 2010]
[26] MLH1: MutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli), [Safran et al., 2010]

with i) chromosomal instability, (CIN), and ii) hypermethylation at CpG islands of different gene promoters, (i.e. the *CpG Island Methylator Phenotype (CIMP)*). The hereditary category includes primarily i) *Familial Adenomatous Polyps (FAP)*, ii) *Lynch syndrome*, (or *Hereditary nonpolyposis colorectal cancer (HNPCC)*), and iii) *MYH-Associated Polyposis, (MAP)*, which has been recently detected and presents similar characteristics to FAP. Major characteristics of the CRC types are briefly summarised also in Table B.2, in Appendix B.

2.6.1 Major intestinal crypt phenomena

Two major phenomena, which are exhibited in the intestinal crypt cycle, are 'crypt fission', (i.e. longitudinal crypt division, producing two daughter crypts, [McDonald et al., 2006]), and the 'bottleneck effect'. *Crypt fission*, (caused by doubling the stem cell number at the crypt base, [Jin et al., 2009]), is considered to be a rare event in normal intestinal tissue, with time to occurrence estimated to be around 25 years for the human colon, [Graham et al., 2011]. Increased crypt fission incidence is considered to facilitate tumour expansion, [Brittan and Wright, 2004], and has been linked to colon adenoma risk, [Preston et al., 2003; Chen et al., 2005; Humphries and Wright, 2008]. Over an intermediate time-interval, (approximately 8.2 years for normal human intestinal tissue), the stem cell number decreases to a single stem cell, which is capable of regenerating the entire crypt, [Yatabe et al., 2001; Brittan and Wright, 2002; Sancho et al., 2003; de Lau et al., 2006; van der Flier and Clevers, 2009]. Denoted as the '*bottleneck effect*', this phenomenon also has been suggested to be a tumour promotion mechanism, given that - if mutated, the original stem cell transmits the abnormal changes to the entire cell population during cell division. Due to a growth advantage bestowed by mutations on an aberrantly changed stem cell, there is a higher probability for a crypt to be populated with cells descending from a mutated stem cell rather than from a normal one, [Leedham and Wright, 2008]. However, if the mother-stem cell is a normal cell, a healthy crypt will be regenerated.

2.6.2 Risk factors for CRC development

Cancer is considered to be a class of diseases that affects mostly older people, [Christensen et al., 2009], (e.g. more than 60% of diagnosed cancer cases are people aged 65 years and more), and is rarely found in children and young adults, [Cancer Research UK, 2014a]. As shown in Figure 2.4, CRC incidence increases during ageing, for both genders. CRC predisposition is also influenced by other characteristics, such as *gender*, [Brenner et al., 2007], *personal and familial cancer history*, [American Cancer Society, 2013], *physical activity*, *diet*, [Cancer Research UK, 2014d], and *environmental factors*, [Hou et al., 2012]. These features are referred as 'risk factors for cancer development', with potential to increase the chance for an individual to develop malignant tumours. Risk factors are typically denoted *personal*, (i.e. can not be changed), and *external* (i.e. lifestyle-related) categories; several of those associated with CRC development are discussed further.

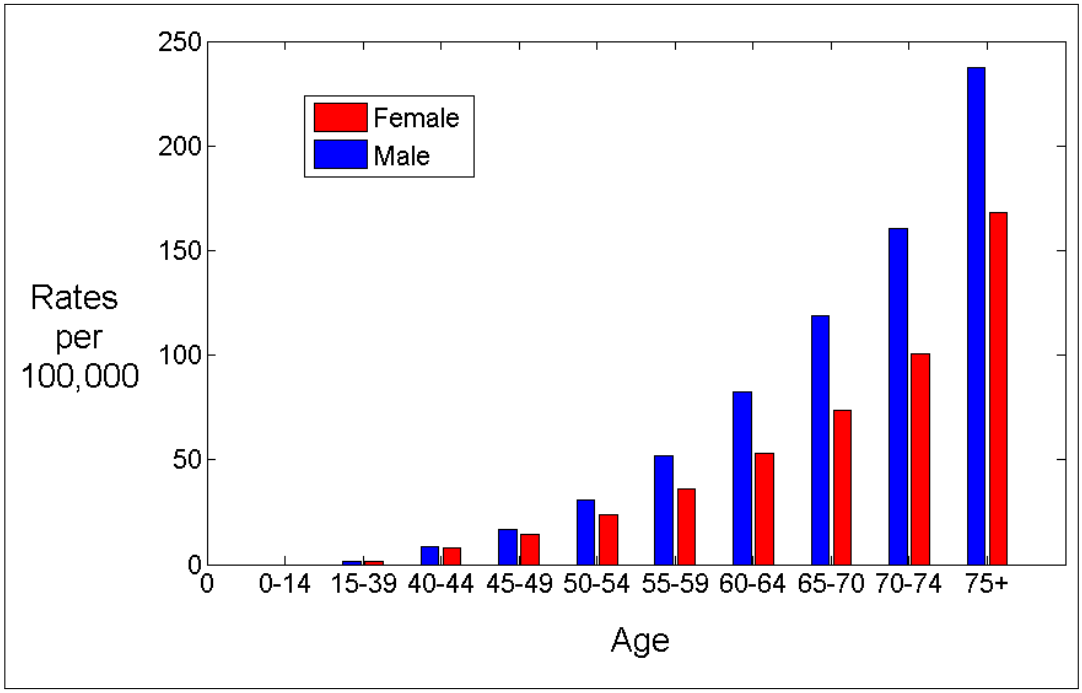


Figure 2.4: CRC incidence rate world-wide for both genders, i.e. males (blue color) and females (red color). The incidence rate of CRC development increases during ageing. In addition, male predisposition to CRC can be observed. Source: GLOBOCAN 2012 - International Agency for Research in Cancer (IARC) - http://globocan.iarc.fr/, [Ferlay et al., 2014].

Personal risk factors Due to influence on molecular signals (genetic and epigenetic), *ageing* is considered to be a major risk factor in cancer initiation and progression, [Fraga and Esteller, 2007]. Studies have reported that DNA methylation and histone modifications accumulate progressively over time, e.g. Fraga et al. [2007]. CRC incidence increases markedly after age 50, with the highest rate observed between the ages 65-75, [SEER, 2013a]. Four age groups, distinguished in cancer statistics, are given, (e.g. [SEER, 2013a; Cancer Research UK, 2014a; Ferlay et al., 2014]), based on common features observed in tumour initiation and progression, (with a focus on CRC **incidence rate˜**). These are age ranges (0 - 14) + (15 - 29): children and young adults; (30 - 49): adults in middle-age; (50 - 74): older adults; (75+): elders. The Table describing characteristics of these groups is given in Appendix B, (see Table B.3), and includes key study references.

Moreover, in recent decades, a range of experiments has been developed to investigate *age-related methylation* leading potentially to abnormal methylation patterns and malignant tumour development. A list of genes, identified as sensitive to age-related methylation in different cancer types, include, for example, *insulin-like growth factor II (IGF2)*, *myogenic differentiation 1 (MYOD1)*, *paired box 6 (PAX6)*, *tumor protein p73 (TP73)*, *secreted frizzled-related protein 1 (SFRP1)* genes in CRC, *Hypermethylated in Cancer 1 (HIC-1)* gene in prostate and brain tumours, and *estrogen receptor (ER)* and *N33* in colon and liver diseases, (Issa and Ahuja [2000]; Teschendorff et al. [2010]).

In addition, *gender* has been reported in the literature as important in colon cancer initiation, [Brenner et al., 2007]. Studies have shown that the risk of developing colon cancer is greater for males than for females, [Ogino et al., 2006b; Cancer Research UK, 2014a], with an approximate 5-year difference between genders suggested for initiation of CRC screening, [Brenner et al., 2007; Frank, 2007], (illustrated in Figure 2.4).

According to Knudson [2001], multiple successive 'hits' are required to transform a normal cell into an abnormal one: fewer mutations are thus needed to produce malignancies if some changes are already inherited. The impact of *heredity* in cancer can be measured by the *familial relative risk ratio*[27] coefficient. Heritable tendencies have been identified

[27]The highest RR was reported for breast cancer (RR=2.02), followed immediately by those for lung (RR

for cancers of the colon, breast, lung and others, [Risch and Plass, 2008], as well as for skin melanomas, [Hemminki et al., 2003]. While familial cancer is less common than spontaneous manifestations, genetic patterns are gradually being established, which may contribute to the identification of those at greater risk.

Finally, individual characteristics such as personal history of CRC and the presence of inflammatory bowel disease, are also associated with an increased risk of CRC development, [Canavan et al., 2006; Burt et al., 2010; Baumgart and Sandborn, 2012].

External risk factors The significant impact of viral and bacterial infections on human tumours has been highlighted in many recent studies, [Carrillo-Infante et al., 2007; Samaras et al., 2010]. Viruses, such as *human papilloma (HPV)*, *hepatitis B*, and *Epstein-Barr* have been associated with cancer initiation, [Carrillo-Infante et al., 2007], due to their capacity to damage normal cell control in the infected organisms by inducing unscheduled cell growth and by avoiding apoptosis. In Perrin et al. [2010], the authors have shown that *Helicobacter pylori* plays a key role in gastric cancer development owing to aberrant increase of DNA methylation level in gastric cells. Further, in Antonic et al. [2013], the authors reviewed several bacteria (e.g. *Streptococcus bovis*, *Helicobacter pylori*) and viruses, (e.g. *John Cunningham virus (JC virus)*), which have been associated with the CRC phenotype. The involvement of the *Fusobacterium* bacteria also in CRC was recently suggested, [Kostic et al., 2012; Antonic et al., 2013; McCoy et al., 2013].

Discussions on the impact of environment and lifestyle on cancer development have also featured in the literature, [Hou et al., 2012; Cancer Research UK, 2014g]. For instance, *tobacco usage* has been associated with a fifth of all cancer cases in the UK, according to a report published in December 2011, [Cancer Research UK, 2014f], and is considered to be a major risk factor in the development of malignancies, [Giovannucci, 2001; Sasco et al., 2004; Mucha et al., 2006; Zisman et al., 2006; Irigaray et al., 2007]. In addition, excessive *alcohol consumption* increases the probability of developing different cancer types, such as oral cavity, pharynx, oesophagus, liver, breast cancer, and is considered to be a high risk

$\approx$2.00) and prostate cancer with RR=1.89 (Risch and Plass, 2008).

factor in colon cancer, [Zisman et al., 2006], where it has been associated with aberrant changes in DNA methylation, [Davis and Uthus, 2004]. Moreover, lifestyle characteristics such as *lack of physical activity*, *high fat diet*, have been associated with an increased risk of CRC development also, [Slattery et al., 2003; Irigaray et al., 2007; Watson and Collins, 2011].

Finally, chemicals such as arsenic and cadmium have been highlighted as *carcinogens*[28] for CRC since exposure to these substances can induce epigenetic alterations in cells, (e.g. global DNA hypomethylation), leading to colon cancer, [Hou et al., 2012]. In consequence, studying the impact of risk factors on different molecular mechanisms can improve understanding of tumour pathways and aid in the search for treatment strategies, but is a major challenge.

2.7 Epigenetic therapies

Although first postulated by Waddington in 1942, epigenetic alterations have only become a focus of interest in cancer research relatively recently, due to their potential use in diagnostics and cancer therapy. Epigenetic drugs may be useful not only for cancer, but also for other human diseases, (e.g. **neurodegenerative disorders˜**, [Abel and Zukin, 2008; Urdinguio et al., 2009]). The aim is to find small chemical compounds that can bind specific epigenetic modifications, inhibit their function and reverse the epigenetic process, [Minucci and Pelicci, 2006]. Based on usage and activity, epigenetic inhibitors are classified into five major groups including: (i) DNMT, (ii) HMT, (iii) HDAC, (iv) HAT and (v) HDM Inhibitors.

Epigenetic drugs already approved by U.S. Food and Drug Administration include inhibitors for DNMT (e.g. vidaza and decitabine) and HDAC, (e.g. vorinostat and romidepsin), which are used for blood cancer, [Kaminskas et al., 2005; Richon, 2006; Rodríguez-Paredes and Esteller, 2011]. Additional drugs, which target for example HDM, HAT, HMT, are currently undergoing clinical and preclinical trials, (Lohse et al. [2011]; Santer et al.

[28]Carcinogen: an agent with direct involvement in causing cancer

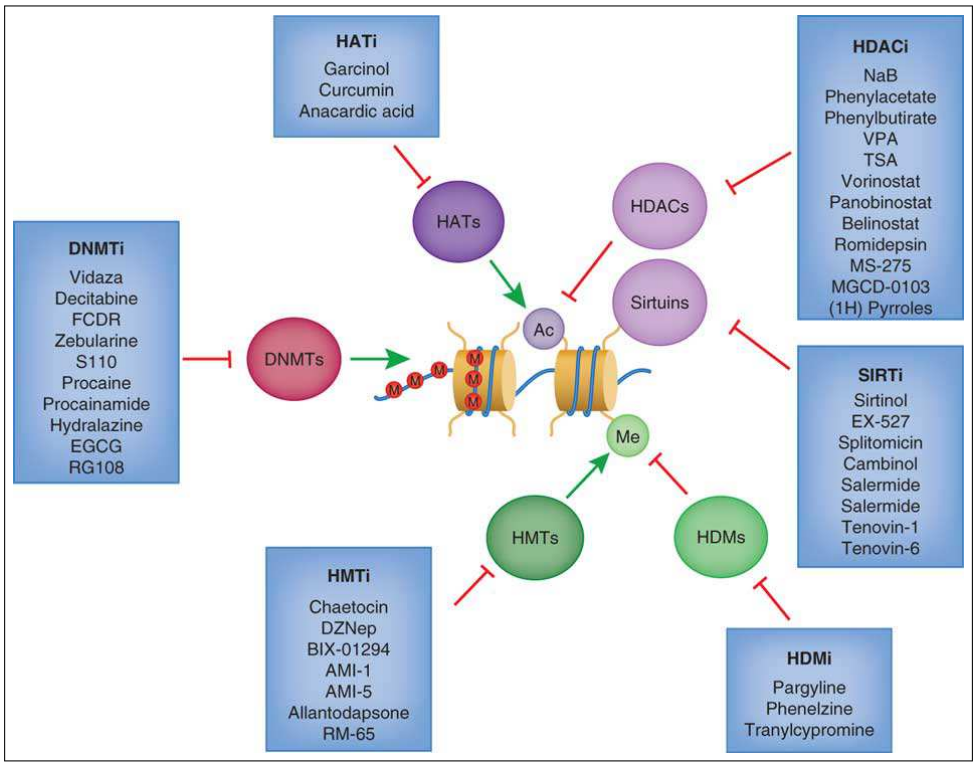


Figure 2.5: Epigenetic drugs and their epigenetic targets in cancer therapy. Note: suffix 'i' refers to 'inhibitor', (e.g. DNMTi = inhibitor for DNA methyltransferases, (DNMT)) and *SIRTi* refer to *sirtuins inhibitors*, a class of HDACi. Reprinted by permission from Macmillan Publishers Ltd: Nature Medicine, (Rodríguez-Paredes and Esteller [2011]), copyright (2011).

[2011]; McCabe et al. [2012], respectively), illustrated in Figure 2.5.

Epigenetic therapy is an active research area in pharmaceutical companies, where the focus is on development of both active formulations and strategies that aim to minimize the potential side-effects of epigenetic drug usage and extend patient survival, [Azad et al., 2013]. In addition, the combination of epigenetic therapy with other cancer treatment approaches, (e.g. standard **chemotherapy**˜), is anticipated to have an high impact on cancer mechanisms, [Azad et al., 2013]. Efforts are also ongoing currently for identification of combined epigenetic drug treatment that can be applied in a large number of malignant tumours as well as to other disorders with epigenetic alterations, [Stein, 2014].

2.8 Summary

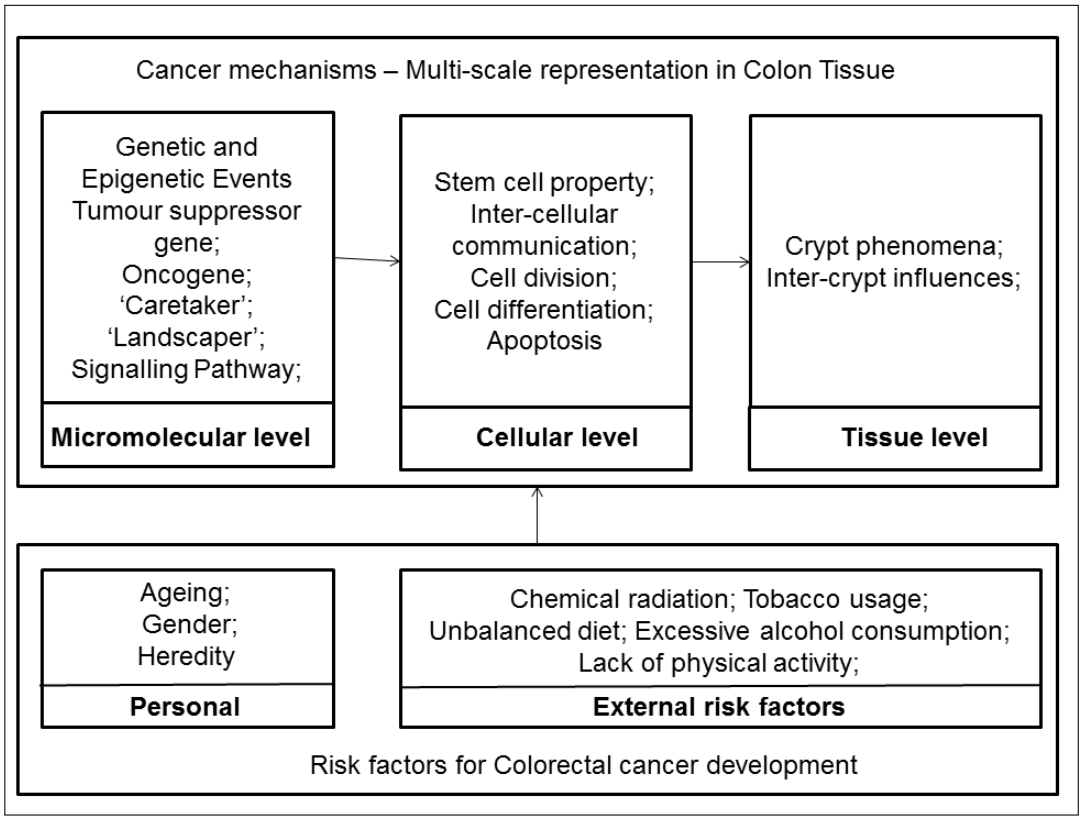


Figure 2.6: Overview of cancer mechanisms discussed

This chapter focused on presenting the biological background (on Epigenetics and CRC dynamics), which is fundamental to understanding the features to be represented in the development of appropriate computational models. Specifically, the aberrant mechanisms leading to CRC development were discussed for three major system layers, (illustrated in Figure 2.6), and including:

- micromolecular, (e.g. genetic and epigenetic events, signalling pathway);
- cellular, (e.g. cell cycle, stem cell property);
- tissue, (e.g. the intestinal crypt).

In addition, the influence of different risk factors, (e.g. ageing, lifestyle characteristics), was briefly reviewed in relation to CRC initiation and progression. The mechanisms de-

scribed in this Chapter can be applied to other cancer types as well. For example, *heredity* has been reported to have a high impact on breast and lung tumours; for lung, the risk of malignancy development can be increased by around 50% for people having a sibling or a parent diagnosed with lung cancer, with the risk marked in the former compared to the latter cases, [Cancer Research UK, 2014e]. Having a first degree female relative affected by breast cancer can double the chance of tumour development. In these conditions, screening tests are recommended to start several years earlier than for groups with no such familial history, [American Cancer Society, 2013; Cancer Research UK, 2014c; NICE - National Institute for Health and Care Excellence, 2014].

Major types of CRC, (e.g. sporadic, FAP), and the cancer staging system were also summarised. The state-of-art with respect to therapies based on epigenetic inhibitors, together with the challenges of epigenetics-based strategies in cancer treatment, were briefly outlined. The overview of major aspects of cancer, its development and treatment, is necessarily limited but aims to provide the platform required to motivate and inform the computational approach.

Chapter 3

Modelling background and context

3.1 Introduction

Cancer research has become a multi-disciplinary area involving collaboration among experts from a large array of scientific fields including biology, medicine, physics, chemistry, psychology, engineering and not least mathematics and computer science. The term of '*P6*' - i.e. Participatory, Personalised, Predictive, Preventive, Psychocognitive and Public[1] - has been proposed for describing the recent perspective of cancer medicine, [Bragazzi, 2013]. Computational and mathematical models can help by crossing the boundaries of *in vivo* and *in vitro* experiments caused by the complexity of the features, specific to malignant systems, and can facilitate understanding of the aberrant mechanisms involved in tumour pathways. Cancer dynamics exhibit a range of spatial and temporal scales; for example, while micromolecular modification can occur at nanosecond scales, cancer development needs several years or decades, [Deisboeck et al., 2011]. Computational simulations can permit exploration of carcinogenesis specifics in real-time, thus, contributing to advances in cancer research.

[1] participatory = keeping the patient continuously informed with regard to their disease status and available treatment options; personalised = the therapeutic program tailored to individual patient characteristics; predictive = predicting disease initiation and progression based on e.g. DNAm level information, (discussed in Section 3.4; preventive = informing on risk of tumour development based on different individual features, (e.g. familial cancer history), and monitoring their healthy state; psycho-cognitive = providing qualified guidance and psychological support to patients; public = sharing information related to diseases and treatments on Internet through specialised websites and organizations.

This chapter provides a review of computational approaches in Cancer Epigenetics and aims to establish the potential for development of hybrid multi-scale models for genetic and epigenetic dynamics in CRC. An overview of the major topics discussed here is illustrated in Figure 3.1, and the structure of this chapter is as follows. A brief description of the Cancer research context, (including recent world-wide initiatives, several commonly-used technologies and methods for DNAm and HM analysis and a set of Genetics and Epigenetics resources), is included in Section 3.2. *Concept formalism* and *examples* are provided for several computational and mathematical approaches, (including *Bayesian Network*, *Agent-based*, *Logistic* and *Multiscale modelling*), used to investigate aberrant modifications at different *scales*, (i.e. *micromolecular*, *cellular* and *tissue* levels), leading to disease development, (Section 3.3). These modelling approaches were considered from the perspective of their applicability to human intestinal systems. In addition, *epigenome-wide associated studies* - a recent trend in Cancer Epigenetics research - is addressed in Section 3.4. Finally, several issues for Computational Epigenetics, which are noted to arise in the context of CRC initiation and progression, are indicated in Section 3.5, and the Thesis focus stated. A brief summary is given in Section 3.6.

3.2 Cancer Research context: brief presentation

3.2.1 World-wide projects on Genetics and Epigenetics

Following complete sequencing of the Human Genome, the investigation of the *Human Epigenome* has become a major area of interest in Cancer research, [Collins et al., 2003; Eckhardt et al., 2004; Jones and Martienssen, 2005; Esteller, 2006]. The *Human Epigenome Pilot Project*[2], which revealed information on DNAm patterns in human chromosomes 6, 20 and 22 in different tissues, [Eckhardt et al., 2006], established the feasibility for developing *Epigenome* projects at large scale. Since then, a considerable number of large-scale initiatives, aimed at detection and interpretation of the epigenetic markers in normal and

[2]The Human Epigenome Pilot Project was launched as a scientific collaboration between the Wellcome Trust Sanger Institute (United Kingdom), Epigenomics AG (Germany) and The Centre National de Gnotypage (France), in early 2000. url: http://www.epigenome.org/index.php?page=pilotproject.

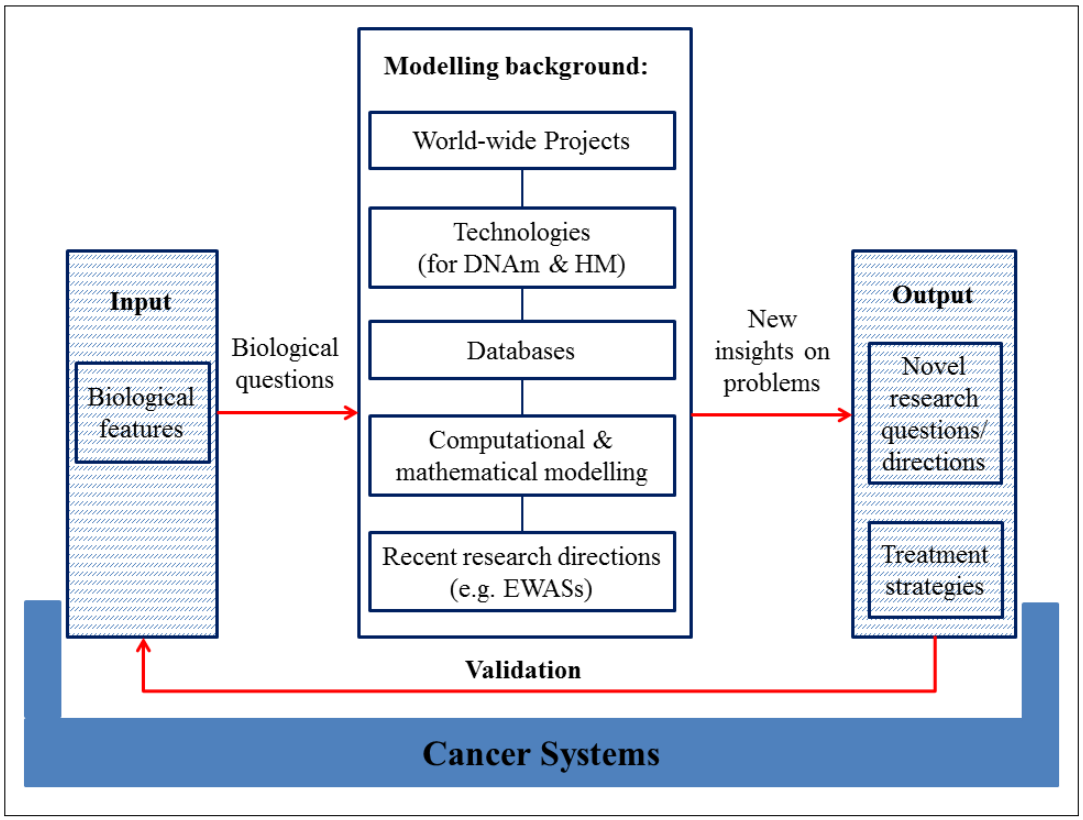


Figure 3.1: Overview of major topics related to Cancer Systems discussed in Chapter 3 The i) *Input* and ii) *Output* components correspond respectively to i) Research Questions & Objectives given the Biological context, (i.e. Chapters 1 and 2), and ii) Main findings, (addressed in Chapter 8).

malignant human systems, has been launched. Several of these world-wide efforts are presented briefly in the next paragraphs and information on a list of Epigenetics resources consulted for this Thesis is included in Table C.1, (Appendix C).

Large amounts of epigenetic data are becoming available and a recent direction in cancer research is represented by the identification of common genetic and epigenetic features among various cancer types and the integration of such micromolecular modifications in strategies for tumour progression, diagnosis and treatment, [Ashworth and Hudson, 2013]. For example, the *Encyclopaedia of DNA Elements* (*ENCODE*)[3] project focuses on the analysis of the functional elements, (e.g. histone modification), in the human genome, while the

[3]ENCODE project started in 2003 funding from the National Human Genome Research Institute; url: http://encodeproject.org/ENCODE/

Roadmap Epigenomics Project[4] and the BLUEPRINT Epigenome Consortium[5] aim to explore the epigenetic landscape in *stem* cells and *blood* tissues, respectively, [Bernstein et al., 2010; Adams et al., 2012; ENCODE Project Consortium, 2012]. In addition, epigenome studies in 250 cell types are being conducted under the direction of the *International Human Epigenome Consortium (IHEC)*[6] and comprehensive analyses on micromolecular modifications in around 50 cancer types have been performed by the *International Cancer Genome Consortium (ICGC)*[7], [Jones et al., 2008; Hudson et al., 2010]. Moreover, through the *The Cancer Genome Atlas* (*TCGA*)[8] efforts, detailed characterisations are available for the genetic and epigenetic markers involved in a wide array of human cancers, including the ovarian, lung, colorectal, breast, endometrial, and renal malignancies, as well as for acute myeloid leukemia, (e.g. Cancer Genome Atlas Research Network [2011, 2012a,b,c, 2013a,b,c]). As an illustration: from these data, a set of 127 significantly mutated genes in malignant systems has been identified in a study on 12 cancer types, (described in Kandoth et al. [2013]).

Translational medicine, which aims to reduce the gap between research and real treatment options and to improve the collaboration between academia and industry in order to achieve this, is another recent direction in Cancer Research. The *Innovative Medicines Initiative (IMI)*[9] is considered the largest initiative between *private* and *public* sectors within Europe and contains around 47 projects, including OncoTrack, among others, [Kamel et al., 2008]. The *OncoTrack*[10] project combines information on genetic and epigenetic events specific to CRC with diverse computational modelling strategies targeted to identification of *biomarkers* that can be used in personalised therapy regimes and also in prognostic strategies for CRC, [Elsner, 2011].

These world-wide initiatives also involve rapid advances in high-throughput technolo-

[4]Roadmap Epigenomics Project: url: http://www.roadmapepigenomics.org/

[5]BLUEPRINT-Epigenome Consortium (2011): url: http://www.blueprint-epigenome.eu.

[6]IHEC: (launched in 2010); url: http://ihec-epigenomes.net/.

[7]ICGC (launched in 2008): url:https://icgc.org/.

[8]TCGA Project (launched in 2005); url: http://cancergenome.nih.gov/.

[9]IMI was launched in 2008 under the European Union and the European Federation of Pharmaceutical Industries and Associations; url: http://www.imi.europa.eu/content/home.

[10]OncoTrack: url: http://www.oncotrack.eu/.

gies, (e.g. *microarrays* and *next-generation sequencing (NGS)*), used for genomics and epigenomics data analysis; thus, information on micromolecular events at genome-wide scale is now more accurate and more efficiently-obtained (from both cost and duration perspectives) than several years ago. However, differences between microarrays and NGS are evident and any choice of technology must consider limitations and advantages of application type. For example, while NGS can explore larger areas in shorter time-periods than microarrays, the former are more expressive than the latter, [Metzker, 2009]. Excellent recent reviews of the features of these technologies are included in Harrison and Parle-McDermott [2011]; Rivera and Ren [2013]; Soon et al. [2013] and references therein, and an overview of NGS-based technologies applied to epigenetic events patterns is illustrated in Figure 3.2. While full discussion of epigenetic technologies is beyond the scope of this Thesis, the next subsection aims to provide a summary of the most common techniques applied to investigate DNA methylation (DNAm) and histone modification (HM).

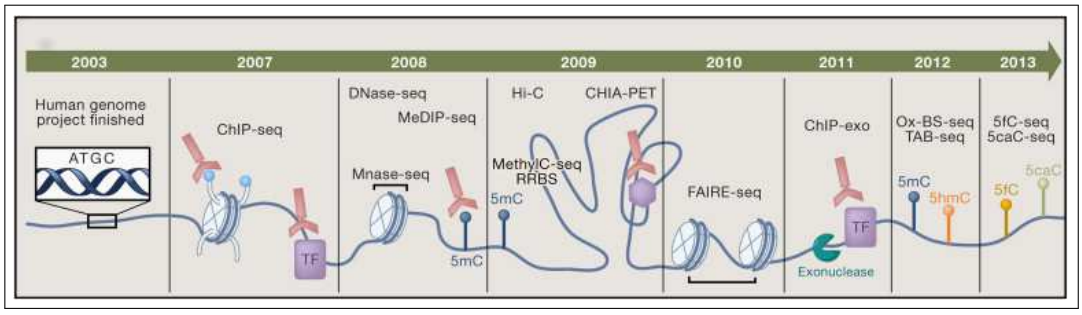


Figure 3.2: A summary of the NGS-based technologies developed for epigenetics analyses over the recent years. Reprinted from Cell, Vol. 155 (1), Rivera, C.M. and Ren, B., Mapping Human Epigenomes, Pages No. 39-55, Copyright (2013), with permission from Elsevier.

3.2.2 Techniques and methods for DNA methylation and histone modification assay

Given their importance both in normal human system operation, (e.g. cell differentiation), and abnormally perturbed systems, (e.g. cancer), DNAm and HM have been intensively studied from both quantitative and qualitative perspectives over recent years, [Coleman and Rivenbark, 2006; Soon et al., 2013]. Several of the principal technologies used to explore

these epigenetic mechanisms are described in the next paragraphs.

Methods for DNA methylation analysis DNAm analysis is highly dependent on the goal of the study, so that methods applied focus on two major directions, namely *gene specific* and *genome-wide methylation*, [Esteller, 2007; Shen and Waterland, 2007]. A classical method for global DNA methylation quantification is *High-performance liquid chromatography, (HPLC)*. Although highly-reproducible, HPLC requires large amounts of DNA, [Oakeley, 1999], so a most important development in DNAm analysis (and in cancer epigenetics), was the 'bisulfite' approach. Specifically, unmethylated cytosine from a DNA sequence is transformed to the corresponding uracil base after sodium bisulfite application and can be separated further by methylated cytosine, [Esteller, 2007]. In addition, the bisulfite treatment combined with e.g. the *polymerase chain reaction, (bisulfite PCR)* can approximate global DNAm levels using relatively little DNA, [Esteller, 2007], and can measure *PMR* (the *percentage of methylated reference* or degree of methylation) for a specific gene in malignant systems, [Ogino et al., 2006a]. A quantitative real-time PCR method (*MethyLight*) was introduced to assay the promoter methylation degree for different genes, (such as MGMT, MLH1, CDKN2A:p16 in colon cancer), [Ogino et al., 2006a]. Alternatively, the *PMR average* (*PMRA*) measurement can be applied in some experiments, where PMR values for a specific gene show variation across multiple runs, [Coleman and Rivenbark, 2006]. Recently, genome-wide DNAm investigations have proved possible through NGS-based methods such as *reduced representation bisulfite sequencing (RRBS)*, *MethylC-seq*, *methylated DNA immunoprecipitation, (MeDIP-seq)*, which benefit from the parallelisation strategies specific to NGS technologies, (reviewed in Rivera and Ren [2013]; Soon et al. [2013]).

Methods for Histone modification analysis Two major approaches used for HM analysis are I) *Mass Spectrometry*, which requires histone digestion, (reviewed in Witze et al. [2007]), and II) *Chromatin immunoprecipitation (ChIP)*, which can be applied for i) *gene-specific* analysis, e.g. ChIP-qPCR, or ii) *genome-wide* analysis, e.g. ChIP-chip, (ChIP

combined with microarrays), ChIP-Seq, (i.e. ChIP combined with NGS technologies). Of fairly recent introduction, ChIP-Seq is considered a powerful and efficient technique given its fast processing-time and genome-accessibility features, (reviewed in Schones and Zhao [2008]).

Large amounts of data on genomics and epigenomics are being generated in Cancer research and collective efforts are being made to integrate this information in specialised resources, with a view to dissemination through the scientific community. Several of these important databases related to cancer mechanisms are reported in the next subsection.

3.2.3 Databases on genetic and epigenetic events

Over the last few decades, databases that focus on aspects of tumour pathways have increasingly been developed and populated. These include major resources such as i) *Gene Expression Omnibus (GEO)*[11], ii) KEGG PATHWAY[12], and smaller dedicated efforts, developed by localised groups, such as iii) Embryonic Stem Cells Database, (ESCDb)[13] and iv) StatEpigen[14], and some of them are briefly described in the next paragraphs as such data constitute the primary support for information on cancer mechanisms, considered for the Colorectal Cancer Model presented in this Thesis. Information on a more comprehensive list of databases consulted for the current project is included in Table C.2, (in Appendix C).

GEO is a large public repository on genomic and epigenomic data generated by high-throughput technologies and contains two major components, namely *GEO DataSets*, (storing both original and curated data), and *GEO Profiles*, (providing only curated information). Data retrieval is facilitated by an advanced searching system that integrates attributes such as *author*, studied *organism*, type of the applied *technology* and targeted *gene* name.

KEGG PATHWAY is a database-component from the *Kyoto Encyclopedia of Genes and Genomes (KEGG)* resource, containing data on pathways involved in different human

[11] GEO is developed by the National Center for Biotechnology Information, US.

[12] KEGG PATHWAY is developed by the Kanehisa Laboratories, Japan.

[13] ESCDb is developed at the Bioinformatics, Algorithmics, and Data Mining Research Group, Estonia, [Jung et al., 2010]. url:http://biit.cs.ut.ee/escd/

[14] StatEpigen was developed by SCI-SYM Research Center, Dublin City University, Ireland.

diseases. KEGG resource includes a collection of databses that provide information at *molecular*, *gene* and *system* levels in addition to data on *drug* and *human diseases*, (e.g. KEGG GENE, KEGG DRUG, KEGG DISEASE, reviewed in Kanehisa et al. [2014]).

StatEpigen is a manually - curated database that has been designed to provide specific information on conditional relationships between genetic and epigenetic events affecting various genes at different pathology phenotype levels, [Barat and Ruskin, 2010]. Although currently specialised to colon cancer research, StatEpigen also contains data on other cancer types, (such as stomach, lung, liver).

In this section, an overview of the context, world-wide of Cancer Epigenetics research in recent years, was provided. In addition to improving technologies and generating micromolecular event data, a number of computational and mathematical models has been/ is being developed in order to explore human cancerous systems. The next section elaborates on some of these modelling approaches, applied to biological systems, but specifically as these can be used to address CRC development.

3.3 Previous Model development and related analyses

Computational and mathematical approaches can be applied to translate biological questions into formal algorithmic languages and to explore various hypotheses related to both normal and abnormal biological systems. This can help understand the aetiology of complex diseases and their evolution and to propose new research directions and therapeutic strategies. Major classes of mathematical and computational models include i) discrete versus continuous (depending on the way of changing variable states, i.e. at specific time points or continuously over time), ii) static versus dynamics, (given by system property of evolving over time), iii) deterministic versus stochastic (probabilistic) models; (if a model does not include any probabilistic component is considered deterministic; otherwise, it is a stochastic model). Stochastic models are used for prediction and estimation, (given that the outcome is random), and are developed using methods such as Markov chain, Monte Carlo,

Bayesian network, (reviewed in Wilkinson [2009]).

Various modelling approaches, (including construction of Boolean and artificial neural networks; support vector machines and K-nearest neighbours; Bayesian methods and decision trees; agent-based and multi-scale modelling), have been applied previously to investigate the complexity of tumour phenomena, (reviewed in e.g. Bock and Lengauer [2008]; Tracqui [2009]; Lowengrub et al. [2010]; Lim et al. [2010]; Rejniak and Anderson [2011]; De Matteis et al. [2013]; Johnson et al. [2013]; Wang et al. [2014]). Details for several of these, (including Bayesian network, agent-based, logistic and multi-scale models), are presented briefly in this section, as these constitute the basis for the computational components developed in this Thesis. Specifically, in targeting the ultimate objective of investigating colorectal *tumorigenesis* by linking microscopic effects to macroscopic outcomes, the Bayesian network is the fundamental layer of the E-G Network Model, (Chapters 4 and 5), the agent-based model approach describes the dynamics of the AgentCrypt model, (Chapter 6), and logistic modelling features have been integrated in LogisticCrypt, (Chapter 7). Examples of models, based on similar principles, that have been applied to other biological systems are also provided.

3.3.1 Bayesian Network

Concept formalism A *Bayesian network, (BN)*, is effectively a probabilistic graphical model that describes joint and conditional probability distributions for a set of variables X = (x_1, x_2, .., x_m). Formally, P($x_i \mid x_j$) is interpreted as the probability of x_i given x_j, (x_i, $x_j \in$ X), [Heckerman, 1998; Lucas et al., 2004]. The network is graphically represented by a *directed acyclic graph*, where each node represents a variable or group of variables. A directed edge, (i.e. *an arrow*), between two nodes signifies conditional dependence between the variables, [Heckerman, 1998; Lucas et al., 2004]. Additionally, the absence of an arrow between two nodes indicates that no reliable information is available on the conditional relationship between those variables, [Lucas et al., 2004]. A node n_i is defined to be the *parent* for node n_j if there is a direct influence between n_i and n_j, illustrated by an arrow $n_i \longrightarrow n_j$ in graph, (with the set of parents for the node n_j denoted as *Parents(n_j)*).

Consequently, the node n_j is referred as a *child* for n_i. The probability values of the conditional dependencies between a node child and its parents are defined by a *conditional probability table*. A BN follows the Markov assumption such that a variable depends only on its parents, i.e. is independent of the rest of the variable set. Thus, the formula used to describe the joint probability among a set of variables, X = (x_1, x_2, .., x_m), can be simplified as follows:

$$P(X) = \prod_{i=1,m} P(x_i \mid Parents(x_i)) \tag{3.1}$$

where the index i = 1..m runs over all *m* nodes in the BN. An example of a (very simple) BN for a set of five variables, (A, B, C, D, E), is illustrated in Figure 3.3. Considering formula 3.1, the probability for the given variable set is:

$$P(A, B, C, D, E) = P(C|\ A) \times P(D|\ A,\ B) \times P(E|\ D) \times P(A) \times P(B) \tag{3.2}$$

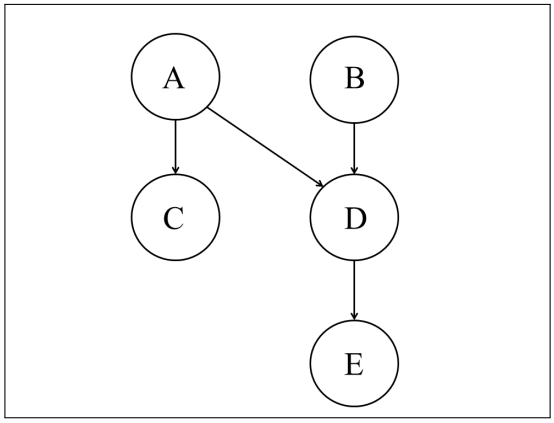


Figure 3.3: An example of a small Bayesian network

The BN approach can encode *conditional relationships* between genes, allowing intuitive representation for input data, as well as handling noisy and incomplete data sets, (common to biological systems). In addition, BN can handle continuous, (non-integer) values in the [0, 1] range, while a Boolean network can integrate only discrete (and binary) values. However, during framework development for any realistic problem, a BN requires initial knowledge of many probabilities, [Heckerman, 1998], and it is limited by exclusion

of cycles, (see definition); so that, during development, methods for cycle avoidance and removal must also be considered. Nevertheless, *ab initio* models based on BN-approaches have been developed to investigate diverse research questions and hypotheses. Some examples follow.

Examples of BN models BN models have been applied successfully in different fields requiring diagnosis or forecasting, e.g. Heckerman et al. [1995]; Lucas et al. [2004], and where prior knowledge on data causality was available. In biomedicine and bioinformatics, such approaches have been used for prediction of i) protein-protein interactions in *yeast* and in other model systems, based on both homogeneous and heterogenous data types, (e.g. Jansen et al. [2003]; Troyanskaya et al. [2003]; Browne et al. [2009]); ii) functionality for a newly-identified set of proteins in the *Drosophila* system, [van Bemmel et al., 2013], and iii) mutation interdependencies in different cancer types, [Gerstung et al., 2009], amongst others. In addition, BNs have been constructed to study conditional relationships between different protein types found in chromatin, (e.g. van Steensel et al. [2010]), in various histone modifications, (e.g. in *S. cerevisiae*, [Cui et al., 2011]), between histone methylation and gene expression, (e.g. Yu et al. [2008]), between gene expression levels, [Friedman et al., 2000; Hartemink et al., 2002; Pe'er et al., 2002], and to compare HM and DNAm patterns at CpG islands, (e.g. Kim and Jung [2009]; Lv et al. [2010]). The BN approach has been applied also for improving protein-protein interaction prediction using Gene Ontology hierarchies, [Wang et al., 2010], for estimating DNAm levels using data from experiments based on the methylated DNA immunoprecipitation technologies, [Down et al., 2008], as well as for identifying miRNA profiles specific to certain cancer types based on comparative analysis of 51 solid cancers and leukaemia, and corresponding samples from normal tissues, [Volinia et al., 2010]. Moreover, BN models have been used to address cancer type-specific hypotheses; for instance, in breast cancer, for predicting *microcalcifications* [15], (e.g. Burnside et al. [2006]), assisting mammography results and investigating cancer risk development, (e.g. Rubin et al. [2004]; Cruz-Ramírez et al. [2007]; Maskery et al.

[15]Tiny abnormal flecks that usually appear in the female human breast and can be a cancer indicator

[2008]).

3.3.2 Cellular Automata & Agent-based Model

Given the relationship between Cellular Automata and Agent-based Modelling, the computational approach is described for both as follows.

Concept formalism The *Cellular automaton (CA)* is a discrete model usually visualised as a set of identical cells, which are distributed on a one, two or many-dimensional grid and can be found in a discrete and finite number of states. At each time-step, every cell changes its state as a result of a transition rule, which takes into account its previous state and information received from its neighbours. One well known and early 2D CA application is Conway's "Game of Life", where cells can be found in two states, namely *alive* and *dead*, and have eight neighbours each, [Conway, 1970]. In this example, the state of a cell is decided based on the number of its neighbours; for example, a live cell with less that two neighbours dies, while one with exactly two lives. Obliviously, there are many interpretations of this Game and the basic concept was used at an early stage to describe e.g. competition among species, different models of propagation and similarly, (e.g. Silvertown et al. [1992]; Sirakoulis et al. [2000]). Abundant examples occur in nature, in seashell patterns, leaf gas exchange and other plant dynamics, in physical phenomena such as avalanche and sandpiles formation, (e.g. Peak et al. [2004]; Kronholm and Birkeland [2005]; Cervelle et al. [2007]; Coombes [2009]). Indeed the question has also been raised at to whether The Universe is a CA and there are a number of supporters of CA-physics. Importantly, the cellular automata concept has also found many applications in information theory and at Biology - Computer Science interface, whether e.g. in database searching, in evolutionary & genetic algorithms, in detection of errors, (reviewed in Ganguly et al. [2003]). CA models are deterministic models and a significant extension was introduction of stochastic CA, where rules are updated based on probabilities. Stochastic CA models have been developed recently to describe e.g. cell dynamics and infection spread, [Garijo et al., 2012; Duan et al., 2013]). In consequence, the flexibility of the CA approach for

multi units/ cells is well-established.

An extension to CA concepts, *agent-based modelling (ABM)* refers to the representation of a system by its constituent basic units to which attributes can be assigned. ABM is a discrete and dynamic approach that describes interactions between individuals, taking into account also their features and environment influences. Specifically, ABM can be applied to heterogeneous complex systems, described by the following typical characteristics: i) large number of different types of entities, (denoted as *agents*); ii) agent *interaction*, (or 'social' ability): agents can exchange different types of information, (e.g. information on their location, internal state); iii) agent *autonomy*, (decision-making capability): agents can operate without intervention from user, i.e. they have a degree of self-determination with respect to their actions; iv) agent *reactivity*: agents can respond to stimuli received from their environment, (e.g. can change their position) and v) agent *adaptation*: agents follow an adaptive process, (or self-learning capability), [Bandini et al., 2009; Macal and North, 2010] and references therein.

In biology, there is often a need to describe basic processes that direct individual actions, (e.g. cell division, migration), and to aggregate these in order to understand higher level dynamics, (e.g. tumour spread to other locations). ABM provides a natural manner to characterise complex heterogeneous system behaviours following a bottom-up approach and can also integrate emergent phenomena. In addition, ABM is flexible as agents can be both 'rational' and 'adaptive', (can evolve) and can generate new hypotheses. Disadvantages of ABM include the need of detailed knowledge of individual behaviour as well as of assumptions on which behaviour aspects are crucial for overall system investigation and which can be ignored. In addition, considerable computational expertise is required for ABM development and significant time has to be allocated for computational simulation run, given that large systems of e.g. thousands or million of agents can be also modelled using this approach.

Examples Over recent decades, CA, (extended form to permit more than two possible states for cells) and ABM methods have been applied to explore dynamics of different biological systems including immune systems response to viral infections including HIV[16] and *M. tuberculosis*, (e.g. Mannion et al. [2000]; Ruskin et al. [2002]; Segovia-Juarez et al. [2004]; Perrin et al. [2006a,b]), drug dissolution, (e.g. Barat et al. [2006]; Bezbradica et al. [2014]), and human tumour growth, (reviewed in Hwang et al. [2009] and references therein). In addition, ABMs have been built to analyse e.g. cell dynamics in systems with different calcium concentrations, [Walker et al., 2004], abnormal stem cell proliferation caused by APC mutation, [Boman et al., 2001], aberrant methylation and gastric cancer initiation induced by bacterial infection, [Perrin et al., 2010], and intestinal crypt dynamics, [Pitt-Francis et al., 2009]. An excellent review for intestinal crypt and CRC applications, drawing on these modelling techniques, is provided in De Matteis et al. [2013].

3.3.3 Logistic Model

The growth rate of a biological population in an 'ideal' environment, (e.g. no limitations of food or space), is proportional to population size, i.e. a certain number of new individuals is produced per unit of time. In general, this rate is described by simple differentiation equations, (first and second order), with population size represented by e.g. exponential forms, (unconstrained natural growth):

$$dP/dt = r \times P \tag{3.3}$$

$$P(t) = P_0 \times e^{rt}$$

where P = population as a function of time t, P_0 = population at time t = 0, and r = the proportionality constant. In a logistic model, the aim is to define the relations between the constituent elements in terms of how these affect e.g. growth or decline, constraints given by e.g. limited space and food resources. *Inter-* and *intra*-specific competitions refer, respectively, to interactions between more than one and within one species using the same

[16] HIV = human immunodeficiency virus

resources. Consequently, in the present instance, we define i) intra-specific competition in a population of cells as relationships between current and maximum cell population size, and ii) inter-specific competition between more cell type populations in the intestinal crypt as relationships between current and maximum cell numbers of the whole crypt, (i.e. between space, currently-occupied by compound cell populations, and crypt capacity). Of interest, from the viewpoint of work presented here, is the potential to use logistic model to explore deregulations of cell population size over time leading to increased incidence of major crypt phenomena such as 'crypt fission' and the 'bottleneck effect', (Section 2.6), associated with CRC development risk.

Concept formalism The population with *intra-specific competition* has been described mathematically by *the density-dependent coefficient, (DDC)*, [Sinclair and Pech, 1996], using equation 3.4

$$DDC = \frac{K - N}{K} \tag{3.4}$$

where N is the population size at a certain time, K is denoted as the *carrying capacity* (and represents the maximum allowed size for a species within a specific environment, i.e. has a constant value in a specific environment), [Hui, 2006]. The rate of population growth in a system with intra-specific competition is given by the product r $\times$ DDC, where r is the population growth rate in a similar system *without intra-specific competition*; DDC is calculated as for Equation (3.4), and takes a value in the [0, 1] range, (given $N \leq K$). Thus, by including the DDC, the population growth rate slows down as population size (N) increases, i.e. the growth rate depends on the density of the population.

Examples Mathematical models have been built to examine abnormal crypt dynamics, (reviewed in Van Leeuwen et al. [2006] and references therein), [Nowak et al., 2002; Johnston et al., 2007], for cell migration mechanisms, (e.g. Simpson et al. [2007] and references therein), and to investigate stem cell growth, determined by de-regulation in methylation patterns over time, [Yatabe et al., 2001]. Moreover, considering cancer as a heterogeneous

complex system (having multiple species), a class of mathematical models, which focuses on both *intra-* and *inter*specific competitions, has been applied in studies for gene therapy, [Tsygvintsev et al., 2013], tumour-immune cell relationship, [Nagy, 2004], carcinogenesis mutations, [Foryś, 2009], and colonic crypt population growth, [Smallbone and Corfe, 2014].

3.3.4 Multi-scale modelling

Concept formalism Multi-scale modelling have attracted increased attention in recent years as further improvements in computing power have offered the potential to view systems rather than just components of these. As such, the Systems Biology approach to cancer research, with the ultimate aim of facilitating investigation and treatment of malignancy is gaining momentum. Multi-scale descriptions span the micromolecular level, (i.e. genetic and epigenetic landscape), through molecular and cellular levels, (e.g. signalling pathway and cell division mechanisms, respectively), to macroscopic scales, (where tumour development can be observed), [Deisboeck et al., 2011]. Thus, major advantages of the multi-scale modelling approach include the possibility of including experimental data at all system levels, exploring inter-scale phenomena interactions and testing drug effects on the whole body or organ. Although considered a powerful approach, multi-scale computational modelling presents many major challenges, including the complexity of different temporal and spatial scales as well as multiple parameters required for system description. Specifically, both rapid dynamics and small spaces, (with alterations that can occur at nanosecond scales inside cell nucleus), and low dynamics and large spaces, (e.g. metastasis affecting more than one tissue, phenotype change over years or even decades), must be accommodated. Given these demands, different modelling strategies have been used to describe biological phenomena at various scales. Specifically, network-based models have been developed to characterise the micromolecular level, ordinary differential equations have been used for investigations at molecular levels, while both partial differential equations and the agent-based modelling approach have been applied for exploring mechanisms at cellular level. Finally, continuous models have been built to describe cell population dynamics at tissue

level. In consequence, “one size fits all” is an unlikely hypothesis.

Examples Multi-scale models have been developed to explore e.g. the therapeutic impact of cytostatic agent use in cancer development [Ribba et al., 2006], tumour *angiogenesis*[17], dynamics of avascular tumour progression over time, alterations during the development of lung, breast and prostate cancers and brain tumours, (reviewed in Mantzaris et al. [2004]; Peirce [2008]; Tracqui [2009]; Deisboeck et al. [2011]; Chakrabarti et al. [2012] and references therein). In CRC research, multi-scale computational models have been developed for exploring e.g. the *aberrant crypt foci*[18] expansion in both normal and abnormal colon systems, or relationships between Wnt signalling pathway, cell neighbour influences and cell cycle activity in intestinal crypts, [Van Leeuwen et al., 2009; Figueiredo et al., 2013]. In a recent multi-scale model for Epigenetics research, [Przybilla et al., 2014], the authors conclude that alterations in gene transcription and epigenetic inheritance mechanisms may be a cause of decline observed during ageing in hematopoietic stem cells. In consequence, developments are exciting, but non-trivial to achieve.

3.3.5 Complexity of Interdependent Epigenetic Signals in Cancer Initiation

The Colorectal Cancer Model, described in this Thesis, was proposed as part of the *Complexity of Interdependent Epigenetic Signals in Cancer Initiation, (CIESCI)*, project, a scientific collaboration between three main partners: the Centre for Scientific Computing and Complex Systems Modelling, (SCI-SYM)[19], School of Computing, Dublin City University, the Bellvitge Institute for Biomedical Research, (IDIBELL)[20], Barcelona, (Spain), and the Bioinformatics, Algorithmics, and Data Mining Research Group (BIIT)[21], Institute of Computer Science and Estonian Biocenter, University of Tartu, (Estonia). The CIESCI project aimed to investigate interdependencies between epigenetic events related to can-

[17] angiogenesis = formation of new blood vessels from vessels that exist already; it has been associated to benign to malign transformation.

[18] Aberrant crypt foci is a colon tissue phenotype prior to adenomas occurrence that detected with screening methods.

[19] SCI-SYM: url: http://sci-sym.dcu.ie/.

[20] IDIBELL: url: http://www.idibell.cat/.

[21] BIIT: url: http://biit.cs.ut.ee/.

cer initiation, incorporating three main elements: i) laboratory experiments (carried out at IDIBELL); ii) data mining (establishing pipeline and analysis of new data at BITT); iii) computational models (developed at SCI-SYM).

The SCI-SYM contribution to this project is also linked to other projects on computational modelling for molecular events in malignant diseases as well as in-depth bioinformatics investigations of curated data from primary and secondary sources, (represented by the StatEpigen database described in Section 3.2.3). In the former, an agent-based model was developed by DCU and collaborators to determine the risk of gastric cancer from consideration of the aberrant DNA methylation level induced in cells, following infection with Helicobacter pylori, [Perrin et al., 2010]. In addition, a framework has been built to give better insights on interdependencies between micromolecular events, (such as DNAm and HM), in abnormal situations, [Raghavan and Ruskin, 2011].

Computational and mathematical modelling has been promoted strongly as a powerful field in Cancer Epigenetics research, [Bock, 2012]. Given recent technological progress, computational models are needed to explore features of biological systems, which incorporate large amounts of data, and to highlight evidence of tumour development. Several modelling approaches were outlined here, (with a focus on intestinal tumours), demonstrating that investigation can be performed at single or multiple scales, for specific or general systems. In additional to offering new insights, results from computational analyses of initial queries can indicate new therapy strategies and research directions.

3.4 Epigenome-Wide Association Studies

One such new direction is that of *EWASs* or *Epigenome-Wide Association Studies*. Major features and limitations of EWASs are briefly reviewed in what follows, providing motivation for the cross-comparative computational analysis on DNAm in intestinal tissues, (Chapter 6).

EWASs aim to detect genome-wide differences in epigenetic marks that can be corre-

lated with predispositions to cancer development, [Rakyan et al., 2011; Verma, 2012; Mill and Heijmans, 2013]. Given that DNAm has been reported to be more stable in nature than histone modification, [Cedar and Bergman, 2009], and that a range of high-throughput technologies has been developed for DNAm analysis, EWASs have focused mainly on identification of DNAm differences that can be associated with disease phenotype, [Verma, 2012]. DNAm patterns specifically characterise tissues, [Rakyan et al., 2011], and variations, reported on measurements performed between tissues within the same individual, (i.e. intra-individual, inter-tissue), are noted to be much larger than those recorded on the same tissues, for a group of individuals, (i.e. inter-individual, intra-tissue), [Mill and Heijmans, 2013]. While considered to be a highly promising development, with regard to epigenetic variation as a cancer *biomarker*, the EWAS approach faces several difficulties, not least *sample selection*. Intrinsic limitations apply to collection of samples from the same group of patients, especially from internal tissues. In consequence, integrated samples from more accessible tissues such as blood or buccal, (e.g. Lowe et al. [2013]; Shenker et al. [2013]; Petersen et al. [2014]), have been used to examine the different biological hypotheses to date. For instance, an association between body-mass index and methylation level variations has been reported in a recent study performed on blood and adipose tissues, [Dick et al., 2014]. Thus, given accessibility, white blood count is one of the most commonly used tissues in EWASs; however, in this case, cell heterogeneity must be considered during interpretation of DNA methylation analysis findings, [Paul and Beck, 2014]. Another major challenge for EWAS viability is selection of the most significant patient-category for a given research question, (i.e. based on personal characteristics such as ageing and gender, as well as lifestyle features such as chemical exposure and tobacco usage), [Verma, 2012]. As an illustration from one such study, white blood cell samples for a group of breast and colon cancer patients, (described in Shenker et al. [2013]), were assessed to investigate the impact of smoking on epigenetic alterations during cancer development.

Computational modelling offers some possibilities of overcoming these restrictions as examination of a range of epigenetic patterns and variations in internal tissues over time can be simulated from available data. The potential for computational comparisons on *intra-*

and *inter*-tissue DNAm variation in the intestinal systems is discussed in the next section.

3.5 Limitations and proposed focus

Although Cancer Epigenetics research is gaining ground, computational models that assimilate information on epigenetic alterations in human malignancies are still relatively few. One reason for this is that *in silico* investigations of the epigenetic mechanisms are performed usually for 'generic' biological systems. While exploration of common features among malignancies provides an overview of the tumour pathways, the advantages of cancer type-specific models would be development of personalised treatment strategies (aimed at improvement in therapy efficiency and reduction of drug adversity). Such extended models would also permit information on individual characteristics (such as ageing and gender) and on environmental features (such as chemical radiation and tobacco usage) to be included. Further, given the EWASs problems with internal tissue sample collections, no comparison on epigenetic variation (in human intestinal systems) is currently available (or known of). Small intestine cancer incidence is considerable lower than that of CRC, with a global 2014 estimate of around one-sixteenth the number of new cases compared with the latter, [SEER, 2013a,b]. While this difference is partially explain by the presence of toxic residuals at colorectal level, as well as morphological differences between intestinal tissues, (e.g. high colon cell division rate), the ways in which tissue-specific mechanisms affect epigenetic patterns leading to disease development are not well understood yet. Finally, while contribution of epigenetic therapy is already recognised in some treatment strategies for blood cancers and, potentially, for solid tumours, (Section 2.7), the analysis of the impact of methylation inhibitors in CRC systems over time is still limited despite preliminary work.

Given these limitations on the state-of-the-art, a computational model following a hybrid approach, (BN, agent-based and logistic), is proposed in this Thesis to investigate human CRC dynamics, with a focus on DNAm level modifications for three main system layers. These are: i) micromolecular, (genetic and epigenetic mechanisms), ii) cell (crypt inter- and intra-communications), and iii) cell population, (crypt phenomena). The influ-

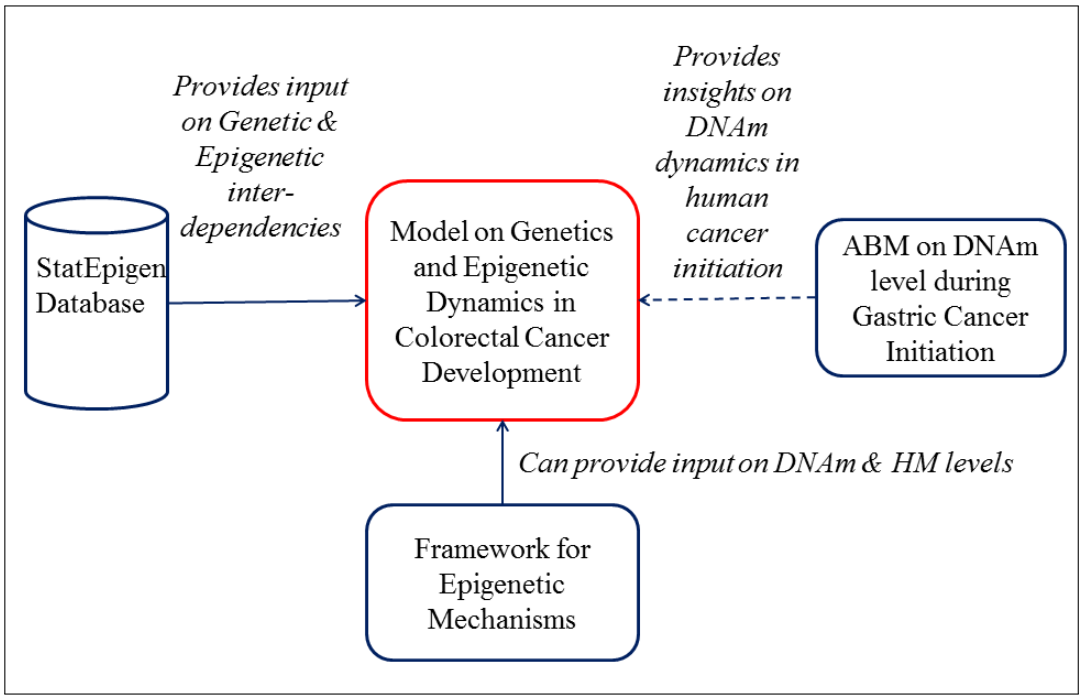


Figure 3.4: Simplified representation of the relationships between the Colorectal Cancer Model developed in this Thesis and other CIESCI-related work developed at SCI-SYM

ences of individual characteristics, environmental features and potential epigenetic inhibitor presence are also considered. Moreover, a theoretical comparison between human colon and another gastrointestinal tract tissue, (small intestine), for DNAm changes in both normal and abnormal systems is also undertaken. The relationship between the dynamic CRC model proposed here and other CIESCI-related projects, developed in SCI-SYM, are illustrated in Figure 3.4. Specifically, input data on micromolecular events affecting different genes in CRC are drawn from the StatEpigen database, while information on DNAm and HM levels can be provided by the micromolecular framework described in Raghavan and Ruskin [2011]. Moreover, given similarities between the stomach and colon crypt structures, the agent-based model on gastric cancer described in Perrin et al. [2010] can be used as a proof of concept for CRC dynamics, (Chapter 6).

Although toolkits for developing BN, agent-based and multi-scale models exist, the computational models for CRC dynamics described within the next chapters follow *ab initio* developmental strategies. Both commercial and free tools, (e.g. *Analytica*, *BayesiaLab*,

Chordalysis, *GeNIe*), have been developed for generating BN-based applications over recent decades. (A larger list of these tools is included in Table C.3, in Appendix C.) The *gRain* package from Bioconductor also provides functionality for generating BNs, [Højsgaard, 2012]. In addition, a comprehensive review of toolkits available to generate ABM, (including major ABM-frameworks such as Repast[22], NetLogo[23] and Swarm[24]), is provided in Nikolai and Madey [2009]. Authors have considered several criteria for classification including i) programming language used for model implementation, ii) operating system (OS) needed for running the framework, iii) licence options, iv) customer/user support availability and v) application domain, (e.g. biological, social, educational systems). However, although such tools are widely-used, (e.g. GeNIe has around 2000 users), their applicability is less good for specialised studies where for instance, integration of heterogeneous data types or different dynamics among the variable sets are needed. An analysis of these software solutions potential, (for both BN and ABM), including known limitations, [Nikolai and Madey, 2009], plus other issues, parallelisation options, were considered. This indicated that such toolkits were not suited for the purpose of this study, since the computational components of the Colorectal Cancer Model need to have the same configuration set with respect to the programming language used, (e.g. C++), to permit simulation running on both Windows and Linux OS, inclusion of additional libraries, (e.g. MPI[25]), and connectivity of developed components. Open-source software for building multi-scale models also exists. For instance, Charste[26] is a platform developed in C++, supported by Linux, and tailored to modelling i) cardiac behaviour and ii) cell population dynamics related to cancer development with an initial focus on CRC, [Mirams et al., 2013]. However, although Charste has been used to develop several multi-scale models for CRC, (e.g. Van Leeuwen et al. [2009]; Fletcher et al. [2012]), it does not provide functionality suitable for describing genetic-epigenetic interdependencies at the micromolecular scale.

[22]Repast url: http://repast.sourceforge.net/

[23]NetLogo: url: https://ccl.northwestern.edu/netlogo/

[24]Swarm: url: http://savannah.nongnu.org/projects/swarm

[25]Message Passing Interface (MPI) is a standardized library for message-passing and includes functions that permit code paralellisation.

[26]Charste stands for **C**ancer, **H**eart and **S**oft **T**issue **E**nvironment, url: http://www.cs.ox.ac.uk/chaste/.

3.6 Summary

This chapter builds on the biological background summarised in Chapter 2 to describe computational and mathematical modelling approaches used to investigate abnormal biological systems. To address gaps, the development of a hybrid computational model for human CRC dynamics is proposed, to be developed for three main layers, (namely micromolecular, cellular and tissue levels). The next chapters describe the components of the model framework, starting with the Bayesian network for genetic and epigenetic events in CRC.

Chapter 4

Epigenetic-Genetic (E-G) Network Model

4.1 Introduction

The E-G Network Model, the development of which is described in this chapter, is a novel framework for genetic and epigenetic events observed at different stages of colorectal cancer development, with a focus on gene relationships and tumour pathways. Specifically, the E-G Network targets the first research objective of our research, i.e. "To quantify the impact of genetic and epigenetic interdependencies on the methylation level in CRC development and to estimate the influence of ageing and gender with respect to methylation level in abnormal colon cells", (Subsection 1.2.1). The E-G Network is the first component of the overall Colorectal Cancer Model (as illustrated in Figure 4.1) and its potential connection with other components developed during this research is discussed in Chapter 8. Given the advantages of Bayesian networks, (subsection 3.3.1), this approach can be used to describe abnormal micromolecular events and their interdependencies in the CRC context. The gene framework integrates empirical data on conditional relationships between molecular signals found in CRC development, obtained from the StatEpigen database, [Barat and Ruskin, 2010]. In addition, information on signalling pathways involved in CRC development, (e.g.

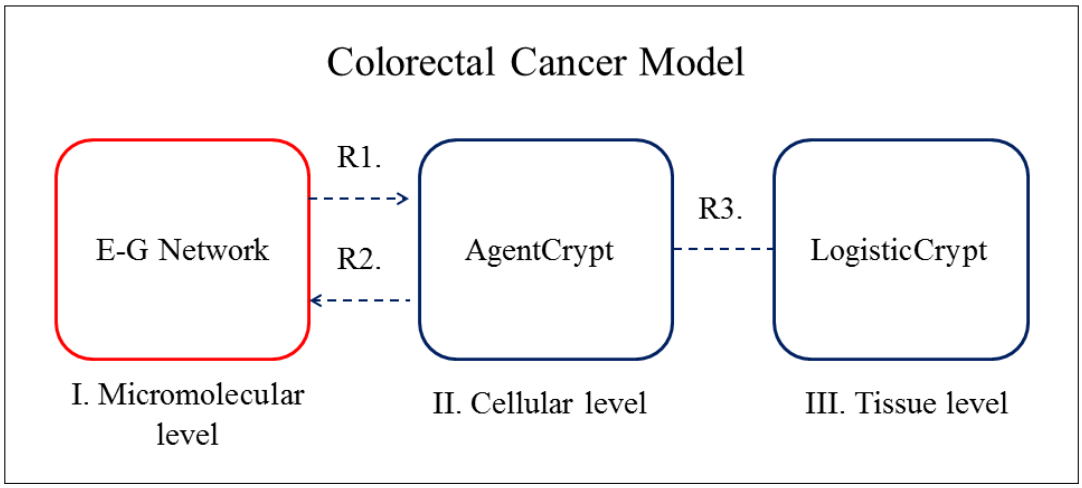


Figure 4.1: Structure of the Colorectal Cancer Model - focus on the E-G Network model The relationships between model components may refer to R1) average methylation level for gene network / intestinal cell; R2) patient features; R3) intestinal crypt dynamics, explored following bottom-up and top-down approaches.

Wntβ catenin, MAPK), is included from the KEGG Pathway database, [Kanehisa et al., 2014], and available literature. Known to have a crucial role in cancer development, DNA methylation and how it changes constitute the main objective of the E-G Network Model.

The base E-G Network is described in this chapter, (with its extensions presented in Chapter 5). Specifically, the data set and gene framework structure, together with an overview of the interdependencies between the gene network layers, are detailed in Sections 4.2 and 4.3. The algorithm proposed for the identification of the *most plausible pathways* in a given gene network is introduced in Section 4.4. The expressions for gene methylation update and major assumptions made in model formulation, (such as the value-ranges considered for DNA methylation for the cancer stages), are specified in Sections 4.5 and 4.6. The network dynamics and the main contributions and limitations of the E-G Network are also discussed.

4.2 Data set

The E-G framework consists of a network built to represent gene relationships, based on empirical statistical data from StatEpigen. The gene relationships are designated: *simple*,

(recording incidence of a micromolecular event affecting a gene, given the phenotype of the analysed samples), and *conditional*, (giving pairwise affected gene dependence at certain cancer stages). These data are used to inform the set of conditional interdependencies for the Bayesian network, (see Figure 4.2). In general, the conditional relationships (CR) from StatEpigen are expressed mathematically by:

$$CR(G2, G1, e2, e1, s, c) = P((G2,\ e2)|(G1,\ e1), s, c) \tag{4.1}$$

i.e. the probability of an event *e2* to occur for gene *G2* given the presence of the event *e1* for gene *G1*, in the stage *s* of cancer type *c*. Similarly, a *simple relationship (SiR)* is the probability that event *e* occurs for gene *G* at stage *s* of cancer type *c*, i.e.

$$SiR(G, e, s, c) = P((G,\ e), s, c) \tag{4.2}$$

Moreover, in the current thesis, a conditional relationship between two molecular events, observed for the same gene, is referred to as a 'self-relationship' (SeR), being expressed mathematically by:

$$SeR(G, e2, e1, s, c) = P((G,\ e2)|(G,\ e1), s, c) \tag{4.3}$$

i.e. the current micromolecular modification *e2* depends on the previous event *e1* observed for the same gene *G*. This is illustrated by an example for the APC gene, a tumour-suppressor gene (TSG) that is found to be frequently mutated and/or hypermethylated in the very early stages of colon cancer, (see Subsection 2.6). From the StatEpigen data, [Barat and Ruskin, 2010]:

$$P(APC\ mutation \mid APC\ H+,\ adenoma,\ colon) = 0.538 \tag{4.4}$$

which determines

$$SeR(APC,\ mutation,\ H+,\ adenoma,\ colon) = 0.538 \tag{4.5}$$

The numerical values are subject to refinement as further data become available and the network itself can be subjected to 'stress-testing' to determine its reliance on given values, (i.e. sensitivity analysis).

4.3 Gene framework structure

Structured in three main layers, (*micromolecular events*, *gene relationships* and *cancer stages*), the E-G Network explores the effect on phenotype of a number of key genetic and epigenetic factors. These include DNA methylation, (both *hyper* and *hypo*methylation), and *histone modifications*, combined with tumour pathway information relating to genetic mutations and gene expression.

While colon cancer is the initial focus, the gene framework is designed to be extended to other types of cancer. Therefore, more general terms, such as '*carcinoma in situ*' or '*invasive carcinoma*' have been used to denote colon cancer stages, instead of more disease-specific usage. Thus, the gene framework recognizes 'healthy' state and three main cancer stages: *carcinoma in situ*, *invasive carcinoma* and *metastasis*. The interdependence between current and extended framework levels is illustrated in Figure 4.2, interpreted as follows:

- The lowest level is that of the *micromolecular events*, (*hyper-/ hypo*methylation, mutations, gene expression, histone modification and histone variants), and contains information seen as 'external' input for the gene network. For example, resources such as GEO, ENCODE, KEGG or COSMIC, (Chapter 3), can be interrogated here for data on DNA methylation, gene expression and histone modification levels and for information on mutation and signalling pathways.

- The mid-layer is that of the *gene relationships* and is represented by the Bayesian

network, (built based on input from StatEpigen database).

- The final layer is that of '*cancer initiation and progression*', giving information on outputs at the network cancer stage. This layer includes three sublevels: *carcinoma in situ*, invasive carcinoma and metastasis.

Information on risk factor influence, (such as ageing, gender, heredity, environmental features, viral and bacterial infection) in cancer development and the stem cell property should be ideally integrated in the extended gene framework.

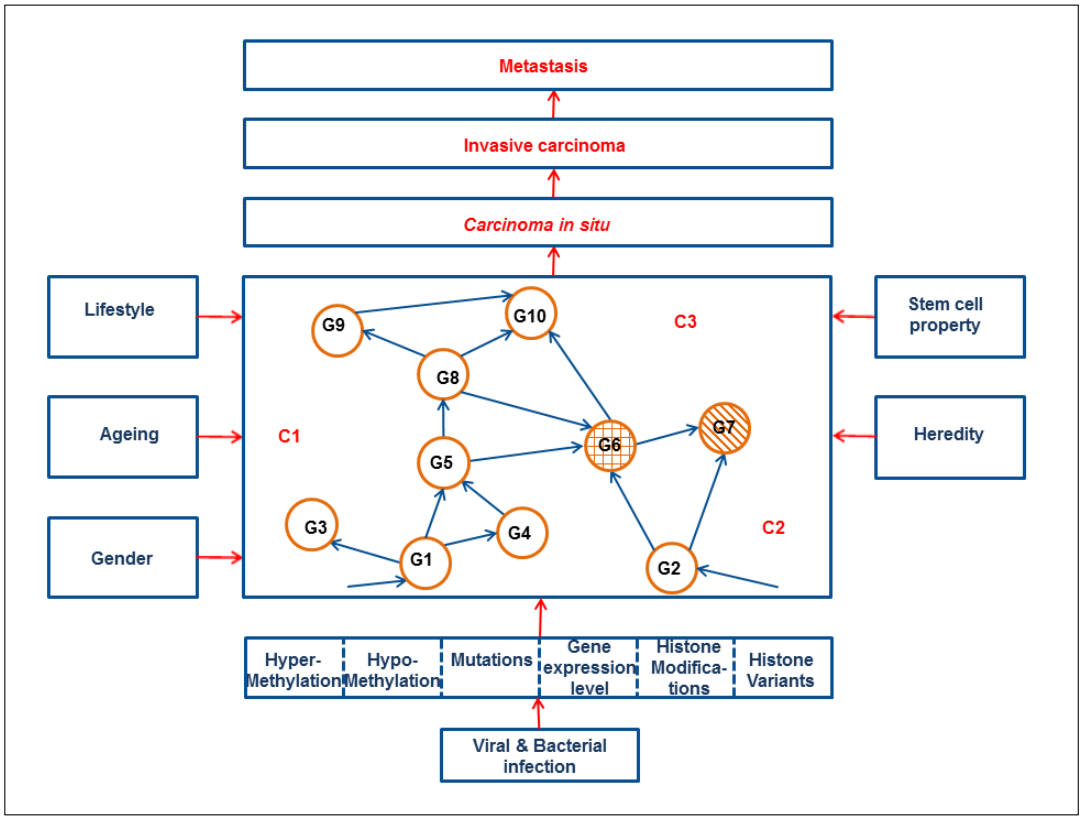


Figure 4.2: Dynamics of the E-G Network Model
The middle layer describes gene relationships associated with three types of cancer, generically labelled C1, C2, and C3, where G1 to G10 substitute for actual gene labels. Risk factor influence is also illustrated. Legend: empty circle: gene found in one cancer type, circle with hatch-shading and grid patterns: gene found in two or three cancer types, respectively. Edge: gene relationship.

According to the construction methods presented in subsection 3.3.1, a connected graph is built of genes as nodes, (represented by circles in Figure 4.2). The conditional rela-

tionship between two genes is denoted by an arrow between the nodes that include those two genes. Information on conditional gene relationships can be taken from e.g. StatEpigen database, which provides data on genetic and epigenetic interdependencies at different stages of various cancer types, [Barat and Ruskin, 2010]. Additionally, the node shading indicates gene affiliation to one or more types of cancer. For example, gene *G3*, (illustrated by an empty circle), has been observed in cancer type *C1*, (e.g. *NR3C1*[1] in colon cancer, [Barat and Ruskin, 2010]). The circle with hatched shading, (e.g. *G7*), indicates a gene seen in two types of cancer, such as *C2* and *C3*, (e.g. *RASSF1A* in lung and stomach cancer), and finally, a gene that has been identified as being abnormally changed in three cancer types, (such as *KRAS* in colon, lung and stomach), is illustrated graphically by a circle with grid pattern, (e.g. *G6*). An example of the biological data from StatEpigen that feeds into the gene layer in Figure 4.2 is shown in Table 4.1; the definition of symbols for these genes are presented in Table C.4, according to information from the GeneCards Database, [Safran et al., 2010].

Table 4.1: Genes involved in different genetic and epigenetic events (hypermethylation (H+), gene expression (GE), mutation (Mut)) in colon, lung and stomach cancer phenotypes

Group from table	**Genes**	**Colon Cancer (C1)**	**Lung Cancer (C2)**	**Stomach Cancer (C3)**
G1	CRABP1	Present (H+)		
G2	NORE1		Present (H+)	
G3	NR3C1	Present (H+)		
G4	MLH1	Present (H+; GE)		
G5	MGMT	Present (H+; GE)		
G6	KRAS	Present (Mut)	Present (Mut)	Present (Mut)
G7	RASSF1A		Present (H+)	Present (H+)
G8	APC	Present (H+; Mut)		
G9	TP53	Present (Mut; GE)		
G10	CDKN2A:P14	Present (H+)		

[1]NR3C1: nuclear receptor subfamily 3, group C, member, [Safran et al., 2010]

4.4 The d-plausible pathways in gene network - different routes of cancer development

With respect to pathways linking gene changes, we refer to *d-plausible* pathways which consist of sets of inter-related gene modifications that can occur in the gene network with higher probabilities than a threshold value, calculated based on 'd' attribute and the network structure, (as shown further). Their detection in the E-G Network Model is of particular interest since they can be studied in association with known *signalling pathways* for CRC development, (subsection 2.5). For example, if some genes from a SP are found to lie on a plausible pathway in a given gene network, then there is a high chance that this SP will be activated.

In the connected gene network, each node has been labelled with one of three position attributes: i.e. as a *start*, *middle* or *final* node, based on its position in the current acyclic graph[2]. An initial definition of the 'parent' and 'child' nodes was given (subsection 3.3.1). Similarly, a gene G_i is defined to be a *parent* for a gene G_j in the gene network if there is a conditional relationship CR(G_j, G_i, e_j, e_i, s, c) in the analysed data, where e_i, e_j are the molecular events observed, respectively, for G_i, and G_j at the stage s of cancer type c, (i, j are indices with value range from 1 to the total number of genes in network). Further, the gene G_j is referred to as a *child* for G_i. Hence, a *start* node is a gene that has no parents in the gene framework, while a *final* node is a gene that does not have any children. Subsequently, a *middle* node illustrates a gene that has both parents and children, i.e. can be either argument in two or more different conditional relationships for the given gene network. Furthermore, if the assumption is made that a pathway is a directional connection of graph edges, beginning at a start node and containing zero or more middle nodes and ending in a final node, [Sedgewick and Wayne, 2011], it is relatively easy to determine all pathways in the gene network. Given this result, the decision on which pathway is more plausible is made by comparing the *pathway score* value and the *d-Network Threshold*

[2]The gene network is a Bayesian network, which is defined as a directed acyclic graph, where edges describe dependencies between a set of variables, represented by nodes. The edge weights are given by the values of conditional probabilities among the set of variables, (subsection 3.3.1).

(dNT). These are computed as follows.

Pathway score calculation The pathway focuses on the relationships between the genetic and epigenetic changes observed for different genes, (i.e. excluding the 'self-gene' relationships, where the molecular event depends on the previous status of the same gene). The total probability of a set of independent events is given by P = $\Pi_{i=1}^{n}$ P_{e_i}, where P_{e_i} = probability of an independent event e_i from the considered set, [Heckerman, 1998; Lucas et al., 2004]. Based on this, the probability of a pathway in the gene network can be given by the product of all compound, conditional gene relationship probabilities and the simple relationship probability for the start gene. In addition, logarithm in base 10 ($\log_{10}$) is applied in order to facilitate calculation, (i.e. use sum instead of product). Therefore, the pathway score is defined as the sum of all logarithms of the compounds, conditional gene relationship probabilities and the simple relationship probability for the start gene, multiplied by the *pathway contribution coefficient*. This last quantity is equal to the ratio between the pathway length, (n = the number of genes found in the current pathway), and the total number of network genes, (V). The score of the k^{th} pathway in the gene network is described by the following expression, (with the variables of interest included in Table 4.2):

$$SCORE(k) = [log_{10}(SiR(G_{St}^{k}, e_{St}^{k}, s, c)) + \sum_{i=1,j=1}^{n} (log_{10}(CR(G_{i}^{k}, G_{j}^{k}, e_{i}^{k}, e_{j}^{k}, s, c))] \times n/V \tag{4.6}$$

Table 4.2: Definition of variables for pathway score calculation, expression (4.6)

Variable name	**Variable description**
G_{St}^{k}	the start gene of the k^{th} pathway from the network;
e_{St}^{k}	the molecular event corresponding to the G_{St}^{k} gene;
G_{i}^{k}, G_{j}^{k}	the i^{th}, j^{th} gene respectively of the k^{th} pathway;
e_{i}^{k}, e_{j}^{k}	the molecular event corresponding respectively to the G_{i}^{k}, to the G_{j}^{k} genes;
n	the number of all genes found in the k^{th} pathway;
V	the total number of the network genes;
s	the stage of cancer for the current pathway;
c	the type of cancer for the k^{th} gene pathway.

D-network definition *A d-network* in the E-G Network Model is defined to be a network that has the same structure with a given gene network, (initially built by the E-G Network Model), with the constraint that all its edges have the same *weight value*, equal to *d*, (d ∈ [0, 1]). These weights indicate that the value of every simple or conditional gene relationship in d-network is equal to d, i.e.:

$$SiR(G_i, e_i, s, c) = d \tag{4.7}$$

$$CR(G_i, G_j, e_i, e_j, s, c) = d \tag{4.8}$$

$\forall$ G_i, = gene in network, G_j = parent of the gene G_i, i.e. $G_j \in$ Parents(G_i), e_i, e_j = events corresponding to G_i and G_j, respectively, i, j $\in$ [1, V], V = total number of genes from network, s = cancer stage, c = cancer type. In order to determine the *dNT* value, the score for every pathway from the d-network is calculated based on expression (4.6). The highest value from all these computed scores is assigned as the dNT parameter, which is used in the decision step.

The *d*-network was defined in this way in order to highlight the impact of gene pathway length rather than the compound edge weight values. In real systems, the plausibility of e.g. three micro-molecular events occurring under specified conditions is higher than that for say 100 events. For example, if two pathways, *Path1* and *Path2*, with 3 and 100 genes, respectively, are identified in a d-network of V genes in total, the pathway score, (calculated with expression (4.6), are:

$$SCORE(Path1) = (log_{10}d + log_{10}d) \times 2/V = 2^2/V \times log_{10}d; \tag{4.9}$$

$$SCORE(Path2) = (log_{10}d + .. + log_{10}d) \times 99/V = 99^2/V \times log_{10}d. \tag{4.10}$$

Because $\log_{10}$ d ≤ 0, $\forall$ d $\in$ [0, 1] $\Rightarrow$ SCORE(Path1) $>$ SCORE(Path2). Thus, the shorter pathway has a higher score (is more plausible) than the longer pathway. Considering this, comparison with a threshold, based on the former, would provide more accurate information

on pathway plausibility than using the latter.

Decision on d-plausible pathway If the score of kth gene pathway, SCORE(k), is higher than the assigned dNT value, then the kth gene pathway is considered a d-plausible pathway in the gene network. We illustrate with an example.

Example on d-plausible pathway calculation A simple network of four genes, found in the carcinoma stage of colon cancer, (consisting of APC, MLH1, MCC , CDKN2A:p16) is represented in Figure 4.3. This network is based on data taken from StatEpigen, [Barat and Ruskin, 2010], on the simple and conditional relationships in which this set of genes is implicated. The weight information is given by the empirical probability value of the gene relationships, (e.g. the weight of the edge between the nodes, MLH1 and APC genes, is 0.688, equal to the probability of finding APC hypermethylated[3] in CRC, given MLH1 hypermethylated, [Barat and Ruskin, 2010]). This set of gene relationships is given by:

$$
\begin{aligned}
&P(CDKN2A:p16\ HpM+ \mid\ carcinoma,\ colon) = 0.249; \qquad (4.11)\\
&P(MLH1\ HpM+ \mid\ carcinoma,\ colon) = 0.140;\\
&P(APC\ HpM+ \mid\ CDKN2A:p16\ HpM+,\ carcinoma,\ colon) = 0.244;\\
&P(APC\ HpM+ \mid\ MLH1\ HpM+,\ carcinoma,\ colon) = 0.688;\\
&P(MCC\ HpM+ \mid\ APC\ HpM+,\ carcinoma,\ colon) = 0.269;\\
&P(MCC\ HpM+ \mid\ CDKN2A:p16\ HpM+,\ carcinoma,\ colon) = 0.794,
\end{aligned}
$$

where CDKN2A:p16, MLH1, APC and MCC are genes involved in CRC development, (Chapter 2), and HpM+ = hyermethylation.

The three distinct pathways that can be identified in this gene network together with their score values[4], (computed for each of these pathways using expression (4.6)), are:

1. *Pathway 1:* CDKN2A:p16 $\rightarrow$ MCC;

[3]Hyper- and hypomethylation refer to increase and decrease, respectively, of methylation.
[4]Given that probabilities $\in$ [0, 1], $\log_{10}$ of these takes values ≤ 0.

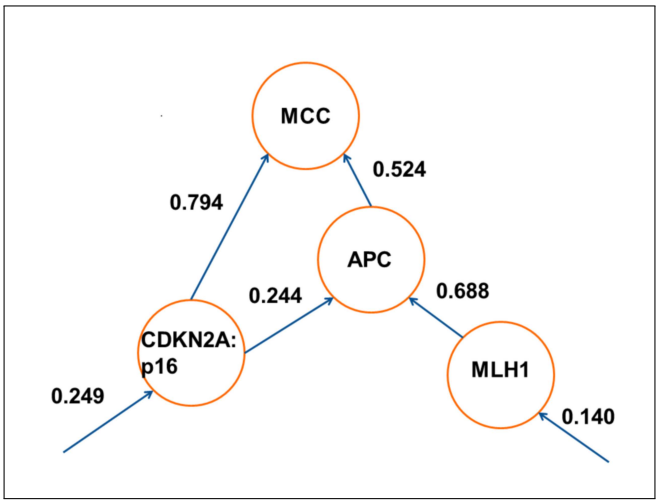


Figure 4.3: Example of a small gene network for carcinoma, colon cancer. Legend: circle: gene in CRC, edge: conditional relationship between micro-molecular events affecting different genes.

SCORE (Pathway 1) = [SiR(CDKN2A:p16, HpM+, carcinoma, colon) + CR(MCC, CDKN2A:p61, HpM+, HpM+, carcinoma, colon)] $\times$ 2/4 =

= [$\log_{10}$(0.249) + $\log_{10}$(0.794)] $\times$ 2/4 = -0.351;

(Given that two genes are involved in Pathway 1, out of four genes in the gene network, the contribution coefficient for Pathway 1 is equal to 2/4.)

2. *Pathway 2:* CDKN2A:p16 $\rightarrow$ APC $\rightarrow$ MCC;

   SCORE (Pathway 2) = [SiR(CDKN2A:p1, HpM+, carcinoma, colon) + CR(APC, CDKN2A:p1, HpM+, HpM+, carcinoma, colon) + CR(MCC, APC, HpM+, HpM+, carcinoma, colon)] $\times$ 3/4 =

   = [$\log_{10}$(0.249) + $\log_{10}$(0.244) + $\log_{10}$(0.524)] $\times$ 3/4 = -1.122;

   (Three genes are found in Pathway 2, out of four genes in the gene network; thus, contribution coefficient = 3/4.)

3. *Pathway 3:* MLH1 $\rightarrow$ APC $\rightarrow$ MCC;

   SCORE (Pathway 3) = [SiR(MLH1, HpM+, carcinoma, colon) + CR(APC, MLH1, HpM+, HpM+, carcinoma, colon) + CR(MCC, APC, HpM+, HpM+, carcinoma, colon)] $\times$ 3/4 =

= [$\log_{10}(0.140) + \log_{10}(0.688) + \log_{10}(0.524)$] $\times$ 3/4 = -0.972.

(Since three genes are included in Pathway 3, out of four genes in the gene network, contribution coefficient = 3/4.)

The *dNT* must then be calculated for these gene pathways based on *d* value. Two examples of dNT calculation are given, where *d* value is considered, sequentially, higher and lower than the average of edge weight values, included in the gene network from Figure 4.3. Given that the average is ≈ 0.44, in the first example, a value of e.g. d = 0.60 was assigned, ($0.60 > 0.44$), and in the second example, e.g. d = 0.40, ($0.40 < 0.44$). Other choices for *d* values are, of course, possible and this parameter needs to be considered for a sensitivity analysis. For example, *d* value can be chosen closer to the maximum and minimum of the edge weight values, (i.e. extreme values) in the gene network. However, if *d* is set to a high value (closer to maximum), the dNT will have a high value, which will reduce the number of plausible pathways (example 3, with d = 0.70). If *d* value is taken to be small, the threshold dNT will be low, allowing a high number of plausible pathways, (example 4, d = 0.10). Thus, an optimal *d* value should be decided based on biological questions - whether the interest is in the behaviour of the majority or in features exposed only by a 'peak' of gene pathways.

Example 1, d = 0.60. In the d-network, with d = 0.60, SCORE values were computed for all three pathways following the arrows in gene network, (Figure 4.3):

1. SCORE(*0.60*-Pathway1) = ($\log_{10}$ 0.60 + $\log_{10}$ 0.60) $\times$ 2/4 = - 0.22;
2. SCORE(*0.60*-Pathway2) = ($\log_{10}$ 0.60 + $\log_{10}$ 0.60 + $\log_{10}$ 0.60) $\times$ 3/4 = -0.49;
3. SCORE(*0.60*-Pathway3) = ($\log_{10}$ 0.60 + $\log_{10}$ 0.60 + $\log_{10}$ 0.60) $\times$ 3/4 = -0.49.

As result, -0.22 is assigned to dNT since it is the highest computed value. Given this example, none of the identified pathways for the gene network from Figure 4.3 are d-plausible pathways, with d = 0.60, because their score values are lower than that of the dNT, (-0.351 $<$ -0.22, -1.122 $<$ -0.22, -0.972 $<$ -0.22).

Example 2, d = 0.40. The score values computed for all identified pathways in a d-network, where d = 0.40, are:

1. SCORE(*0.40*-Pathway1) = ($\log_{10}$ 0.40 + $\log_{10}$ 0.40) $\times$ 2/4 =-0.39;

2. SCORE(*0.40*-Pathway2) = ($\log_{10}$ 0.40 + $\log_{10}$ 0.40 + $\log_{10}$ 0.40) $\times$ 3/4 = -0.89;

3. SCORE(*0.40*-Pathway3) = ($\log_{10}$ 0.40 + $\log_{10}$ 0.40 + $\log_{10}$ 0.40) $\times$ 3/4 =- 0.89.

Given that -0.39 is the highest value from all computed, it is assigned to dNT, (when d = 0.40). Considering the values identified for Figure 4.3, the *Pathway 1* is detected as d-plausible, (d = 0.40), because its score value is higher than that of dNT, (i.e. -0.351 $>$ -0.39).

Example 3, d = 0.70. The score values are computed for all identified pathways in a d-network, where d = 0.70:

1. SCORE(*0.70*-Pathway1) = ($\log_{10}$ 0.70 + $\log_{10}$ 0.70) $\times$ 2/4 =-0.15;

2. SCORE(*0.70*-Pathway2) = ($\log_{10}$ 0.70 + $\log_{10}$ 0.70 + $\log_{10}$ 0.70) $\times$ 3/4 = -0.34;

3. SCORE(*0.70*-Pathway3) = ($\log_{10}$ 0.70 + $\log_{10}$ 0.70 + $\log_{10}$ 0.70) $\times$ 3/4 =- 0.34.

Given that -0.15 is the highest value from all computed, it is assigned to dNT, (when d = 0.70, higher than average = 0.44). Considering the values identified for Figure 4.3, none of the pathways shown are d-plausible pathways.

Example 4, d = 0.10. The score values are computed for all identified pathways in a d-network, where d = 0.10:

1. SCORE(*0.10*-Pathway1) = ($\log_{10}$ 0.10 + $\log_{10}$ 0.10) $\times$ 2/4 = -1;

2. SCORE(*0.10*-Pathway2) = ($\log_{10}$ 0.10 + $\log_{10}$ 0.10 + $\log_{10}$ 0.10) $\times$ 3/4 = -2.25;

3. SCORE(*0.10*-Pathway3) = ($\log_{10}$ 0.10 + $\log_{10}$ 0.10 + $\log_{10}$ 0.10) $\times$ 3/4 = -2.25.

Given that -1 is the highest value from all those computed, it is assigned to dNT, (when d = 0.10). Considering the values identified for Figure 4.3, *Pathway 1* and *Pathway3* are

detected as d-plausible, (d = 0.10), since -0.351 $>$ -1 and -0.972 $>$ -1, i.e. their score values higher than that of dNT, (d = 0.10).

This section presented an algorithm for identification of the d-plausible pathways in the gene network, which can describe qualitatively pathways in gene networks, (i.e. to decide whether they are plausible). We need to consider the initialization and update steps for methylation level inside the gene network in order to investigate quantitatively DNA methylation differences that can be induced by associated micromolecular events, including mutation, *hyper-*/ *hypo*methylation, histone acetylation/ methylation during cancer development.

4.5 DNA methylation

In the E-G Network model, gene methylation[5] level *update* is a complex process, influenced by a combination of at least five factors: cancer stage, gene relationship type, the nature of the micromolecular events, signalling pathways and histone modifications. The next two subsections describe how initial gene methylation levels are set and the expressions applied to update the methylation level in the gene framework.

4.5.1 Methylation initialization step

*Hyper*methylation of CpG islands found in TSG *promoter*[6] is typically associated with gene silencing leading to tumour progression, where methylation increase depends also on the cancer stage. For example, gene TP53 is thought to have a higher promoter methylation level in CRC metastasis than in an incipient stage such as polyps. In E-G Network, the gene methylation level is expressed as the PMRA[7] value for that gene, i.e. $\in$ [0, 100], (Subsection 2.3.1). For example, if the PMRA value for the MGMT gene in a specific CRC stage is reported to be equal to 33.13, [Ogino et al., 2006b], then, the MGMT methylation

[5]Note: In E-G Network, we refer to 'gene promoter methylation level' using the simplified formulation of 'gene methylation level', i.e. the word 'promoter' is implicit. Analogous, 'methylation level' is used instead of 'promoter methylation level. However, when confusions can occur, the word 'promoter' is explicitly specified.

[6]Promoters are DNA sequences that indicate the *locus* where gene transcription begins, (usually located upstream from transcript start sites).

[7]PMRA = average of PMR, where PMR = degree of methylation.

level corresponding to the same CRC stage inside the gene framework is taken to be equal to 33.13. In an ideal gene framework, initial methylation level values are known for every compound gene at each cancer stage. However, such data are not always available; therefore, in the E-G Network, four value-ranges have been chosen to assign initial methylation levels based on the given stages, (included in Table 4.3).

Table 4.3: Initial methylation ranges (PMRA values) for cancer stages

Raw	**Cancer stages**	**Methylation ranges**
1	Healthy	0 - 0.50
2	*Carcinoma in situ*	8.00 - 8.50
3	Invasive carcinoma	18.00 - 18.50
4	Metastasis	65.00 - 65.50

These value ranges were chosen based on the assumption that variation in DNAm level is higher as tumour progresses. Specifically, from Table 4.3, it can be inferred that methylation level in the healthy phenotype is characterised by range [0, 8), (i.e. difference $\simeq$ 8), in *carcinoma in situ* by range [8, 18), (with difference $\simeq$ 10), in invasive carcinoma by range [18, 65), (i.e. difference $\simeq$ 47) and finally, in metastasis by range [65, 100], (i.e. difference $\simeq$ 35). Larger methylation variation seems to be permitted for invasive carcinoma than for metastasis. However, 'invasive carcinoma' term is used to refer to three cancer stages within the Overall Stage Grouping System[8], namely Stage I, II and III, (Section 2.5). Thus, methylation level difference for each of these stages is around 15 - 16, i.e. higher than that considered for *carcinoma in situ*, and obviously, lower than that for metastasis. However, other values can be considered as well, and a sensitivity analysis is planned for these value ranges, (Chapter 8). In addition, the gene framework provides functionality for handling external input to gene methylation level. Thus, the values specified in Table 4.3 can be adjusted if additional or modifying information on methylation level becomes available.

During this step, the methylation level value for each gene is randomly-generated within the determined ranges, according to the cancer stage information provided by gene relationships in StatEpigen, [Barat and Ruskin, 2010]. For example, suppose the following two

[8]Overall Stage Grouping is a system where cancer development is described by Roman numbers, (from I to IV), in addition to 0, which is considered the least advanced cancer stage, (Section 2.5).

gene relationships occur in the network:

i) P(MGMT H+ | KRAS mutation, carcinoma in situ, colon) = 0.282;

ii) P(APC H+ | TP53 mutation, invasive carcinoma, colon) = 0.688.

where both relation values are indicated by StatEpigen. Thus, the initial methylation level for the MGMT and KRAS genes will be a value in the 8.00 - 8.50 range, since this relationship was observed in *carcinoma in situ*, (row 2 in Table 4.3). Additionally, for APC and TP53 genes, the methylation level is initially higher in this case because the second relationship was observed in invasive carcinoma, (so value within the range given by row 3 in Table 4.3).

In conclusion, values within ranges from Table 4.3 are assigned to methylation level only when new genes are introduced in the gene network. However, it is also important to know how methylation level changes for genes that already exist in the gene network, i.e. how gene methylation level is updated over time.

4.5.2 Methylation update step

In nature, modifications of the gene promoter methylation level do not follow a clear pattern of occurrence. Given this, the assumption that methylation modifications appear at every gene promoter in every cell during every cell division seems unrealistic. Therefore, in the E-G Network, the update of the gene promoter mehtylation level is considered to be a probabilistic event. In model terms, the methylation update step for a gene G is performed only if a probability value R, (randomly-generated within [0, 1] range), is higher than the value of the threshold 'TH_DNAM(G)', calculated using expression (4.12). Of course, every gene relationship in which a gene G is involved in the gene network can contribute to gene G methylation updates. For example, if there are three gene relationships containing the APC gene, three independent methylation update events for the APC gene are possible; in consequence, three sets of R_k and TH_DNAM$_k$(G), k$\in \{1, 3\}$, values will be calculated. In addition, if the gene G is involved in a signalling pathway, whose deregulations contribute to CRC development, there is a higher chance for G to be affected in the given network; thus, the threshold value TH_DNAM$_k$(G) decreases. Specifically, TH_DNAM$_k$(G) value is

given by 1 minus the value of the k^{th} gene relationship in which gene G is involved in the given network together with the value of a model input parameter, denoted as *the signalling pathway coefficient, (SIG)*, which indicates that the gene G has been observed in a signalling pathway.

$$TH\_DNAM_k(G) = 1 - P_k(G) - SIG(G) \tag{4.12}$$

where:

$$SIG(G) = \begin{cases} \textbf{model input value},\ (constant\ for\ all\ genes\ in\ network)\ , with\ non-zero \\ value,\ if\ the\ gene\ G\ is\ known\ to\ form\ part\ from\ a\ signalling\ pathway; \\ \textbf{0},\ otherwise. \end{cases}$$

P_k(G) is equal to the value of gene relationships where gene G is involved and can be given by StatEpigen database. Specifically, P_k(G) represents:

- The probability of the k^{th} conditional relationship that includes the gene G, if this is a middle or an end gene. It represents the incidence of a molecular event for the current gene, given the molecular events observed for its k^{th} parent within the gene network.
- The simple relationship probability of the gene G, if this is a start gene;
- The 'self-relationship' probability if the gene G is inside a 'self-updating' methylation step.

So, if the R_k value is higher than that of the TH_DNAM$_k$(G), then the methylation level of a gene G is updated, using the following expression, (with variables of interest included in Table 4.4):

$$M'(G) = M(G) \pm W_k(G) \times \frac{Parent(G)}{(E+S)} \tag{4.13}$$

In expression (4.13), the sign '+' or '-' is assigned according to the observed molecular event present in the gene relationship: '+' corresponds to *hyper*methylation, '-' to *hypo*methylation. For mutation, a random variable, internal to the network, decides whether

Table 4.4: Definition of variables used in the update methylation level step, Expression (4.13))

Parameter name	Parameter description
M'(G)	DNAm of gene G in the next Time-step;
M(G)	DNAm of gene G in the current Time-step;
Parent(G)	the number of the gene G parents (considered equal to 1 if G is a start gene or belongs to a 'self-relationship').
E	total number of edges from network;
S	total number of 'self-relationships' from network, considered zero in a non 'self-relationship' update step;
W_k	a random value within range [10^{-3}, 10^{-2}] for *hyper* and *hypo*methylation, or [5×10^{-4}, 10^{-3}] for mutations; these ranges differ between *hyper*/*hypo*methylation and mutations, because the former events affect methylation level directly, while the latter can induce subsequent changes in methylation level. Therefore, the W_k coefficient represents the contribution of the k^{th} parent to the new gene methylation level, based on the molecular event history observed for the current gene.

'+' or '-' is chosen, i.e. as to whether mutation causes an increase or decrease in methylation level. Based on expression (4.13), methylation level is updated at time T+1 based on its value at time T and the several other factors. The number of parent genes, (i.e. Parent(G)), indicates how a gene is connected to the rest of the genes in the network. This information can be useful to identify the *highly-connected* genes, ('hub genes'), and in analysis of their disease contribution, in comparison with the other genes. Additionally, the ratio 1 / (E+S) shows the *strength* of the gene network. For example, in a small network, (with fewer edges and 'self-relationships' in the graph), the methylation level will be updated to a higher value than in a larger network. As a result, the former network will register a *faster progression* in tumour development than the latter.

4.6 Histone modification and the relationship with DNA methylation

Similarly to DNAm, the histone acetylation or methylation levels are also considered to be dependent on cancer stage, [Yoo and Jones, 2006]. In order to achieve similar representa-

tion of DNAm and HM levels in the gene framework, histone modification levels are also expressed as average methylation/ acetylation percentage values, (Section 2.3). Initially, randomly-generated values within the ranges given in Table 4.5 are assigned to HM level, according to the cancer stage of the entire gene network. Histone modification is less stable in nature than DNam level, [Cedar and Bergman, 2009]. Considering this, lower differences between the value-ranges are taken for HM than for DNAm level initialization based on cancer stages, (Table 4.3). The E-G Network provides also functionality for integration of initial HM levels from external resources and the value-ranges from Table 4.5 can be refined if information becomes available. However, this model requires prior information on DNAm level in order to estimate HM level. A sensitivity analysis for these value-ranges is also considered, (Chapter 8).

Table 4.5: Initial histone acetylation and methylation ranges (average percentage values) based on cancer stage

Cancer stages	**Initial histone ranges**
Healthy	0.010 - 0.019
Carcinoma in situ	0.020 - 0.032
Invasive carcinoma	0.033 - 0.075
Metastasis	0.076 - 0.090

The interplay between DNAm and HM, (discussed in Subsection 2.3.4), has also been considered in the E-G Network Model. Thus, in addition to the data drawn from StatEpigen, [Barat and Ruskin, 2010], the methylation level of a gene G is updated using expression (4.14), (adapted from that given by Raghavan and Ruskin [2011]), which links the methylation and acetylation of the core histones H3 and H4 to DNAm:

$$M'(G) = \begin{cases} M(G) \pm Mean \times R, & \text{if } (M(G) \pm Mean \times R) \geq 0 \\ 0, & \text{otherwise.} \end{cases} \tag{4.14}$$

where: M'(G) = DNAm of the gene G in the next Time-step; M(G) = DNAm of the gene G in the current Time-step; R = random number ($\in$ [10^{-5}, 10^{-4}]). The sign '+' or '-',

corresponding to methylation level increase or decrease, respectively, is decided by a random variable, internal to the network. *Mean* is the average of acetylation and methylation of the histones H3 and H4. Given that acetylation is less stable in nature than methylation, the E-G model assumes that a percentage of the histone acetylation level can be lost during a cell cycle, (i.e. between two successive model iterations); thus, only the remainder is considered for *Mean* calculation:

$$Mean = \frac{p \times Ace + Me}{2} \tag{4.15}$$

where Ace, Me = acetylation, methylation level of histones H3 and H4; p $\in$ [0, 1] is a randomly-generated value.

The 'feedback' from DNAm to HM, i.e. the DNAm contribution to histone evolution inside a gene G, is described by expression (4.16). Histone modification level at time T+1 is updated based on i) histone acetylation/ methylation level at time T, ii) changes occurring during this time interval and iii) methylation level of gene G, (at promoter *locus*):

$$H'_E(G) = \begin{cases} H_E(G) \pm [W_E(G) + M(G) \times R], & \text{if } H_E(G) \pm [W_E(G) + M(G) \times R] \geq 0; \\ 0, & \text{otherwise.} \end{cases} \tag{4.16}$$

where: E = acetylation or methylation for the current histone; H_E'(G) = the histone acetylation or methylation level (governed by event E) in the next Time-step; H_E(G) = the histone acetylation or methylation level (due to event E) in the current Time-step; R = random number ($\in [10^{-5}, 10^{-4}]$); M(G) = DNAm of the gene G; W_E(G) = the weight ratio for the event E based on the cancer stage where gene G was indicated and is given by the product of two randomly-generated values Perc and R_W, i.e. W_E(G) = Perc $\times$ R_W, with Perc $\in$ [0, 1] (i.e. a percentage), and $R_W \in$ Table 4.5, which indicates that HM update rate is also dependent on cancer stages). Again, '+' or '-' sign is randomly determined.

Similar to the gene methylation updates based on the relationships data, (subsection 4.5.2), the updates of DNAm and HM (based on their interdependency) are also considered

to be probabilistic events. Thus, expressions (4.14) and (4.16) are applied only if the value of a randomly-generated probability *Rand* is higher than that of a threshold TH_HM, which is provided as a model input parameter, (expression 4.17). A sensitivity analysis will include also the TH_HM parameter.

$$Rand \geq TH\_HM \tag{4.17}$$

Following the objective to investigate genetic and epigenetic interdependencies during cancer development, DNAm and HM update steps are studied combined in the *methylation cycle*, which represents the process of updating network methylation level over time.

4.7 Network dynamics

Given its overall importance, the DNAm update step is taken as the fundamental time-step for the gene framework, as a whole. The gene methylation level change, from simple and conditional relationships, is referred to as the STATIC_UPDATE. Similarly, the methylation update, based on any 'self-relationship', is labelled as the DYNAMIC_UPDATE. The DNAm updates, based on HM, are labelled as the DNA_METH_UPDATE step, while the evolution of histones, given the DNAm level, is referred to as the HIST_MODIF_UPDATE step. Chronologically, a network *methylation cycle* starts with one DNA_METH_UPDATE step, followed by a STATIC_UPDATE phase, which is combined with one DYNAMIC_UPDATE iteration. Since DNAm has been reported to be more stable in nature than HM, [Cedar and Bergman, 2009], different dynamics are needed to handle HM events in the E-G Network Model. Hence, the methylation cycle ends with 'm' HIST_MODIF_UPDATE iterations, (m > 1). The steps of the methylation cycle are briefly illustrated in the simplified Figure 4.4.

In the figure, M_G denotes the methylation level of a generic gene G and H3,4 refers to the methylation and acetylation of the histones 3 and 4 corresponding to gene G. The dual influences between HM and DNAm are specified by the relationships (D1) and (D4). Additionally, the methylation level of the G gene is updated according to its conditional relationships with other network genes (G'), (relationship (D2), Figure 4.4). A 'self-relationship' is also considered to contribute to the new methylation level of the gene G, (relationship

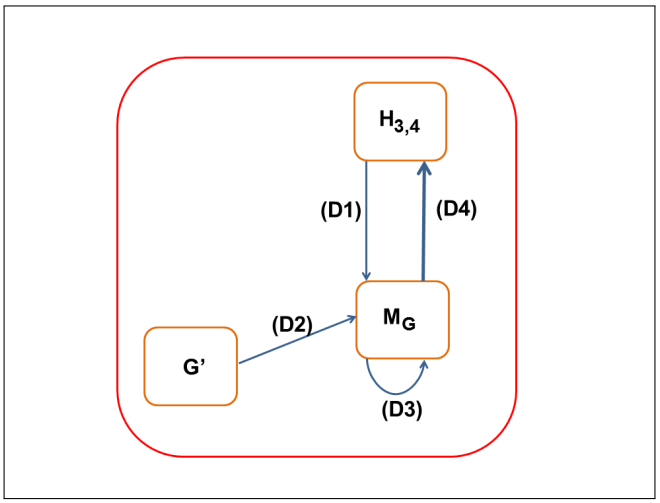


Figure 4.4: Dynamics of the methylation cycle of a generic gene G, in the E-G Network Model
Specifically, the D1 relationship refers to the DNA_METH_UPDATE step, D2 and D3 indicate STATIC_UPDATE and DYNAMIC_UPDATE phases respectively, and finally, D4 illustrates HIST_MODIF_UPDATE iteration. Legend: M_G: methylation of the gene G promoter; G': the genes from the network which are found in conditional relationship with the gene G; H3,4: average methylation and acetylation of the histones 3 and 4 corresponding to gene G.

(D3), Figure 4.4).

The gene network is allowed to evolve over time for a number of methylation cycles, ($\in [10^2, 10^4]$), and at the end of the simulation, the gene network is checked to ascertain if it is still in the initial cancer stage or has advanced, in which case the new stage is identified. The length of the simulation is determines by methylation cycle duration. Specifically, the gene network mimics dynamics of human colon stem cells, for which, methyation cycle was approximated to stem cell division time, i.e. around seven days, (Section 2.4). Thus, while a number of 10^2 iterations can be approximated to 2 years, ($\frac{100\ iterations}{52\ iterations/year} \approx 1.92$ years), a number of 10^4 iterations indicates $\approx$ 190 years. This extreme value is considered in order to analyse conditions when cancer evidence is not marked. Moreover, the E-G Network has been developed with potential for extension to other cell types, (such as progenitor and differentiated cells), or other tissue types, (such as the small intestine). Lifetime of both progenitor and differentiated cells is shorter than that of the stem cell; for example, if progenitor cycle is approximated to 2 days, a number of 10^4 would correspond to $\approx$ 55

years, (10^4 iterations $\times \frac{2days}{1iteration}$ = 20000 days $\approx$ 55 years), which is realistic with respect to human lifetime. Or, for example, if the gene network describes stem cell dynamics in a small intestine system, (with stem cell division cycle $\sim$5 days), the iteration number range [10^2, 10^4], can be approximated to 1.37 - 137 years. Thus, two approaches can be used to set simulation time in the E-G framework. Specifically, these refer to: a) when gene networks are allowed to evolve over time until cancer progresses to a specific stage and b) when an iteration number, (calculated based on information on stem cell lifecycle duration) is taken and cancer development assessed at the end of simulation.

Implementation of the E-G Network model The E-G Network is an object-oriented framework written using C++. It also uses Visual Leak Detector, [CodePlex, 2014], to detect plausible memory leaks in the code. System entities specific to the gene network, including histones, genes, micromolecular events, gene relationships and the gene network itself, are described by individual classes. In addition, functionality is provided for integration of information on patient features and for processing biological data from external resources, which are considered model input. Details on the E-G Network implementation is given in Section 5.3.3, following presentation of the extended E-G Network, (including the decision methods considered for cancer development and ageing/ gender influences into the model).

4.8 Discussion

The E-G Network, presented in this chapter, itself has a multi-layered structure, which considers three important elements involved in cancer development: molecular signals (genetic and epigenetic), gene relationships and the transitions between cancer stages.

The main contribution of the model is the introduction of an integrated framework to analyse different micromolecular events and their interdependencies, with respect to their impact on DNAm level measured at gene promoter location. An algorithm has been proposed for detecting the most *plausible* tumour pathways in a given gene network based on

genetic and epigenetic interdependencies observed at different CRC stages. This information can be further combined with signalling pathway data and integrated into an extended analysis on tumour routes. For example, given alterations of the generic gene G, what is the likelihood that the Wnt signalling pathway will be affected? An answer can be given using the 'plausible pathway' algorithm. If information on pathways between the gene G and any genes involved in the Wnt pathway exists, then these genes can be connected in a network, and the score of the pathway between them can be calculated. The immediate question is clearly how to interpret the significance of this score value in the given network: for instance, how different is a score of -0.33 from -0.31 or from -0.35? Secondly, the 'methylation cycle' has been introduced to represent the process of updating the gene promoter methylation level based on information on five component elements: network cancer stage, gene network structure, signalling pathways, HM and nature of the molecular event observed for specific genes. For a given gene network, the methylation level of a gene G can be updated according to information on micromolecular interdependencies affecting the gene itself and the genes to which G is connected in the network. Information on signalling pathway can also be considered. Thus, if two genes, known to be part of the same signalling pathway, are connected in a gene network, abnormalities in one gene increase the chance of the second gene being affected by genetic and epigenetic mechanisms involved. In addition, both DNAm and HM levels are taken to be influenced by cancer stages. Based on the dual relationships between these epigenetic patterns, (reported in literature), a set of expressions has been proposed for updating DNAm and HM levels over time.

A limitation of basic E-G Network is the absence of methods to evaluate malignant tumour development in the gene network. In addition, information on the influence of risk factors, such as ageing and gender, in CRC initiation and progression is also lacking. Considered a major risk factor in cancer development due to its influence on genetic and epigenetic events, (subsection 2.6.2), the argument for inclusion of *ageing*, in any realistic gene framework, is strong. In model terms, this can be linked to methylation level updates, given the progressive accumulation of DNA methylation and histone modifications over time, [Fraga et al., 2007]. Moreover, specific genes (such as ER, IGF2) that have been iden-

tified to be more *sensitive to age-related methylation* must be considered in this context. Given its reported influence on CRC initiation, specific account should also be taken of *gender* in any extended genetic/epigenetic framework. In consequence, the extended E-G Network Model, (described in the next chapter), integrates methods to asses cancer initiation and progression based on genetic and epigenetic information and focuses on analysis of ageing/gender impact on micromolecular events during CRC development.

Chapter 5

Extensions to E-G Network Model

5.1 Introduction

Following on from the limitations, notated in the prototype E-G Network Model, (Chapter 4), analysis methods for tumour progression, (as recorded by the E-G Network), and the way in which different risk factors, (such as ageing and gender), influence the dynamics of tumour initiation, are the main topics addressed in this chapter. Considered a major risk factor in cancer development, we report on the incorporation of *ageing* in the E-G Network model, as well as *gender*, as the influence of which is also implicit in colon cancer initiation, (Subsection 2.6.2). In addition, information on a set of genes that have been reported to be sensitive to age-related methylation in CRC development, (e.g. IGF2, ER), [Issa and Ahuja, 2000; Teschendorff et al., 2010], is also explored and integrated into the gene framework. The work presented aims to investigate the abnormal modifications from *carcinoma in situ* to invasive carcinoma, colon cancer, by addressing questions, such as: i) What differences are to be expected in tumour progression for males and females of similar age? ii) What is the impact of aberrant modification in the promoter of a gene, such as IGF2, that is considered to be sensitive to age-related methylation, (Subsection 2.6.2)? We address these questions through case studies which aim to exploit the potential of extended E-G Network Model. These case studies are based on synthetic patients, with profiles generated by recourse to methods for methylation level initialization based on cancer stages, (Section

4.5).

The methods, proposed for assessment of tumour progression in the gene network, are presented in Section 5.2 and two case studies, which illustrate the contribution of ageing and gender in tumourigenesis, are described in Section 5.3. Implementation details for the extended E-G Network and ongoing parallelisation development are also presented. The results obtained for each case study are reported in Section 5.4. Finally, a summary on the inclusion of the ageing and gender influences in the E-G Network and any limitations are given and the need for cell and tissue level modelling is highlighted.

5.2 Cancer progression evaluation

This subsection introduces the methods used to analyse tumour progression in the E-G Network. The gene network is rated as being in 'healthy' state or in one of the three possible cancer stages (namely *carcinoma in situ*, invasive carcinoma and metastasis), using two methods:

A. Calculation of the average network methylation[1] level, (*network score*);

B. Calculation of the percentage of highly methylated genes (due to influence of current genetic and epigenetic events).

Firstly, the average network methylation level, *network score*, (Method A from above), is tested against predefined methylation values (PMRA[2] values) based on the cancer stage, (Table 5.1), in order to assess the initial cancer stage of the gene network. Further, the network score is computed after every methylation cycle and tested against these value-ranges. For example, a network score of 18 indicates that the network is in *carcinoma in situ* stage. If its score is 54 after a number of methylation cycles, then the gene network is considered to be in *invasive carcinoma*, meaning that the tumour has progressed over the time period, (mimicked by simulation). The value-ranges from Table 5.1 are inferred from

[1] Similarly to the base E-G Network, (Chapter 4), the word 'promoter' is implicitly considered in formulation 'gene methylation level', which stands for 'gene promoter methylation level'. When confusions can occur, the word 'promoter' is specified explicitly.

[2] PMRA: Average percentage of methylated reference

those used for DNAm level initialization step[3], (Section 4.5), and considered for sensitivity analysis, (Chapter 8).

Table 5.1: Cancer stage decision based on average promoter methylation level (PMRA values) of the entire network

Average of network methylation level	**Cancer stages**
0 - (<) 8	Healthy
8 - (<) 18	*Carcinoma in situ*
18 - (<) 65	Invasive carcinoma
65 - 100	Metastasis

Method (A) is, however, limited by its sensitivity to outlier genes; a *single* gene with a high promoter methylation level can significantly increase the network score. Based on this observation, method (B), the *gene percentage* method, was derived to evaluate cancer stage. Here, the decision on network cancer stage is made in two steps: first, the highest promoter methylation level (PMRA values) is determined in the gene network and second, the percentage of genes that contain this methylation level is compared with two reference (percentage) values, e.g. 20% and 80%, (Table 5.2).

Table 5.2: The relationship between percentage of highly methylated genes and cancer stage (methylation level is given by PMRA values)

DNAm level	**0 - (<) 20%**	**20 - (<) 80%**	**80% - 100%**
0 - (<) 8	Healthy	Healthy	Healthy
8 - (<) 18	Healthy	*Carcinoma in situ*	Invasive carcinoma
18 - (<) 65	*Carcinoma in situ*	Invasive carcinoma	Metastasis
65 - 100	Invasive carcinoma	Metastasis	Metastasis

According to information from Table 5.2, if the highest methylation level in a gene network is between 8 and 18, observed in less than 20% of total gene number, the gene network is still considered to be in a 'healthy' stage, even if this range is accounted to be specific to *carcinoma in situ*. However, if the same highest methylation level is recorded for more than 80% of the total gene number, the respective gene network is in the invasive carci-

[3]The value-ranges used to initialize DNA methylation level are included in Table 4.3. In choosing these values, the major assumption was that higher variation in DNAm level is observed as tumour progresses, e.g. from 0 - 8 in 'healthy' phenotype, to 65 - 100 in metastasis.

noma stage already, and does not represent *carcinoma in situ*. Similarly, for example, if the highest methylation level ≈ 50 is 'measured' for around 12% and 82%, respectively, of all genes included in two gene networks, the first network is the *carcinoma in situ* stage, while the second network is already marked into metastasis. The values-ranges from DNAm level are similar to those included in Table 5.1. A sensitivity analysis is necessary to investigate these value-ranges; for example, methylation level intervals can be equilibrated, e.g. 0 - 33%, 33 - 67%, 67 - 100% or extended more, e.g. 0 - 5%, 5 - 95%, 95 - 100%.

The 'network score' and 'percentage of highly methylated genes' methods can be combined if needed. For example, if the focus is on the epigenetic, (or methylation), landscape of the entire cell, the first method may have a higher weight on the tumour progression decision. Contrary, if the interest is on a specific set of genes, the second method may be considered more applicable for purposes of analysis.

Discussion on cancer progression

The flow chart from Figure 5.1 is a schematic representation of the way in which tumour progression is decided inside the E-G Network Model.

Initially, the network includes gene relationships within the lowest cancer stage. If several gene relationships containing the same genes, but in different cancer stages, are found in StatEpigen, the ones from the most advanced cancer stage are kept in a repository for further refinement. After each methylation cycle, the network is rated, using the methods described in the previous subsection. At this step, the gene network can be updated if it has progressed to a more advanced cancer stage. Some of the actual gene relationships may be overwritten, (by updated information for similar relationships from the repository). Additionally, new gene relationships corresponding to the current cancer stage can be included. The gene framework is allowed to evolve over time, in order to register its 'potential' for cancer progression.

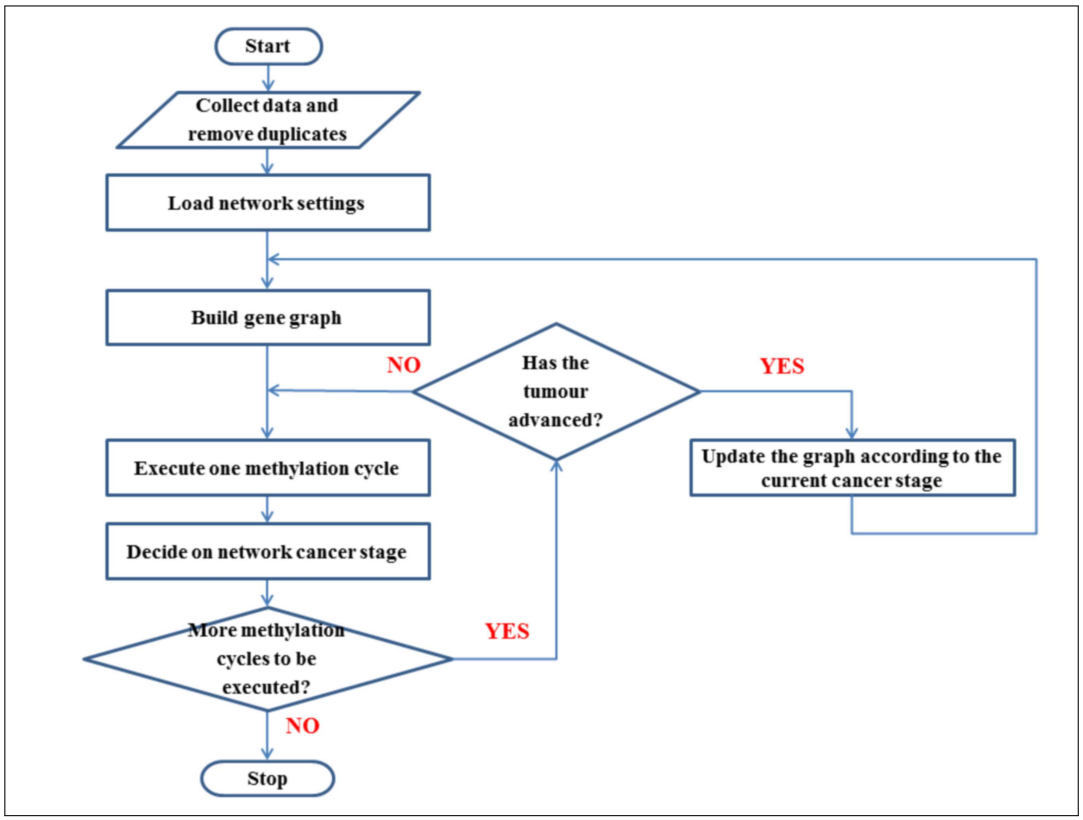


Figure 5.1: Schematic view of tumour progression possibilities in the E-G Network model

5.3 Implementation

The E-G Network enables analysis of methylation level changes for the entire gene network over time and gives information on cancer initiation and progression. Given that epigenetic patterns can be inherited through cell generations, [Probst et al., 2009], the DNAm level was analysed after every stem cell division step, [Boman and Wicha, 2008; Clevers, 2011; Papailiou et al., 2011]. Therefore, for the case studies described in this chapter, the major considerations were that the gene network should mimic the molecular modifications inside an (adult) colon stem cell and that the time step was realistic: for a methylation cycle is the interval between two successive stem cell divisions, which was estimated to be equal to around seven days, [Potten et al., 2003; Frank, 2007], (Section 2.4).

The ageing and gender impact on micromolecular events

The influence of an individual's age and gender with respect to genetic and epigenetic modifications over time was handled as follows. An *AG* parameter was proposed to describe the impact of these factors on DNAm dynamics, with values inferred from available literature, (e.g. Ferlay et al. [2014]). Given that a *mutation* event is considered to have lower impact on DNAm level than *hypo-* and *hyper*methylation events, the ageing and gender influences on mutation occurrence probabilities are described by a randomly-generated value, R_{Mut}, lower than the corresponding AG parameter values. The AG parameter set has the same value for a given gene network in its entirely, (as individual characteristics are constant for a given cell), and is integrated into the DNAm level update step, (Section 4.5). Specifically, the AG parameter affects the *threshold* of the micromolecular event probabilities, (introduced in expression (4.12)), as shown below:

$$TH\_DNAM_k(G) = 1 - P_k(G) - SIG(G) - AG(ageing, gender) \tag{5.1}$$

Thus, in addition to its dependence on the gene relationship probability values and to the gene characteristics forming part of a signalling pathway, (i.e. P_k(G) and SIG(G)), the probability threshold that permits micromolecular modifications on a given gene G, TH_DNAM$_k$(G), is also influenced by individual's age and gender.

Similarly, the influence of individual characteristics has been also considered with respect to the DNAm-HM relationships, (Section 4.6). In model terms, the constant value of the TH_HM threshold, (used in expression (4.17) for updating DNAm and HM level based on their interdependencies), has been replaced by a parameter based on ageing and gender. Thus, the dual-relation HM and DNAm updates are now permitted only if the randomly-generated *Rand* value is higher than that given by the TH_HM parameter for a specific ageing-gender group, i.e.:

$$Rand \geq TH\_HM(ageing,\ gender) \tag{5.2}$$

Genes sensitive to age-related methylation

Research, [Issa and Ahuja, 2000; Teschendorff et al., 2010], has also identified a specific group of genes known to be sensitive to age-related methylation and this information can be incorporated into E-G Network. In model terms, a coefficient, denoted as '*sensitive to an age-related methylation*', *(SAM)*, was considered to affect the value of the probability threshold that permits DNAm level modifications in the gene network. Thus, methylation modification probability of a gene sensitive to age-related methylation is higher than that of a gene without this feature. In addition, a coefficient, denoted as '*age-sensitive gene*', *(ASG)*, was introduced to target to the individual age at which such age-related methylation modifications can be observed. Therefore, the methylation update step expression becomes:

$$TH\_DNAM_k(G) = 1 - P_k(G) - SIG(G) - AG(ageing, gender) - SAM \quad (5.3)$$

where

$$SAM \begin{cases} \in (0,1), & if\ Ageing \geq ASG; \\ 0, & otherwise. \end{cases}$$

Network settings common to both case studies

In section 2.6, four major age groups were identified and described based on the risk of CRC development. For both case studies described in this chapter, ten core age groups were taken, (starting at age 38 going in 5 year steps to age 83: 38, 43, 48, 53, 58, 63, 68, 73, 78, 83). In addition, an example of AG parameter values, defined for the *hypo-* and *hyper*methylation events, is provided in Table 5.3. Both age groups and value-ranges were chosen based on literature which reported that CRC risk starts increasing at around 40 (in hereditary CRC, [Cunningham et al., 2010]) and has its peak between 65 and 75, [Scholefield, 2002]. The age-range was chosen in this way in order to span and slightly extend this age range (from 40 to 75+), with the network 'age 38' viewed as one control sample, where tumour progression is not influenced by ageing impact on genetic and epigenetic modifications, and the network for 'age 83' is another control sample, where ageing damps cell dynamics, and consequently, aberrant changes involved in tumour development, [Campisi,

2003].

Authors have suggested that the 5-year interval is important in analysis of gender differences reported in CRC development, [Frank, 2007; Brenner et al., 2007], and we have followed this for the exploration here although in practice, of course, any time step could have been chosen to investigate age profile overall. Indeed, in choosing a slightly extended upper and lower bound for the age range considered, we have deliberately allowed for some fuzziness in terms of whether age-difference is specific to a particular age-range or observed more generally.

Table 5.3: Example of AG parameter values for different groups of individuals, based on age-gender characteristics

Age-groups	**Male**	**Female**
<40	0	0
40 - 44	0.10	0
45 - 49	0.10	0.10
50 - 54	0.17	0.10
55 - 59	0.25	0.17
60 - 64	0.35	0.25
64 - 69	0.35	0.35
70 - 74	0.27	0.35
75 - 79	0.23	0.27
80+	0.23	0.23

Note: The values from Table 5.3 are inferred from different statistical reports on CRC incidence, e.g. GLOBOCAN 2012, [Ferlay et al., 2014], Cancer Research UK, [Cancer Research UK, 2014b]. In model terms, the gene relationship probabilities in the gene network are affected during the DNAm update step.

Analogously, an example for the TH_HM parameter is included in Table 5.4, with age/gender patterns taken for TH_HM parameter value choice similar to those followed for AG parameter value selection. The ASG value was set to 50 years, given that the sporadic CRC incidence was reported to increase after this age, [Issa et al., 1994]. Research has indicated that abnormalities of signalling pathways can have high impact on cancer development, (decided in the gene network based on methylation level, (Section 5.2)). Values of 0.05 and 0.10 were taken for SAM and SIG based on tuning parameters: the probability

of an event occurring for a gene G found in a signalling pathway was taken to increase with maximum 0.50 the initial probability taken from StatEpigen database, (expression 5.3). This extreme value is obtained when the gene network represents a male, age 60 - 70, (or female, age 65- 75), e.g. AG = 0.35, and gene G is also sensitive to age-related methylation, (e.g. SAM = 0.05), i.e. AG(age, gender) + SAM + SIG = 0.50. Other values are, of course, also possible and sensitivity analysis for these parameters is planned, (Chapter 8).

Table 5.4: An example of the TH_HM parameter for different groups of individuals, based on age-gender characteristics

Age-groups	**Male**	**Female**
<40	0.60	0.60
40 - 44	0.55	0.60
45 - 49	0.55	0.55
50 - 54	0.48	0.55
55 - 59	0.39	0.48
60 - 64	0.39	0.39
64 - 69	0.28	0.39
70 - 74	0.28	0.28
75 - 79	0.37	0.28
80+	0.37	0.37

Note: Similar to the values in Table 5.3, the values from Table 5.4 are inferred from different statistical reports on CRC incidence, e.g. GLOBOCAN 2012, [Ferlay et al., 2014], Cancer Research UK, [Cancer Research UK, 2014b]. In model terms, they describe the probability thresholds for permitting HM updates, based on DNAm - HM interdependencies.

Initially, tumours of all patients were considered to be in *carcinoma in situ*. The goal of the simulations was the investigation of instances when the gene networks moved to the next cancer stage, invasive carcinoma. The focus was on promoter hypermethylation information, and in order to decide on tumour progression, the average methylation level for the entire gene network has been analysed, (using *the network score* method described in Section 5.2). Therefore, an increase in the average methylation level of the gene network can be associated with tumour development. The features of each case study are described below.

5.3.1 Case study 1 on 'Combined ageing and gender influence'

Genes such as APC, MGMT, MLH1, KRAS, TP53 are considered to be *key*-genes in CRC, given the impact on tumour development of any changes, (Section 2.6). Based on this information, four different gene networks were built according to the following criteria:

1. The first network, ('APC-TP53' network), was created including data on conditional relationships between micromolecular events affecting APC, TP53 and other genes.

2. The second network, ('KRAS-BRAF' network), included the conditional relationships between KRAS and BRAF and other genes.

3. The third network, ('APC-MGMT' network), utilised the conditional relationships between APC, MGMT and other genes.

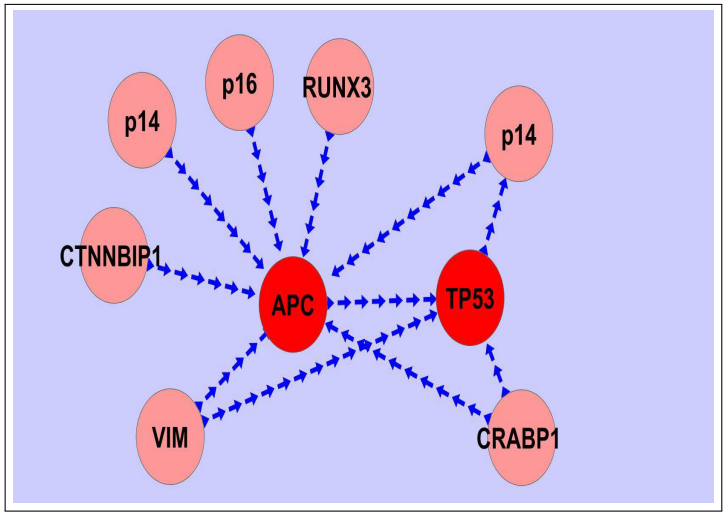
4. The fourth network, ('APC-MLH1' network), incorporated the conditional relationships between APC, MLH1 and other genes.

The network size in each case was similar in terms of the number of nodes and edges, which were in the range 9 - 15 and 10 - 18 respectively. The graphical data visualisation was created using *Cytoscape*, [Smoot et al., 2011], and network structures are illustrated in Figure 5.2. The study was designed to analyse 100 patient profiles for each gender, (with 10 male and 10 female profiles generated for every age subgroup) and for every gene network.

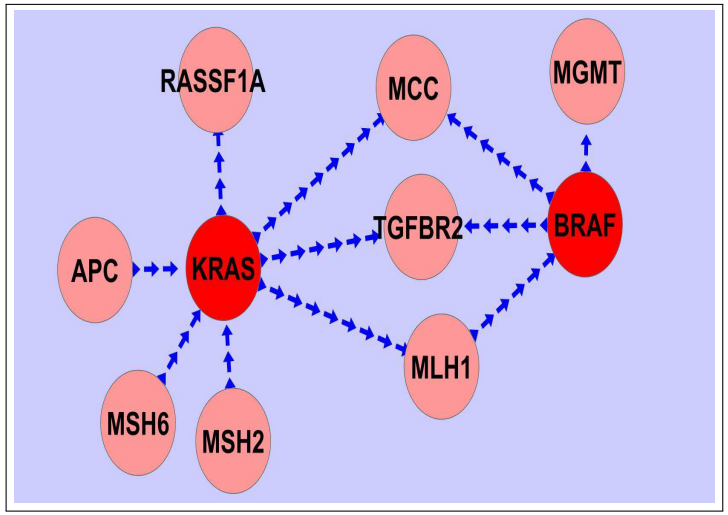
These patients are wholly synthetic, generated from set of possible parameter values given in Table 4.3, (Section 4.5). Specifically, this case study focused on the following questions:

1. What is the relationship between younger and older patients regarding rapidity of tumour progression, (advance to the next cancer stage)?

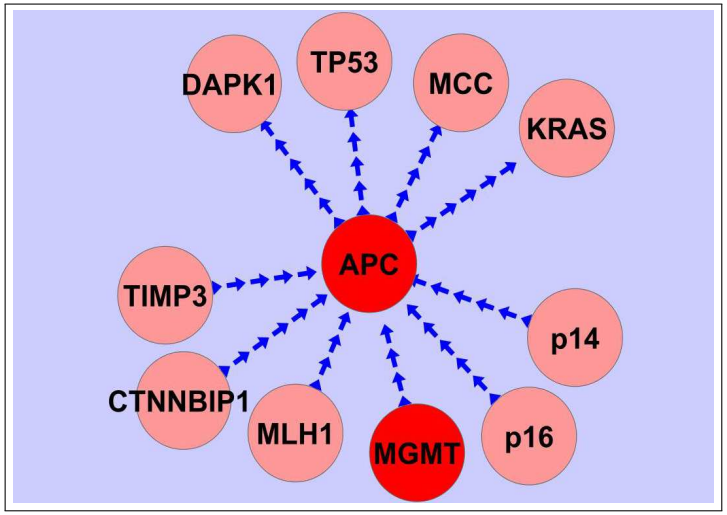
2. Is there a marked difference between males and females in terms of methylation level observed over time?



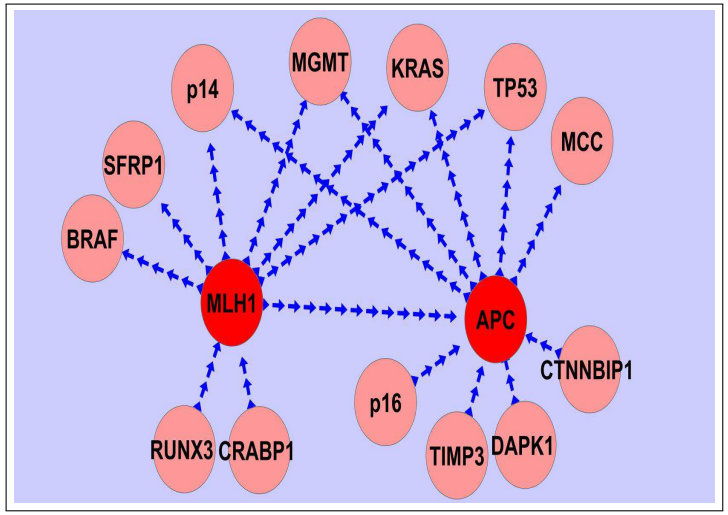
a) APC - TP53 network



b) KRAS - BRAF network;



c) APC - MGMT network

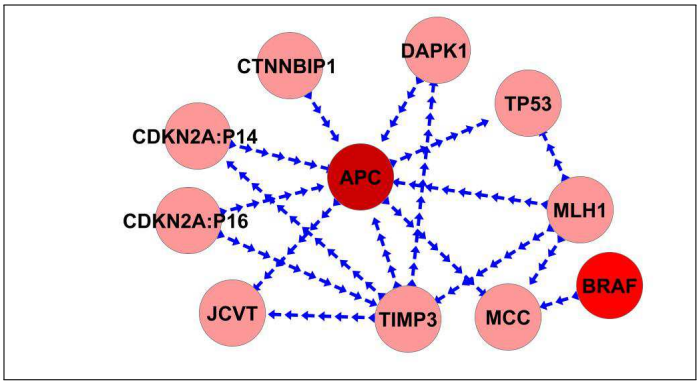


d) APC - MLH1 network.

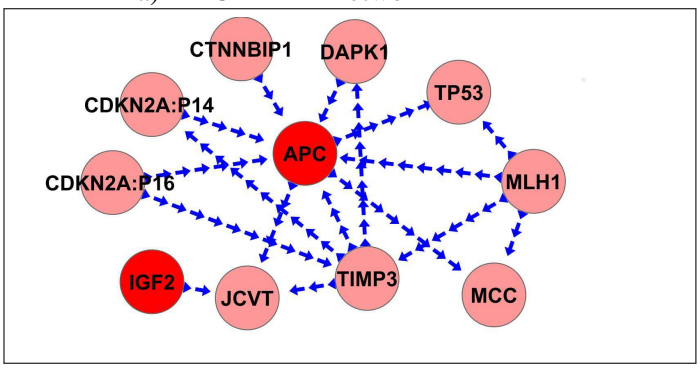
Figure 5.2: The structure of the a) 'APC-TP53', b) 'KRAS-BRAF', c) 'APC-MGMT' and d) 'APC-MLH1' gene networks investigated for the analysis of the ageing/gender impact on micromolecular events. The networks were built using conditional gene relationships data from StatEpigen database. Legend: Circle - gene in network, Edge - gene relationship.

5.3.2 Case study 2 on 'Gene sensitive to age-related methylation'

In order to analyse the impact of aberrant methylation of gene sensitive to age-related methylation, two gene networks, namely 'APC-BRAF' and 'APC-IGF2', were built from empirical data on relationships between APC and other genes, (e.g. MLH1, TP53), with a focus on promoter hypermethylation information. The major difference between the two was, thus, the replacement of the BRAF gene with the IGF2 gene, (since the latter is considered as being potentially sensitive to age-related methylation, [Issa and Ahuja, 2000]). The conditional relationships between these key genes and others in the 'APC-BRAF' and 'APC-IGF2' networks were also taken into account.



a) 'APC-BRAF' network



b) 'APC-IGF2' network

Figure 5.3: Structure of the 'APC-BRAF' (a) and the 'APC-IGF2' (b) gene networks. The IGF2 gene is considered to be sensitive to age-methylation, therefore the focus was on the ageing impact on the methylation level deregulations inside a cell, (mimicked by the network), with aberrant modifications of the IGF2 gene. The networks were built using conditional gene relationships data from the StatEpigen database. Legend: Circle - gene in network, Edge - gene relationship.

The network structures are illustrated in Figure 5.3. Similar to the previous case study, 100 patients profiles have been generated for each gender for both gene networks, (i.e. synthetic patients).

5.3.3 Implementation details of the E-G Network

Class description and diagram

The E-G Network model is an object oriented framework (written using C++), where the major classes are the following, (with the class diagram illustrated in Figure 5.4):

- *BayesianNetwork* - This is considered the main class of the E-G Network model as it encodes the gene network functionality and assures the communication between the model core and external applications. For instance, preprocessing of the biological data, (i.e. removing duplicates from the initial data and those relationships reported to have zero probability), loading an individual's characteristics, building the gene graph, resolving the potential circuits, evaluating the gene network, calculating the gene pathway scores, processing 'cellular divisions', are several of the actions performed by the BayesianNetwork objects. Among other elements, a BayesianNetwork object contains a NetworkSettings object and a collection of GeneRelationship objects.

- *NetworkSettings* - All characteristics presented by an individual, (e.g. ageing, gender, heredity, viral and bacterial infection), are loaded into a NetworkSettings object and transmitted to the BayesianNetwork in one step.

- *GeneRelationships* class - connects information on gene relationships from StatEpigen database, [Barat and Ruskin, 2010], (encoded by the GeneEvent class), and gene characteristics, (e.g. gene name, gene sensitivity to ageing, methylation level), which are encoded by the Gene class. In addition, the class initializes the gene methylation level based on cancer stage information.

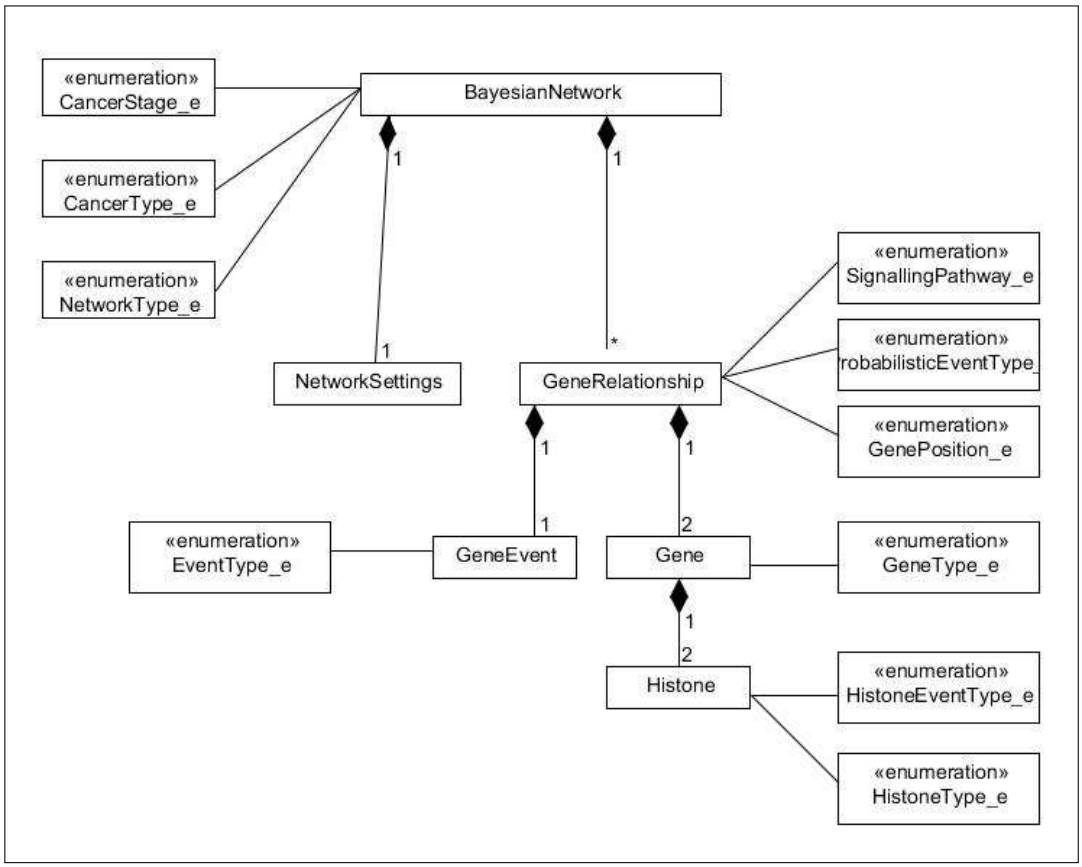


Figure 5.4: Class diagram for the E-G Network model. The dark-shaded diamond shape indicates the composition relationship between classes, (i.e. the component lifecycle is controlled by the container-class: if the latter is destroyed, the former is destroyed as well). For example if a BayesianNetwork object is destroyed, the instances representing the compound graph elements (nodes and edges) are also destroyed. In addition, the relationship between the numbers of instances of classes is represented, when information on precise ratio between those instances is known (e.g. Gene and GeneRelationship classes), by precise integer values; otherwise, when a variable number of compound instances is contained within a class, (e.g. GeneRelationship and BayesianNetwork classes), the '*' sign is used.

- *GeneEvent* - This handles causal relationships between two molecular events, (e.g. hyper/hypomethylation, mutation), affecting two genes of a certain 'cancer stage, cancer type' pair. It represents the information from StatEpigen database.

- *Gene* - This encodes information on gene characteristics, (e.g. name, type - tumour suppressor, oncogene, caretaker; methylation level), and information on 'self-relationship', (Section 4.2), and on signalling pathways that include it. Moreover, it 'communicates' with H3 and H4 histone regarding modifications.

- *Histone* - This class handles information on H3 and H4 histones and provided functionality for the initialization and update of their modifications, (e.g. acetylation and methylation), also based on the dual relationship between the DNAm and HM, (Section 4.7).

Finally, the doxygen tool, [van Heesch, 2008], was used to generate code documentation, which can be accessed online, at the following url:
http://www.computing.dcu.ie/ iroznovat/documentation/html/index.html.

Parallelisation strategies for the E-G Network

The case studies for the E-G Network integrated gene networks of modest size (nodes in range 9 - 15 and edges in 10 - 18, respectively) and analysed ageing and gender influences for a total of 1000 'patients' with respect to cancer initiation, (i.e. transition between *carcinoma in situ* and invasive carcinoma). Given these settings, serial simulation took around 114 min. However, the aim is to evaluate risk factor impact on methylation level changes in larger system size, (e.g. $\sim$ 100 genes), for larger patient number, (e.g. 10^3 for each age/gender group per network), for longer time-periods, (e.g. to explore modifications between *carcinoma in situ* and metastasis), and for larger gene network set, (i.e. to mimic cell population dynamics). In addition, given the main objective of our research to investigate CRC dynamics in multi-scale systems, (Section 1.2.1), parallelisation implementation is clearly needed to reduce simulation running time. However, development of parallelisation strategies is not trivial and the main challenge is minimization of compute node communication overhead.

The performance of a parallel algorithm can be evaluated by *speedup* (S_P) and *efficiency* (E_P) values:

$$S_P = T_S/T_P \tag{5.4}$$

$$E_P = S_P/P \tag{5.5}$$

where P = processor number, T_S = execution time for serial algorithm, T_P = execution

time of the parallel algorithm with P processors. While *linear* speedup, (achieved when S_P = P), or efficiency = 1 is ideal, this is not generally obtained in complex systems, (where high communication between compute nodes is needed to represent relationships between compound entities). Thus, rigorous analysis of system dynamics is necessary prior implementation of parallelisation algorithm. In the E-G Network, communication is handled using MPI paradigm, (*Message Passing Interface*) and two parallelisation strategies have been identified. These target i) the 'patient' groups analysis and ii) methylation level update step, (Section 4.5).

In the 'age/ gender' case studies, given that no communication was needed among systems, the main thread allocated a node to each ageing/gender group, (i.e. in total, 20 nodes/ network), and collected results at the end of the simulation. In this case, *temporal* parallelisation was used, (i.e. iterations for the gene networks run simultaneously), and efficiency close to 1 was achieved. Specifically, efficiency ranked from 0.87 for the 'KRAS-BRAF' networks to 0.99 for the 'APC-BRAF' systems, (speedup and efficiency values for each gene network are included in Table 5.5). This strategy can be used also for testing a group of identical gene networks, i.e. for cell population, (discussed in Chapter 8).

Table 5.5: Speedup and Efficiency values for each gene network included in the 'age/ gender' case studies

Gene network	Serial running time	Parallelisation running time	Speedup (20 nodes)	Efficiency
APC - TP53	807 sec	43 sec	18.76	0.93
KRAS - BRAF	2275 sec	128 sec	17.77	0.88
APC - MGMT	758 sec	38 sec	19.94	0.99
APC - MLH1	900 sec	48 sec	18.75	0.93
APC - BRAF	1109 sec	56 sec	19.80	0.99
APC - IGF2	1029 sec	54 sec	19.05	0.95

Note: Results from Table 5.5 are based on tests run on 20 nodes/network, with 10 'patients'/ node.

The second parallelisation method targets division of the gene network into subnetworks, (i.e. *spatial* parallelisation). The challenge here is characterisation of the optimal subnetwork size, (with respect to both node (gene) and edge numbers). A solution can be

given by detection of 'hub'[4] genes and splitting the gene network in the way in which every 'hub' gene determines a subnetwork. An immediate derived approach is running every gene on a compute node, given that no communication between genes is needed during methylation level update step. However, in both cases, communication master/ slave is necessary after every methylation cycle in order to inform on the new gene methylation level, (required for calculation of e.g. the average network mehtylation level, used in cancer decision methods, (Section 5.2)). In this case, a solution for overhead reduction can be given by extending time-interval for information exchange between master/ slave, (e.g. at every 4, 10, .. iterations). This work is ongoing and given communication in the gene network, linear speedup is again not expected. However, parallelisation implementation in the E-G Network is possible and would facilitate extended model testing.

5.4 Results and discussion

Methylation level was seen to increase for every patient during the simulations. This is due, in part, to the fact that the current study included mostly data on promoter hypermethylation. According to the first decision method (described Section 5.2) on methylation level increase in the gene network, any tumour therefore advanced to the next cancer stage for all patients over the time period of the simulation. For an individual patient, the time taken to tumour advance to invasive carcinoma was given by the *methylation cycle number*, measured during the simulation. Further, for each age/gender group, the average of the methylation cycle numbers, denoted as *methylation cycle number average, (MCA)*, was calculated. The results obtained are discussed below.

5.4.1 Results for case study 1 on 'Combined ageing and gender influence'

The MCA number at which the tumour advanced to invasive carcinoma is illustrated in Figure 5.5 for each of the four networks.

[4] A 'hub' gene is considered to be a strongly-connected gene in the network, (subsection 4.5.2).

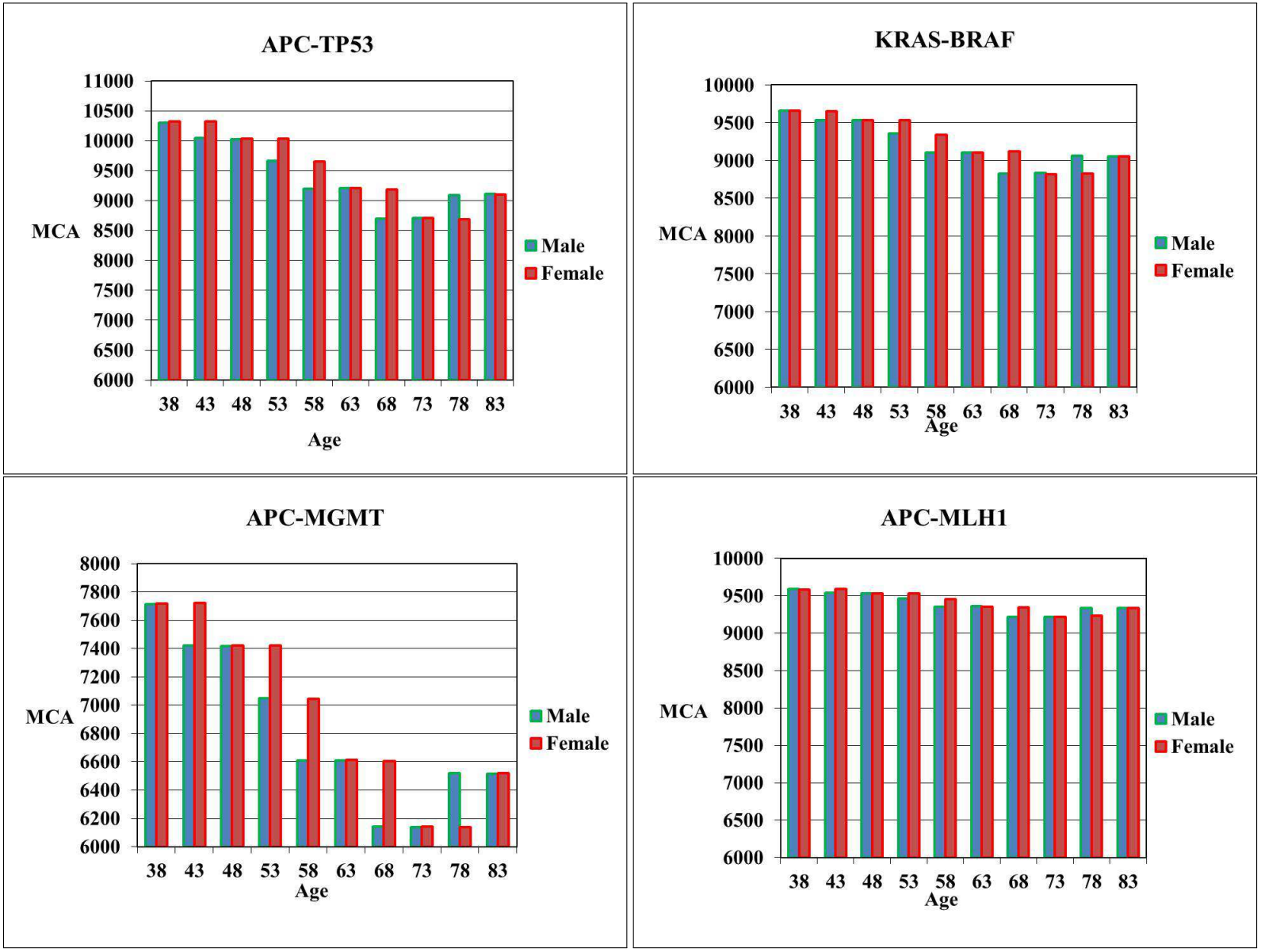


Figure 5.5: The average network Methylation Cycle Number, threshold at which tumours advance to Invasive Carcinoma for every age/gender patient group; Legend: male (blue), female (red)

Gender influence

Some gender differences can be observed in the analysed networks, where MCA numbers recorded for males from one age subgroup are similar to those of females from the next age subgroup. For example, results from the 'APC - TP53' network show that the time taken for tumour progression in a female of age 53 is closer to that for a male of age 48 than to that of a male of age 53, (i.e. MCA numbers corresponding to the fourth female and third male ageing subgroups, (10,032 and 10,024), are closer than those for the fourth female and fourth male age subgroups, (10,032 and 9,662)). The same remark can be made, for example, for the MCA numbers seen for males of 63 years and females of 68 years from the 'KRAS - BRAF' network, (the 6^{th} male and 7^{th} female subgroup respectively), where $\text{MCA}_{Female,68} \approx 9.188$, $\text{MCA}_{Male,63} \approx 9.206$ and $\text{MCA}_{Male,68} \approx 8.694$.

Ageing influence

For an analysis on the age influence, the results shown in Figure 5.5, were regrouped into four major age groups: less than 50 years, (< 50), between 50 and 64 years, (50-64), between 65 and 74 years, (65-74), and higher than 75 years, (75+), as shown in Figure 5.6. This rearrangement was carried out to reflect age impact on tumour development, specifically reported in the cancer literature: CRC risk increases significantly after age 50, has its peak between 65 and 75 years and decreases slowly after age 75, due to age-related influence on cell dynamics, [Christensen et al., 2009; SEER, 2013a; Cancer Research UK, 2014a; Ferlay et al., 2014].

The age influence can be seen in cancer development for all gene networks. The MCA numbers are lower for the '65-75' and the '75+' age groups showing that progression to invasive carcinoma is faster for those age categories than for younger patients (i.e. of '$<$50 years' and '50 - 64 years'). For example, in 'APC-TP53' networks, while the MCA numbers for the '$<$50 years' group were $\approx$ 10,200, those specific to '65-75' age group were $\approx$ 8,800 for both genders.

However, the age difference was far less marked in the 'APC-MLH1' network. One plausible cause is the network structure, (Figure 5.2). (The 'APC - TP53', 'KRAS - BRAF'

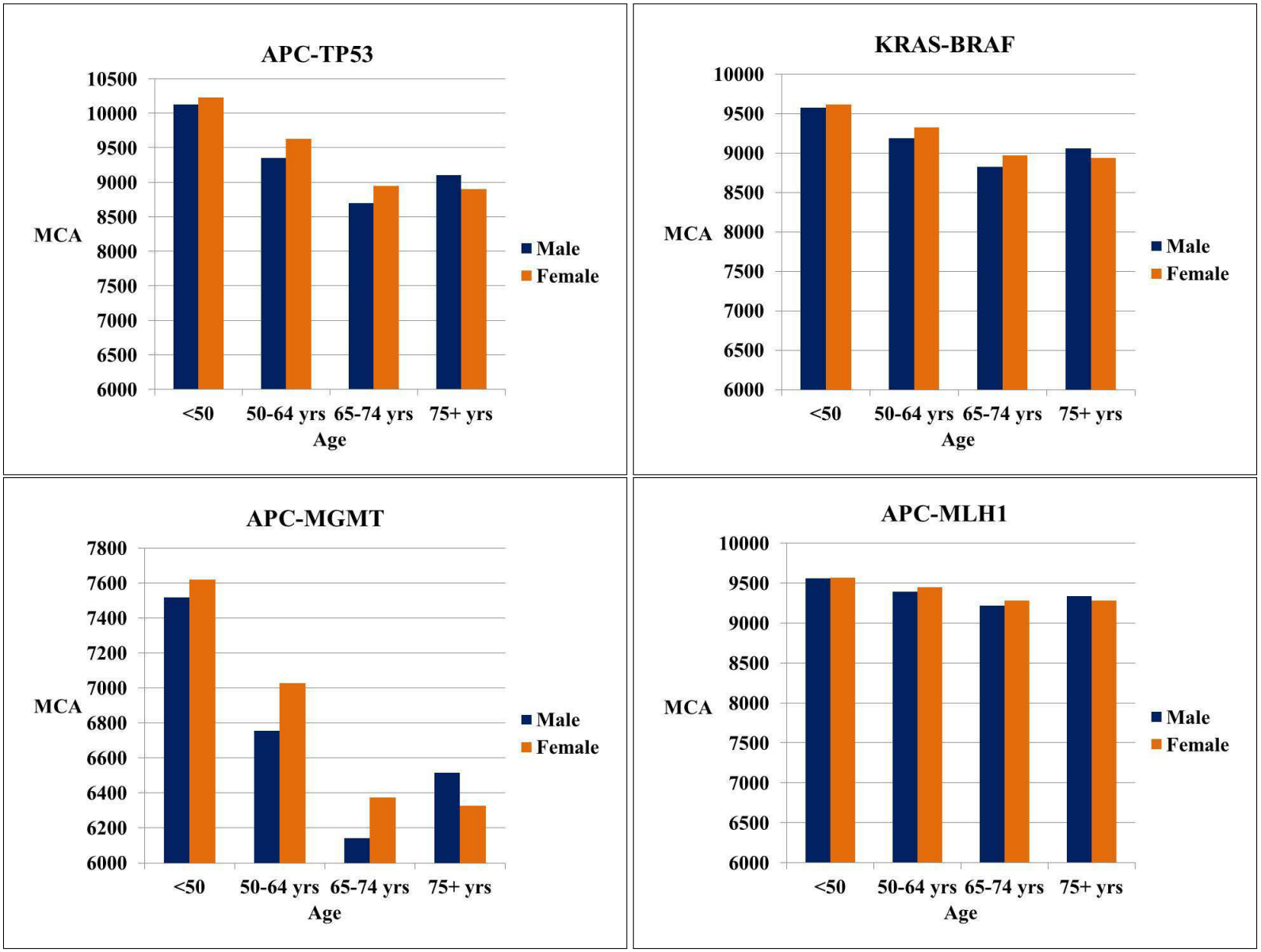


Figure 5.6: The average network Methylation Cycle Number after which a tumour advance to Invasive Carcinoma categorised by four major age groups: less than 50 years, (< 50), between 50 and 64 years, (50-64), between 65 and 74 years, (65-74), and higher than 75 years, (75+). Legend: male (blue), female (orange).

and 'APC - MGMT' networks are represented by a star structure with only one 'hub' gene, while the 'APC - MLH1' has a star structure with two 'hub' genes). Additionally, network size and gene relationships that depend on the 'hub' gene can impact on these results. Clearly, this is a preliminary analysis and further tests are needed to analyse other dependent networks (e.g. APC-KRAS, APC-MCC), with e.g. three 'hub' genes (such as APC, MGMT and MLH1) and similar number for connected nodes, as well as those networks where no. nodes is increased, (e.g. to 100+).

5.4.2 Results for case study on 'Gene sensitivity to age-related methylation'

The results obtained do indicate that time taken for tumour progression in 'APC-IGF2' was shorter than that for the 'APC-BRAF' network. Since the IGF2 gene is sensitive to age-related methylation, its abnormal promoter methylation makes a higher contribution to the total network methylation level. Similar to the analysis of age influence from the previous case study, results obtained were grouped into the four major age-groups, suggested by the literature (i.e. $<$50, 50-64, 65-74 and 75+ years, respectively), for each gender. The overall comparison on the MCA numbers corresponding to the 'APC-BRAF' and 'APC-IGF2' networks is illustrated in Figure 5.7, and detailed information is provided in Table 5.6.

Specifically, the 'APC-IGF2' networks moved to invasive carcinoma during shorter time-periods than the 'APC-BRAF' networks, where time differences recorded for these networks were around 725, 620, 550 and 620 methylation cycles for the four major age-groups respectively. 'Gender difference' was illustrated also by results of 'APC-BRAF' and 'APC-IGF2' networks, where the lowest MCA number, (corresponding to the shortest time of tumour to progress from Stage 0 to Stage 1), was found for the '65-74 years' and '75+ years' groups in males and females, respectively.

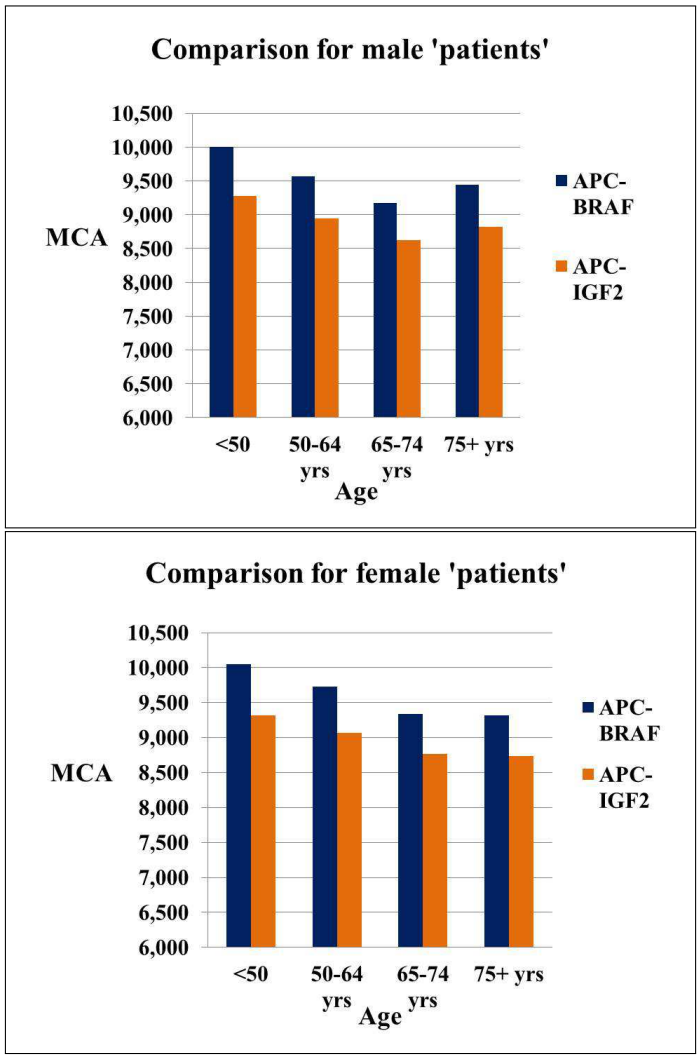


Figure 5.7: Comparison of time, (MCA number), at which the 'APC-BRAF' and 'APC-IGF2' gene networks move to Invasive Carcinoma. Results refer to male (top), and female 'patients', (bottom). Legend: Results for the 'APC-BRAF' network are shown in blue and those for the 'APC-IGF2' network in yellow.

While results provide some support for age-sensitive gene contributing to cancer growth, further tests on other genes exhibiting age-related methylation are clearly also necessary to assess relative impact on disease progression. In addition, further tests including larger network samples should facilitate investigation not only of weather these differences are significant in the statistical sense between these age groups but biologically supported, given deregulations of age-sensitive genes such as IGF2.

Table 5.6: Detail on MCA number results corresponding to gender and age-groups: <50 years, 50-64 years, 65-74 years and 75+ years, for both 'APC-BRAF' and 'APC-IGF2' gene networks.

Gender	**Age-Group**	**APC - BRAF**	**APC - IGF2**
Male	<50 years	10,001	9,277
	50 - 64 years	9,571	8,948
	65 - 74 years	9,177	8,622
	75+ years	9,445	8,825
Female	<50 years	10,049	9,320
	50 - 64 years	9,734	9,069
	65 - 74 years	9,334	8,761
	75+ years	9,316	8,738

5.5 Summary

Two methods have been proposed for deciding whether cancer progression occurs, based on methylation level: i) average network methylation level and ii) percentage of highly methylated genes. In addition, parallelisation strategies have been identified for the extended E-G Network and the need of parallelisation implementation in the context of more complex scenarios for the gene network was highlighted. Moreover, the influence of ageing and gender with respect to genetic and epigenetic modifications induced over time has been modelled using the following steps:

- Definition and incorporation of an Ageing-Gender (AG) parameter in the methylation level update step of the basic E-G Network Model in order to represent the impact of an individual's ageing and gender on DNAm level.

- Definition of the TH_HM function accounting for ageing and gender, to accommodate the update of the histone modification patterns based on DNAm - HM interdependencies.

- Increased probability of methylation modification occurrence permitted for those genes, identified as sensitive to age-related methylation for individuals $>$ e.g. 50 years of age.

Obviously, values of these quantities can be further refined as new data become available. In order to investigate the impact of ageing and gender on DNAm level in CRC initiation, a case study was performed and four gene networks were considered. These included information on a set of *key*-genes in CRC development and a spectrum of age for both genders. In a second case study, abnormal changes of the IGF2 gene were also considered in terms of the effect on gene network. Results obtained to date showed that tumour progression rate does appear to differ across patient groups and is influenced by ageing, gender and associated gene characteristics. This rate was found to be higher for older, (i.e. '65+ years'), than for younger patients, (i.e. '$<$ 65 years'), for men than for women, and for those presenting with an aberrant modification in the IGF2 gene compared to those without. These conclusions are based on methylation cycle number analysis. The model can be refined to include incorporation of these features as can be shown to reproduce effects of ageing and gender on methylation for some genes in agreement with clinical studies. Not all genes seemed to be similarly age-influenced despite being treated similarly, which argues for further refinement being needed in the way in which the model accounts for age. Of course, experimented data are limited to date on these features also, but the model does permit exploration of 'what if' scenarios, which may help in identifying possible outcomes for mapping also onto real patient data.

A clear limitation of the E-G network model is the lack of precise information on the *real length* of the methylation cycle and consequently, this component is approximated to the colon stem cell cycle, which lasts around seven days, [Potten et al., 2003]. For example, based on current results, the average methylation cycle number is equal to $\sim$8,690 for the 75+ years old female in 'APC - TP53' network. According to the stated relation between real time and methylation cycle, this represents $\sim$167 years, which is obviously highly unrealistic as the time required by a tumour to progress between *carcinoma in situ* to invasive carcinoma for the 75+ years old female group or represents essentially infinite latency. This result is clearly due to simplistic information on abnormal molecular modifications observed in a (small) gene network, (i.e. less than 20 genes), in which only the influence of ageing and gender factors was considered. However, CRC predisposition is

influenced by other characteristics also, such as physical activity, diet, and environmental factors, (Subsection 2.6.2). Exposure to chemicals such as arsenic and cadmium have been reported to induce global DNA hypomethylation and histone modifications in cells leading to colon cancer, [Hou et al., 2012]; in consequence, relevant environmental data should be included in an extended analysis of tumour progression dynamics.

Another limitation is that the gene networks have mimicked only one stem cell, which is obviously unrealistic for predicting cancer development. In addition, a cell communicates with its neighbours through *gap junctions*, (regulated by the *connexin* family), and reduced inter-cell communication has been reported also in CRC development, (associated with the silencing of different connexin genes such as GJC1, due to promoter hypermethylation), [Sirnes et al., 2011]. Thus, any realistic analysis of the DNAm levels must be extended to cell populations and different cell types.

Recognition of the shortcoming in network-based methods of this type has motivated the development of the AgentCrypt model, (described in the next chapter). This permits DNAm variation to be investigated in the intestinal crypt over time and enables focus on aberrant modifications within abnormal systems, exposed to potential carcinogens.

Chapter 6

AgentCrypt - an Agent-based Model of Intestinal Crypt Dynamics

6.1 Introduction

The AgentCrypt is an agent based-model which has been developed to investigate methylation variation in the human intestinal tissues under the influence of different environmental conditions. These include e.g. carcinogen and methylation inhibitors and methylation patterns are thus of interest during both cancer initiation and progression. The AgentCrypt model addresses, therefore, the second major research objective proposed in this thesis, namely "To evaluate quantitatively methylation level variation in intestinal tissues under different external conditions during CRC initiation and progression", (Subsection 1.2.1), and its position in the overall Colorectal Cancer Model is highlighted in Figure 6.1. In Chapters 4 and 5, methylation level modifications over time was investigated for a human colon stem cell, (represented by the gene network in E-G Network Model). The AgentCrypt adds to this work on the analysis of the methylation modifications at tissue level. The remainder of the chapter is structured, as follows. Initially, deregulation of the crypt dynamics, which can induce aberrant methylation patterns over time are explored for the colon crypt systems, (Section 6.2). Subsequently, and motivated by the need of computational analysis in

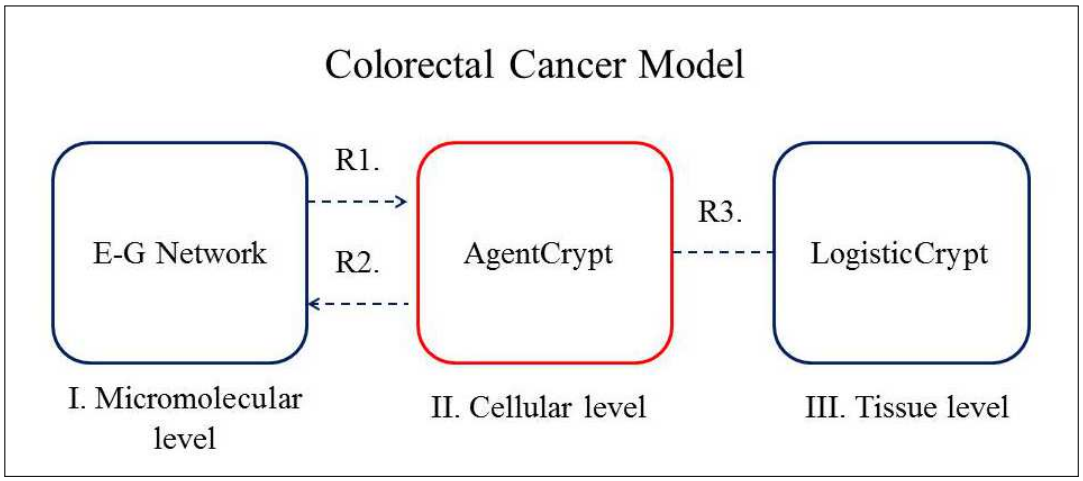


Figure 6.1: Structure of the Colorectal Cancer Model - focus on the AgentCrypt model
The relationships between model components may refer to R1) average methylation level for gene network / intestinal cell; R2) patient features; R3) intestinal crypt dynamics, explored following bottom-up and top-down approaches.

the context of EWAS research, (Section 3.4), AgentCrypt is extended to small intestine tissue. This extension takes account of the similarities in and differences between the structure and dynamics of the colon and small intestine crypts, (Section 2.4). A comparative analysis is performed on methylation differences characteristic of intestinal tumour initiation, (Section 6.3). The impact of external factors, such as carcinogens and epigenetic inhibitors on methylation level is also investigated, (Section 6.4), and an overview of AgentCrypt implementation is provided in Section 6.5. Two case studies and the results obtained are described in Section 6.6, and model findings are summarised in the final section. The case studies aim i) to quantify the methylation variation in normal and abnormal crypts, (both carcinogen-affected and unaffected), within the colon and small intestine tissues; and ii) to investigate the effect of a set of potential methylation inhibitors with respect to methylation modifications in intestinal crypt groups over time.

6.2 AgentCrypt for Colon Crypt Dynamics during Cancer Initiation

As seen in Chapter 2, intestinal crypts contain several different cell types[1] and have somewhat different structure dependent on tissue type. The crypt cells follow different cycles of generation to death, and the number of cells of each type and the state achieved in the cell cycle with respect to other cells determine the 'health' of the crypt system. The AgentCrypt is a computational model, which employs the agent-based paradigm, and is used to the structure and dynamics of the intestinal crypts, which include colon and small intestine type. AgentCrypt focuses on the analysis of the DNAm modifications in intestinal systems, induced by deregulations in cellular mechanisms during cancer initiation. The ABM features have been presented (Section 3.3.2), and the composition and dynamics of the intestinal crypt have been described (Subsection 2.4). Thus, the colon crypt can be modelled using the ABM approach as follows:

1. the colon crypt consists of a large number of cells of different types: stem, progenitor and fully-differentiated. It can be viewed, therefore, as a complex system with heterogeneous components. In model terms, the cells are the *agents* and the crypt itself, together with its neighbours are the *environment* of the agents.

2. inter-cell influences are represented by agents interactions;

3. the cell ability to perform e.g. division, differentiation or apoptosis at a specific time point is modelled as autonomously for each agent;

4. the influences of the mother-crypt neighbour, carcinogens and methylation inhibitor are represented as agent reactivity, (i.e. its ability to respond to a stimulus received from its environment);

5. finally, cell ageing is modelled in terms of the adaptive/ 'self-learning' ability of an agent.

[1] Stem, progenitor and (fully-)differentiated cells are present in both colon and small intestine crypts. In addition, Paneth cell population is located at the base of the small intestine crypt, (Figure 2.2).

6.2.1 Colon Crypt Structure and Dynamics

Agent communication within the colon crypt mimics the interdependencies between three cell types, namely, (a) *stem cells, (Stem)*, (b) *progenitor cells, (Prog)*, (c) *fully-differentiated cells, (Diff)* , (Figure 6.2). Stem cells are located at the crypt centre, progenitor cells (i.e. the transit cells) are present in stem cell group proximity, at both left and right sides and finally, fully differentiated cells occupy the top of the crypt. Both progenitor and differentiated cells migrate upwards during their lifetime. Additionally, due to the crypt structure, interdependencies between cell-agent neighbours are considered in AgentCrypt and the '*coefficient of influence*', ICoef(C', C), is introduced as a model input parameter (to provide characterisation of these cell inter-relations).

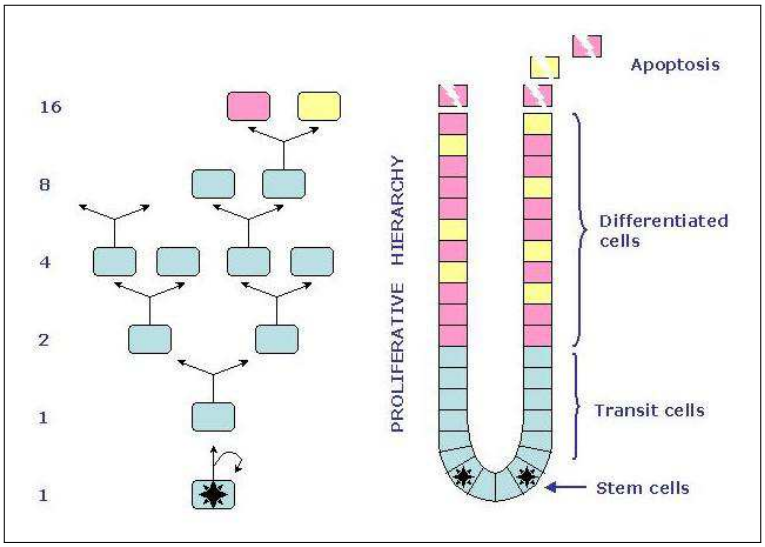


Figure 6.2: Structure and dynamics of a normal colon crypt - Simplified representation
Stem cells, (located at the crypt centre), can undergo symmetrical or asymmetrical division, (when two new stem cells or a stem and a progenitor cells, respectively, are born, and mother-stem cell is replaced by a daughter stem cell). Progenitor cells, (i.e. the transit cells), can also divide, (with a higher rate than the stem cells), but relatively few times. Finally, fully differentiated cells occupy the top of the crypt and can undergo apotosis, i.e. effective removal from the crypt.
Image adapted from the Integrative Biology Project Office, Oxford University Computing Laboratory. url: http://www.integrativebiology.ac.uk/colon-cryptic-dynamics.html

Cell Cycle *Cell cycle* is defined to be the set of possible actions performed by agents over time and generally includes *updates of methylation level*, cellular *division*, *differentiation* and *apoptosis*. However, the cellular actions performed during a cycle depend on the cell type. Specifically, in order to describe the way that cell methylation level evolves, the 'methylation level update' step is introduced into the crypt framework as a cell action performed by all cell types within the colon crypt. In addition, different cell actions are defined for each cell type, based on its characteristics, (summarised in Table 6.1). Thus, after the methylation level step, which is mandatory for every cell, a stem cell can undergo division or apoptosis, a progenitor cell can divide or differentiate, and the cycle of a differentiated cell can terminate through apoptosis.

Table 6.1: Cell actions specific to each cell type within the colon crypt

Cell Type	**Cell Action Name**	**Action in AgentCrypt Model**
Stem cell	*symmetrical* / *asymmetrical* division	Two stem cells / a stem and a progenitor cell, respectively, are added to the colon crypt. The mother stem cell is replaced by one daughter stem cell, while the other daughter cell is located to the left or right of the mother cell's initial position, within its cell type groups.
	apoptosis, (i.e. 'programmed death')	The stem cell is removed from the colon crypt.
Progenitor cell	cell division	The progenitor cell is replaced by one of its daughters after division; the two new progenitors are inserted on the same side of the crypt as the mother cell, but migrate to the end of the branch.
	differentiation	The progenitor cell loses its ability to divide further, becoming a *differentiated* cell. Relocated to the differentiated cell group.
Differentiated cell	apoptosis	The differentiated cell is removed from the colon crypt.

6.2.2 Cell Ageing

Cell ageing, refers to the number of cell cycles since a cell was born and has also been taken into account in the crypt framework. While a stem cell can continue to divide over an

unlimited time period, a progenitor cell has time-limited proliferation capability, which decreases as the cell ages and eventually ceases, [Masutomi et al., 2003; Sharpless et al., 2004; Liu and Rando, 2011; Schepers et al., 2011]. As a corollary, cell's differentiation capability increases over time. Cell ageing also influences the dynamics of apoptosis such that an older differentiated cell will clearly have a higher chance of dying and being removed from the crypt than a younger cell, [Sharpless et al., 2004; Liu and Rando, 2011; Schepers et al., 2011]. In model terms, over a cell cycle, each cell *C* from a crypt generates random values, (i.e. $P_C \in [0, 1]$), for every specific action *A* that relates to that cell. (P_C(A) thus represents the probability of the cell *C* executing the action *A*, where *A* can be cell division, differentiation or apoptosis). Threshold values, ($T_{CellType} \in [0, 1]$), are considered for each cell type to control actions taking place and are determined by the logistic functions introduced in expressions (6.2) - (6.3), (calculation shown further). Thus, an action *A* can occur only if $P_C(A) > T_{CellType}(A)$. Given that a stem cell is replaced by one of its daughter cells in normal systems during symmetrical division, cell ageing is considered to have no influence on stem cell-agent actions and threshold values for stem cell division, (either symmetrical or asymmetrical), are defined as model input parameters. However, given that progenitor and differentiated cells have limited lifetime, their actions, (i.e. division, differentiation and apoptosis), are considered to be influenced by cell ageing. Since *logistic regression*, (described by *logistic* or *sigmoidal* functions), can estimate the probability of an event to occur based on a set of independent variables, [Cramer, 2002], in model terms, a family of sigmoidal functions, with the following general form, has been proposed to compute the threshold values for progenitor and differentiated cell actions over time:

$$f(t) = 1/[1 + e^{\pm a(t-c)}] \tag{6.1}$$

where: t = current cell age (i.e. the number of cycles since the cell was born), sign (+/-) indicates an increase/decrease in cell action probability over time. The parameter 'c' is associated with cell cycle duration, which is considered to be, in average, around 2.5 days for progenitor, (c = 2.5), and one day for differentiated cells, (c = 1), for the test cases

reported here. However, other values are also possible and a sensitivity analysis for 'cell lifetime' parameter is planned, (Chapter 8). Additionally, the parameter 'a' is linked to the number of actions performed by cells. Thus, given that a progenitor can undergo both division and differentiation, a = 2, whereas for differentiated cells, a = 1, (as apoptosis is the single specific action for this cell type). Considering these parameter values and expression (6.1), the threshold values for cell division and apoptosis in the test cases are calculated using expressions (6.2) - (6.3), and are illustrated in Figure 6.3.

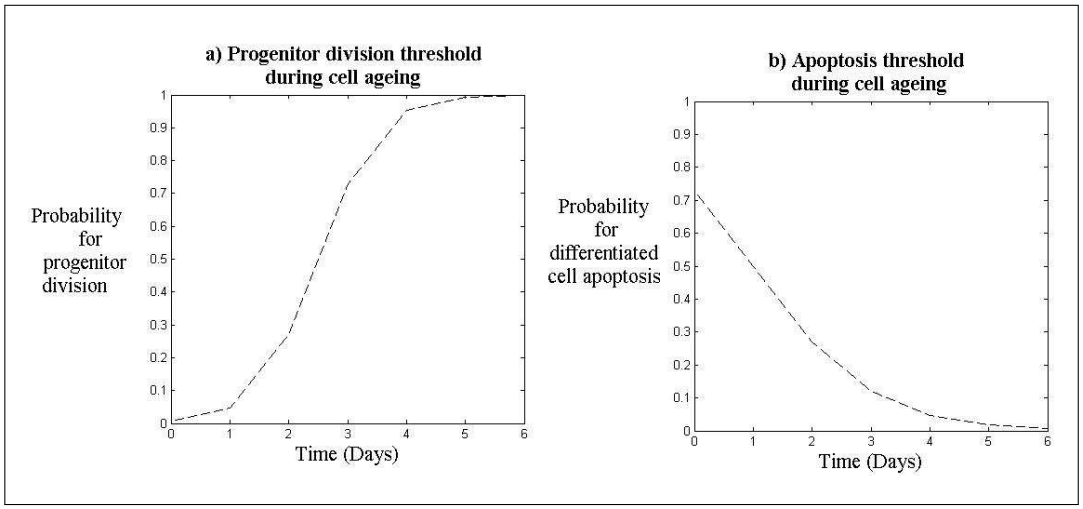


Figure 6.3: a) Progenitor division and b) Differentiated cell apoptosis over time
The progenitor division ability decreases over time, as the threshold increases with age. In addition, the probability of a cell death increases given that the apoptosis threshold decreases during cell ageing.

For example, if a progenitor cell C1 generates a value P_{C1}(Division) = 0.70 in the second day since it was born, given that P_{C1}(Division) $> T_{PROG}$(Division), (with T_{PROG}(Division) $\simeq$ 0.27, calculated with expression (6.2)), cell C1 will perform division. However, if the same value P_{C1}(Division) = 0.70 is generated in e.g. the fourth day of cell C1 lifecycle, given that P_{C1}(Division) $< T_{PROG}$(Division), (with T_{PROG}(Division) $\simeq$ 0.95), cell C1 will be unable to divide, so undergoes differentiation.

$$ProgDiv(t) = \frac{1}{1 + e^{(-2)\times(t-2.5)}} \tag{6.2}$$

$$DiffApoptosis(t) = \frac{1}{1 + e^{(t-1)}} \tag{6.3}$$

6.2.3 The Colon Crypt Group

For a real biological system, such as the colon, crypts are part of tissue structure and clearly do not exist in isolation. Hence, observation on cancer initiation, made based on information regarding abnormal modifications within a *single* crypt, would be unrealistic. In consequence, the analysis of aberrant changes is extended to the *Colon Crypt Group*, which considers a large number of crypts and can provide information on an increase in 'tissue area susceptibility' to tumour development. The *Colon Crypt Group* is defined as a collection of *n* crypts found in the colon neighbourhood, where n $>$ 0 is a model input parameter.

The main characteristics of the Colon Crypt Group are i) *crypt intra-* and *inter-dependencies* and ii) *carcinogen* and *methylation inhibitor influences*, (which target equally every crypt within the group). The main assumption made for the AgentCrypt model applied to these groups is that crypt intra- and inter-dependencies, together with cell type, dictate methylation level dynamics within cells. As noted in Subsection 2.4.2, abnormal changes in *connexin* gene expression can lead to deregulation of the cell cycle and influence on cell methylation patterns. Crypt inter-influences have been also reported, [Humphries and Wright, 2008], and these are also, therefore, taken into account during the methylation level update step in AgentCrypt. In addition, carcinogens are considered to have a direct impact on crypt dynamics as these relate to cell division, differentiation and apoptosis; consequently, they have indirect influence on methylation level modifications within cells. Very recently, possibilities for methylation inhibition have been explored, [Lohse et al., 2011; Santer et al., 2011; McCabe et al., 2012], and AgentCrypt also proposes a mechanism for exploring the methylation inhibitor influence in the human intestinal tissue. Outputs of the model are information on the average DNAm level within cells in a Colon Crypt Group. A simplified overview on the entities and relationships involved in this version of the AgentCrypt model is illustrated in Figure 6.4.

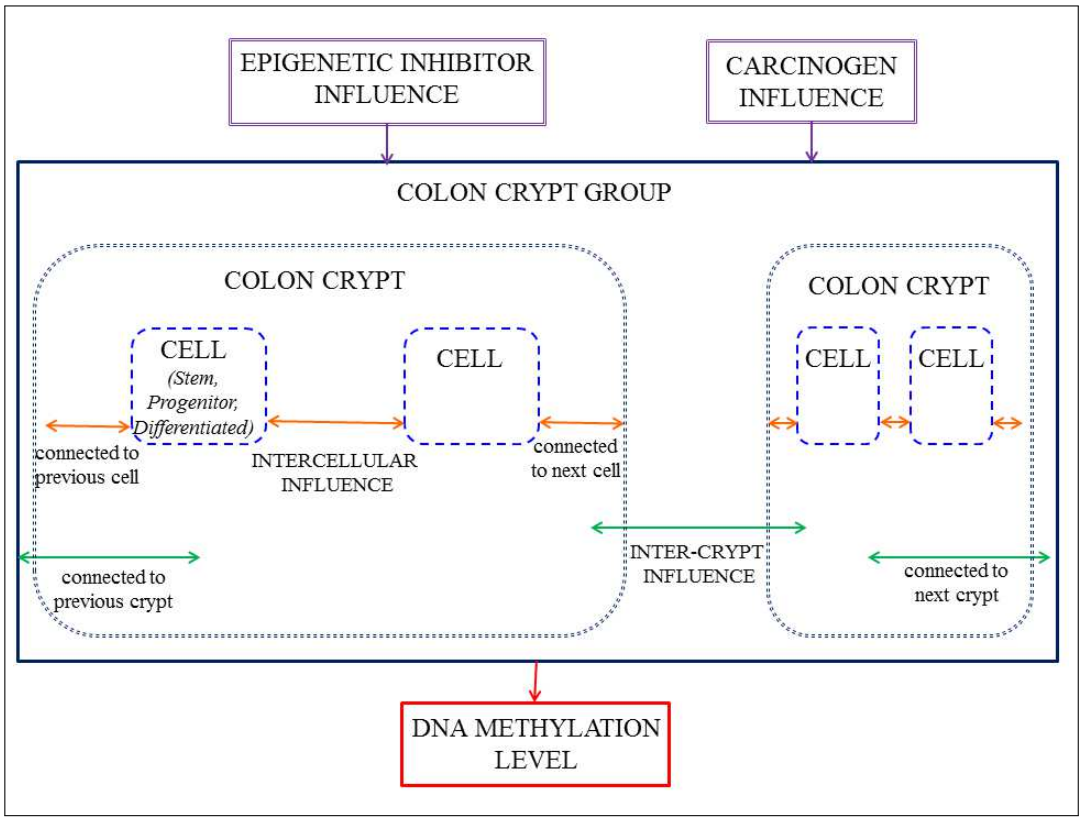


Figure 6.4: Colon Crypt Framework - Simplified Structure Representation
The focus of the AgentCrypt model is on the aberrant DNAm level within cells from the colon crypt group. The dynamics of DNAm are dictated by cell and crypt interdependencies and cell type. In addition, DNAm level is influenced indirectly by carcinogens, which have a direct impact on cell dynamics. Different potential methylation inhibitors can be also applied and their effect on DNAm can be explored.

6.2.4 Methylation level update within the Colon Crypt Group

Methylation levels in the model are taken to be values within the range $[0, 1]$[2], where the healthy phenotype is characterised by a high average methylation level in the cell, (i.e. ≈ 1), but decreases as cancer progresses. Initially, the crypt group is considered to be in a healthy state and the methylation level for every cell is initialized as 'healthy', using data from a study on genome-wide DNA methylation analysis of normal colon and colon adenocarcinoma, performed in 22 pairs of colorectal cancer and adjacent tissues and 19 samples from healthy individuals, (GEO[3] accession GSE42752, [Naumov et al., 2013]).

[2]Values from range [0, 1] correspond to the percentage of methylation level measured in a given sample.
[3]GEO = Gene Expression Omnibus database, [Barrett et al., 2013]

Cancer is considered to have progressed if there is decrease in the average methylation level calculated for the entire group.

Given that global *hypomethylation* has been associated with ageing, [Pogribny and Vanyushin, 2010], cell methylation level in AgentCrypt is considered to decrease over time. In model term, cell methylation level at time t+1 is given by cell methylation level at time t, minus methylation level lost during this time interval, i.e.:

$$M'(C) = M(C) - ML(C) \tag{6.4}$$

where: M(C), M'(C) = methylation level of cell C in the current and next Time-step, respectively. ML(C) represents methylation level value removed during the current update step for cell C and is calculated based on three major elements, (given by crypt *intra-* and *inter*-influences, (Section 2.4)):

1. methylation decrease specific to cell C at a given time, due to ageing, (a random value DM $\in [10^{-5}, 10^{-4}]$);

2. the influence of the cell neighbours, (from both left and right sides), computed based on neighbour cell methylation level, the 'coefficient of influence' (subsection 6.2.1), and a randomly generated value, (R $\in [5 \times 10^{-2}, 10^{-1}]$), which indicates that the influences between same two cells can be variable over time. However, if the crypt is normal, no neighbour-cell influences have been considered with respect to methylation level modifications. (Expression (6.6) is proposed to calculate the cell neighbours influences.)

3. the influence of the Cr crypt neighbours, *CNI(Cr)*. This is calculated based on the average methylation level of crypt neighbours, AverMethyl(Cr'), and the **I**nter-**C**rypt **C**oefficient, (model input parameter, ICC $\in [10^{-3}, 10^{-2}]$), which indicates individual predisposition for intestinal tumour development. For example, ICC parameter has a higher value for an individual presenting the *inflammatory bowel syndrome*[4] than for an individual with no such disease. The CNI(Cr) is computed using expression (6.7) and

[4]Inflammatory bowel syndrome has been associated with an increased risk of CRC development, (Section 2.4).

has the same value for all cells of a given crypt Cr, i.e. every cell is affected in the same way by the mother crypt neighbours.

Specifically, methylation level differences between two successive time-steps in colon crypt group are described by expressions (6.5) - (6.7), (with variable defined in Table 6.2):

$$ML(C) = DM + Infl(Left, C) + Infl(Right, C) + CNI(Cr) \tag{6.5}$$

$$Infl(C', C) = \begin{cases} R \times M(C') \times ICoef(C', C), & \text{if the colon crypt is abnormal;} \\ 0, & \text{otherwise.} \end{cases} \tag{6.6}$$

$$CNI(Cr) = ICC \times \sum_{CrNabr \in \{NeighboursofCr\}} AverMethyl(CrNabr) \tag{6.7}$$

The dominant term here is that for Infl(C', C), which can lead to considerable methylation level differences between abnormal and normal cells. This would imply that the environment of the cell influences methylation modification dynamics inside the cell.

Table 6.2: Variable definitions for methylation level update step

Parameter name	**Parameter description**
ML	the methylation value that will be removed during the current update step;
DM	ageing-related methylation decrease specific to cell C at a given time, (a random value $\in [10^{-5}, 10^{-4}]$);
M(C')	methylation level of cell C' in the current Time-step, ($\in$ [0, 1]);
ICoef (C', C)	coefficient of cell C' influence on cell C; model input parameter with value $\in [10^{-3}, 10^{-2}]$
Infl(Left, C), Infl(Right, C)	the methylation value that is added to cell C due to left/right neighbour influences;
CNI(Cr)	the neighbour-crypt influences on the Cr crypt;
R	random value from $[5 \times 10^{-2}, 10^{-1}]$;
ICC	the **I**nter-**C**rypt **C**oefficient, (model input parameter, with a value from $[10^{-3}, 10^{-2}]$);
AverMethyl(Cr')	the average methylation level of the Cr' crypt, (computed based on the methylation level of every compound cell);

While methylation level is conserved during cell division, methylation level changes occur during cell differentiation, [Klug et al., 2010; Meissner, 2010; Kaaij et al., 2013], and the parameter *DiffMethCoef* is introduced to indicate such modifications. Thus, methylation level after cell C differentiation can be computed based on methylation level of cell C before undergoing differentiation:

$$M_{DIFF}(C) = M_{PROG}(C) \times DiffMethCoef \tag{6.8}$$

where: M_{PROG}(C), M_{DIFF}(C) = methylation level of cell C before and after differentiation step, respectively, and DiffMethCoef = the methylation value removed during cell differentiation step, (model input parameter, with a value $\in$ [0, 1]).

6.3 Extension to the small intestine tissue

The small intestine and colon tissues share a number of features, which can be considered also in terms of crypt structure (Section 2.4). Thus, AgentCrypt has been extended to mimic and inspect the dynamics of the small intestine crypt during cancer initiation. The methylation level in every cell within the small intestine crypt is initialized from study data on genome-wide DNA methylation profiling, (GEO accession GSE50475, [Lay et al., 2014]). In model terms, the major change is the inclusion of a *Paneth* cell group, where maximum and minimum Paneth cell numbers, (i.e. *MaxPanethNo* and *MinPanethNo*), are additional input parameters. Thus, the Paneth cell-agent, which can either migrate to the crypt base or undergo apoptosis, is included as a fourth type in AgentCrypt. The progenitor cell cycle is modified also to permit differentiation into Paneth cells and the relationship describing cell-agent interdependencies is extended to integrate the Stem-Paneth and Paneth-Paneth influences in addition to those considered for the colon crypt, (subsection 6.2.1). Moreover, based on expression (6.1) and given that Paneth cell lifetime has been approximated to 20 days, [Sancho et al., 2003; Clevers and Bevins, 2013], the Paneth cell probability to undergo

apoptosis during its cycle is described by:

$$PanethApoptosis(t) = \frac{1}{1 + e^{(t-20)}} \tag{6.9}$$

6.4 External influences

Influences of the external factors such as carcinogenes and methylation inhibitors have been considered also in AgentCrypt and their impact on methylation level is explored through intestinal crypt groups.

6.4.1 Carcinogens influences

Carcinogen influence permits molecular abnormalities to accumulate, (with faster dynamic), and facilitates cell growth, [Papailiou et al., 2011]. In model terms, carcinogen presence is associated with deregulations in crypt dynamics rather than in terms of direct impact on methylation level. Thus, carcinogens affect threshold values characteristic of each cellular action within AgentCrypt, by addition/subtraction of a percentage $\delta 1$ from the threshold value, $\delta 1 \in$ [0, 100], (an input parameter). Given that stem cell division and apoptosis are rare events, threshold values characteristic to this cell group are fractionally modified by the product fStem $\times \delta 1$ value, where fStem $\in$ [0, 1], (an input parameter). Hence, under the influence of carcinogens, the threshold value for cell division is *lower* than that within a normal system, with complementary probability of cell proliferation increased. Similarly, threshold values for apoptosis are *increased* in affected systems, leading to decrease in the normal rate of cell death.

6.4.2 Methylation inhibitor activity

Epigenetic drugs are considered a promising option in cancer therapy and several methylation and acetylation inhibitors are already being applied in blood cancer, (discussed in Section 2.7). Therefore, the AgentCrypt model was extended also to investigate the influence of potential methylation inhibitors in the intestinal crypt over time.

Major drug features are *efficacy*, (which represents the maximum response achieved after drug application), and *potency*, (which measures drug amount needed to produce a specific effect). These can be characterised by a *drug response curve*, which is represented by a sigmoidal (or logistic) function and indicates relationship between drug dose and effect, (i.e. response to treatment). A drug response curve is described typically by a *slope* factor and EC_{50}, or *half maximal effective concentration*, (representing drug concentration that determines 50% efficacy). As a percentage, therefore, $EC_{50} \in [0, 0.50]$; thus, efficacy has a value in [0, 1] range. If slope = 1, the drug response curve has a *standard slope*; otherwise, it has a 'variable slope'. In general, inhibitors are considered to have a standard slope. In addition, information on time to '*inflection point*' (IP), (when drug curve changes its concavity, i.e. effect starts increasing or decreasing) is typically part of the drug response curve description. Efficacy is illustrated therefore, by drug response curve peak, (the higher the peak, the greater the efficacy), and potency is displayed by dose-response curve level at a point on x-axis (a lower value of x for given response suggests greater potency), (Figure 6.5). The general form of logistic function, (expression (6.10)), can be adapted to describe drug response curve, (expression (6.11)).

$$f(x) = \frac{c}{1 + e^{-Slope \times x}} \tag{6.10}$$

$$DRC(t) = \frac{EFF}{1 + e^{-Slope \times (t - IP)}} \tag{6.11}$$

where: t = time since drug/ inhibitor was applied, EFF is drug efficacy and EFF = $2 \times EC_{50}$, Slope = slope of drug response curve, IP = inflection point of drug response curve, EC_{50} = half maximal effective concentration.

In AgentCrypt, the presence of an inhibitor affects the methylation level update step during a cell cycle, (given by expression (6.4)), and methylation modifications are gradually blocked, since the inhibitor effect is considered dependent on the time elapsed since the drug has been introduced to the system. Specifically, the inhibitor influence increases initially

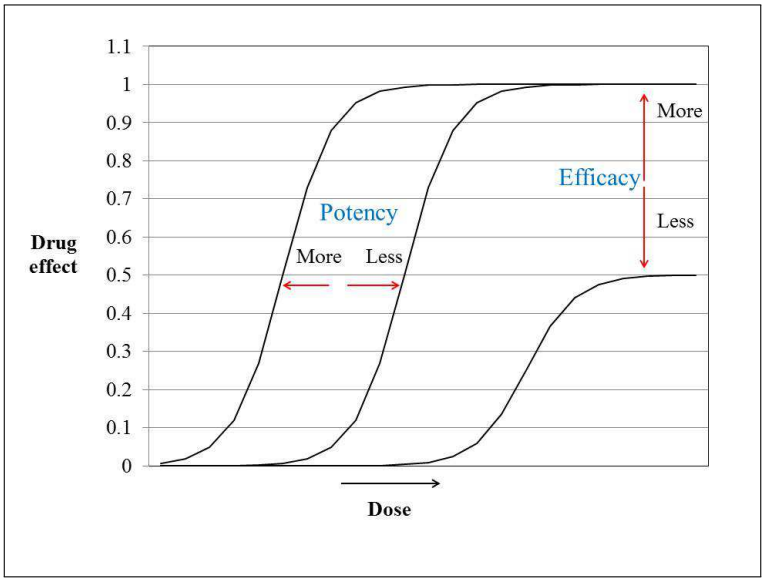


Figure 6.5: Drug Efficacy and Potency indicated by Drug response curve
High peak and left position of a drug reposnse curve indicate high efficacy and potency
Image adapted from:
http://classconnection.s3.amazonaws.com/370/flashcards/641370/png/untitled1321511264457.png

up to a maximum value, (i.e. reaches inhibitor efficacy), and maintains a constant level for a specific time-period, (denoted as the *Inhibitor Maximum Efficacy Time* (*IMET*)), after which, inhibitor effect decreases. In the model, inhibitors are considered to have standard slope, (i.e. Slope = 1), as noted earlier. Based on these and expression (6.11), the effect of a methylation inhibitor is characterised by a combination of two logistic functions, denoted as *the inhibitor response curve, (IRC)*, (expression 6.12), and IMET.

$$IRC(t) = \begin{cases} \frac{EFF}{1+e^{-(t-IP_{INCR})}}, & \text{if t} \leq \text{IMET, i.e. 'increase' phase ;} \\ 1 - \frac{EFF}{1+e^{-(t-IP_{DCR})}}, & \text{otherwise, ('decrease' phase).} \end{cases} \tag{6.12}$$

where t is time (days) since the inhibitor was applied, *EFF* represents drug efficacy, calculated based on EC_{50}, (with EFF = 2 $\times$ EC_{50}), and IP_{INCR} and IP_{DCR} are inflection points of drug effect (i.e. times at which inhibitor effect *increases* and *decreases*, respectively). EC_{50}, IP_{INCR} and IP_{DCR} are model input parameters and are also dominant terms here.

Thus, IRC(t) value is closer to 0 if $EC_{50} \approx 0$ and increases to 1 as EC_{50} increases. In addition, a higher IP_{INCR} value implies later IRC(t) increase, i.e. lower potency. Finally, higher IP_{DCR} implies prolonged effect of the drug.

Example: An example of methylation inhibitor influence is given for a set of three inhibitors during a time-period of around 90 days, ($\approx$ one quarter of year). The effect of the first inhibitor, I1, is taken to start increasing around seven days after application, (i.e. $IP1_{INCR}$ = 7); for the second and third inhibitors, I2 and I3, the increase is taken around fourteen and twenty one days, respectively, after application, (i.e. $IP2_{INCR}$ = 14, $IP3_{INCR}$ = 21). In addition, while the same efficacy applies for inhibitors I1 and I2, a relatively low value is assigned to efficacy of inhibitor I3, (e.g. $EC1_{50}$ = $EC2_{50}$ = 50%, $EC3_{50}$ = 40%). Finally, IMET was taken $\approx$ 50 days. The inhibitors described by these settings are given by expression (6.13), and their effect over time-period of 90 days is illustrated in Figure 6.6. Considering both efficacy and potency features, inhibitors *I1* and *I3* can be seen as the most and least efficient, respectively, within this set. (Inhibitors I1 and I2 have similar efficacy, but I1 has higher potency than I2. I3 has both efficacy and potency lower than I1).

$$
\begin{aligned}
&Inhibitor1,\ (I1):\ IP1_{INCR} = 7\ days,\ IP1_{DCR} = 90\ days, EC1_{50} = 50\%; \\
&Inhibitor2,\ (I2):\ IP2_{INCR} = 14\ days,\ IP2_{DCR} = 90\ days, EC2_{50} = 50\%; \\
&Inhibitor3,\ (I3):\ IP3_{INCR} = 21\ days,\ IP3_{DCR} = 90\ days, EC3_{50} = 40\%.
\end{aligned}
\tag{6.13}
$$

Methylation modifications, ML, occur over time, and in intestinal systems with no inhibitors, methylation level is updated using expression (6.4), (subsection 6.2.4). In crypts for which methylation inhibition applies, methylation level changes are reduced by a value depending on strength of inhibitor effect: higher effect determines higher methylation reduction. Thus, in inhibitor-influenced systems, methylation level variation is given by $ML \times (1 - IRC(t))$ and based on this and on expression (6.4), the expression for methy-

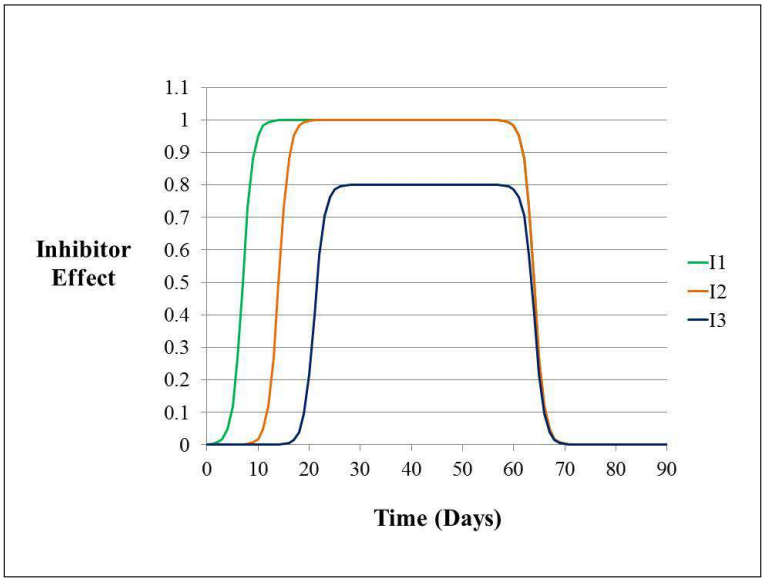


Figure 6.6: Comparison between the effect of three potential inhibitors on methylation level over a time-period of 90 days defined by the following settings: i) Inhibitor 1: IP1$_{INCR}$ = 7 days, IP1$_{DCR}$ = 90 days, (green); ii) Inhibitor 2: IP2$_{INCR}$ = 14 days, IP2$_{DCR}$ = 90 days, (orange); iii) Inhibitor 3: IP3$_{INCR}$ = 21 days, IP3$_{DCR}$ = 90 days, (blue); iv) IMET $\approx$ 50 days. *Inhibitor1* and *Inhibitor3* can be considered the most and less efficient, respectively, within this set.

lation level updates between two successive time-steps in the intestinal crypt becomes:

$$M'(C) = M(C) - ML \times (1 - IRC(t)) \tag{6.14}$$

where M'(C), M(C) = methylation level of cell C in the current and next Time-step, respectively, and ML = methylation modification between current and next Time-step; t = time since the inhibitor was applied and *IRC(t)* = inhibitor effect at time t. For example, if the inhibitor effect at time t is maximum, i.e. IRC(t) = 1, methylation variation between times t and t+1 is equal to zero, (expression 6.15). However, if no inhibitors are applied, i.e. IRC(t) = 0, methylation variation achieves a maximum, (given by expression 6.4).

$$M'(C) = M(C) - ML \times (1 - 1) = M(C) - ML \times 0 = M(C) \Rightarrow M'(C) - M(C) = 0 \tag{6.15}$$

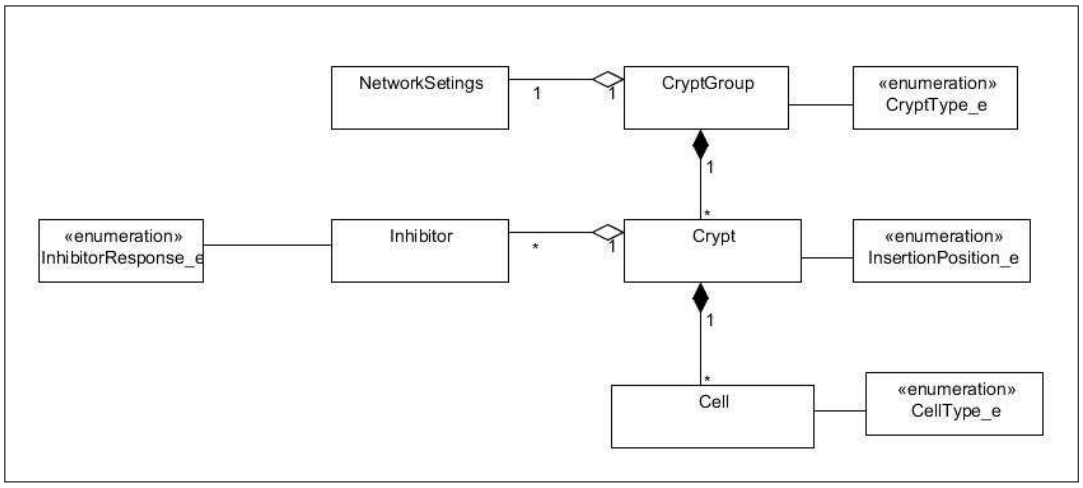


Figure 6.7: Class diagram for the AgentCrypt model. The dark-shaded diamond shape indicates the composition relationship between classes, (similar to the E-G Network Model, Chapter 5), while the empty diamond indicates the aggregation relationship between classes, i.e. the component lifecycle is not controlled by the container class: if the container class is destroyed, aggregated instances are not destroyed - only the association between container class and compound instances is affected.

6.5 Implementation details of the AgentCrypt

Similar to the E-G Network, (Chapter 4), AgentCrypt code is object-oriented, implemented using C++. Using the same programming language for developing both models facilitates their connection and integration into a single multi-scale framework for CRC dynamics.

The AgentCrypt structure is represented by the following major classes, with their relationships shown in the class diagram (Figure 6.7):

- *Cell* - This class describes cell-agent characteristics, including the division, differentiation and apoptosis functionality.

- *Crypt* - This class handles the functionality of the intestinal (both small intestine and colon) crypts; a Crypt object contains a large number of Cell objects and can contain a set of Inhibitor objects.

- *CryptGroup* - This class represents the tissue level in the AgentCrypt model and mimics characteristics of the Intestinal Crypt Group. A CryptGroup object contains a collection of Crypt objects. The CryptGroup class also offers functionality for

handling information on individual characteristics, (i.e. the NetworkSettings class from the E-G Network Model, (Section 5.3.3)).

- *Inhibitor* - This class encodes methylation inhibitor characteristics and describes functionality for calculating inhibitor response curve values over time.

Similar to the E-G Network, the Mersenne Twister is used as the random number generator and Visual Leak Detector, [CodePlex, 2014], is integrated in AgentCrypt for detecting plausible memory leaks in code. Finally, the code documentation is generated again with the *doxygen* tool, [van Heesch, 2008], and can be accessed online at the following url: http://www.computing.dcu.ie/ iroznovat/documentation/html/index.html.

6.6 Results and discussion

AgentCrypt was applied to analysis of DNA methylation modifications within both small intestine and colon crypt groups, in healthy and malignant systems, (and under assumed carcinogen influence). This case study was used to explore evolution of intestinal crypt mechanisms. In addition, two 'plausible' methylation inhibitors, described in subsection 6.4.2, were considered in intestinal crypt systems and their impacts on methylation level were investigated during cancer initiation. Input parameters are grouped as follows: a) Table 6.3 contains information on cell numbers specific to colon and small intestine crypts, respectively; and b) Table 6.4 contains information on intestinal crypt dynamics, such as *coefficient of influence* introduced in subsection 6.2.1. The data are organized such that information from Table 6.3 is tissue-specific, while data contained in Table 6.4 are common to both colon and small intestine crypt groups. These values are derived from literature (Section 2.4), or considered as potential input data sets; other values are possible as well and a sensitivity analysis is certainly needed for every parameter, (discussed in Chapter 8).

Tissue-specific parameter values

The total cell number is approximated to 2000 cells and 260 cells in colon and small intestine crypts, respectively, informing on *InitCellNumber* parameter value in intestinal sys-

Table 6.3: Input parameter values on cell number for small intestine and colon crypts

Parameter name	**Value in the colon**	**Value in the small intestine**
InitCellNumber	2000	260
InitStemCellNumber	16	6
MaxTotalCellNo (2 × InitCellNumber)	4000	520
MaxStemNo (2 × InitStemCellNumber)	32	12
MinStemCellNumber	1	1
MinCellNumber (Perc$_{MIN}$ × InitCellNumber, Perc$_{MIN}$ = 50%)	1000	130
InitPanethCellNumber	0	10
MaxPanethCellNumber (2 × InitPanethCellNumber)	0	20
MinPanethCellNumber (Perc$_{MIN}$ × InitPanethCellNumber, Perc$_{MIN}$ = 50%)	0	5
InitProgCellNumber (Perc$_{PROG}$ × InitCellNumber, Perc$_{PROG}$ = 25%)	500	65
MaxProgNo (2 × InitProgCellNumber)	1000	130
MinProgCellNumber (Perc$_{MIN}$ × InitProgCellNumber, Perc$_{MIN}$ = 50%)	250	32

tems. In addition, stem cell group consists of around 16 - 19 cells in the colon, (e.g. InitStemCellNumber = 16 in colon), and 1 - 6 cells in small intestine crypt, (e.g. InitStemCellNumber = 6). Major crypt phenomena, such as 'crypt fission and the 'bottleneck effect', were reported to occur over long time-periods in normal crypts. 'Crypt fission' is characterised by cell population augmentation and was associated with doubling stem cell number. The 'bottleneck effect' is represented by cell group reduction and decrease in stem cell number up to a single stem cell. Based on these, it can be considered that MaxStemNo = 2 × InitStemCellNumber, MaxTotalCellNo = 2 × InitCellNumber and MinStemCellNumber = 1. Paneth cells are present in a relatively small number at the base of small intestine crypt and are absent from normal colon crypt. Given these, Paneth cell number in normal systems, InitPanethCellNumber, is considered to be equal to zero in colon and close to stem population size in small intestine crypt, (e.g. InitPanethCellNumber = 10 cells). Given no precise information available on progenitor group size, the progenitor cell number in the model is taken equal to a percentage of crypt cell population size, e.g. Perc$_{PROG}$ = 25%. Finally, similar to maximum cell numbers which are double of normal cell numbers, the

minimum cell numbers are calculated as a percentage of normal cell population sizes, (e.g. Perc$_{MIN}$ = 50%).

Common parameter values

The parameter values from Table 6.4 are chosen based on several criteria, such as: i) the system computational performance, (e.g. cell numbers in crypt), ii) a framework-defined 'inter-cellular influence hierarchy' based on cell type, (according to which, for example, the influence of a stem cell on a progenitor is higher than the reverse and the influences between cells of the same type are represented by a common value for the whole system), (i.e. coefficient of influence), iii) 'middle' changes induced by carcinogens on cell dynamics, (e.g. carcinogen influence $\approx$ 50%) and iv) low methylation level differences occurring during cell differentiation.

Table 6.4: Input parameter values of intestinal crypt dynamics used in the AgentCrypt for both small intestine and colon crypt

Parameter group	**Parameter name & value**
Crypt number per group	n = 10.
Threshold values	T$_{Stem}$(SymDivision) = 0.99; T$_{Stem}$(AsymDivision) = 0.94.
Cell cycle duration	ProgCellCycle = 2.5; DiffCellCycle = 1.
Methylation level change due to cell differentiation	DiffMethCoef = 90%;
Carcinogen influence	δ1 = 50%; fStem = 0.1;
Coefficient of influence	ICoef(Stem, Stem)= 0.002; ICoef(Paneth, Paneth) = 0.002; ICoef(Prog, Prog) = 0.002; ICoef(Diff, Diff) = 0.002; ICoef(Stem, Prog) = 0.003; ICoef(Stem, Paneth) = 0.003; ICoef(Prog, Stem) = 0.001; ICoef(Prog, Diff) = 0.003; ICoef(Diff, Prog) = 0.001; ICoef(Paneth, Stem) = 0.001; CYPT_INTER_COEF = 0.001.

Given that probabilities for stem cell division and apoptosis to occur are considerable

low in normal systems, significant high values are assigned for stem cell action thresholds, (e.g. T_{Stem}(SymDivision) = 0.99, T_{Stem}(AsymDivision) = 0.94). Stem cell cycle was approximated to seven and five days in human colon and small intestine, respectively, and intestinal tissues are characterised by high cell turnover rate, (Section 2.4). Based on these, lifetime was considered around 2.5 days for progenitors and one day for differentiated cells in both crypt types. Given limited precise information on intestinal tissue dynamics, parameter values included in Table 6.4 are proposed for the test cases presented in this chapter. However, alternative values are also possible and a sensitivity analysis is required to investigate system behaviour with various value-ranges for input parameters, (Chapter 8).

6.6.1 Case study on DNA methylation level variation during intestinal cancer development

Initially, DNAm variation in the small intestine and colon crypt over time was considered. Specifically, three major systems, namely i) the *healthy* crypt, ii) the *aberrant* crypt and iii) the *carcinogen* crypt, were defined and incorporated in this test case. DNAm level changes are permitted for all three crypt types, but intra-crypt communication was not considered to affect cell methylation level in the healthy crypt case. Under carcinogen influence, cell action threshold values are modified ($\delta 1$), as described in Section 6.4.1. Three groups of 10 crypts were defined for each system type, (i.e. healthy, aberrant and carcinogen crypt), and a comparison of DNA methylation variation was performed for both small intestine and colon tissues. The set of input parameters, that describe these systems, is given in Table 6.3 and 6.4.

The crypt groups were allowed to evolve over time, (simulation time of 550 iterations, approximated to 1.5 years in colon tissue) and the analysis considered both *intra-* and *inter-*tissue comparisons, (addressing elements of Research Objective 3, (Section 1.2.1)). Based on individual crypt methylation level (given by the average methylation level of the compound cells), the *average methylation level* was calculated for each crypt group, (illustrated by Figure 6.8). This shows that although average methylation level values decreased for

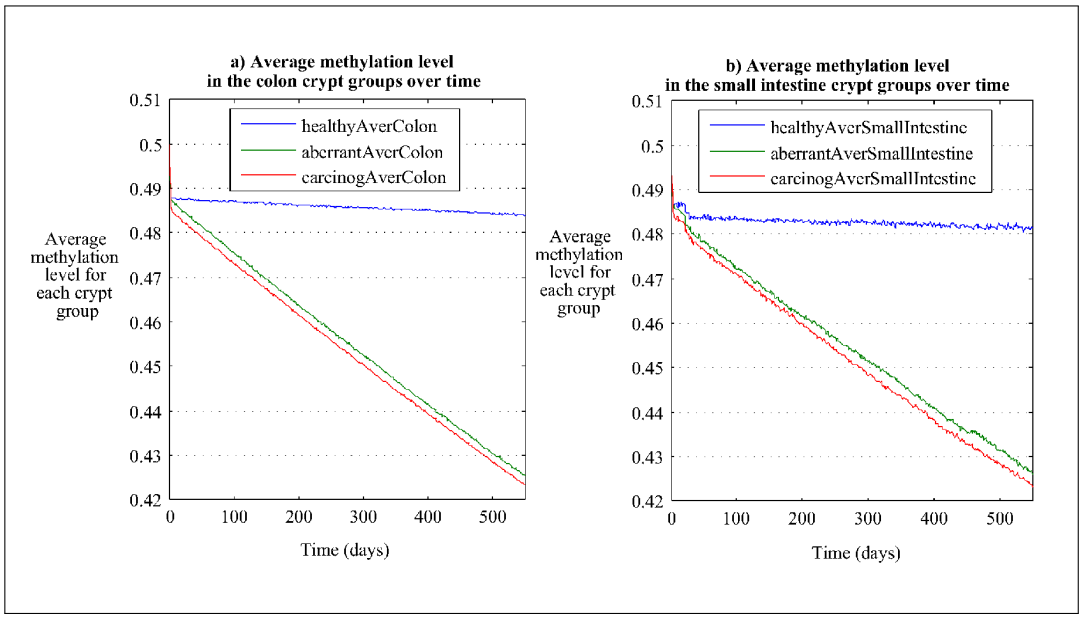


Figure 6.8: Comparison between methylation levels recorded in a) colon and b) small intestine
The average methylation level is calculated for each of the healthy, aberrant and carcinogen-influenced crypt groups within both the small intestine and colon over a 550-day period. The highest methylation level is recorded in the healthy crypt groups for both tissue types, while the lowest methylation level (i.e. the highest decrease) is observed in the 'carcinogen' crypt groups.

each crypt group during the simulation, updates followed different patterns for healthy, aberrant and carcinogen-influenced crypts. Differences between final and initial average methylation levels are given in Table 6.5, for both colon and small intestine crypt groups. Specifically, methylation differences observed are ≈1.58% in healthy, ≈7.44% in aberrant, and ≈7.64% in carcinogen-affected systems in colon tissue, and ≈1.83%, ≈7.39% and ≈7.64% in corresponding crypt groups in the small intestine. The lowest methylation level decrease over time was recorded for healthy crypts, (interpreted as changes in methylation level due to ageing), while the highest *hypo*methylation difference was observed in the 'carcinogens' groups. The average methylation level within aberrant crypt groups decreased by ≈5.85% in the colon and with ≈5.55% in the small intestine in comparison with that recorded for healthy systems. Differences in average methylation level decrease between carcinogen and healthy crypt groups were ≈6.05% in the colon and ≈5.81% in the

small intestine. These results on the methylation level loss between the 'aberrant - healthy' and 'carcinogen - healthy' group-pairs indicate that while carcinogens do not have a direct impact on epigenetic mechanisms, methylation patterns are differently affected in the systems where these influences are present. In addition, results indicate an increase in average methylation level of the small intestine during the first week, ($\sim$ 30 days, equivalent to Paneth cell cycle duration, (Section 2.4)), a pattern not shown in the colon. A cause of this phenomenon can be related to Paneth cell group presence in the small intestine and the way in which crypt methylation level is influenced by cell dynamics. Specifically, although cells initially have similar methylation level, the differentiated cell population will, after a period of time, will contain lower methylation level than stem and progenitor cells, (due to demethylation during differentiation). In addition, an increase in crypt methylation level can occur between two successive time points when differentiated cells undergo apoptosis and stem/ progenitor cells divide, (transmitting accurate methylation information). In the colon, given that the differentiated cell population is extended during every progenitor differentiation, the decrease in methylation level is maintained over time. However, this is not exhibited by the small intestine, where progenitors can differentiate also into Paneth cells, resulting in no changes in differentiated cell population. In consequence, if several differentiated cells die, an increase in crypt methylation level can be recorded.

Finally, relatively higher values were recorded in the colon, (though broadly comparable), suggesting that colon may be more sensitive than small intestine tissue with respect to methylation level dynamics induced by abnormal conditions. This is also an interesting observation given that simulation time was relatively short in comparison with cancer development time, (i.e. $\approx$1.5 years versus several decades). Thus, further tests should consider longer time-periods for intra- and inter-tissue analysis of methylation level modifications.

6.6.2 Case study on methylation inhibitors influence

The impact of methylation inhibitors on intestinal crypt methylation level was analysed in a second case study, in small intestine tissue, using inhibitors with the highest and lowest effect from the set defined in expression (6.13), i.e. I1, ($IP1_{INCR}$ = 7 days, $IP1_{DCR}$ =

Table 6.5: Difference between the Final and Initial Average Methylation Level, recorded for Healthy, Aberrant and Carcinogen Crypts within the Colon and Small Intestine Groups

System Type	**Row**	**Measurement**	**Colon**	**Small intestine**
Healthy	R1	Initial level	0.49997;	0.49988;
	R2	Final level	0.48411;	0.48150;
	R3	δ_{R2-R1}	-0.01586;	-0.01838;
Aberrant	R4	Initial level	0.49988;	0.49983;
	R5	Final level	0.42546;	0.42589;
	R6	δ_{R5-R4}	-0.07442;	-0.07394;
Carcinogen	R7	Initial level	0.49990;	0.49986;
	R8	Final level	0.42349;	0.42338;
	R9	δ_{R8-R7}	-0.07641;	-0.07648.

90 days, $EC1_{50}$ = 50%), and I3, ($IP3_{INCR}$ = 21 days, $IP3_{DCR}$ = 90 days, $EC3_{50}$ = 40%). Similar to the previous case study, methylation levels in the respective crypt groups was initialized using data from GEO accession GSE50475 dataset. In addition, four crypt groups were defined as follows:

- S1: normal crypt group, with no inhibitor applied;
- S2: abnormal crypt group, with no inhibitor applied;
- S3: abnormal crypt group with inhibitor I1 applied;
- S4: abnormal crypt group with inhibitor I3 applied;

These test systems were allowed to evolve over 90 days, (i.e. $\sim$ quarter year) and four synthetic individuals were comprised in this case study. The average methylation level was recorded for every system, (illustrated in Figure 6.9).

Results indicate that methylation level decreased in every system defined; however, different patterns of variation were observed, (average methylation changes for the simulation are shown in Table 6.6). Specifically, the differences recorded between the final and initial methylation level for each system are i) $\approx$ 1.65% in S1, (healthy crypt group), ii) $\approx$2.66% in S2, (abnormal system with no inhibitor applied), iii) $\sim$ 1.75% in S3, (abnormal system with inhibitor I1 applied) and iv) $\approx$2.06% in S4, (abnormal system with inhibitor I3 applied). Lowest and highest differences can be seen in systems S1 and S2, i.e. the normal

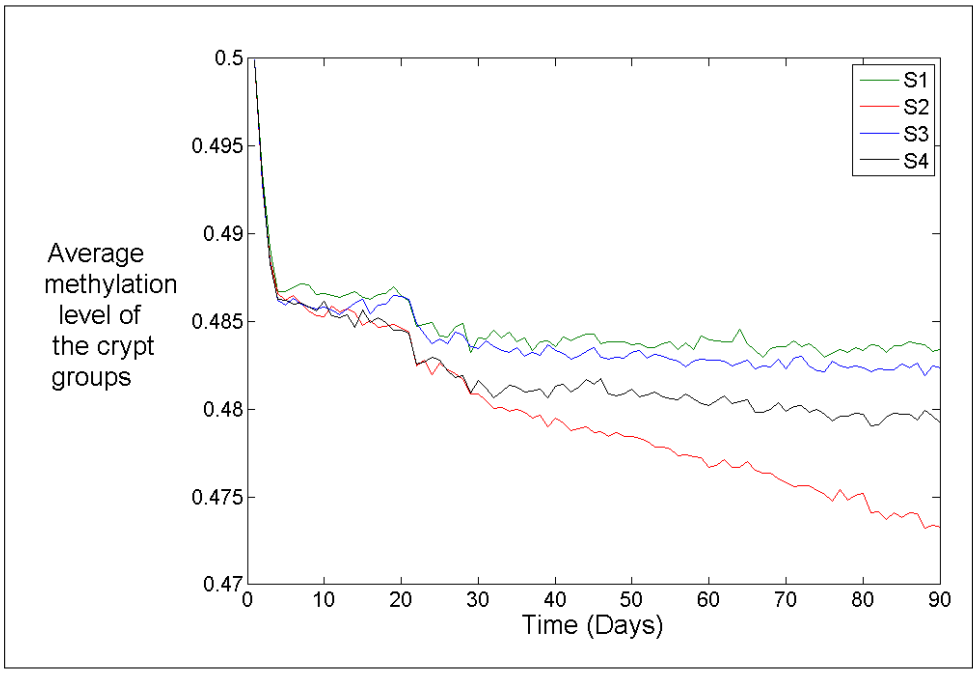


Figure 6.9: Comparison between average methylation level in four different intestinal crypt systems, over 90 'days', where different methylation inhibitor patterns were applied. S1 (green) and S2 (red), refer to normal and abnormal systems, respectively, and are considered to be 'control groups', given that no inhibitors were applied. Inhibitors I1 and I3 have been applied in S3 (blue) and S4 (black), respectively.

and the abnormal systems with no inhibitor applied, respectively. Given that inhibitor I1 was observed to have a higher effect than inhibitor I3 over time, (Figure 6.6), the difference for system S3 (with I1 applied), is less than that for S4, (with I3). Differences of ≈ 0.60 and 0.90% in average methylation level decrease over time can be seen among abnormal systems, with and without inhibitors applied, i.e. S2 - S3: 2.66% - 1.75% $\approx$ 0.90% and S2 - S4: 2.66% - 2.06% $\approx$ 0.60%. As expected, methylation inhibition does result in lower average methylation variation, which implies that some of epigenetic 'control' is possible and this has clear therapeutic value. However, % changes are still small and one issue for future work is how much deviation can be tolerated by the system before significant effects are observed. In addition, future work is planned to target methylation analysis in systems, where inhibitors are applied either sequentially, e.g. I1 then I3, or combined, e.g. I1+I3.

Table 6.6: Difference between the Final and Initial Average Methylation Level in the set of crypt groups

Crypt system/Methylation level	**S1 (healthy)**	**S2 (abnormal, no inhibitors)**	**S3 (abnormal + I1)**	**S4 (abnormal + I3)**
Initial ($\mathbf{M}_i$)	0.4999333	0.49989	0.4998907	0.499836
Final ($\mathbf{M}_f$)	0.4834276	0.4732609	0.4823474	0.4792197
Difference ($\Delta = \mid \mathbf{M}_f - \mathbf{M}_i)\mid$	0.0165057	0.0266291	0.0175433	0.0206163

6.7 Summary

AgentCrypt focuses on DNAm level variation, induced during cancer initiation and progression, in cells within the small intestine and colon crypt groups. The main contribution of AgentCrypt is to highlight methylation variation with respect to potential carcinogen and inhibitor influence. The main entities in AgentCrypt consists of *stem*, *progenitor* and *differentiated* cells in the colon, with the addition of the *Paneth* cell group in the small intestine. AgentCrypt integrates rules for every cell type in order to describe specific actions related to methylation update, cell division, differentiation and apoptosis. Considered to have a major impact on the cell cycle, *ageing* was also included in AgentCrypt, with mathematical expressions proposed to describe how progenitor, differentiated and Paneth cells undergo division and apoptosis. The main assumption of AgentCrypt is that cell methylation level can be influenced by both cell and mother-crypt neighbours in abnormal systems, due to intra- and inter-crypt influences. However, no inter-cell influences on methylation level were considered in 'healthy' crypts, (characterised by no alterations of inter-cellular gaps).

AgentCrypt implementation also considered carcinogen and methylation inhibitor influences. While methylation inhibitors have a direct impact on methylation modifications over time, carcinogens influence cell division, differentiation and apoptosis, without affecting methylation patterns directly. Fluctuations of the methylation inhibitor effect were also considered, and the *inhibitor response curve* expression proposed to describe the dynamics

of blocking methylation modifications over time. Two case studies were built to analyse the impact of such factors on intestinal cancer development.

Experimental data, (GEO accession GSE42752 and GSE50475), were used to initialize the model which was allowed to evolve over a period of 550 days. Crypt groups were observed and average methylation level was recorded. Differences were higher for carcinogen influenced systems and overall methylation level were more stable in the small intestine than in the colon crypt group. In addition, methylation level changes were inhibited under different patterns in abnormal systems, during a 90 days - period. Inhibitor effect was marked for abnormal crypt groups, with largest average methylation differences observed $\sim$ 0.60 - 0.90% lower when inhibitor was present. An interesting observation is that the final methylation level in abnormal systems with inhibitors applied, (measured after 90 days), is similar to methylation level measured after around 20 - 30 days in abnormal crypts, with no inhibitors. Thus, a difference of around 60 - 70 days can be observed between abnormal systems, implying that inhibitor presence can shrink tumour progression. However, no inhibitor considered was able to restore methylation to normal levels. Employment of multiple, (combined), inhibitors, which complement each another with respect to effect over time, may offer a solution, but clearly, additional tests are required to determine the appropriate inhibitor characteristics.

AgentCrypt described dynamics of different intestinal crypt systems focusing on individual cell activity. However, it is also important to investigate overall cell population deregulations in order to understand abnormal changes within intestinal crypt leading to cancer development. This specific analysis at intestinal tissue level is performed using the LogisticCrypt model, presented in the next chapter.

Chapter 7

LogisticCrypt - a Logistic Model of Intestinal Crypt Structure

7.1 Introduction

Based on the need to interpret base level features of a system as these impact on system behaviour as a whole, the LogisticCrypt model has been developed to perform comparative analysis between colon and small intestine tissues with regard to cell number deregulations observed in abnormal systems, (i.e. to target our third research objective, namely "To assess the effect of deregulation on intestinal crypt dynamics with regard to tumour development", (Section 1.2.1)). Specifically, these mechanisms refer to both cell proliferation and incidence of major intestinal crypt phenomena[1], which have been associated with increased risk of cancer development, (Chapter 2). The LogisticCrypt model describes crypt structure with regard to number of each cell type, (i.e. stem, progenitor, fully-differentiated and Paneth cells), at given time. Study focus is on two elements:

1. cell *division*, *differentiation* and *apoptosis* rates within the intestinal crypt;

2. *cell competition* related to 'crypt space'.

[1]Major crypt phenomena are 'crypt fission' (longitudinal crypt division) and the 'bottleneck effect' (crypt regeneration from a common ancestor-cell). The occurrence of both phenomena has been reported in normal intestinal crypts only over long time periods, approximated with 25 years for 'crypt fission' and 8.2 years for the 'bottleneck effect', (Chapter 2).

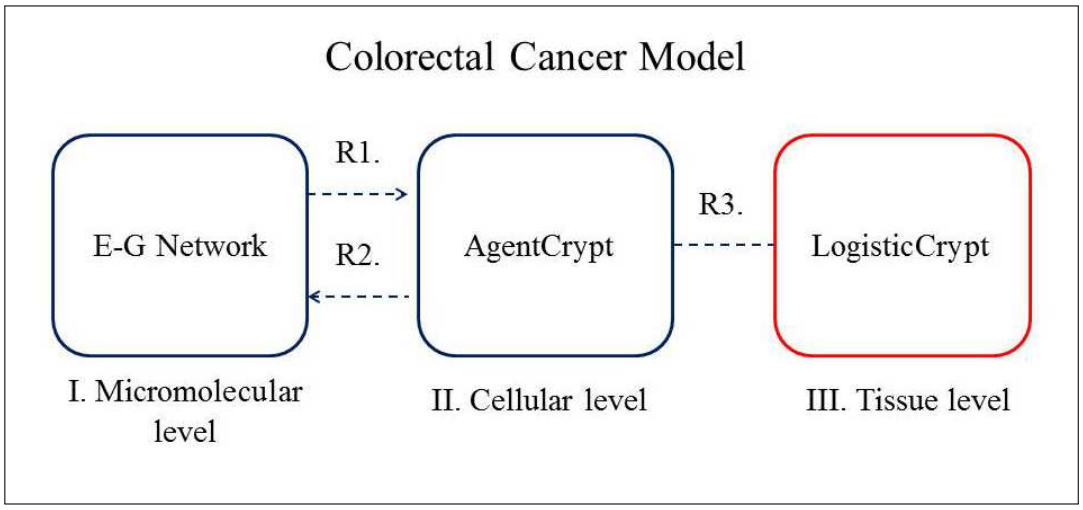


Figure 7.1: Structure of the Colorectal Cancer Model - focus on the LogisticCrypt model The relationships between model components may refer to R1) average methylation level for gene network / intestinal cell; R2) patient features; R3) intestinal crypt dynamics, explored following bottom-up and top-down approaches.

The position of the LogisticCrypt model in the Colorectal Cancer Model is highlighted in Figure 7.1. Thus, while AgentCrypt, (Chapter 6), characterises intestinal crypt dynamics, following a bottom-up approach, with information passed from cellular to tissue level, LogisticCrypt aims to complement this information through a top-down approach, where the prevalence of each cell type is determined by cell division, differentiation and apoptosis rates and restricted by intestinal crypt capacity. In order to achieve cross-comparison between intestinal tissues, LogisticCrypt was built initially to mimic crypt structure in the colon, (Section 7.2), and subsequently extended to the small intestine, (Section 7.3). In addition, similar to AgentCrypt, *carcinogen* influence is considered in order to investigate intestinal crypt modifications during tumour development. Implementation details are provided in Section 7.4 and the results obtained from the analysis of crypt deregulations, (e.g. aberrant increase in stem cell division), are described in Section 7.5. Finally, main contributions and limitations of the top-down model are presented in Section 7.6.

7.2 Modelling the colon crypt

In LogisticCrypt, the colon crypt consists of three main cell populations, namely *stem*, *progenitor* and *differentiated* cell groups, for which, specific actions are considered, citepReya2005. Similarly to the AgentCrypt model, (Chapter 6), stem cells can undergo both symmetrical and asymmetrical[2]) division or apoptosis, progenitors can divide or differentiate, and differentiated cells are removed from crypt following apoptosis, (Figure 7.2), [Potten et al., 2003; Frank, 2007; Humphries and Wright, 2008; Khalek et al., 2010; Vaiopoulos et al., 2012]. While in AgentCrypt, the focus was on individual cell behaviour, LogisticCrypt aims to investigate colon crypt dynamics at cell population level, i.e. to explore variations of the cell population size over time, as these can indicate tumour development, [Humphries and Wright, 2008; Vaiopoulos et al., 2012], (Chapter 2). In colon crypt, stem cell number increases only during symmetrical division, (by addition of new stem cells), and decreases during apoptosis. The progenitor cell group enlarges following both progenitor and asymmetrical stem cell divisions and diminishes during differentiation. Augmentation in differentiated cell set is determined by progenitor differentiation, while reduction of differentiated cell population size is caused by apoptosis.

The LogisticCrypt describes intestinal crypts based on stem, progenitor and differentiated cell number and rates for both symmetrical and asymmetrical stem cell division, (Rate$^{Stem}_{SymDiv}$ and Rate$^{Stem}_{AsymDiv}$), for stem and differentiated cell apoptosis, (Rate$^{Stem}_{Apoptosis}$ and Rate$^{Diff}_{Apoptosis}$, respectively) and for progenitor division and differentiation rates, (Rate$^{Prog}_{Div}$ and Rate$^{Prog}_{FullDiff}$, respectively), considered with values $\in$ [0, 1] (fractions). For example, the number of stem cells that undergo apoptosis at a given time t is equal to Rate$^{Stem}_{Apoptosis} \times$ Stem(t). In addition, given that three possible cell actions are defined for stem cell group, (namely symmetrical and asymmetrical divisions and apoptosis), the sum of these cell action rates is $\leq$ 1, (shown in expression (7.1)). The inequality indicates that a stem cell subgroup performs no action during a time-interval, (i.e. neither division or apoptosis), as stem cell division or apoptosis are rare events in normal colon crypts, (which are triggered

[2]Two new stem cells or a stem and a progenitor cells are born during symmetrical and asymmetrical divisions , respectively. In both situations, the mother stem cell is replaced by one daughter stem cell.

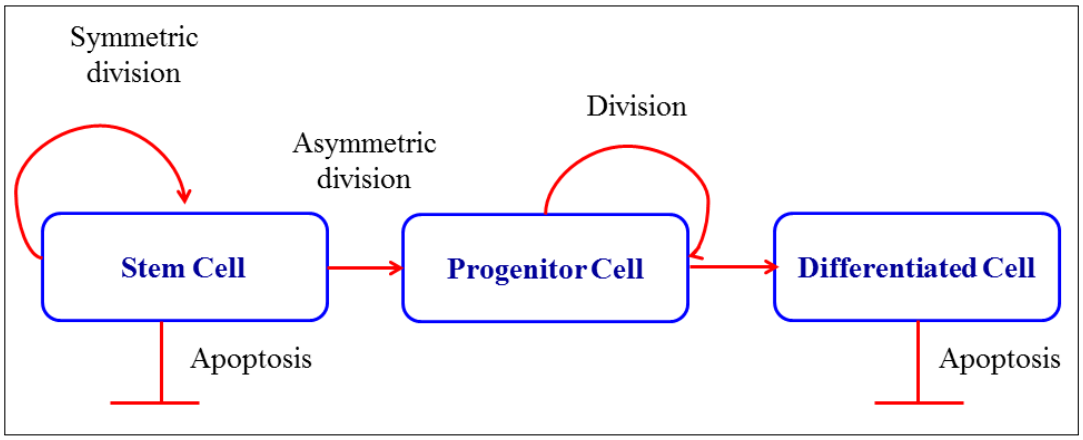


Figure 7.2: Intestinal Cell Actions in the LogisticCrypt model
A stem cell can divide symmetrically or asymmetrically, resulting into two new stem cells or a stem and a progenitor cells, respectively. In addition, a stem cell can undergo apoptosis, following which, it is removed from crypt. A progenitor can divide into two new progenitor cells or differentiate into a fully-differentiated cell. A differentiated cell can undergo apoptosis, being removed from the crypt, [Potten et al., 2003; Humphries and Wright, 2008].

by abnormal changes that affect crypt homestasis, [Bach et al., 2000]). Similar, the sum of the progenitor cell division and differentiation rates is ≤ 1, (expression 7.2).

$$Rate^{Stem}_{SymDiv} + Rate^{Stem}_{AsymDiv} + Rate^{Stem}_{Apoptosis} \leq 1 \tag{7.1}$$

$$Rate^{Prog}_{Div} + Rate^{Prog}_{FullDiff} \leq 1 \tag{7.2}$$

At every time-step, each cell type group is updated according to its specific actions. Thus, the stem cell population at time t+1 is composed from stem cells found in the colon crypt at the previous time-step, t, in addition to the new stem cells, (born through stem cell symmetrical division during this interval), minus dead stem cells that underwent apoptosis during the same time-period. The numbers of new and removed stem cells are calculated based on stem cell symmetrical division and apoptosis rates, respectively. Based on these, the expression for stem cell number updates is:

$$Stem(t+1) = Stem(t) + BornStem(t+1) - DeadStem(t+1) \tag{7.3}$$

$$BornStem(t+1) = Stem(t) \times Rate^{Stem}_{SymDiv}$$

$$DeadStem(t+1) = Stem(t) \times Rate^{Stem}_{Apoptosis} \Rightarrow$$

$$Stem(t+1) = Stem(t) + Stem(t) \times (Rate^{Stem}_{SymDiv} - Rate^{Stem}_{Apoptosis})$$

where BornStem, DeadStem = new and removed stem cell numbers, respectively, at time t+1. Analogously, the progenitor cell population at time t+1 is given by the total of i) progenitors found in crypt at time t and ii) new progenitors, (born through both progenitor and asymmetrical stem cell division), from which, the group of progenitors that underwent differentiation is removed. The size of the differentiated cell group at time t+1 is computed from the differentiated cell number at time t, plus the number of new differentiated cells, (originated from progenitors), minus differentiated cells that underwent apoptosis. In conclusion, the progenitor and differentiated cell updates are given by:

$$Prog(t+1) = Prog(t) + Stem(t) \times Rate^{Stem}_{AsymDiv} + Prog(t) \times (Rate^{Prog}_{Div} - Rate^{Prog}_{FullDiff}) \tag{7.4}$$

$$Diff(t+1) = Diff(t) + Prog(t) \times Rate^{Prog}_{FullDiff} - Diff(t) \times Rate^{Diff}_{Apoptosis} \tag{7.5}$$

with variables defined as given in Table 7.1.

Table 7.1: Variable definitions for LogisticCrypt in the colon crypt

Prog(t), Diff(t)	the numbers of progenitor and differentiated cells, respectively, observed in crypt at time t;
Prog(t+1), Diff(t+1)	numbers of above cell types at time t+1;
$Rate^{Prog}_{Div}$ and $Rate^{Prog}_{FullDiff}$	the rates of progenitor cell division and differentiation, respectively;
$Rate^{Diff}_{Apoptosis}$	the rate differentiated cell apoptosis.

Intra- and *inter*-specific competitions are also incorporated into LogisticCrypt to de-

scribe crypt dynamics. In such systems, an increase in one species' population restricts the expansion of the other populations, (due to inter-specific competition), and an increase in population size leads to decrease in population growth rate, (due to intra-specific competition), (subsection 3.3.3).

7.2.1 Modelling cell intra- and interspecific competition

Intra-specific competition, indicated by the **D**ensity-**D**ependence **C**oefficient, DDC, (Subsection 3.3.3), is represented for every cell type within the colon crypt, where the DDC values for stem, progenitor and differentiated cells, (i.e. DDC_{STEM}, DDC_{PROG} and DDC_{DIFF}, respectively), are computed based on expression (3.4), as discussed further.

The $DDC_{STEM}(t)$ value is calculated using expression (7.6), based on the stem cell carrying capacity, i.e. maximum stem cell number defined for colon crypt, (MaxStemNo), and stem cell number found in colonic system at time t, Stem(t). $DDC_{STEM}(t)$ takes a value between 0 and 1, which is closer to 1 when Stem(t) is much smaller than MaxStemNo, and closer 0 when Stem(t) reaches stem cell carrying capacity, (i.e. Stem(t) $\simeq$ MaxStemNo). Thus, the $DDC_{STEM}(t)$ value declines as stem cell population size increases, and the addition of the $DDC_{STEM}(t)$ coefficient to stem cell population growth rate, (i.e. $Rate^{Stem}_{SymDiv} - Rate^{Stem}_{Apoptosis}$), indicates that the latter decreases, as well, as stem cell population extends. While expression (7.3) describes stem cell population growth in systems with *no* competition, expression (7.7), (which integrates $DDC_{STEM}(t)$), is proposed for stem cell number updates in colon crypt with intra-specific competition over time.

$$DDC_{STEM}(t) = \frac{[MaxStemNo - Stem(t)]}{MaxStemNo} \tag{7.6}$$

$$Stem(t+1) = Stem(t) + Stem(t) \times (Rate^{Stem}_{SymDiv} - Rate^{Stem}_{Apoptosis}) \times DDC_{STEM}(t) \tag{7.7}$$

where: MaxStemNo = stem cell *carrying capacity*, i.e. maximum stem cell number defined for colon crypt; Stem(t+1), Stem(t) = stem cell numbers at time t+1 and t, respectively; $Rate^{Stem}_{SymDiv}$, $Rate^{Stem}_{Apoptosis}$ = stem cell symmetrical division and apoptosis rates, respec-

tively. Similarly, the density-dependence coefficient is calculated for progenitor cells, using expression (7.8), and is considered during the progenitor cell update step. Thus, the initial expression (7.4), used for progenitor cell number dynamics, changes to expression (7.9).

$$DDC_{PROG}(t) = \frac{[MaxProgNo - Prog(t)]}{MaxProgNo} \tag{7.8}$$

$$\begin{aligned} Prog(t+1) = Prog(t) + [Stem(t) \times Rate^{Stem}_{AsymDiv} + \\ + Prog(t) \times (Rate^{Prog}_{Div} - Rate^{Prog}_{FullDiff})] \times DDC_{PROG}(t) \end{aligned} \tag{7.9}$$

where MaxProgNo is maximum progenitor cell number in colon crypt, $DDC_{PROG}(t)$ is the density-dependent coefficient for progenitor cell group and the rest of variables are defined in Table 7.1. Analogous, density-dependence coefficient and update step for differentiated cells are given by expressions (7.10) and (7.11), respectively. Considering crypt structure, the carrying capacity for differentiated cells is calculated based on the maximum numbers for i) total, ii) stem and iii) progenitor cells allowed in colon crypt, i.e. MaxTotalCellNo, MaxStemNo, and MaxProgNo, respectively.

$$DDC_{DIFF}(t) = \frac{[MaxTotalCellNo - MaxStemNo - MaxProgNo - Diff(t)]}{(MaxTotalCellNo - MaxStemNo - MaxProgNo)} \tag{7.10}$$

$$Diff(t+1) = Diff(t) + [Prog(t) \times Rate^{Prog}_{FullDiff} - Diff(t) \times Rate^{Diff}_{Apoptosis}] \times DDC_{DIFF}(t) \tag{7.11}$$

where $DDC_{DIFF}(t)$ = density-dependent coefficient for differentiated cells at given time t and the other variables are defined as given in Table 7.1.

Given that the stem cell number range is known, (i.e. 16 - 19 in colon and 1-6 in small intestine, (Section 2.4)), **the inter-specific competition** describes the relationship between

the progenitor and differentiated cells within the colon crypt. Specifically, this refers to the total number of progenitor and differentiated cells from the crypt that must always be less than or equal to the crypt capacity without stem cell number, (expression 7.12). Thus, given that the crypt capacity is limited, an increase in progenitor cell number determines a decrease in differentiated cell population size.

$$Prog(t) + Diff(t) \leq MaxTotalCellNo - MaxStemNo \tag{7.12}$$

Considering expressions (7.6) - (7.12), the cell population changes in colon crypt over time are determined based on a set of 12 parameters, namely i) stem, progenitor and differentiated cell numbers at time t, ii) the maximum sizes of total, stem and progenitor cell populations, and iii) the rates of stem and progenitor cell divisions, stem and differentiated cell apoptosis and progenitor differentiation. A major phase in the LogisticCrypt model development is minimization of the set of the input parameters needed to describe the colon crypt dynamics. This can be achieved using relationships between model parameters and the steps made for parameter group reduction are discussed in the next subsection.

7.2.2 Relationships between cell action rates in the healthy colon crypt system

The relationship between model dependent and independent parameters relies for plausibility on knowledge of the mechanisms which characterise the healthy colon crypt. This is a system of different cell types, where cell action rates ensure low fluctuations in any cell type number in successive time-steps, [Van Leeuwen et al., 2006; Humphries and Wright, 2008; Leedham and Wright, 2008]. Thus, phenomena such as 'crypt fission' and the 'bottleneck effect', (introduced in Chapter 2), can occur in normal systems only over long time-periods.

In order to compute the relationship between cell division, differentiation and apoptosis rates in the healthy[3] colon crypt, an 'abstract' system, with *no* variations on cell number over time, is initially considered. In such system, the cell number differences between two

[3]Given that cell cycle is affected during tumour development, the relationships between cell action rates in abnormal systems are calculated based on corresponding relationships in healthy crypts, but taking account also of deregulations of cell division, differentiation and apoptosis, [Van Leeuwen et al., 2006; Humphries and Wright, 2008; Leedham and Wright, 2008].

successive time-steps, i.e. *t* and *(t+1)*, are equal to zero for each cell type:

$$Stem(t+1) - Stem(t) = 0 \tag{7.13}$$

$$Prog(t+1) - Prog(t) = 0 \tag{7.14}$$

$$Diff(t+1) - Diff(t) = 0 \tag{7.15}$$

The steps performed to identify a relationship between stem cell division and apoptosis rates in such a system are presented further. For a complete summary of how other relationships, (including progenitor division and differentiation, differentiated cell apoptosis rates), are calculated, see Appendix D.

Stem cell group - calculation of cell action rate relationships Relationships between stem cell action rates are calculated based on expressions for i) stem cell number updates in colon crypt over time, ii) sum of all stem cell action rates and iii) 'no fluctuation in stem cell population size' condition. The calculation procedure is given by:

Input: Expression (7.1) for sum of all stem cell action rates:

$Rate^{Stem}_{SymDiv}$ + $Rate^{Stem}_{AsymDiv}$ + $Rate^{Stem}_{Apoptosis} \leq 1$

Expression (7.7) for stem cell number updates in colon crypt over time:

Stem(t+1) = Stem(t) + Stem(t) $\times$ (Rate$^{Stem}_{SymDiv}$ - Rate$^{Stem}_{Apoptosis}$) $\times$ DDC$_{STEM}$(t), with DDC$_{STEM}$(t) $\neq 0$

Expression (7.13) for no variations in stem cell compartment:

Stem(t+1) - Stem(t) = 0

Output: Rate$^{Stem}_{SymDiv}$? Rate$^{Stem}_{Apoptosis}$? Rate$^{Stem}_{AsymDiv}$

Calculation: Given expression (7.13), Stem(t+1) - Stem(t) = 0 $\overset{+expr.(7.7)}{\Longrightarrow}$

Stem(t) + Stem(t) $\times$ (Rate$^{Stem}_{SymDiv}$ - Rate$^{Stem}_{Apoptosis}$) $\times$ DDC$_{STEM}$(t) - Stem(t) = 0 $\Rightarrow$

(Rate$^{Stem}_{SymDiv}$ - Rate$^{Stem}_{Apoptosis}$) $\times$ DDC$_{STEM}$(t) = 0 $\Rightarrow$

Rate$^{Stem}_{SymDiv}$ - Rate$^{Stem}_{Apoptosis}$ = 0, (given DDC$_{STEM}$(t) $\neq$ 0) $\Rightarrow$

$$Rate^{Stem}_{Apoptosis} = Rate^{Stem}_{SymDiv} \tag{7.16}$$

From expressions (7.1) and (7.16), $\Rightarrow$

$$Rate^{Stem}_{SymDiv} \leq \frac{(1 - Rate^{Stem}_{AsymDiv})}{2} \tag{7.17}$$

According to expression (7.17), stem cell population dynamics can be described by only two parameters, namely symmetrical and asymmetrical stem cell division rates.

Progenitor and differentiated cell groups - summary of relationship calculation between cell action rates

Further, the relationships between progenitor cell action rates are based on expressions for i) progenitor cell updates, ii) sum of progenitor cell action rates in colon crypt and iii) 'no variations in progenitor cell compartment' condition, (with calculation details provided in Appendix D):

Input: Expression (7.2) on sum of progenitor cell action rates in colon crypt:

$Rate^{Prog}_{Div} + Rate^{Prog}_{FullDiff} \leq 1$

Expression (7.9) on progenitor cell updates:

Prog(t+1) = Prog(t) + [Stem(t) $\times Rate^{Stem}_{AsymDiv}$ + Prog(t) $\times (Rate^{Prog}_{Div} - Rate^{Prog}_{FullDiff})] \times$ $DDC_{PROG}(t)$

Expression (7.14) for no variations in progenitor cell compartment:

Prog(t+1) - Prog(t) = 0

Output: $Rate^{Prog}_{Div}$? $Rate^{Prog}_{FullDiff}$

Solution: Given expressions (7.9) and (7.14) $\Rightarrow$

$$Rate^{Prog}_{FullDiff} = \frac{Stem(t_0)}{Prog(t_0)} \times Rate^{Stem}_{AsymDiv} + Rate^{Prog}_{Div} \tag{7.18}$$

In addition, using expression (7.2) $\Rightarrow$

$$Rate_{Div}^{Prog} \leq \frac{[1 - \frac{Stem(t_0)}{Prog(t_0)} \times Rate_{AsymDiv}^{Stem}]}{2} \tag{7.19}$$

According to expression 7.18, the progenitor differentiation rate can be calculated based on i) the initial stem and progenitor cell number and ii) the progenitor and asymmetrical stem cell division rates. Progenitor population growth over time is controlled by expression (7.19), which indicates that an increase in stem cell symmetrical division rate determines a decrease in progenitor division rate.

Finally, the differentiated cell apoptosis rates can be calculated based on i) initial size of each cell type population and ii) the rates of progenitor and symmetrical stem cell divisions, as indicated by expression (7.20); (again, complete calculation procedure is provided in Appendix D).

Input: Expression (7.11) on differentiated cell updates:

Diff(t+1) = Diff(t) + [Prog(t) $\times Rate_{FullDiff}^{Prog}$ - Diff(t) $\times Rate_{Apoptosis}^{Diff}$] $\times DDC_{DIFF}(t)$

Expression (7.15) for no variations in differentiated cell compartment:

Diff(t+1) - Diff(t) = 0

Expression (7.18) for progenitor differentiated rate:

$Rate_{FullDiff}^{Prog} = \frac{Stem(t_0)}{Prog(t_0)} \times Rate_{AsymDiv}^{Stem} + Rate_{Div}^{Prog}$

Output: $Rate_{Apoptosis}^{Diff}$? $Rate^{Prog}$? $Rate^{Stem}$

Solution: From expressions (7.11) and (7.15) $\Rightarrow$

$Rate_{Apoptosis}^{Diff} = \frac{Prog(t_0)}{Diff(t_0)} \times Rate_{FullDiff}^{Prog}$, (complete calculation steps provided in Appendix D).

In addition, using expression (7.18) to replace $Rate_{FullDiff}^{Prog}$, $\Rightarrow$

$$Rate_{Apoptosis}^{Diff} = \frac{Prog(t_0)}{Diff(t_0)} \times [\frac{Stem(t_0)}{Prog(t_0)} \times Rate_{SymDiv}^{Stem} + Rate_{Div}^{Prog}] \tag{7.20}$$

Relationships between cell action rates in healthy systems

Given the evidence of 'crypt fission' and the 'bottleneck effect' in a normal intestinal crypt, low variation in stem cell number exists over time, [Yatabe et al., 2001; Jin et al., 2009], (Subsection 2.6.1). Thus, the relationship between the symmetrical division and apoptosis rate of the stem cell in 'no variation' systems), (expression (7.16)), should be extended to permit small fluctuations in stem cell population size over time, i.e.

$$Rate^{Stem}_{Apoptosis} = Rate^{Stem}_{SymDiv} \pm \delta_{COLON},\ with\ \delta_{COLON} \in (0, 1) \tag{7.21}$$

where $\pm\delta_{COLON}$ indicates low changes in stem cell number over time. In case of '+', stem cell apoptosis rate is higher than stem cell symmetrical division rate, i.e. stem cell population size decreases (more stem cells die than are born). In case of '-', division rate is higher than apoptosis rate and stem cell number increases.

Stem cell symmetrical division and apoptosis rates can be deduced based on time needed for major crypt occurrence in healthy human crypts, which was approximated to 8.2 years and 25 years for the 'bottleneck effect' and 'crypt fission', respectively, [Yatabe et al., 2001; Graham et al., 2011]. The crypt cycle within a normal system is defined to start with a decrease in stem cell number, (inducing the 'bottleneck effect'), followed by an expansion of the stem cell compartment, (which precedes 'crypt fission'). For example, a set of values, with $Rate^{Stem}_{Apoptosis}$ = 0.01 and $Rate^{Stem}_{SymDiv}$ = 0.003 during 'bottleneck effect' (i.e. $Rate^{Stem}_{Apoptosis} > Rate^{Stem}_{SymDiv}$) and $Rate^{Stem}_{SymDiv}$ = 0.0115 during 'crypt fission' (i.e. $Rate^{Stem}_{Apoptosis} < Rate^{Stem}_{SymDiv}$), was found to be plausible for describing both major crypt phenomena in reported time, (i.e. 8.2 and 25 years respectively, as illustrated in Figure 7.3).

A stem cell cycle is given by stem cell division, approximated to seven days in human colon, [Potten et al., 2003; Frank, 2007]. Based on these, time for the 'bottleneck effect' to occur is $\approx$ 1 cycle/week $\times$ 52 weeks/year $\times$ 8.2 years = 426.4 cycles. Similarly, time period for 'crypt fission' incidence is $\approx$ 25 years $\times$ 52 cycles/year = 1300 cycles. Other values for these rates are also possible and there is certainly a need to perform a sensitivity analysis for colon crypt parameters, (phase discussed in Chapter 8).

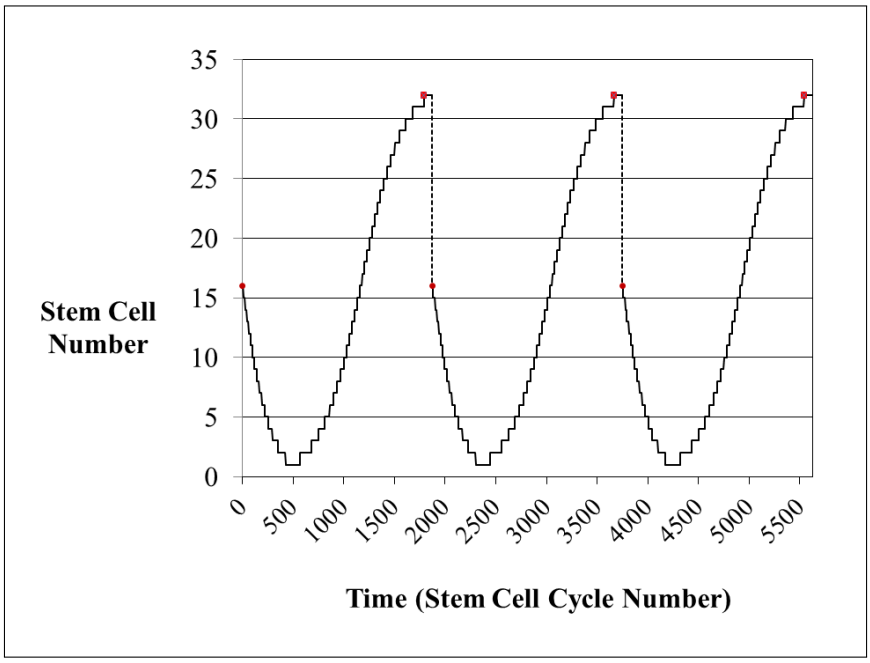


Figure 7.3: Stem cell numbers over three successive Colon Crypt cycles
A crypt cycle is considered to start with a decrease in stem cell number until the 'bottleneck effect' is observed, after which, the stem cell number increases until it reaches a maximum value (fission). Red points are used as markers for the crypt cycle. The interrupted line shows the decrease in stem cell number recorded after crypt fission.

In conclusion, the set of independent variables needed to describe colon crypt dynamics is given by i) stem cell and progenitor division rates and ii) the initial size for each cell type population. Other parameter values can be calculated, (e.g. for progenitor differentiation rate) using expressions (7.16), (7.18), and (7.20). Thus, the size of colon crypt parameter group is reduced from initial twelve to six. In addition, major crypt phenomena were simulated by the LogisticCrypt model based on a set of plausible values for stem cell symmetrical division and apoptosis rates, (e.g. $Rate^{Stem}_{Apoptosis}$ = 0.01 and $Rate^{Stem}_{SymDiv}$ = 0.003 during 'bottleneck effect' and $Rate^{Stem}_{SymDiv}$ = 0.0115 during 'crypt fission').

7.3 LogisticCrypt - Extensions

7.3.1 LogisticCrypt for the small intestine tissue

Given the similarities[4] between the structure of colon and small intestine crypts, (Chapter 2), the LogisticCrypt model is extended to investigate cell population dynamics in small intestine. Information on Paneth cell group is integrated, as follows. Similarly to *(fully-)* differentiated cells, 'apoptosis' is considered to be the single action performed by Paneth cells, given that Paneth cells are differentiated cells, [Umar, 2010]; in addition, progenitor differentiation into Paneth cells is added to progenitor action set, [Kim et al., 2005]. In consequence, two new parameters, namely rates for *progenitor differentiation into Paneth cells* and *Paneth apoptosis*, (i.e. $Rate^{Prog}_{Paneth}$ and $Rate^{Paneth}_{Apoptosis}$, respectively), are included in LogisticCrypt, in addition to those used for the colon crypt. Intra-specific competition features are considered also for Paneth cell population, where the the density-dependence coefficient for Paneth cell group at a given time t, $DDC_{PANETH}(t)$, is calculated based on Paneth cell carrying capacity, MaxPanethNo, and Paneth cell number at time t, Paneth(t), (expression (7.22)). Based on these, Paneth cell population at time t+1 is given by i) Paneth cell group at time t, ii) new Paneth cells, minus iii) the set of Paneth cells that underwent apoptosis during the last time interval, (expression (7.23), with variables defined in Table 7.2).

$$DDC_{PANETH}(t) = \frac{[MaxPanethNo - Paneth(t)]}{MaxPanethNo} \tag{7.22}$$

$$Paneth(t+1) = Paneth(t) + [Prog(t) \times Rate^{Prog}_{Paneth} - \tag{7.23}$$
$$- Paneth(t) \times Rate^{Paneth}_{Apoptosis}] \times DDC_{PANETH}(t)$$

[4]Paneth cell type is the fourth cell type specific to small intestine crypt, in addition to stem, progenitor and differentiated cells, found also in the colon crypt. Paneth cells are differentiated cells, (derived from progenitors), reported in relative small number in comparison with differentiated and progenitor population size, and are located at crypt base in small intestine crypt. Paneth cell presence in colon crypt is considered a marker for tumour development, [Sancho et al., 2003; Kim et al., 2005; Nicolas et al., 2007; Humphries and Wright, 2008; Umar, 2010; Clevers and Bevins, 2013], (Section 2.4), (Chapter 2).

Table 7.2: Variable definitions for LogisticCrypt for the small intestine crypt

Parameter name	**Parameter description**
$DDC_{PANETH}(t)$	The density-dependence coefficient related to Paneth cell at time t;
MaxPanethNo	the carrying capacity for the Paneth cells;
Stem(t), Prog(t), Prog(t+1), $Rate^{Stem}_{SymDiv}$, $Rate^{Stem}_{Apoptosis}$, $Rate^{Stem}_{AsymDiv}$, $Rate^{Prog}_{Div}$, $Rate^{Prog}_{FullDiff}$, $Rate^{Diff}_{Apoptosis}$	parameters as defined for colon crypt in 7.1;
Paneth(t), Paneth(t+1)	the Paneth cell numbers observed in small intestine crypt at t and t+1, respectively;
$Rate^{Prog}_{Paneth}$	the rate of progenitor differentiation into Paneth cells, (model input parameter with value in the interval [0, 1]);
$Rate^{Paneth}_{Apoptosis}$	the rate of Paneth cell apoptosis in the small intestine.

Given the inclusion of 'progenitor differentiation into Paneth cells' to progenitor action set, expressions for progenitor cell dynamics are also changed. Specifically, progenitor cell population size at time t+1 is now calculated based on i) population size at time t, ii) new progenitors, (born through progenitor and asymmetrical stem cell divisions), from which, iii) the number of progenitors that underwent differentiation into both (fully-)differentiated and Paneth cells, is subtracted, (extending, thus, expression (7.9), used for progenitor group updates in colon crypt):

$$Prog(t+1) = Prog(t) + [Stem(t) \times Rate^{Stem}_{AsymDiv} + \\ + Prog(t) \times (Rate^{Prog}_{Div} - Rate^{Prog}_{FullDiff} - Rate^{Prog}_{Paneth})] \times DDC_{PROG}(t) \tag{7.24}$$

In addition, the sum of all progenitor action rates is given by:

$$Rate^{Prog}_{Div} + Rate^{Prog}_{FullDiff} + Rate^{Prog}_{Paneth} \leq 1 \tag{7.25}$$

The relationships between cell action rates in the healthy small intestine crypt are calculated similarly to those from the healthy colon crypt, where the 'no - variation' small intestine system is described by:

$$Paneth(t+1) - Paneth(t) = 0 \tag{7.26}$$

$$Stem(t+1) - Stem(t) = 0 \tag{7.27}$$

$$Prog(t+1) - Prog(t) = 0 \tag{7.28}$$

$$Diff(t+1) - Diff(t) = 0 \tag{7.29}$$

Since in the model, the inclusion of Paneth cell group was not linked to changes in stem cell population dynamics, (expression (7.23)), the relationship between stem cell action rates in the healthy small intestine crypt is similar to expression (7.21), (used in the colon crypt), i.e.:

$$Rate^{Stem}_{Apoptosis} = Rate^{Stem}_{SymDiv} \pm \delta_{SI}, with\ \delta_{SI} \in (0,1) \tag{7.30}$$

Given the modifications introduced in progenitor dynamics set, the relationships between progenitor action rates can be summarised by expressions (7.31) and (7.32), (again, full calculation provided in Appendix D). These are calculated based on i) progenitor cell updates, ii) sum of all progenitor cell action rates in small intestine and iii) 'no variations in progenitor cell compartment' condition:

Input: Expression (7.14) for no variations in progenitor cell compartment:

Prog(t+1) - Prog(t) = 0

Expression (7.24) on progenitor cell updates:

Prog(t+1) = Prog(t) + [Stem(t) $\times Rate^{Stem}_{AsymDiv}$ + Prog(t) $\times (Rate^{Prog}_{Div} - Rate^{Prog}_{FullDiff} - Rate^{Prog}_{Paneth})] \times DDC_{PROG}(t)$

Expression (7.25) on sum of all progenitor cell action rates in small intestine:

$Rate^{Prog}_{Div} + Rate^{Prog}_{FullDiff} + Rate^{Prog}_{Paneth} \leq 1$

Output: $Rate^{Prog}_{Div}$? $Rate^{Prog}_{FullDiff}$? $Rate^{Prog}_{Paneth}$

Solution: Given expressions (7.14) and (7.24) $\Rightarrow$

$$Rate^{Prog}_{FullDiff} = \frac{Stem(t_0)}{Prog(t_0)} \times Rate^{Stem}_{AsymDiv} + Rate^{Prog}_{Div} - Rate^{Prog}_{Paneth} \tag{7.31}$$

In addition, using expression (7.25) $\Rightarrow$

$$Rate^{Prog}_{Div} \leq \frac{[1 - \frac{Stem(t_0)}{Prog(t_0)} \times Rate^{Stem}_{AsymDiv}]}{2} \tag{7.32}$$

Thus, the rate of progenitor differentiation into (fully-) differentiated cells, $Rate^{Prog}_{Full_Diff}$, depends on the rates of i) asymmetrical stem cell and progenitor divisions, $Rate^{Stem}_{AsymDiv}$ and $Rate^{Prog}_{Div}$ respectively, (similar to the colon crypt), and ii) progenitor differentiation into Paneth cells, $Rate^{Prog}_{Paneth}$. An increase in $Rate^{Prog}_{Paneth}$ determines a decrease in $Rate^{Prog}_{Full_Diff}$. In addition, $Rate^{Prog}_{Div}$ is independent by $Rate^{Prog}_{Paneth}$, but depends on $Rate^{Stem}_{AsymDiv}$, (expression 7.32).

Finally, relationships between cell action rates are calculated for differentiated and Paneth cells. The differentiated cell apoptosis rate is given by i) initial stem, progenitor and differentiated cell numbers, ii) stem and progenitor division rates and also iii) progenitor differentiation into Paneth cell rate, (expression (7.33)). Based on these, it can be observed that while an increase of either stem or progenitor division rate causes an increase in differentiated cell apoptosis rate, an increase of $Rate^{Prog}_{Paneth}$ induces a decrease of $Rate^{Diff}_{Apoptosis}$, (in order to maintain the cell number equilibrium in normal systems). In addition, Paneth cell apoptosis rate depends on progenitor differentiation in Paneth cell rate, (expression (7.34)), and, in healthy small intestine crypts, an increase in the latter determines an increase in the former.

$$Rate^{Diff}_{Apoptosis} = \frac{Prog(t_0)}{Diff(t_0)} \times [\frac{Stem(t_0)}{Prog(t_0)} \times Rate^{Stem}_{AsymDiv} + Rate^{Prog}_{Div} - Rate^{Prog}_{Paneth}] \tag{7.33}$$

$$Rate^{Paneth}_{Apoptosis} = \frac{Prog(t_0)}{Paneth(t_0)} \times Rate^{Prog}_{Paneth} \tag{7.34}$$

7.3.2 Carcinogen influence

The carcinogen influence in LogisticCrypt is associated with modifications in the cell action rates that determine cell population growth. Specifically, since cell proliferation is given by both increased division and decreased apoptosis rates, the stem and progenitor cell division rates are increased by a value $\delta 2$, where $\delta 2 \in [0, 1]$, (input), while the apoptosis rates, for both fully-differentiated and Paneth cells, are decreased by the same $\delta 2$ value. For example, the expression for carcinogen impact on symmetrical stem cell division rate is:

$$Rate^{Stem}_{SymDiv} \leftarrow Rate^{Stem}_{SymDiv} \times (1 + \delta 2) \tag{7.35}$$

The $\delta 2$ parameter was considered for a sensitivity analysis, with $\delta 2 \in \{5\%, 10\%, , 50\%\}$; the results are discussed in subsection 7.5.2.

7.4 Implementation details of the LogisticCrypt model

Adopting the same rationale as before, for E-G Network & AgentCrypt models, the LogisticCrypt was also developed as an objected oriented model using C++. The LogisticCrypt class is the main class and provides functionality for handling crypt dynamics based on fixed rate values of cell division, differentiation and apoptosis. The intestinal crypt characteristics, e.g. the stem cell number, are stored in a LogisticModelSettings object and transmitted in a single step to the LogisticCrypt object. The class diagram for this model is simple and is illustrated in Figure 7.4.

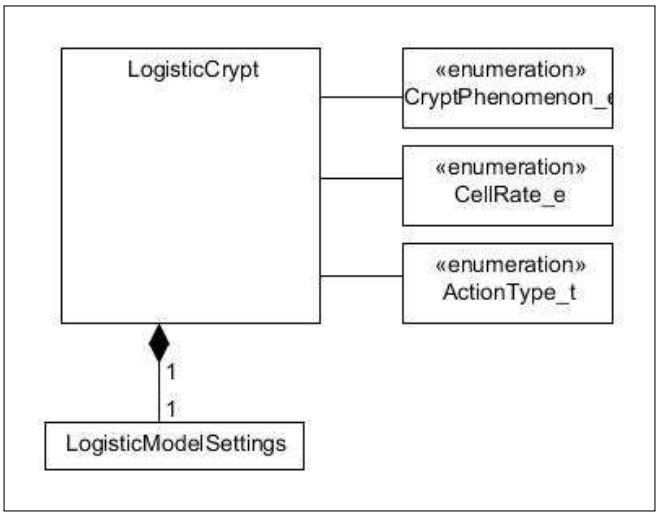


Figure 7.4: Class diagram for the LogisticCrypt model

7.5 Results and discussion

LogisticCrypt function was investigated through test cases which explored crypt dynamics following modifications in stem and progenitor cell groups and consideration of carcinogen influence in intestinal crypts, (Subsection 7.3.2). Specifically, the focus was on a comparison between 'healthy' and 'affected' crypts with regard to the occurrences of i) *crypt fission*, (doubling of the no. of stem cells), [McDonald et al., 2006; Jin et al., 2009], and the '*bottleneck effect*', (reduction of the no. of stem cell to 1), [Yatabe et al., 2001]; and ii) *tumour growth*, (i.e. total cell group at maximum), in carcinogen-affected and unaffected systems. Both *intra-* and *inter*-tissue comparisons on crypt phenomena are discussed.

Cell number parameter values A set of input parameters, which describes cell population size in a normal system for each intestinal tissue, is included in Table 7.3. These values were considered a potential setting group for the systems included in the simulations presented here, (discussion on parameter value choice follows). However, other values for cell number parameters, (e.g. Paneth cell number), should be considered also, and the need of a sensitivity analysis of this parameter set is anticipated, (discussed in Chapter 8). Given that cell population size was approximated to around 2000 cells and 260 cells in healthy colon and small intestine crypt, respectively, [Sancho et al., 2003; Nicolas et al., 2007; Khalek

et al., 2010; Vaiopoulos et al., 2012], (Section 2.4), these values were assigned to the 'initial cell number' parameter, i.e. *InitCellNumber*, in each intestinal tissue. Similarly, the stem cell number at time t_0, *Stem(t_0)*, was set to be 16 - 19 cells in colon and 1 - 6 cells in small intestine crypt, [Sancho et al., 2003; Brittan and Wright, 2004; Khalek et al., 2010], (Section 2.4).

Table 7.3: Input parameter values used in LogisticCrypt for the cell numbers in the intestinal crypts

Parameter name	**Parameter value in the colon crypt**	**Parameter value in the small intestine crypt**
Stem(t_0)	16	6
InitCellNumber	2000	260
Paneth(t_0)	0	10
Prog(t_0) (0.25 $\times$ InitCellNumber)	500	65
MaxStemNo (2 $\times$ InitStemCellNumber)	32	12
MinStemCellNumber	1	1
MaxProgNo	1000	130
MaxPanethCellNumber	0	20
MaxTotalCellNo (2 $\times$ InitCellNumber)	4000	520

Precise information on progenitor and Paneth cell numbers in healthy human intestinal crypts was not found to be available. In consequence, the progenitor cell compartment was considered to occupy initially a percentage of crypt space, (e.g. Perc_{PROG} = 25%). Thus, the initial progenitor cell number, *Prog(t_0)*≈500 cells in colon crypt, (25% $\times$ 2000 cells), and 65 cells in the small intestine crypt. The Paneth cell group at the bottom of the small intestine crypt is relatively small, so the initial Paneth cell number, *Paneth(t_0)*, was approximated to 10 cells in small intestine, and no cells in the colon crypt, (as absent for healthy colon tissue). Finally, the maximum stem cell number, *MaxStemNo*, was set to twice the initial stem cell number Stem(t_0), (since 'crypt fission' leads to doubling stem cell population size in intestinal crypts, [McDonald et al., 2006; Jin et al., 2009], (Section 2.6)). Similarly, the maximum numbers for i) total, ii) progenitor and iii) Paneth cells were taken to be double initial numbers of these cell groups. Moreover, given the evidence of the

'bottleneck effect' in normal systems, (associated with decrease in stem cell population to a single stem cell, [Yatabe et al., 2001], (Section 2.6)), the minimum stem cell number was set to 1:

$$MaxStemNo = 2 \times Stem(t_0) \tag{7.36}$$

$$MaxTotalCellNo = 2 \times InitCellNumber \tag{7.37}$$

$$MaxProgNo = 2 \times Prog(t_0) \tag{7.38}$$

$$MaxPanethNo = 2 \times Paneth(t_0) \tag{7.39}$$

$$MinStemCellNumber = 1 \tag{7.40}$$

Cell action rate parameter values In addition to cell number parameters, a set of input parameters, (included in Table 7.4), is proposed to describe cell action rates; (with a sensitivity analysis for this parameter group discussed in Chapter 8). The values of stem cell division and apoptosis rates in normal systems, included in the test cases presented here, are the same as those of the stem cell action rates, used to simulate the occurrence of major crypt phenomena in healthy colon crypt, (Subsection 7.2.2). Specifically, $Rate^{Stem}_{Apoptosis}$ was considered equal to 0.01, $Rate^{Stem}_{SymDiv}$ to 0.003, (during the 'bottleneck effect' phase) and $Rate^{Stem}_{SymDiv}$ to 0.0115 per week, (during 'crypt fission' phase). Both symmetrical and asymmetrical stem cell divisions are rare events; however, the latter was reported to have a higher occurrence rate than the former. In consequence, asymmetrical stem cell division was assigned to a higher value than those considered for symmetrical cell division rate, (e.g. $Rate^{Stem}_{AsymDiv}$ = 0.02 per week, which indicates that around 2% from stem cell population perform asymmetrical division every week). While differentiated cell life-cycle was approximated to several hours in intestinal crypt, the lifetime for Paneth cells was reported to be around 20 - 30 days, (Section 2.4). Based on these, the rate of progenitor differentiation into Paneth cells, was approximated with a small value, e.g. 0.05, i.e. $Rate^{Prog}_{Paneth}$ = 0.05 per day. In addition, given that three possible actions were defined for progenitors, a value of 0.33, was considered for $Rate^{Prog}_{FullDiff}$. Thus, the progenitor division rate could be

Table 7.4: Input parameter values used in LogisticCrypt for the intestinal crypt

Parameter name	**Parameter description**
Cell action rates	$Rate^{Stem}_{SymDiv}$ = 0.003 per week; (value used for the symmetrical stem cell division during the 'bottleneck effect' phase). $Rate^{Stem}_{SymDiv}$ = 0.0115 per week; (value used for the symmetrical stem cell division during the 'crypt fission' phase). $Rate^{Stem}_{Apoptosis}$ = 0.01 per week; $Rate^{Stem}_{AsymDiv}$ = 0.02 per week; $Rate^{Prog}_{Div}$ = 0.38 per day; $Rate^{Prog}_{Paneth}$ = 0.05 per day;
Carcinogen influence	$\delta 2 \in \{0.05, 0.10, .. 0.50\}$.

calculated using expression (7.31) and approximated to 0.38, i.e. $Rate^{Prog}_{Div}$ = 0.38 per day. Finally, 'the carcinogen impact on cell action rates' parameter, $\delta 2$, was integrated in a sensitivity analysis, with respect to tumour growth occurrence, i.e. changes of cell population size in crypt systems over time.

7.5.1 'Crypt fission' and the 'bottleneck' effect

Comparison on the crypt fission time-point between the small intestine and colon was achieved by considering a group of eleven crypts for each tissue, where the rate of symmetrical stem cell division was increased from 0.015, (healthy crypt), to 0.039 per week, in the tenth abnormal crypt, (i.e. 'the most affected' one), by increments of $\sim 10\%$ between consecutive crypts. Similarly, for the 'bottleneck effect' simulations, the stem cell division rate, considered to be 0.003 in a 'healthy crypt', was decreased across a group of ten abnormal crypts for each tissue to around 0.001 per week (in the 10^{th} abnormal crypt), by $\sim 10\%$ between each consecutive crypt. Given that colon stem cell cycle was approximated to one week and the time-period of 'crypt fission' occurrence to 25 years, the simulation time was taken equal to 1300 iterations or stem cell cycles, (i.e. 25 years $\times$ 52 weeks/ year $\approx$ 1300 weeks). At the end of these simulations, both 'bottleneck effect' and 'fission' phenomena were recorded in normal and all abnormal crypts, with time to occurrence shown in Table 7.5 for both intestinal crypt groups.

Table 7.5: Time-steps (Weeks) for 'Crypt Fission' and 'Bottleneck Effect in the colon and small intestine crypt groups

Crypt Phe-nomenon	**Tissue Type**	**Cr 1** Healthy	**Cr 2** (Ab 1)	**Cr 3** (Ab 2)	**Cr 4** (Ab 3)	**Cr 5** (Ab 4)	**Cr 6** (Ab 5)	**Cr 7** (Ab 6)	**Cr 8** (Ab 7)	**Cr 9** (Ab 8)	**Cr 10** (Ab 9)	**Cr 11** (Ab 10)
Crypt Fission	Colon	1388	1068	852	697	582	492	421	362	314	275	241
	Small Intestine	710	547	436	357	298	252	215	186	161	140	123
Bottleneck Effect	Colon	487	467	451	437	424	414	406	398	391	386	381
	Small Intestine	334	228	220	212	208	203	199	195	192	189	186

Intra-tissue comparison The average time-period, measured to 'crypt fission' occurrence, (i.e. time taken to reach the maximum limit of 12 and 32 cells in the small intestine and colon crypts, respectively in each 11-crypt group), was found to be $\approx$ 311 weeks in the small intestine crypt group, (with *standard deviation*[5] $StDev^{SI}_{Fission} \approx 178$ weeks, over 1300 iterations), and $\approx$ 608 weeks in the colon crypt group, (with $StDev^{Colon}_{Fission} \approx 364$ weeks). The longest and the shortest times correspond to the first and the 11^{th} crypt, respectively, in each group. Specifically, these are approximated to 710 weeks and 123 weeks for the small intestine, and to 1388 weeks and 241 weeks for colon crypt. While the first crypt represents the *normal* system, defined by small changes of stem cell population size over time, (induced by low difference between stem cell symmetrical division and apoptosis rates, e.g. $\sim$ 0.005, calculated based on $Rate^{Stem}_{SymDiv1}$ = 0.0115 and $Rate^{Stem}_{Apoptosis}$ = 0.01 per week), the 11^{th} crypt mimics *the most 'affected'* system, represented by the highest difference between stem cell action rates from the whole crypt group, i.e. 0.029, (with $Rate^{Stem}_{SymDiv10}$ = 0.039 and $Rate^{Stem}_{Apoptosis}$ = 0.01 per week). Thus, the differences between stem cell action rates, (which increased from 0.005 in normal system to 0.029 in the 10^{th} abnormal system), reduced the time-period for 'crypt fission' manifestation in the most 'affected' crypt when compared with healthy crypts, with $\sim$ 1147 weeks in colon, (i.e. 1388

[5]Standard deviation is estimated based on crypt samples, using the following formula:

$$s = \sqrt{\frac{1}{n-1} \times \Sigma_{i=1}^{n}(x_i - \bar{x})^2} \tag{7.41}$$

where $\bar{x}$ = sample mean, N = sample size.

weeks - 241 weeks), and with $\sim$ 587 weeks in small intestine, (i.e. 710 weeks - 123 weeks), (Table 7.5).

Similarly, the average time-period to occurrence of the 'bottleneck effect' was measured in eleven crypts for both tissues, with values $\approx$ 206 weeks in the small intestine crypt group, (with $StDev^{SI}_{BottleNeck} \approx$ 17 weeks over 1300 iterations), and $\approx$ 422 weeks in the colon crypt group, (with $StDev^{Colon}_{BottleNeck} \approx$ 35 weeks). The extreme values of the time to the 'bottleneck effect' occurrence were around 334 weeks and 186 weeks for the small intestine crypt and 487 weeks and 381 weeks for the colon crypt, corresponding to the first crypt, (i.e. normal system), and to the 11^{th} crypt, (i.e the most 'affected' system), respectively. The most 'affected' crypt was characterised by the highest difference between stem cell apoptosis and symmetrical division rates, estimated to 0.009, (with $Rate^{Stem}_{SymDiv11}$ = 0.001 and $Rate^{Stem}_{Apoptosis}$ = 0.01 per week), when compared with other crypts; for instance, the difference between the same rates was around 0.008 in the 6^{th} crypt, (with $Rate^{Stem}_{SymDiv6}$ = 0.002 and $Rate^{Stem}_{Apoptosis}$ = 0.01 per week), and $\sim$ 0.007 (with $Rate^{Stem}_{SymDiv1}$ = 0.003 and $Rate^{Stem}_{Apoptosis}$ = 0.01 per week) in the first crypt.

In conclusion, alterations of stem cell population dynamics can increase major crypt phenomena incidence, leading to higher predisposition to tumour development in intestinal systems.

Inter-tissue comparison While similar alterations were considered for stem cell dynamics in both small intestine and colon crypt groups, results indicate that the time to occurrence of major crypt phenomena is more variable in the latter case. Specifically, the standard deviation[6] for the time-periods recorded for 'crypt fission' occurrence was $StDev^{Colon}_{Fission} \approx$ 364 weeks and $StDev^{SI}_{Fission} \approx$ 178 weeks, over 1300 iterations among the 11-crypt group in colon and small intestine, respectively. Similarly, the standard deviation for the 'bottleneck effect' incidence time was approximated to $StDev^{Colon}_{BottleNeck} \approx$ 35 weeks and $StDev^{SI}_{BottleNeck} \approx$ 17 weeks over 1300 iterations, among colon and small intestine crypt groups, respectively. Thus, given that variance of time for major crypt phenomena for-

[6]It refers to sample standard deviation estimated with formula (7.41).

mation was higher for colon that for small intestine crypt groups, the colon crypt group seems more sensitive to *deregulations* in the stem cell compartment than the small intestine crypt group. Similar findings on this difference between intestinal tissues have been reported from a study on epidermal growth factor in rats with regard to crypt fission and cell proliferation, [Berlanga-Acosta et al., 2001].

7.5.2 Tumour growth in systems with and without carcinogen influence

A second comparison between small intestine and colon crypts was performed to investigate the time taken to reach maximum crypt capacity, i.e. 520 and 4000 cells in small intestine and colon crypts, respectively, since this may signal abnormalities associated with tumour growth. Specifically, the variation in total cell number was recorded in two systems, namely *S1* and *S2*, represented by groups of eleven crypts for each tissue, where different external influences were applied to cell action rates, (as follows):

1. in S1, the progenitor cell division rate was increased by 5% between successive crypts, in order to facilitate cell proliferation, while other cell action rates were maintained at initial values;

2. in S2, "carcinogen influence" was considered and every cell action rate was changed according to specifications (subsection 7.3.2), with $\delta 2 \in \{0.05, 0.10, .. 0.50\}$.

At the end of the simulation, the total cell number reached the maximum limit in all intestinal crypts in both systems, with time-periods for each crypt shown in Table 7.6

The average time taken for the S1 group was $\approx$ 318 days in small intestine (with $StDev^{SI}_{ProgDereg} \approx$ 22 days), and $\approx$ 1495 days in colon crypts, ($StDev^{Colon}_{ProgDereg} \approx$ 20 days). In the system with carcinogen influence applied, i.e. S2, the average time taken for cell number extension was $\approx$ 70 days in small intestine and $\approx$ 116 days in colon, (with $StDev^{SI}_{Carcinogen} \approx$ 40 days and $StDev^{Colon}_{Carcinogen} \approx$ 68 days, respectively). Therefore, the average time taken for doubling of the initial cell number was around thirteen times shorter in the colon crypt group subject to carcinogen influence, S2, (with changes in all cell division and apoptosis rates), in comparison to the S1 group, (with only progenitor cell division

Table 7.6: Time-steps (Days) to reach Maximum Crypt Capacity in systems with and without Carcinogen influence

Group type	**Cr 1**	**Cr 2**	**Cr 3**	**Cr 4**	**Cr 5**	**Cr 6**	**Cr 7**	**Cr 8**	**Cr 9**	**Cr 10**
Colon (S1)	1543	1514	1502	1494	1490	1487	1485	1482	1481	1478
Carcinogen Colon (S2)	279	181	138	113	97	86	77	71	65	61
Small Intestine (S1)	372	337	324	316	308	308	306	304	303	302
Carcinogen Small Intestine (S2)	166	109	83	69	52	52	47	44	40	38

rate affected). Similarly, the total cell number in the small intestine crypt group increased faster in S2 than in S1, where the ratio between these systems was ≈ 4.50. Although similar cell action rate alterations were considered for both colon and small intestine crypt groups, it appears that the impact on crypt dynamics was considerably higher for the colon than for the small intestine in terms of tumour development. These results support the conclusion stated in the previous test case, which suggests that the small intestine is less sensitive than the colon with respect to cell cycle changes over time. In addition, results are in agreement with an earlier experimental study on cell proliferation in intestinal tissues in rats, described in Berlanga-Acosta et al. [2001].

7.6 Summary

LogisticCrypt is a discrete model on crypt structure, (represented by stem, progenitor and differentiated cells in colon, with addition of Paneth cell population in small intestine), which is monitored during cancer development. LogisticCrypt considered both *intra-* and *inter*specific competitions among cell populations within the intestinal crypt. The 'healthy crypt', which permitted low variations in cell numbers over time, was described by a set of values on stem and progenitor cell division rates in the colon, and, additionally, by differentiation rate into Paneth cells in the small intestine. Other cell actions rates, e.g. stem cell

apoptosis, were deduced from equilibrium equations and calculated based on relationships between all cell types within normal systems, (a step that facilitates reduction of the set of input parameters needed to describe the intestinal crypt).

A sensitivity analysis for a) stem cell division rate and b) 'carcinogen impact' on crypt dynamics was performed with respect to time-scales needed for i) 'crypt fission' and 'bottleneck effect', and ii) tumour growth occurrence. When comparing between tissues, shorter average time-intervals for both 'fission' and 'bottleneck effect' were recorded for the small intestine than for the colon crypt group. However, higher variance for these times of occurrence was recorded for the latter compared to the former, suggesting that the colon crypt is more sensitive to alterations in stem cell compartment than the small intestine crypt. Modifications on progenitor division rate were also considered as well as carcinogen influence for each tissue type. Results for carcinogen impact on cell proliferation (tumour growth), indicated that the extension of the cell population to crypt capacity was accelerated by a factor of 13 for the colon and 4.50 for the small intestine in carcinogen-influenced crypt groups compared to systems with increase in progenitor cell division rates only. This appears to indicate that the colon crypt is more sensitive to deregulations in cell action rates than the small intestine crypt, a conclusion in agreement with an earlier experimental study on crypt fission and cell proliferation in small intestine and colon crypts in rats, [Berlanga-Acosta et al., 2001]. However, LogisticCrypt results are inevitably tentative, given limited information on crypt phenomena in small intestine tissue. In LogisticCrypt, the cell action rate was considered to be equal for small intestine and colon crypts, although cell proliferation was reported to be more rapid in the former case, [Berlanga-Acosta et al., 2001]. Thus, further work on LogisticCrypt must include scenarios which consider differences in cell action rates between tissue types.

Finally, sensitivity analysis is certainly needed for the rest of unknown parameters, (e.g. maximum progenitor cell number, progenitor cell division rate), and is discussed in the next chapter.

Chapter 8

Concluding discussion and future work

8.1 Summary and conclusions

Epigenetics has emerged only recently as a fundamentally important area of biological and medical research that has implications for our understanding of human diseases including cancer, autoimmune and neuropsychiatric disorders. The motivation of this work is the considerable interest generated in investigation of genetic and epigenetic features in terms of how these contribute to disease, in this case colorectal cancer development. In this Thesis, a cross-scale computational model was presented for CRC dynamics, which aims to help in analysis of aberrant changes observed at *micro-molecular*, *cellular* and *tissue* levels in malignant intestinal systems. Novel components were required as follows. A network-based framework, i.e. the E-G Network Model, was built initially to study interdependencies between genetic and epigenetic signals recorded at different CRC stages. This was extended to include and assess the impact of different risk factors, (such as *ageing* and *gender*) on micro-molecular events in tumour pathways. Subsequently, an agent-based model, AgentCrypt, was developed to describe the dynamics of the human intestinal system, (within both colon and small intestine tissues). This was used to evaluate the aberrant methylation variation

induced by deregulations of the crypt mechanisms in malignant systems, influenced by various factors. In addition, several potential inhibitors were considered for methylation modifications in abnormal intestinal crypts and their effect evaluated with respect to tumour initiation. Finally, a logistic model, LogisticCrypt, was built to provide information on the intestinal crypt structure at specific time points and to facilitate the investigation of aberrant changes in cell number with respect to tumour initiation and progression.

Main findings

The role of *epigenetic events* in cancer development was outlined and the importance of understanding interdependencies between micro-molecular signals, (genetic and epigenetic), was demonstrated with a view to potential use as *targets* for *personalized cancer therapies*. Model types and major databases on cancer dynamics were reviewed and in some case, used as primary data resources. The background research indicated that a single model formulation has limited capability to investigate malignant changes, since large sets of parameters are needed to characterise human biological system. In consequence, *hybridised* models are required to represent the aberrant molecular mechanisms in tumour development, from cellular to whole-body scales. This motivated development of the component models reported in this Thesis.

The first computational component of the multi-scale model, i.e. *the genetic and epigenetic signals network*, integrated empirical data on conditional relationships between micro-molecular events observed at different CRC stages. Information on signalling pathways involved in CRC development, (e.g. Wntβ catenin, MAPK), was included from KEGG Pathway database, [Kanehisa et al., 2014], and other sources. The framework permitted analysis of the impact on *DNA methylation level* of interdependencies between aberrant modifications. The framework enabled introduction of the '*methylation cycle*' to denote the process of updating gene methylation level inside the network, based on five *key*-elements. In addition, an algorithm was developed for identification of the *most plausible pathways* in gene networks based on the conditional relationships between genetic and epigenetic events. A limitation at this model developmental stage, however, was lack of precise infor-

mation on *real time* methylation cycle and consequently, this was approximated by *stem cell cycle* for the human colon. The novelty of the framework formulation rests on proposing two criteria to evaluate the tumour progression inside a gene network: a) average network methylation level, (*network score*), and b) percentage of genes highly methylated due to the influence of abnormal genetic and epigenetic changes.

The genetic and epigenetic events framework was extended to enable '*customisation*' of malignant systems, (represented by the gene network), by incorporating information on patient characteristics, (e.g. ageing, gender), considered to have a crucial impact on cancer mechanisms. Information on a key-group of genes, (e.g. IGF2, ER, [Issa and Ahuja, 2000]), reported to be sensitive to *age-related methylation*, was integrated into the gene network. For these specific genes, a *coefficient* was inferred from reported studies to denote the relationship between ageing and methylation patterns and was included in the methylation level update algorithm. This enhancement offered potential for improved overall investigation of *colorectal tumour-genesis* through 'patient typing' in terms of malignancy development features.

A *comparative analysis* on *DNAm variation* over time was performed in normal and abnormal systems, for *both* colon and small intestine crypts using an agent-based model to mimic major influences in intestinal systems. Global methylation level decrease was marked for the colon compared with the small intestine, although similar alterations were induced in both tissues. These results suggest that the methylation patterns inside the colon crypt are more sensitive to deregulations in the proximal tissue than those recorded inside the small intestine. In addition, methylation level decrease was accelerated by *carcinogen* influence. This result is important in guiding further analysis on the effect of environmental factors on epigenetic mechanisms. Finally, a set of potential methylation *inhibitors* with different potencies over time was proposed and applied in abnormal intestinal systems and the impact on methylation level investigated.

A second *comparative analysis* between the colon and small intestine was performed to determine time taken to occurrence of recognised *crypt phenomena* in the intestinal crypt. Results were more variable for the colon crypt, compared with the small intestine crypt,

suggesting that the former is more sensitive to alterations in cell division, differentiation and apoptosis rates than the latter. Similar observations were concluded from an experimental study on crypt fission and cell proliferation in rat intestinal tissues, described in Berlanga-Acosta et al. [2001].

8.2 Future work

Connecting model-components in a multi-scale computational model for CRC dynamics

An obvious limitation of the AgentCrypt model is the absence of information on patient characteristics, (e.g. ageing, gender). Given the overall aim of a multi-scale computational model for CRC dynamics, (i.e. representing CRC specifics at micro-molecular, cellular and tissue layers), the final step is to connect AgentCrypt to the E-G Network Model. Specifically, there is a need to link information on (i) micro-molecular layer (gene relationship, genetic and epigenetic mechanisms), (ii) cellular/ tissue level (cell/ crypt interdependencies), (iii) personal characteristics (e.g. ageing, gender, heredity) and (iv) carcinogen influences (e.g. tobacco usage, chemical exposure), in order to analyse tumour pathways and to identify potential targets for epigenetic inhibitors that can be explored in cancer therapy. A link between the gene network and the agent-based model is feasible, as initialization and update of the methylation level status inside a cell is driven by AgentCrypt, (which relies on the strength of gene dependencies to set some of the input parameters). This E-G Network - AgentCrypt combined model can be extended to integrate information on inhibitors for methylation on specific key-genes from tumour pathways in order to analyse their effect at cellular/ tissue level. This approach may improve understanding of aberrant molecular events observed for genes such as APC, KRAS, which are known to have a crucial role in CRC initiation and progression, [Jones and Baylin, 2002]. In addition, given interest in studying the combined effect of two or more epigenetic inhibitors in solid tumours, [Hatzimichael and Crook, 2013], the E-G Network - AgentCrypt model can be extended to investigate the consequences of using epigenetic inhibitor combinations in the colon tissue

during cancer development.

Sensitivity analysis for the parameter sets integrated in the E-G Network, AgentCrypt & LogisticCrypt

Given the lack of precise biological information available on intestinal crypt systems during cancer development, extensive sensitivity analysis is necessary for the set of parameters included in the Colorectal Cancer Model. This analysis is necessary for further validation as parameter changes should lead to model outputs which can be marked for known patterns. In addition, parameter variation can indicate more sensitive factors in cancer development, providing new hypotheses to be tested.

Tests should integrate alternative values for the gene network parameters, including network size, simulation time, the d-value used for identification of d-plausible pathways, the coefficients for the genes that were i) identified in signalling pathways, or ii) reported to be sensitive to age-related methylation, i.e. SAM and SIG coefficients, (E-G Network). Some adjustments are likely also for the value-ranges considered for the relationships between epigenetic event level and cancer stages, incorporated in i) DNA methylation and histone modification initialization and update steps and ii) cancer progression decision methods, (Tables 4.3 - 5.2). Resources such as GEO database can be also consulted for gene promoter methylation level in CRC, (similar as for the global methylation level in intestinal cells that was used in the AgentCrypt for colon and small intestine systems). Moreover, the E-G Network model can be extended to include also average gene methylation level and to analyse global *hypo*methylation over time. Analogously, different value-ranges must be considered for the AG and TH_HM parameters, (Tables 5.3 - 5.4), which describe ageing and gender impact on micro-molecular events. The sensitivity analysis will be carried out systematically, for larger sample sizes, (e.g. 1000 synthetic patients), with focus on the assessment of cancer progression within gene networks.

In addition, sensitivity analysis should address the parameter set used for the intestinal crypt structure and dynamics. Specifically, i) the crypt intra- and interdependencies and their weights in updating DNAm level, (i.e. ICoef, CNI and DiffMethCoef); ii) carcinogen

contribution coefficients, (i.e. $\delta 1$); iii) the value - ranges for every cell type number, (e.g. MaxProgNo); iv) cell cycle duration, (e.g. for progenitor cells); as well as v) the cell rates of division, differentiation and apoptosis, (LogisticCrypt).

Parallel implementation for E-G Network and AgentCrypt

Parallel implementation is anticipated for extension of the E-G Network and AgentCrypt models to accommodate the complexity of intestinal crypt dynamics during cancer development. For example, the time for a single simulation, run using AgentCrypt for 550 iterations, in an 'i7-2600, 3.40 GHz, 8 GB RAM' system, was around 12 - 13 the minutes for small intestine and around 10 hours for the colon. However, the further developmental phase is not trivial, and in complex systems, with high communication overhead, the *speedup*, (Section 5.3), of such approach can be difficult to estimate. Several strategies already identified during model development target parallelisation for

i. the group of 'individuals', (where an 'individual' is mimicked by AgentCrypt using the *Intestinal Crypt Group*). Given that every individual is independent with respect to intestinal tissue dynamics from every other individual, every Intestinal Crypt Group can run on a distinct compute node and no communication is needed between nodes. The speedup in this case is almost linear given that the Intestinal Crypt Group are analysed concurrently.

ii. a single 'individual', (i.e. one Intestinal Crypt Group in AgentCrypt). The Intestinal Crypt Group can be split between a group of compute nodes, with every Crypt running on a single node. However, given 'inter-crypt influences' considered in AgentCrypt, communication between nodes does exist in this case and linear speedup is not anticipated.

iii. 'chunks of cells', (i.e. inside the Crypt in AgentCrypt). Given the parallelisation features, (illustrated in Figure 6.4), a Crypt can be divided in cell sets of same of similar sizes, with every group allocated to a distinct compute node. Slave compute nodes can calculate cell methylation level and decide on cell division, differentiation and apop-

tosis and once cell methylation level and actions are determined, the master node is informed and updates crypt structure. Limitations at this level are clearly given by communication between nodes and linear speedup is not possible. However, communication is reduced by using 'chunks of cells' instead of running every cell on distinct node. In addition, algorithm efficiency seems also influenced by crypt phenomena. While is a crypt with a large cell number, all allocated nodes run, (are 'busy'), and the overall job is divided, in a crypt with a small cell number, only a part of considered nodes work and the remainder is in 'waiting' state, causing decrease in algorithm efficiency. Thus, of high importance is choosing node number and cell group size.

iv. the gene network, (E-G Network Model). Parallelisation at this level involves gene network division into subnetworks with similar node and edge numbers, in order to homogenize time needed for updates among the set of subnetworks. However, communication between sub-networks is obviously required and speedup is again, not linear.

Parallelisation is available for the ageing/gender case study (Chapter 5). Other components are in the preliminary development stage, (e.g. for methylation level update in the AgentCrypt), but this work is ongoing.

Development of a graphical interface and visualisation

The code documentation file mentioned in Chapter 5 was provided in order to facilitate understanding of technical details on model implementation. The future plan is to develop a graphical interface for the Colorectal Cancer Model, in order to help non-computational scientists use the framework. Thus, input information such as *patient characteristics*, (e.g. ageing, gender), *tumour specifics*, (e.g. mutation of a specific gene), or *time-period* considered for prediction, (e.g. 3-5 years), could be easily imported through a graphical interface into the model, in order to be used for analysis. Additionally, a visual representation of the output data, (e.g. the way in which crypt dynamics change under carcinogen influence), would also be of value.

Other Inclusions / Extensions:

Inclusion of nutrition and physical activity influence in the Colorectal Cancer Model

Studies have shown that, while a diet containing more vegetables and fruits and a more active program of physical exercises can help reduce CRC risk, a high intake of red meat and alcohol and a limited amount of physical activity have been associated with increased risk, [Ryan-Harshman and Aldoori, 2007; Harriss et al., 2009]. In addition, the influence of nutrition with regard to epigenetic modifications was recently reported also, (reviewed in Parle-McDermott and Ozaki [2011] and references therein). Ideally, therefore, the influence of lifestyle characteristics should be taken into account as an extension of the model presented here. In terms of such an extension, the aim would be to determine a *quantitative coefficient* to represent the impact of these risk factors on gene methylation level, (based on data from reported statistical studies), and to link this information to the update step.

Inclusion of miRNAs

MiRNAs are involved in cell cycle phases and their deregulation has been observed for different cancer types, (Subsection 2.3.3). MiRNA potential for CRC diagnosis and prognosis has been recently reported, (reviewed in Luo et al. [2011]; Schetter et al. [2011]; Menéndez et al. [2013] and references therein). Therefore, the plan would be to integrate relevant data on miRNAs into the Colorectal Cancer Model, to link these changes to information on methylation level from the gene network and to analyse their influence at tissue level, (i.e. on crypt dynamics).

Extension to other types of cancer, (e.g. liver and lung cancer)

While colorectal cancer was the initial focus, the model presented here has been designed to be generalizable also to other types of cancer. Therefore, more general terms, such as '*carcinoma in situ*', have been used to denote colon cancer stages in the framework, instead of more disease-specific usage. Based on similar mechanisms observed in cancer development among different tissues at molecular or cell level, (e.g. common genes involved in tumour pathways, high cell division rate), a future long-term plan would be to extend the current framework to other types of *solid* cancer, such as *liver* and *lung*. The difficulty, as

usual, is parametrisation and this extension is non-trivial. The task would involve refining existing metrics from the Colorectal Cancer Model and integrating information on tissue-specific malignant phenomena, (such as the high impact of *tobacco* usage in lung cancer, [Hecht, 2003], or viral infection in liver cancer, (reviewed in Herceg and Vaissière [2011]; Dandri and Locarnini [2012] and references therein).

8.3 Final remarks

The Colorectal Cancer Model aims to investigate the tumour mechanisms at three main layers, namely the micro-molecular, cellular and tissue levels. The E-G Network Model can be used as a prototype-tool to aid prediction of CRC development based on information on micro-molecular interdependencies, and towards personalised medicine. AgentCrypt can be seen as proof of concept to be used in epigenetic epidemiology, to accommodate in *vivo* and in *vitro* experiments on tissue-specific methylation modifications. Finally, LogisticCrypt can be considered a prototype tool for investigating the dynamics of a multi-species system characterised by *intra-* and *inter*specific competition. Full scale-up, in computational terms, remains a challenge for the future.

Bibliography

Abel, T. and Zukin, R. S. (2008). Epigenetic targets of HDAC inhibition in neurodegenerative and psychiatric disorders. *Current Opinion in Pharmacology*, 8(1):57–64.

Adams, D., Altucci, L., Antonarakis, S. E., Ballesteros, J., Beck, S., Bird, A., Bock, C., Boehm, B., Campo, E., Caricasole, A., et al. (2012). BLUEPRINT to decode the epigenetic signature written in blood. *Nature Biotechnology*, 30(3):224–226.

Aguilar, C. A. and Craighead, H. G. (2013). Micro-and nanoscale devices for the investigation of epigenetics and chromatin dynamics. *Nature Nanotechnology*, 8(10):709–718.

Ahmed, D., Eide, P., Eilertsen, I., Danielsen, S., Eknæs, M., Hektoen, M., Lind, G., and Lothe, R. (2013). Epigenetic and genetic features of 24 colon cancer cell lines. *Oncogenesis*, 2(9):e71.

Al-Sohaily, S., Biankin, A., Leong, R., Kohonen-Corish, M., and Warusavitarne, J. (2012). Molecular pathways in colorectal cancer. *Journal of Gastroenterology and Hepatology*, 27(9):1423–1431.

Alegría-Torres, J. A., Baccarelli, A., and Bollati, V. (2011). Epigenetics and lifestyle. *Epigenomics*, 3(3):267–277.

Allis, C. D., Jenuwein, T., Reinberg, D., and Caparros, M.-L. (2007). *Epigenetics*. CSHL Press.

American Cancer Society (2013). American Cancer Society - Heredity and Cancer.

Available at: http://www.cancer.org/cancer/cancercauses/geneticsandcancer/heredity-and-cancer, [Accessed 07/05/2014].

Antonic, V., Stojadinovic, A., Kester, K. E., Weina, P. J., Brücher, B. L., Protic, M., Avital, I., and Izadjoo, M. (2013). Significance of infectious agents in colorectal cancer development. *Journal of Cancer*, 4(3):227 – 240.

Arnaud, C., Sebbagh, M., Nola, S., Audebert, S., Bidaut, G., Hermant, A., Gayet, O., Dusetti, N. J., Ollendorff, V., Santoni, M.-J., Borg, J., and Lécine, P. (2009). MCC, a new interacting protein for Scrib, is required for cell migration in epithelial cells. *FEBS Letters*, 583(14):2326–2332.

Ashworth, A. and Hudson, T. J. (2013). Genomics: Comparisons across cancers. *Nature*, 502(7471):306–307.

Azad, N., Zahnow, C. A., Rudin, C. M., and Baylin, S. B. (2013). The future of epigenetic therapy in solid tumours - lessons from the past. *Nature Reviews Clinical Oncology*, 10(5):256–266.

Bach, S. P., Renehan, A. G., and Potten, C. S. (2000). Stem cells: the intestinal stem cell as a paradigm. *Carcinogenesis*, 21(3):469–476.

Bandini, S., Manzoni, S., and Vizzari, G. (2009). Agent based modeling and simulation: An informatics perspective. *Journal of Artificial Societies and Social Simulation*, 12(4):4.

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

Barat, A., Ruskin, H. J., and Crane, M. (2006). Probabilistic models for drug dissolution. Part 1. Review of Monte Carlo and stochastic cellular automata approaches. *Simulation Modelling Practice and Theory*, 14(7):843–856.

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., , Yefanov, A., Lee, H., Zhang,

N., Robertson, C., Serova, N., Davis, S., and Soboleva, A. (2013). NCBI GEO: archive for functional genomics data sets - update. *Nucleic Acids Research*, 41(D1):D991–D995.

Baumgart, D. C. and Sandborn, W. J. (2012). Crohn's disease. *The Lancet*, 380(9853):1590–1605.

Baylin, S. B. and Jones, P. A. (2011). A decade of exploring the cancer epigenome–biological and translational implications. *Nature Reviews Cancer*, 11(10):726–734.

Berdasco, M. and Esteller, M. (2013). Genetic syndromes caused by mutations in epigenetic genes. *Human Genetics*, 132(4):359–383.

Berlanga-Acosta, J., Playford, R. J., Mandir, N., and Goodlad, R. A. (2001). Gastrointestinal cell proliferation and crypt fission are separate but complementary means of increasing tissue mass following infusion of epidermal growth factor in rats. *Gut*, 48(6):803–807.

Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M. A., Beaudet, A. L., Ecker, J. R., , Farnham, P., Hirst, M., Lander, E., Mikkelsen, T., and Thomson, J. (2010). The NIH roadmap epigenomics mapping consortium. *Nature Biotechnology*, 28(10):1045–1048.

Bestor, T. H. (2000). The DNA methyltransferases of mammals. *Human Molecular Genetics*, 9(16):2395–2402.

Bezbradica, M., Ruskin, H. J., and Crane, M. (2014). Probabilistic Pharmaceutical Modelling: A Comparison Between Synchronous and Asynchronous Cellular Automata. In *Parallel Processing and Applied Mathematics*, pages 699–710. Springer.

Bird, A. (2002). DNA methylation patterns and epigenetic memory. *Genes & Development*, 16(1):6–21.

Bjornsson, H. T., Fallin, M. D., and Feinberg, A. P. (2004). An integrated epigenetic and genetic approach to common human disease. *TRENDS in Genetics*, 20(8):350–358.

Bock, C. (2012). Analysing and interpreting DNA methylation data. *Nature Reviews Genetics*, 13(10):705–719.

Bock, C. and Lengauer, T. (2008). Computational epigenetics. *Bioinformatics*, 24(1):1–10.

Bollag, G., Tsai, J., Zhang, J., Zhang, C., Ibrahim, P., Nolop, K., and Hirth, P. (2012). Vemurafenib: the first drug approved for braf-mutant cancer. *Nature Reviews Drug Discovery*, 11(11):873–886.

Boman, B. M., Fields, J. Z., Bonham-Carter, O., and Runquist, O. A. (2001). Computer modeling implicates stem cell overproduction in colon cancer initiation. *Cancer Research*, 61(23):8408–8411.

Boman, B. M. and Wicha, M. S. (2008). Cancer stem cells: a step toward the cure. *Journal of Clinical Oncology*, 26(17):2795–2799.

BowelScreen (2014). BowelScreen for Health professionals - Lung cancer risks and causes. Available at: http://www.bowelscreen.ie/healthprofessional, [Accessed 18/09/2014].

Bragazzi, N. L. (2013). From P0 to P6 medicine, a model of highly participatory, narrative, interactive, and augmented medicine: some considerations on Salvatore Iaconesis clinical story. *Patient Preference and Adherence*, 7:353 – 359.

Brenner, H., Hoffmeister, M., Arndt, V., and Haug, U. (2007). Gender differences in colorectal cancer: implications for age at initiation of screening. *British Journal of Cancer*, 96(5):828–831.

Brittan, M. and Wright, N. A. (2002). Gastrointestinal stem cells. *The Journal of Pathology*, 197(4):492–509.

Brittan, M. and Wright, N. A. (2004). Stem cell in gastrointestinal structure and neoplastic development. *Gut*, 53(6):899–910.

Browne, F., Zheng, H., Wang, H., and Azuaje, F. (2009). An integrative bayesian approach

to supporting the prediction of protein-protein interactions: A case study in human heart failure. *World Academy of Science, Engineering and Technology*, 53:457–463.

Burnside, E. S., Rubin, D. L., Fine, J. P., Shachter, R. D., Sisney, G. A., and Leung, W. K. (2006). Bayesian network to predict breast cancer risk of mammographic microcalcifications and reduce number of benign biopsy results: initial experience 1. *Radiology*, 240(3):666–673.

Burt, R. W., Barthel, J. S., Dunn, K. B., David, D. S., Drelichman, E., Ford, J. M., Giardiello, F. M., Gruber, S. B., Halverson, A. L., Hamilton, S. R., et al. (2010). Colorectal cancer screening. *Journal of the National Comprehensive Cancer Network*, 8(1):8–61.

Campisi, J. (2003). Cancer and ageing: rival demons? *Nature Reviews Cancer*, 3(5):339–349.

Canavan, C., Abrams, K., and Mayberry, J. (2006). Meta-analysis: colorectal and small bowel cancer risk in patients with Crohn's disease. *Alimentary Pharmacology & Therapeutics*, 23(8):1097–1104.

Cancer Genome Atlas Research Network (2011). Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609–615.

Cancer Genome Atlas Research Network (2012a). Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, 489(7417):519 – 525.

Cancer Genome Atlas Research Network (2012b). Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407):330 – 337.

Cancer Genome Atlas Research Network (2012c). Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61 – 70.

Cancer Genome Atlas Research Network (2013a). Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature*, 499(7456):43–49.

Cancer Genome Atlas Research Network (2013b). Genomic and Epigenomic Landscapes of Adult *De Novo* Acute Myeloid Leukemia. *The New England Journal of Medicine*, 368(22):2059–2074.

Cancer Genome Atlas Research Network (2013c). Integrated genomic characterization of endometrial carcinoma. *Nature*, 497(7447):67–73.

Cancer Research UK (2014a). Cancer Research UK - Bowel cancer incidence statistics. Available at: http://www.cancerresearchuk.org/cancer-info/cancerstats/incidence/age/, [Accessed 07/05/2014].

Cancer Research UK (2014b). Cancer Research UK - Bowel cancer statistics. Available at: http://www.cancerresearchuk.org/cancer-info/cancerstats/types/bowel/incidence/, [Accessed 19/06/2014].

Cancer Research UK (2014c). Cancer Research UK - Definite breast cancer risks. Available at: http://www.cancerresearchuk.org/about-cancer/type/breast-cancer/about/risks/definite-breast-cancer-risks, [Accessed 18/09/2014].

Cancer Research UK (2014d). Cancer Research UK - Diet causing cancer. Available at: http://www.cancerresearchuk.org/cancer-help/about-cancer/causes-symptoms/causes/diet-causing-cancer, [Accessed 07/05/2014].

Cancer Research UK (2014e). Cancer Research UK - Lung cancer risks and causes. Available at: http://www.cancerresearchuk.org/about-cancer/type/lung-cancer/about/lung-cancer-risks-and-causes, [Accessed 18/09/2014].

Cancer Research UK (2014f). Cancer Research UK - Smoking and cancer. Available at: http://www.cancerresearchuk.org/cancer-info/healthyliving/smokingandtobacco/, [Accessed 19/06/2014].

Cancer Research UK (2014g). Cancer Research UK - Your environment and cancer. Available at: http://www.cancerresearchuk.org/cancer-help/about-cancer/causes-symptoms/causes/your-environment-and-cancer, [Accessed 07/05/2014].

Carey, N., Marques, C. J., and Reik, W. (2011). DNA demethylases: a new epigenetic frontier in drug discovery. *Drug Discovery Today*, 16(15):683–690.

Carrillo-Infante, C., Abbadessa, G., Bagella, L., and Giordano, A. (2007). Viral infections as a cause of cancer (Review). *International Journal of Oncology*, 30(6):1521.

Carthew, R. W. and Sontheimer, E. J. (2009). Origins and mechanisms of miRNAs and siRNAs. *Cell*, 136(4):642–655.

Cedar, H. and Bergman, Y. (2009). Linking DNA methylation and histone modification: patterns and paradigms. *Nature Reviews Genetics*, 10(5):295–304.

Cervelle, J., Formenti, E., and Masson, B. (2007). From sandpiles to sand automata. *Theoretical Computer Science*, 381(1):1–28.

Chakrabarti, A., Verbridge, S., Stroock, A. D., Fischbach, C., and Varner, J. D. (2012). Multiscale models of breast cancer progression. *Annals of Biomedical Engineering*, 40(11):2488–2500.

Chen, R., Rabinovitch, P. S., Crispin, D. A., Emond, M. J., Bronner, M. P., and Brentnall, T. A. (2005). The initiation of colon cancer in a chronic inflammatory setting. *Carcinogenesis*, 26(9):1513–1519.

Cho, N.-Y., Choi, M., Kim, B.-H., Cho, Y.-M., Moon, K. C., and Kang, G. H. (2006). Braf and kras mutations in prostatic adenocarcinoma. *International Journal of Cancer*, 119(8):1858–1862.

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., 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.

Clevers, H. (2006). Wnt/β-catenin signaling in development and disease. *Cell*, 127(3):469–480.

Clevers, H. (2011). The cancer stem cell: premises, promises and challenges. *Nature Medicine*, pages 313–319.

Clevers, H. C. and Bevins, C. L. (2013). Paneth cells: maestros of the small intestinal crypts. *Annual Review of Physiology*, 75:289–311.

CodePlex (2014). Visual Leak Detector for Visual C++ 2008/2010/2012. [Free, open-source system for memory leaks detection for Visual C++]. Available at: http://vld.codeplex.com/, [Accessed 07/05/2014].

Coleman, W. B. and Rivenbark, A. G. (2006). Quantitative DNA methylation analysis: the promise of high-throughput epigenomic diagnostic testing in human neoplastic disease. *The Journal of Molecular Diagnostics: JMD*, 8(2):152.

Collins, F. S., Green, E. D., Guttmacher, A. E., and Guyer, M. S. (2003). A vision for the future of genomics research. *Nature*, 422(6934):835–847.

Conway, J. (1970). The game of life. *Scientific American*, 223(4):4.

Coombes, S. (2009). The Geometry and Pigmentation of Seashells. *Techn. Ber. Department of Mathematical Sciences*, pages 1 – 4.

Cramer, J. S. (2002). The origins of logistic regression. *Tinbergen Institute Working Paper*.

Crick, F. (1970). Central dogma of molecular biology. *Nature*, 227(5258):561–563.

Cruz-Ramírez, N., Acosta-Mesa, H. G., Carrillo-Calvet, H., Alonso Nava-Fernández, L., and Barrientos-Martínez, R. E. (2007). Diagnosis of breast cancer using Bayesian networks: A case study. *Computers in Biology and Medicine*, 37(11):1553–1564.

Cui, X.-J., Li, H., and Liu, G.-Q. (2011). Combinatorial patterns of histone modifications in Saccharomyces cerevisiae. *Yeast*, 28(9):683–691.

Cunningham, D., Atkin, W., Lenz, H.-J., Lynch, H. T., Minsky, B., Nordlinger, B., and Starling, N. (2010). Colorectal cancer. *The Lancet*, 375(9719):1030 – 1047.

Dammann, R., Schagdarsurengin, U., Strunnikova, M., Rastetter, M., Seidel, C., Liu, L., Tommasi, S., and Pfeifer, G. (2003). Epigenetic inactivation of the Ras-association domain family 1 (RASSF1A) gene and its function in human carcinogenesis. *Histology and Histopathology*, 18(2):665–677.

Dandri, M. and Locarnini, S. (2012). New insight in the pathobiology of hepatitis B virus infection. *Gut*, 61(Suppl 1):i6–i17.

Davis, C. D. and Uthus, E. O. (2004). DNA methylation, cancer susceptibility, and nutrient interactions. *Experimental Biology and Medicine*, 229(10):988–995.

de Lau, W., Barker, N., and Clevers, H. (2006). Wnt signaling in the normal intestine and colorectal cancer. *Frontiers in Bioscience: A virtual library of medicine*, 12:471–491.

De Matteis, G., Graudenzi, A., and Antoniotti, M. (2013). A review of spatial computational models for multi-cellular systems, with regard to intestinal crypts and colorectal cancer development. *Journal of Mathematical Biology*, 66(7):1409–1462.

Deisboeck, T. S., Wang, Z., Macklin, P., and Cristini, V. (2011). Multiscale Cancer Modeling. *Annual Review of Biomedical Engineering*, 13(1):1 – 30.

Derynck, R., Akhurst, R. J., and Balmain, A. (2001). TGF-β signaling in tumor suppression and cancer progression. *Nature Genetics*, 29(2):117–129.

Desper, R., Jiang, F., Kallioniemi, O.-P., Moch, H., Papadimitriou, C. H., and Schäffer, A. A. (1999). Inferring tree models for oncogenesis from comparative genome hybridization data. *Journal of Computational Biology*, 6(1):37–51.

Dick, K. J., Nelson, C. P., Tsaprouni, L., Sandling, J. K., Aïssi, D., Wahl, S., Meduri, E., Morange, P.-E., Gagnon, F., Grallert, H., et al. (2014). DNA methylation and body-mass index: a genome-wide analysis. *The Lancet*, 383(9933):1990 – 1998.

Down, T. A., Rakyan, V. K., Turner, D. J., Flicek, P., Li, H., Kulesha, E., Graef, S., Johnson, N., Herrero, J., Tomazou, E. M., et al. (2008). A Bayesian deconvolution strat-

egy for immunoprecipitation-based DNA methylome analysis. *Nature Biotechnology*, 26(7):779–785.

Duan, W., Qiu, X., Cao, Z., Zheng, X., and Cui, K. (2013). Heterogeneous and Stochastic Agent-Based Models for Analyzing Infectious Diseases' Super Spreaders. *IEEE Intelligent Systems*, 28(4):18–25.

Dworkin, A. M., Huang, T. H.-M., and Toland, A. E. (2009). Epigenetic alterations in the breast: Implications for breast cancer detection, prognosis and treatment. In *Seminars in cancer biology*, volume 19, pages 165–171. Elsevier.

Eckhardt, F., Beck, S., Gut, I. G., and Berlin, K. (2004). Future potential of the human epigenome project. *Expert Review of Molecular Diagnostics*, 4(5):609 – 618.

Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V. K., Attwood, J., Burger, M., Burton, J., Cox, T. V., Davies, R., Down, T. A., et al. (2006). DNA methylation profiling of human chromosomes 6, 20 and 22. *Nature Genetics*, 38(12):1378–1385.

Eisenhoffer, G. T., Loftus, P. D., Yoshigi, M., Otsuna, H., Chien, C.-B., Morcos, P. A., and Rosenblatt, J. (2012). Crowding induces live cell extrusion to maintain homeostatic cell numbers in epithelia. *Nature*, 484(7395):546–549.

Elsner, M. (2011). Oncotrack tests drugs in virtual people. *Nature Biotechnology*, 29(5):378–378.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74.

Esteller, M. (2006). The necessity of a human epigenome project. *Carcinogenesis*, 27(6):1121–1125.

Esteller, M. (2007). Cancer epigenomics: DNA methylomes and histone-modification maps. *Nature Reviews Genetics*, 8(4):286–298.

Ferlay, J., Soerjomataram, I., Ervik, M., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D., Forman, D., and Bray, F. (2014). GLOBOCAN 2012 v1. 0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11. Lyon, France: International Agency for Research on Cancer; 2013. Available at: http://globocan.iarc.fr, [Accessed on 28/05/2014].

Figueiredo, I. N., Romanazzi, G., Leal, C., and Engquist, B. (2013). A multiscale model for Aberrant Crypt Foci. *Procedia Computer Science*, 18:1026–1035.

Fingerman, I. M., McDaniel, L., Zhang, X., Ratzat, W., Hassan, T., Jiang, Z., Cohen, R. F., and Schuler, G. D. (2011). NCBI Epigenomics: a new public resource for exploring epigenomic data sets. *Nucleic Acids Research*, 39(suppl 1):D908–D912.

Fletcher, A. G., Breward, C. J., and Jonathan Chapman, S. (2012). Mathematical modeling of monoclonal conversion in the colonic crypt. *Journal of Theoretical Biology*, 300:118–133.

Fodde, R., Smits, R., and Clevers, H. (2001). APC, signal transduction and genetic instability in colorectal cancer. *Nature Reviews Cancer*, 1(1):55–67.

Foryś, U. (2009). Multi-dimensional Lotka–Volterra systems for carcinogenesis mutations. *Mathematical Methods in the Applied Sciences*, 32(17):2287–2308.

Fraga, M. F., Agrelo, R., and Esteller, M. (2007). Cross-Talk between Aging and Cancer. *Annals of the New York Academy of Sciences*, 1100(1):60–74.

Fraga, M. F. and Esteller, M. (2007). Epigenetics and aging: the targets and the marks. *Trends in Genetics*, 23(8):413–418.

Frank, S. A. (2007). *Dynamics of cancer: incidence, inheritance, and evolution*. Princeton University Press.

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, 7(3-4):601–620.

Fritsch, H. and Kühnel, W. (2008). *Color Atlas of Human Anatomy: Internal organs. Volume 2*, volume 2. Thieme.

Fukuyama, R., Niculaita, R., Ng, K. P., Obusez, E., Sanchez, J., Kalady, M., Aung, P. P., Casey, G., and Sizemore, N. (2008). Mutated in colorectal cancer, a putative tumor suppressor for serrated colorectal cancer, selectively represses β-catenin-dependent transcription. *Oncogene*, 27(46):6044–6055.

Ganguly, N., Sikdar, B. K., Deutsch, A., Canright, G., and Chaudhuri, P. P. (2003). A survey on Cellular Automata. *Technical Report, Centre for High Performance Computing, Dresden University of Technology*.

Gardiner-Garden, M. and Frommer, M. (1987). CpG islands in vertebrate genomes. *Journal of Molecular Biology*, 196(2):261–282.

Garijo, N., Manzano, R., Osta, R., and Perez, M. (2012). Stochastic cellular automata model of cell migration, proliferation and differentiation: Validation with in vitro cultures of muscle satellite cells. *Journal of Theoretical Biology*, 314:1–9.

Gerstung, M., Baudis, M., Moch, H., and Beerenwinkel, N. (2009). Quantifying cancer progression with conjunctive Bayesian networks. *Bioinformatics*, 25(21):2809–2815.

Ghildiyal, M. and Zamore, P. D. (2009). Small silencing RNAs: an expanding universe. *Nature Reviews Genetics*, 10(2):94–108.

Giovannucci, E. (2001). An updated review of the epidemiological evidence that cigarette smoking increases risk of colorectal cancer. *Cancer Epidemiology Biomarkers & Prevention*, 10(7):725–731.

Goecks, J., Nekrutenko, A., Taylor, J., and The Galaxy Team (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, 11(8):R86 – R89.

Grady, W. M. (2004). Genomic instability and colon cancer. *Cancer and Metastasis Reviews*, 23(1-2):11–27.

Grady, W. M. and Markowitz, S. D. (2002). Genetic and epigenetic alterations in colon cancer. *Annual Review of Genomics and Human Genetics*, 3(1):101–128.

Graham, T. A., Humphries, A., Sanders, T., Rodriguez-Justo, M., Tadrous, P. J., Preston, S. L., Novelli, M. R., Leedham, S. J., McDonald, S. A., and Wright, N. A. (2011). Use of methylation patterns to determine expansion of stem cell clones in human colon tissue. *Gastroenterology*, 140(4):1241–1250.

Half, E., Bercovich, D., and Rozen, P. (2009). Familial adenomatous polyposis. *Orphanet Journal of Rare Diseases*, 4(1):22 – 45.

Harrison, A. and Parle-McDermott, A. (2011). DNA methylation: a timeline of methods and applications. *Frontiers in Genetics*, 2:74–87.

Harriss, D., Atkinson, G., Batterham, A., George, K., Tim Cable, N., Reilly, T., Haboubi, N., and Renehan, A. G. (2009). Lifestyle factors and colorectal cancer risk (2): a systematic review and meta-analysis of associations with leisure-time physical activity. *Colorectal Disease*, 11(7):689–701.

Hartemink, A. J., Gifford, D. K., Jaakkola, T. S., and Young, R. A. (2002). Combining location and expression data for principled discovery of genetic regulatory network models. In *Pacific symposium on biocomputing*, volume 7, pages 437–449.

Hatzimichael, E. and Crook, T. (2013). Cancer epigenetics: new therapies and new challenges. *Journal of Drug Delivery*, 2013:1–9.

Hecht, S. S. (2003). Tobacco carcinogens, their biomarkers and tobacco-induced cancer. *Nature Reviews Cancer*, 3(10):733–744.

Heckerman, D. (1998). *A tutorial on learning with Bayesian networks*. Springer.

Heckerman, D., Mamdani, A., and Wellman, M. P. (1995). Real-world applications of Bayesian networks. *Communications of the ACM*, 38(3):24–26.

Hemminki, K., Zhang, H., and Czene, K. (2003). Familial and attributable risks in cutaneous melanoma: effects of proband and age. *Journal of Investigative Dermatology*, 120(2):217–223.

Herceg, Z. and Vaissière, T. (2011). Epigenetic mechanisms and cancer: an interface between the environment and the genome. *Epigenetics*, 6(7):804–819.

Herman, J. G. and Baylin, S. B. (2003). Gene silencing in cancer in association with promoter hypermethylation. *New England Journal of Medicine*, 349(21):2042–2054.

Højsgaard, S. (2012). Graphical independence networks with the gRain package for R. Available at: http://www.jstatsoft.org/v46/i10/, [Accessed 19/06/2014].

Hou, L., Zhang, X., Wang, D., and Baccarelli, A. (2012). Environmental chemical exposures and human epigenetics. *International Journal of Epidemiology*, 41(1):79–105.

Hudson, T. J., Anderson, W., Aretz, A., Barker, A. D., Bell, C., Bernabé, R. R., Bhan, M., Calvo, F., Eerola, I., Gerhard, D. S., et al. (2010). International network of cancer genome projects. *Nature*, 464(7291):993–998.

Hughes, L. A., Khalid-de Bakker, C. A., Smits, K. M., van den Brandt, P. A., Jonkers, D., Ahuja, N., Herman, J. G., Weijenberg, M. P., and van Engeland, M. (2012). The CpG island methylator phenotype in colorectal cancer: progress and problems. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, 1825(1):77–85.

Hui, C. (2006). Carrying capacity, population equilibrium, and environment's maximal load. *Ecological Modelling*, 192(1):317–320.

Humphries, A. and Wright, N. A. (2008). Colonic crypt organization and tumorigenesis. *Nature Reviews Cancer*, 8(6):415–424.

Hutchinson, L. (2011). Personalized cancer medicine: era of promise and progress. *Nature Reviews Clinical Oncology*, 8(3):121–121.

Hwang, M., Garbey, M., Berceli, S. A., and Tran-Son-Tay, R. (2009). Rule-based simulation of multi-cellular biological systemsa review of modeling techniques. *Cellular and Molecular Bioengineering*, 2(3):285–294.

Ikegami, K., Ohgane, J., Tanaka, S., Yagi, S., and Shiota, K. (2009). Interplay between DNA methylation, histone modification and chromatin remodeling in stem cells and during development. *International Journal of Developmental Biology*, 53(2):203 – 214.

Irigaray, P., Newby, J., Clapp, R., Hardell, L., Howard, V., Montagnier, L., Epstein, S., and Belpomme, D. (2007). Lifestyle-related factors and environmental agents causing cancer: an overview. *Biomedicine & Pharmacotherapy*, 61(10):640–658.

Irish Cancer Society (2014). Irish Cancer Society - Bowel (colon and rectum) cancer. Available at: http://www.cancer.ie/cancer-information/bowel-colon-rectum-cancer, [Accessed 18/09/2014].

Issa, J. P. and Ahuja, N. (2000). Aging, methylation and cancer. *Histology and Histopathology*, 15(3):835–842.

Issa, J.-P. J., Ottaviano, Y. L., Celano, P., Hamilton, S. R., Davidson, N. E., and Baylin, S. B. (1994). Methylation of the oestrogen receptor CpG island links ageing and neoplasia in human colon. *Nature Genetics*, 7(4):536–540.

Ito, S., Shen, L., Dai, Q., Wu, S. C., Collins, L. B., Swenberg, J. A., He, C., and Zhang, Y. (2011). Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science*, 333(6047):1300–1303.

Ito, T. (2007). Role of histone modification in chromatin dynamics. *Journal of Biochemistry*, 141(5):609–614.

Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S., Emili, A., Snyder, M., Greenblatt, J. F., and Gerstein, M. (2003). A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, 302(5644):449–453.

Jenuwein, T. and Allis, C. D. (2001). Translating the histone code. *Science*, 293(5532):1074–1080.

Jin, G., Ramanathan, V., Quante, M., Baik, G. H., Yang, X., Wang, S. S., Tu, S., Gordon, S. A., Pritchard, D. M., Varro, A., Shulkes, A., and Wang, T. C. (2009). Inactivating cholecystokinin-2 receptor inhibits progastrin-dependent colonic crypt fission, proliferation, and colorectal cancer in mice. *The Journal of Clinical Investigation*, 119(9):2691–2701.

Johannes, F., Porcher, E., Teixeira, F. K., Saliba-Colombani, V., Simon, M., Agier, N., Bulski, A., Albuisson, J., Heredia, F., and Audigier, P. (2009). Assessing the impact of transgenerational epigenetic variation on complex traits. *PLoS Genetics*, 5(6):e1000530.

Johnson, D., McKeever, S., Stamatakos, G., Dionysiou, D., Graf, N., Sakkalis, V., Marias, K., Wang, Z., and Deisboeck, T. S. (2013). Dealing with diversity in computational cancer modeling. *Cancer Informatics*, 12:115 – 124.

Johnston, M. D., Edwards, C. M., Bodmer, W. F., Maini, P. K., and Chapman, S. J. (2007). Mathematical modeling of cell population dynamics in the colonic crypt and in colorectal cancer. *Proceedings of the National Academy of Sciences*, 104(10):4008–4013.

Jones, P. A., Archer, T. K., Baylin, S. B., Beck, S., Berger, S., Bernstein, B. E., Carpten, J. D., Clark, S. J., Costello, J. F., Doerge, R. W., et al. (2008). Moving AHEAD with an international human epigenome project. *Nature*, 454(7205):711–715.

Jones, P. A. and Baylin, S. B. (2002). The fundamental role of epigenetic events in cancer. *Nature Reviews Genetics*, 3(6):415–428.

Jones, P. A. and Martienssen, R. (2005). A blueprint for a human epigenome project: the AACR human epigenome workshop. *Cancer Research*, 65(24):11241–11246.

Jung, M., Peterson, H., Chavez, L., Kahlem, P., Lehrach, H., Vilo, J., and Adjaye, J. (2010). A data integration approach to mapping OCT4 gene regulatory networks operative in embryonic stem cells and embryonal carcinoma cells. *PLoS One*, 5(5):e10709.

Kaaij, L., van de Wetering, M., Fang, F., Decato, B., Molaro, A., van de Werken, H. J., van Es, J. H., Schuijers, J., de Wit, E., de Laat, W., , Hannon, G., Clevers, H., Smith, A., and Ketting, R. (2013). DNA methylation dynamics during intestinal stem cell differentiation reveals enhancers driving gene expression in the villus. *Genome Biology*, 14:R50 – R65.

Kamel, N., Compton, C., Middelveld, R., Higenbottam, T., and Dahlen, S. (2008). The Innovative Medicines Initiative (IMI): a new opportunity for scientific collaboration between academia and industry at the European level. *European Respiratory Journal*, 31(5):924–926.

Kaminskas, E., Farrell, A. T., Wang, Y.-C., Sridhara, R., and Pazdur, R. (2005). FDA drug approval summary: azacitidine (5-azacytidine, Vidaza) for injectable suspension. *The Oncologist*, 10(3):176–182.

Kandoth, C., McLellan, M. D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J. F., Wyczalkowski, M. A., , Leiserson, M. D. M., Miller, C. A., Welch, J. S., Walter, M. J., Wendl, M. C., Ley, T. J., Wilson, R. K., Raphael, B. J., and Ding, L. (2013). Mutational landscape and significance across 12 major cancer types. *Nature*, 502(7471):333–339.

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2014). Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Research*, 42(D1):D199–D205.

Khalek, F. J. A., Gallicano, G. I., and Mishra, L. (2010). Colon Cancer Stem Cells. *Gastrointestinal Cancer Research : GCR*, (Supplement 1):S16 – S23.

Kim, D. and Jung, I. (2009). Regulatory patterns of histone modifications to control the DNA methylation status at CpG islands. *Interdisciplinary Bio Central*, 1(1):4 – 12.

Kim, J. Y., Siegmund, K. D., Tavar, S., and Shibata, D. (2005). Age-related human small intestine methylation: evidence for stem cell niches. *BMC Medicine*, 3(1):10 – 21.

Klug, M., Heinz, S., Gebhard, C., Schwarzfischer, L., Krause, S. W., Andreesen, R., and Rehli, M. (2010). Active DNA demethylation in human postmitotic cells correlates with activating histone modifications, but not transcription levels. *Genome Biology*, 11(6):R63 – R74.

Knudson, A. G. (2001). Two genetic hits (more or less) to cancer. *Nature Reviews Cancer*, 1(2):157–162.

Koch, C. M., Andrews, R. M., Flicek, P., Dillon, S. C., Karaöz, U., Clelland, G. K., Wilcox, S., Beare, D. M., Fowler, J. C., Couttet, P., James, K., Lefebvre, G., Bruce, A., Dovey, O., Ellis, P., Dhami, P., Langford, C., Weng, Z., Birney, E., Carter, N., Vetrie, D., and Dunham, I. (2007). The landscape of histone modifications across 1% of the human genome in five human cell lines. *Genome Research*, 17(6):691–707.

Kohonen-Corish, M., Sigglekow, N., Susanto, J., Chapuis, P., Bokey, E., Dent, O., Chan, C., Lin, B., Seng, T., Laird, P., Young, J., Leggett, B., Jass, J., and Sutherland, R. (2007). Promoter methylation of the mutated in colorectal cancer gene is a frequent early event in colorectal cancer. *Oncogene*, 26(30):4435–4441.

Kostic, A. D., Gevers, D., Pedamallu, C. S., Michaud, M., Duke, F., Earl, A. M., Ojesina, A. I., Jung, J., Bass, A. J., Tabernero, J., Baselga, J., Liu, C., Shivdasani, R., Ogino, S., Birren, B., Huttenhower, C., Garrett, W., and Meyerson, M. (2012). Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. *Genome Research*, 22(2):292–298.

Kouzarides, T. (2007). Chromatin modifications and their function. *Cell*, 128(4):693–705.

Kriaucionis, S. and Heintz, N. (2009). The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science*, 324(5929):929–930.

Kronholm, K. and Birkeland, K. W. (2005). Integrating spatial patterns into a snow avalanche cellular automata model. *Geophysical Research Letters*, 32(19):1 – 4.

Lay, F. D., Triche, T. J., Tsai, Y. C., Su, S.-F., Martin, S. E., Daneshmand, S., Skinner, E. C., Liang, G., Chihara, Y., and Jones, P. A. (2014). Reprogramming of the human intestinal epigenome by surgical tissue transposition. *Genome Research*, 24(4):545–553.

Leedham, S. J. and Wright, N. A. (2008). Expansion of a mutated clone: from stem cell to tumour. *Journal of Clinical Pathology*, 61(2):164–171.

Lim, S. J., Tan, T. W., and Tong, J. C. (2010). Computational Epigenetics: the new scientific paradigm. *Bioinformation*, 4(7):331 – 337.

Liu, L. and Rando, T. A. (2011). Manifestations and mechanisms of stem cell aging. *The Journal of Cell Biology*, 193(2):257–266.

Lohse, B., Kristensen, J. L., Kristensen, L. H., Agger, K., Helin, K., Gajhede, M., and Clausen, R. P. (2011). Inhibitors of histone demethylases. *Bioorganic & Medicinal Chemistry*, 19(12):3625–3636.

Lowe, R., Gemma, C., Beyan, H., Hawa, M. I., Bazeos, A., Leslie, R. D., Montpetit, A., Rakyan, V. K., and Ramagopalan, S. V. (2013). Buccals are likely to be a more informative surrogate tissue than blood for epigenome-wide association studies. *Epigenetics*, 8(4):445–454.

Lowengrub, J. S., Frieboes, H. B., Jin, F., Chuang, Y., Li, X., Macklin, P., Wise, S., and Cristini, V. (2010). Nonlinear modelling of cancer: bridging the gap between cells and tumours. *Nonlinearity*, 23(1):R1 – R91.

Lucas, P. J., van der Gaag, L. C., and Abu-Hanna, A. (2004). Bayesian networks in biomedicine and health-care. *Artificial Intelligence in Medicine*, 30(3):201–214.

Luo, X., Burwinkel, B., Tao, S., and Brenner, H. (2011). MicroRNA signatures: novel biomarker for colorectal cancer? *Cancer Epidemiology Biomarkers & Prevention*, 20(7):1272–1286.

Lv, J., Qiao, H., Liu, H., Wu, X., Zhu, J., Su, J., Wang, F., Cui, Y., and Zhang, Y. (2010). Discovering cooperative relationships of chromatin modifications in human T cells based on a proposed closeness measure. *PloS ONE*, 5(12):e14219.

Macal, C. M. and North, M. J. (2010). Tutorial on agent-based modelling and simulation. *Journal of Simulation*, 4(3):151–162.

Mannion, R., Ruskin, H., and Pandey, R. B. (2000). Effect of mutation on helper T-cells and viral population: A computer simulation model for HIV. *Theory in Biosciences*, 119(1):10–19.

Mantzaris, N. V., Webb, S., and Othmer, H. G. (2004). Mathematical modeling of tumor-induced angiogenesis. *Journal of Mathematical Biology*, 49(2):111–187.

Maskery, S. M., Hu, H., Hooke, J., Shriver, C. D., and Liebman, M. N. (2008). A Bayesian derived network of breast pathology co-occurrence. *Journal of Biomedical Informatics*, 41(2):242–250.

Masutomi, K., Yu, E. Y., Khurts, S., Ben-Porath, I., Currier, J. L., Metz, G. B., Brooks, M. W., Kaneko, S., Murakami, S., DeCaprio, J. A., et al. (2003). Telomerase maintains telomere structure in normal human cells. *Cell*, 114(2):241–253.

Mattick, J. S., Amaral, P. P., Dinger, M. E., Mercer, T. R., and Mehler, M. F. (2009). RNA regulation of epigenetic processes. *Bioessays*, 31(1):51–59.

McCabe, M. T., Ott, H. M., Ganji, G., Korenchuk, S., Thompson, C., Van Aller, G. S., Liu, Y., Graves, A. P., Diaz, E., LaFrance, L. V., , Mellinger, M., Duquenne, C., Tian, X., Kruger, R., McHugh, C., Brandt, M., Miller, W., Dhanak, D., Verma, S., Tummino, P., and Creasy, C. (2012). EZH2 inhibition as a therapeutic strategy for lymphoma with EZH2-activating mutations. *Nature*, 492(7427):108–112.

McCoy, A. N., Araujo-Perez, F., Azcarate-Peril, A., Yeh, J. J., Sandler, R. S., and Keku, T. O. (2013). Fusobacterium is associated with colorectal adenomas. *PloS ONE*, 8(1):e53653.

McCubrey, J. A., Steelman, L. S., Abrams, S. L., Lee, J. T., Chang, F., Bertrand, F. E., Navolanic, P. M., Terrian, D. M., Franklin, R. A., DAssoro, A. B., et al. (2006). Roles of the raf/mek/erk and pi3k/pten/akt pathways in malignant transformation and drug resistance. *Advances in Enzyme Regulation*, 46(1):249–279.

McDonald, S. A., Preston, S. L., Lovell, M. J., Wright, N. A., and Jankowski, J. A. (2006). Mechanisms of disease: from stem cells to colorectal cancer. *Nature Clinical Practice Gastroenterology & Hepatology*, 3(5):267–274.

Meissner, A. (2010). Epigenetic modifications in pluripotent and differentiated cells. *Nature Biotechnology*, 28(10):1079–1088.

Meissner, A., Mikkelsen, T. S., Gu, H., Wernig, M., Hanna, J., Sivachenko, A., Zhang, X., Bernstein, B. E., Nusbaum, C., Jaffe, D. B., Gnirke, A., Jaenisch, R., and Lander, E. (2008). Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, 454(7205):766–770.

Menéndez, P., Villarejo, P., Padilla, D., Menéndez, J. M., and Montes, J. A. R. (2013). Diagnostic and prognostic significance of serum MicroRNAs in colorectal cancer. *Journal of Surgical Oncology*, 107(2):217–220.

Merriam-Webster (2014). Merriam-webster online dictionary. Available at: http://www.merriam-webster.com/medlineplus/, [Accessed 07/05/2014].

Metzker, M. L. (2009). Sequencing technologiesthe next generation. *Nature Reviews Genetics*, 11(1):31–46.

Michor, F., Iwasa, Y., and Nowak, M. A. (2004). Dynamics of cancer progression. *Nature Reviews Cancer*, 4(3):197–205.

Mill, J. and Heijmans, B. T. (2013). From promises to practical strategies in epigenetic epidemiology. *Nature Reviews Genetics*, 14(8):585–594.

Minucci, S. and Pelicci, P. G. (2006). Histone deacetylase inhibitors and the promise of epigenetic (and more) treatments for cancer. *Nature Reviews Cancer*, 6(1):38–51.

Mirams, G. R., Arthurs, C. J., Bernabeu, M. O., Bordas, R., Cooper, J., Corrias, A., Davit, Y., Dunn, S.-J., Fletcher, A. G., Harvey, D. G., et al. (2013). Chaste: an open source C++ library for computational physiology and biology. *PLoS Computational Biology*, 9(3):e1002970.

Molina-Pea, R. and lvarez, M. M. (2012). A simple mathematical model based on the cancer stem cell hypothesis suggests kinetic commonalities in solid tumor growth. *PloS ONE*, 7(2):e26233.

Mucha, L., Stephenson, J., Morandi, N., and Dirani, R. (2006). Meta-analysis of disease risk associated with smoking, by gender and intensity of smoking. *Gender Medicine*, 3(4):279–291.

Nagy, J. D. (2004). Competition and natural selection in a mathematical model of cancer. *Bulletin of Mathematical Biology*, 66(4):663–687.

National Cancer Registry Ireland (2014). Online cancer incidence. Available at: http://www.ncri.ie/client/choose-stats, [Accessed 18/09/2014].

Naumov, V. A., Generozov, E. V., Zaharjevskaya, N. B., Matushkina, D. S., Larin, A. K., Chernyshov, S. V., Alekseev, M. V., Shelygin, Y. A., and Govorun, V. M. (2013). Genome-scale analysis of DNA methylation in colorectal cancer using Infinium HumanMethylation450 BeadChips. *Epigenetics*, 8(9):921–934.

Negrini, M., Nicoloso, M. S., and Calin, G. A. (2009). MicroRNAs and cancer - new paradigms in molecular oncology. *Current Opinion in Cell Biology*, 21(3):470–479.

NICE - National Institute for Health and Care Excellence (2014). Familial breast cancer: Classification and care of people at risk of familial breast cancer and management of breast cancer and related risks in people with a family history of breast cancer. Available at: http://www.nice.org.uk/guidance/CG164/chapter/1-Recommendations, [Accessed 18/09/2014].

Nicolas, P., Kim, K.-M., Shibata, D., and Tavar, S. (2007). The stem cell population of the human colon crypt: analysis via methylation patterns. *PLoS Computational Biology*, 3(3):e28.

Nikolai, C. and Madey, G. (2009). Tools of the trade: A survey of various agent based modeling platforms. *Journal of Artificial Societies and Social Simulation*, 12(2):1 – 2.

Niv, Y. (2007). Microsatellite instability and MLH1 promoter hypermethylation in colorectal cancer. *World Journal of Gastroenterology*, 13(12):1767 – 1769.

Nowak, M. A., Komarova, N. L., Sengupta, A., Jallepalli, P. V., Shih, I.-M., Vogelstein, B., and Lengauer, C. (2002). The role of chromosomal instability in tumor initiation. *Proceedings of the National Academy of Sciences*, 99(25):16226–16231.

Oakeley, E. J. (1999). DNA methylation analysis: a review of current methodologies. *Pharmacology & Therapeutics*, 84(3):389–400.

Ogino, S., Kawasaki, T., Brahmandam, M., Cantor, M., Kirkner, G. J., Spiegelman, D., Makrigiorgos, G. M., Weisenberger, D. J., Laird, P. W., Loda, M., et al. (2006a). Precision and performance characteristics of bisulfite conversion and real-time PCR (MethyLight) for quantitative DNA methylation analysis. *The Journal of Molecular Diagnostics*, 8(2):209–217.

Ogino, S., Kawasaki, T., Kirkner, G. J., Loda, M., and Fuchs, C. S. (2006b). CpG Island Methylator Phenotype-Low (CIMP-Low) in Colorectal Cancer: Possible Associations with Male Sex and KRAS Mutations. *The Journal of Molecular Diagnostics*, 8(5):582–588.

Papailiou, J., Bramis, K. J., Gazouli, M., and Theodoropoulos, G. (2011). Stem cells in colon cancer. a new era in cancer theory begins. *International Journal of Colorectal Disease*, 26(1):1–11.

Parle-McDermott, A. and Ozaki, M. (2011). The impact of nutrition on differential methy-

lated regions of the genome. *Advances in Nutrition: An International Review Journal*, 2(6):463–471.

Paul, D. S. and Beck, S. (2014). Advances in epigenome-wide association studies for common diseases. *Trends in Molecular Medicine*.

Peak, D., West, J. D., Messinger, S. M., and Mott, K. A. (2004). Evidence for complex, collective dynamics and emergent, distributed computation in plants. *Proceedings of the National Academy of Sciences of the United States of America*, 101(4):918–922.

Pe'er, D., Regev, A., and Tanay, A. (2002). Minreg: Inferring an active regulator set. Number Supplement 1, pages 258–267.

Peirce, S. M. (2008). Computational and mathematical modeling of angiogenesis. *Microcirculation*, 15(8):739–751.

Perrin, D., Ruskin, H. J., Burns, J., and Crane, M. (2006a). An agent-based approach to immune modelling. In *Computational Science and Its Applications-ICCSA 2006*, pages 612–621. Springer.

Perrin, D., Ruskin, H. J., and Crane, M. (2006b). An agent-based approach to immune modelling: priming individual response. *Trans Eng Comput Technol*, 17:80–86.

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

Petersen, A. K., Zeilinger, S., Kastenmuller, G., Romisch-Margl, W., Brugger, M., Peters, A., Meisinger, C., Strauch, K., Hengstenberg, C., Pagel, P., Huber, F., Mohney, R. P., Grallert, H., Illig, T., Adamski, J., Waldenberger, M., Gieger, C., and Suhre, K. (2014). Epigenetics meets metabolomics: an epigenome-wide association study with blood serum metabolic traits. *Human Molecular Genetics*, 23(2):534–545.

Pfaffeneder, T., Hackner, B., Truß, M., Münzel, M., Müller, M., Deiml, C. A., Hagemeier, C., and Carell, T. (2011). The discovery of 5-formylcytosine in embryonic stem cell DNA. *Angewandte Chemie*, 123(31):7146–7150.

Pitt-Francis, J., Pathmanathan, P., Bernabeu, M. O., Bordas, R., Cooper, J., Fletcher, A. G., Mirams, G. R., Murray, P., Osborne, J. M., Walter, A., et al. (2009). Chaste: a test-driven approach to software development for biological modelling. *Computer Physics Communications*, 180(12):2452–2471.

Pogribny, I. P. and Vanyushin, B. F. (2010). Age-related genomic hypomethylation. In *Epigenetics of Aging*, pages 11–27. Springer.

Potten, C. S., Booth, C., and Hargreaves, D. (2003). The small intestine as a model for evaluating adult tissue stem cell drug targets. *Cell Proliferation*, 36(3):115–129.

Preston, S. L., Wong, W. M., Chan, A. O., Poulsom, R., Jeffery, R., Goodlad, R. A., Mandir, N., Elia, G., Novelli, M., Bodmer, W. F., Tomlinson, I. P., and Wright, N. A. (2003). Bottom-up histogenesis of colorectal adenomas: origin in the monocryptal adenoma and initial expansion by crypt fission. *Cancer Research*, 63(13):3819–3825.

Probst, A. V., Dunleavy, E., and Almouzni, G. (2009). Epigenetic inheritance during the cell cycle. *Nature Reviews Molecular Cell Biology*, 10(3):192–206.

Przybilla, J., Rohlf, T., Loeffler, M., and Galle, J. (2014). Understanding epigenetic changes in aging stem cells–a computational model approach. *Aging Cell*, 13(2):320–328.

Raghavan, K. and Ruskin, H. J. (2011). Computational epigenetic micromodel-framework for parallel implementation and information flow. In *Proceedings of The Eighth International Conference on Complex Systems, Boston, USA*, pages 340–353.

Rakyan, V. K., Down, T. A., Balding, D. J., and Beck, S. (2011). Epigenome-wide association studies for common human diseases. *Nature Reviews Genetics*, 12(8):529–541.

Rejniak, K. A. and Anderson, A. R. (2011). Hybrid models of tumor growth. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 3(1):115–125.

Reya, T. and Clevers, H. (2005). Wnt signalling in stem cells and cancer. *Nature*, 434(7035):843–850.

Ribba, B., Saut, O., Colin, T., Bresch, D., Grenier, E., and Boissel, J.-P. (2006). A multiscale mathematical model of avascular tumor growth to investigate the therapeutic benefit of anti-invasive agents. *Journal of Theoretical Biology*, 243(4):532–541.

Richon, V. (2006). Cancer biology: mechanism of antitumour action of vorinostat (suberoylanilide hydroxamic acid), a novel histone deacetylase inhibitor. *British Journal of Cancer*, 95:S2–S6.

Risch, A. and Plass, C. (2008). Lung cancer epigenetics and genetics. *International Journal of Cancer*, 123(1):1–7.

Rivera, C. M. and Ren, B. (2013). Mapping human epigenomes. *Cell*, 155(1):39–55.

Rodríguez-Paredes, M. and Esteller, M. (2011). Cancer epigenetics reaches mainstream oncology. *Nature Medicine*, pages 330–339.

Roznovăţ, I. A. and Ruskin, H. J. (2013a). A Computational Model for Genetic and Epigenetic Signals in Colon Cancer. *Interdisciplinary Sciences: Computational Life Sciences*, 5(3):175–186.

Roznovăţ, I. A. and Ruskin, H. J. (2013b). Methylation Inhibitors and Carcinogens in an Agent-Based Model for Colon Crypt Dynamics during Cancer Development. In *Modelling Symposium (EMS), 2013 European*, pages 152–157. IEEE.

Roznovăţ, I. A. and Ruskin, H. J. (2013c). Modelling the Genetic and Epigenetic Signals in Colon Cancer Using a Bayesian Network. In *Proceedings of the European Conference on Complex Systems 2012*, pages 1059–1062. Springer.

Rubin, D. L., Burnside, E. S., and Shachter, R. (2004). A Bayesian Network to assist mammography interpretation. In *Operations Research and Health Care*, pages 695–720. Springer.

Ruskin, H., Pandey, R. B., and Liu, Y. (2002). Viral load and stochastic mutation in a Monte Carlo simulation of HIV. *Physica A: Statistical Mechanics and its Applications*, 311(1):213–220.

Ryan-Harshman, M. and Aldoori, W. (2007). Diet and colorectal cancer Review of the evidence. *Canadian Family Physician*, 53(11):1913–1920.

Safran, M., Dalah, I., Alexander, J., Rosen, N., Stein, T. I., Shmoish, M., Nativ, N., Bahir, I., Doniger, T., Krug, H., Sirota-Madi, A., Olender, T., Golan, Y., Stelzer, G., Harel, A., and Lancet, D. (2010). GeneCards Version 3: the human gene integrator. *Database*, 2010:1 – 16.

Samaras, V., Rafailidis, P. I., Mourtzoukou, E. G., Peppas, G., and Falagas, M. E. (2010). Chronic bacterial and parasitic infections and cancer: a review. *Journal of Infection in Developing Countries*, 4(5).

Sancho, E., Batlle, E., and Clevers, H. (2003). Live and let die in the intestinal epithelium. *Current Opinion in Cell Biology*, 15(6):763–770.

Santer, F. R., Höschele, P. P., Oh, S. J., Erb, H. H., Bouchal, J., Cavarretta, I. T., Parson, W., Meyers, D. J., Cole, P. A., and Culig, Z. (2011). Inhibition of the acetyltransferases p300 and CBP reveals a targetable function for p300 in the survival and invasion pathways of prostate cancer cell lines. *Molecular Cancer Therapeutics*, 10(9):1644–1655.

Sarma, K. and Reinberg, D. (2005). Histone variants meet their match. *Nature Reviews Molecular Cell Biology*, 6(2):139–149.

Sasco, A., Secretan, M., and Straif, K. (2004). Tobacco smoking and cancer: a brief review of recent epidemiological evidence. *Lung Cancer*, 45:S3–S9.

Schepers, A. G., Vries, R., van den Born, M., van de Wetering, M., and Clevers, H. (2011). Lgr5 intestinal stem cells have high telomerase activity and randomly segregate their chromosomes. *The EMBO Journal*, 30(6):1104–1109.

Schetter, A. J., Okayama, H., and Harris, C. C. (2011). The role of microRNAs in colorectal cancer. *Cancer Journal (Sudbury, Mass.)*, 18(3):244–252.

Scholefield, J. H. (2002). Screening for colorectal cancer. *British Medical Bulletin*, 64(1):75–80.

Schones, D. E. and Zhao, K. (2008). Genome-wide approaches to studying chromatin modifications. *Nature Reviews Genetics*, 9(3):179–191.

Sedgewick, R. and Wayne, K. (2011). *Algorithms. 4th*. Pearson Education, Inc.

SEER (2013a). SEER Stat Fact Sheets: Colon and Rectum Cancer, June 2013. Available at: http://seer.cancer.gov/statfacts/html/colorect.html, [Accessed on 07/05/2014].

SEER (2013b). SEER Stat Fact Sheets: Small Intestine Cancer, June 2013. Available at: http://seer.cancer.gov/statfacts/html/smint.html, [Accessed on 07/05/2014].

Segovia-Juarez, J. L., Ganguli, S., and Kirschner, D. (2004). Identifying control mechanisms of granuloma formation during M. tuberculosis infection using an agent-based model. *Journal of Theoretical Biology*, 231(3):357–376.

Sharpless, N. E., DePinho, R. A., et al. (2004). Telomeres, stem cells, senescence, and cancer. *Journal of Clinical Investigation*, 113(2):160–168.

Shen, L. and Waterland, R. A. (2007). Methods of DNA methylation analysis. *Current Opinion in Clinical Nutrition & Metabolic Care*, 10(5):576–581.

Shenker, N. S., Polidoro, S., van Veldhoven, K., Sacerdote, C., Ricceri, F., Birrell, M. A., Belvisi, M. G., Brown, R., Vineis, P., and Flanagan, J. M. (2013). Epigenome-wide association study in the European Prospective Investigation into Cancer and Nutrition (EPIC-Turin) identifies novel genetic loci associated with smoking. *Human Molecular Genetics*, 22(5):843–851.

Silvertown, J., Holtier, S., Johnson, J., and Dale, P. (1992). Cellular automaton models of

interspecific competition for space–the effect of pattern on process. *Journal of Ecology*, pages 527–533.

Simpson, M. J., Zhang, D. C., Mariani, M., Landman, K. A., and Newgreen, D. F. (2007). Cell proliferation drives neural crest cell invasion of the intestine. *Developmental Biology*, 302(2):553–568.

Sinclair, A. and Pech, R. P. (1996). Density dependence, stochasticity, compensation and predator regulation. *Oikos*, 75(2):164–173.

Sirakoulis, G. C., Karafyllidis, I., and Thanailakis, A. (2000). A cellular automaton model for the effects of population movement and vaccination on epidemic propagation. *Ecological Modelling*, 133(3):209–223.

Sirnes, S., Honne, H., Ahmed, D., Danielsen, S. A., Rognum, T. O., Meling, G. I., Leithe, E., Rivedal, E., Lothe, R. A., and Lind, G. E. (2011). DNA methylation analyses of the connexin gene family reveal silencing of GJC1 (Connexin45) by promoter hypermethylation in colorectal cancer. *Epigenetics*, 6(5):602–609.

Slattery, M., Edwards, S., Curtin, K., Ma, K., Edwards, R., Holubkov, R., and Schaffer, D. (2003). Physical activity and colorectal cancer. *American Journal of Epidemiology*, 158(3):214–224.

Smallbone, K. and Corfe, B. M. (2014). A mathematical model of the colon crypt capturing compositional dynamic interactions between cell types. *International Journal of Experimental Pathology*, 95(1):1–7.

Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L., and Ideker, T. (2011). Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, 27(3):431–432.

Soon, W. W., Hariharan, M., and Snyder, M. P. (2013). High-throughput sequencing for biology and medicine. *Molecular Systems Biology*, 9(1 - 14).

Stein, R. (2014). Epigenetic therapies–a new direction in clinical medicine. *International Journal of Clinical Practice*, 68(7):802 – 811.

Suehiro, Y., Wong, C. W., Chirieac, L. R., Kondo, Y., Shen, L., Webb, C. R., Chan, Y. W., Chan, A. S., Chan, T. L., Wu, T.-T., Rashid, A., Hamanaka, Y., Hinoda, Y., Shannon, R., Wang, X., Morris, J., Issa, J., Yuen, S., Leung, S., and Hamilton, S. (2008). Epigenetic-genetic interactions in the APC/WNT, RAS/RAF, and P53 pathways in colorectal carcinoma. *Clinical Cancer Research*, 14(9):2560–2569.

Suganuma, T. and Workman, J. L. (2008). Crosstalk among histone modifications. *Cell*, 135(4):604–607.

Tarakhovsky, A. (2010). Tools and landscapes of epigenetics. *Nature Immunology*, 11(7):565 – 568.

Teschendorff, A. E., Menon, U., Gentry-Maharaj, A., Ramus, S. J., Weisenberger, D. J., Shen, H., Campan, M., Noushmehr, H., Bell, C. G., Maxwell, A. P., , Savage, D., Mueller-Holzner, E., Marth, C., Kocjan, G., Gayther, S., Jones, A., Beck, S., Wagner, W., Laird, P., Jacobs, I., and Widschwendter, M. (2010). Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome Research*, 20(4):440–446.

Timp, W. and Feinberg, A. P. (2013). Cancer as a dysregulated epigenome allowing cellular growth advantage at the expense of the host. *Nature Reviews Cancer*, 13(7):497–510.

Tracqui, P. (2009). Biophysical models of tumour growth. *Reports on Progress in Physics*, 72(5):056701.

Troyanskaya, O. G., Dolinski, K., Owen, A. B., Altman, R. B., and Botstein, D. (2003). A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). *Proceedings of the National Academy of Sciences*, 100(14):8348–8353.

Tsygvintsev, A., Marino, S., and Kirschner, D. E. (2013). A mathematical model of gene therapy for the treatment of cancer. In *Mathematical Methods and Models in Biomedicine*, pages 367–385. Springer.

Turner, B. M. (2005). Reading signals on the nucleosome with a new nomenclature for modified histones. *Nature Structural & Molecular Biology*, 12(2):110–112.

Umar, S. (2010). Intestinal stem cells. *Current Gastroenterology Reports*, 12(5):340–348.

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

Vaiopoulos, A. G., Kostakis, I. D., Koutsilieris, M., and Papavassiliou, A. G. (2012). Colorectal cancer stem cells. *Stem Cells*, 30(3):363–371.

Vaissière, T., Sawan, C., and Herceg, Z. (2008). Epigenetic interplay between histone modifications and DNA methylation in gene silencing. *Mutation Research/Reviews in Mutation Research*, 659(1):40–48.

van Bemmel, J. G., Filion, G. J., Rosado, A., Talhout, W., de Haas, M., van Welsem, T., van Leeuwen, F., and van Steensel, B. (2013). A Network Model of the Molecular Organization of Chromatin in Drosophila. *Molecular Cell*, 49(4):759–771.

van der Flier, L. G. and Clevers, H. (2009). Stem cells, self-renewal, and differentiation in the intestinal epithelium. *Annual Review of Physiology*, 71:241–260.

van Heesch, D. (2008). Doxygen: Source code documentation generator tool. URL: http://www. doxygen.org, [Accessed 19/06/2014].

Van Leeuwen, I., Byrne, H., Jensen, O., and King, J. (2006). Crypt dynamics and colorectal cancer: advances in mathematical modelling. *Cell Proliferation*, 39(3):157–181.

Van Leeuwen, I., Mirams, G., Walter, A., Fletcher, A., Murray, P., Osborne, J., Varma, S., Young, S., Cooper, J., Doyle, B., Pitt-Francis, P., Momtahan, L., Pathmanathan, P.,

Whiteley, J., Chapman, S., Gavaghan, D., Jensen, O., King, J., Maini, P., Waters, S., and Byrne, H. (2009). An integrative computational model for intestinal tissue renewal. *Cell Proliferation*, 42(5):617–636.

van Steensel, B., Braunschweig, U., Filion, G. J., Chen, M., van Bemmel, J. G., and Ideker, T. (2010). Bayesian network analysis of targeting interactions in chromatin. *Genome Research*, 20(2):190–200.

Verma, M. (2012). Epigenome-Wide Association Studies (EWAS) in Cancer. *Current Genomics*, 13(4):308 – 313.

Vinken, M., Rop, E. D., Decrock, E., Vuyst, E. D., Leybaert, L., Vanhaecke, T., and Rogiers, V. (2009). Epigenetic regulation of gap junctional intercellular communication: more than a way to keep cells quiet? *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, 1795(1):53–61.

Vogelstein, B. and Kinzler, K. W. (2004). Cancer genes and the pathways they control. *Nature Medicine*, 10(8):789–799.

Volinia, S., Galasso, M., Costinean, S., Tagliavini, L., Gamberoni, G., Drusco, A., Marchesini, J., Mascellani, N., Sana, M. E., Jarour, R. A., et al. (2010). Reprogramming of miRNA networks in cancer and leukemia. *Genome Research*, 20(5):589–599.

Walker, D., Southgate, J., Hill, G., Holcombe, M., Hose, D., Wood, S., Mac Neil, S., and Smallwood, R. (2004). The epitheliome: agent-based modelling of the social behaviour of cells. *Biosystems*, 76(1):89–100.

Wang, H., Zheng, H., Browne, F., Glass, D. H., and Azuaje, F. (2010). Integration of Gene Ontology-based similarities for supporting analysis of protein–protein interaction networks. *Pattern Recognition Letters*, 31(14):2073–2082.

Wang, Z., Butner, J. D., Kerketta, R., Cristini, V., and Deisboeck, T. S. (2014). Simulating cancer growth with multiscale agent-based modeling. In *Seminars in Cancer Biology*. Elsevier.

Watson, A. J. and Collins, P. D. (2011). Colon cancer: a civilization disorder. *Digestive Diseases*, 29(2):222–228.

Wilkinson, D. J. (2009). Stochastic modelling for quantitative description of heterogeneous biological systems. *Nature Reviews Genetics*, 10(2):122–133.

Witze, E. S., Old, W. M., Resing, K. A., and Ahn, N. G. (2007). Mapping protein post-translational modifications with mass spectrometry. *Nature Methods*, 4(10):798–806.

Yang, L., Belaguli, N., and Berger, D. H. (2009). MicroRNA and colorectal cancer. *World Journal of Surgery*, 33(4):638–646.

Yatabe, Y., Tavaré, S., and Shibata, D. (2001). Investigating stem cells in human colon by using methylation patterns. *Proceedings of the National Academy of Sciences*, 98(19):10839–10844.

Yoo, C. B. and Jones, P. A. (2006). Epigenetic therapy of cancer: past, present and future. *Nature Reviews Drug Discovery*, 5(1):37–50.

Yoon, J.-H., Dammann, R., and Pfeifer, G. P. (2001). Hypermethylation of the CpG island of the RASSF1A gene in ovarian and renal cell carcinomas. *International Journal of Cancer*, 94(2):212–217.

Yu, H., Zhu, S., Zhou, B., Xue, H., and Han, J.-D. J. (2008). Inferring causal relationships among different histone modifications and gene expression. *Genome Research*, 18(8):1314–1324.

Zisman, A. L., Nickolov, A., Brand, R. E., Gorchow, A., and Roy, H. K. (2006). Associations between the age at diagnosis and location of colorectal cancer and the use of alcohol and tobacco: implications for screening. *Archives of Internal Medicine*, 166(6):629–634.

Appendices

Appendix A

Glossary

5hmC, **5fC** and **5caC** are DNA methylation variants resulted sequentially from methylation cytosine (5mC) oxidation processes, which are regulated by the Tet (ten eleven translocation) protein group, as illustrated by the following relationships:
5mC $\xrightarrow{oxidation}$ 5hmC $\xrightarrow{oxidation}$ 5fC $\xrightarrow{oxidation}$ 5caC.

Acetylation is addition of an acetyl group to a chemical compound.

Agent-based model (ABM) consists of units or entities at basis level, to which attributes are assigned.

Age Sensitive Gene (ASG parameter) indicates age when methylation level changes of a gene sensitive to age-related methylation are observed.

Apoptosis is the process of programmed cell death.

Bayesian network is a network of genes in which change of a given gene is conditional on change of one or more other genes.

Benign (tumour) does not invade nearby tissues or spread throughout the body, i.e. not cancerous.

Biomarker is an entity that characterizes specifically abnormal systems and can be used to differentiate tumour from normal tissues.

Cell is the the smallest structural and functional unit of an organism.

Cell cycle contains the set of events between two successive cell divisions.

Cell differentiation is the process during which a less specialized cell, (e.g. stem cell), becomes a more specialized cell type; in some cases, (e.g. colon), a fully-differentiated cell loses the ability of performing further divisions.

Cell division is the process during which a parent cell divides into two daughter cells, transmitting genetic information.

Cell proliferation is the process of increasing cell numbers, through cell division.

Cellular automaton is a discrete model consisting of large number of identical cells found in a discrete number of states. All cells change state simultaneously, based on transition rules with their neighbours.

Chemotherapy is a standard cancer therapy, where cancerous cells are targeted by different drugs.

Chromatin immunoprecipitation (ChIP) is a method for histone modification analysis, which can be applied at both gene specific and genome-wide levels, also in combination with microarray and NGS technologies, (e.g. ChIP-chip, ChIP-Seq).

Chromatin is the combination of DNA and proteins that makes up the contents of the cell nucleus. Chromatin is present in two forms: **heterochromatin** (highly condensed and characterised by low levels of gene transcription), and **euchromatin** (less compacted and characterised by high levels of gene transcription).

Chromosome is an organized arrangement or package of DNA in cell.

Chromosomal instability is a class of genomic instability characterised by deletions or duplications of parts of chromosomes; it has been associated with malignant tumour development.

Colonoscopy is the screening procedure that permits examination of the entire colon and removal of detected polyps in the same session.

Colorectal cancer develops in the colon, rectum or in the appendix. It is also referred as colon cancer, rectal cancer, bowel cancer or colorectal adenocarcinoma.

CpG Island Methylator Phenotype represents the malignant phenotype given by CpG island hypermethylation of promoters in a group of genes.

Density-Dependent Coefficient determine the relationships between growth rate within a population at a given time t based on population size at time t and its maximum capacity permitted in a specific environment.

DNA, or deoxyribonucleic acid, is a double-stranded macromolecule that encodes the genetic information in almost all living organisms and is composed from base pairs of nucleotides. The nucleotides contain the bases adenine (A), guanine (G), cytosine (C), and thymine (T). In nature, base pair affinities are A-T and G-C.

DNA methylation is a molecular process that involves the addition of a methyl group to a cytosine or adenine ring. (Adenine methylation has been reported recently in prokaryotes, [Aguilar and Craighead, 2013] and references therein.)

DNA methyltransferase family represents enzymes catalysing methyl group transfer to DNA.

DNA repair is the process used by a cell to identify and repair abnormal changes in DNA.

d-Network Threshold (dNT) is a model coefficient that determines if gene pathways found in a gene network are d-plausible, based on the value of *d-* model parameter.

Drug response curve is a curve measuring level of response of drug present over time.

Epigenetic events within a cell are heritable and reversible modifications that affect gene expression but not the DNA sequence.

Epigenome refers to the totality of epigenetic markers in an organism.

Epithelium is one of the four major animal tissue types, (along connective, muscle and nervous), characterised by high cell division rate and little intercellular gaps. Epithelium assures functions of e.g. tissue protection, hormone secretion, nutrient absorption, sensation detection.

Exon is the coding part of a gene that is used in translation to protein.

Familial adenomatous polyposis represents an incipient or initial stage of adenoma, characterised by large numbers of polyps in both colon and rectal tissues, which can progress to CRC if are not treated.

Gene is the basic physical and functional hereditary unit, comprises part of the DNA sequence and determines a specific characteristic of an organism.

Gene expression is the process of synthesising the gene product, (protein or RNA).

Gene Expression Omnibus is a database of genetic and epigenetic data generated by high throughput technologies. technologies

Genome consists of all the genetic material found in an organism.

Genotype is the gene collection for an individual. It refers also to the alleles inherited for a specific gene.

Histone is a protein package of the DNA nucleosome sequence. There are five known families of histones H1/H5, H2A, H2B, H3 and H4, grouped in two categories: the *core*, (H2A, H2B, H3, H4) and the *linker* histones, (H1 and H5).

Histone acetyltransferases represent enzyme family catalysing acetyl group transfer to histones.

Histone deacetylases represent enzyme family removing acetyl group from histones.

Histone demethylases represent enzyme family removing methyl group from histones.

Histone modification refers to addition of different chemical compounds (e.g. acetyl, methyl) to histone 'tails', forming e.g. histone acetylation, methylation.

Histone methyltransferases represent enzyme family catalysing methyl group transfer to histones.

High-performance liquid chromatography is method for global DNA methylation analysis.

Histone variant is a variant form of a major histone modification. For example, H2A.X and H2A.Z are histone H2A variants, while the H3.3 is a variant for histone H3.

Homeostasis is the balance or stable state that characterises the internal environment of a normal biological system; this is usually altered in malignant tumour.

Hypermethylation is an increase in DNA methylation level.

Hypomethylation is a decrease in DNA methylation level.

Inhibitor Maximum Efficiency Time represents the time period with the highest inhibitor effect.

Incidence rate of cancer is the rate per unit time of cancer occurrence in a population scaled to the population size.

Inflammatory bowel disease represents a group of disorders involving inflammation of different parts of the intestine. Its presence is considered to be a risk factor in bowel cancer development.

Intron is a non-coding DNA/ RNA segment.

Leukaemia is a malignant disease of the blood-forming organs characterised by an aberrant increase in the white blood cells number.

Lymph is a transparent liquid, (composed from white blood cells, mainly lymphocytes), which circulates through the lymphatic vessels and intercellular gaps and facilitates substance exchange between blood and solid tissues.

Malignant implies ability to invade nearby tissues or spread throughout the body, i.e. cancerous.

Metastasis characterises the tumour extension from the initial (primary) tissue to other parts of the body.

Methylation is addition of a methyl group to a chemical compound.

The **methylation cycle number average (MCA)** is the average time period after which all gene networks within a given group advance to the next cancer stage.

Micro RNAs are small sequences of non-coding RNA molecule with role in cell cycle regulation.

Microsatellite instability is the condition of increased mutation rate caused by deregulations in the DNA mismatch repair mechanism.

Mortality rate = number of deaths per unit of time caused by a disease in a population scaled to the population size.

Mutation is a modification in a DNA sequence. Mutations can be i) germline, (which can be transmitted to offspring), and ii) somatic, (which are not passed on).

Neurodegenerative disorder refers to a group of diseases that primarily include affections of the neurons in the human brain, (e.g. Parkinson's, Alzheimer's disease).

Next-Generation Sequencing is a collective name for more recent high throughput methods of determining e.g. variation at genetic and epigenetic levels.

Nucleotide is a compound consisting of a base, (A, G, C, T), sugar, and phosphate.

Oncogene is a gene that has potential to cause cancer.

Percentage of methylated reference measures the degree of methylation for a given gene. When PMR values show variation across multiple runs, average PMR can be calculated based on reported PMR values.

Phenotype is composed from the observable characteristics of a living organism. A simplified mathematical formula that describes the phenotype is: Phenotype = Genotype

+ Environment + Epigenetics, incorporating both environment and Epigenetics influences in gene expression.

The **polymerase chain reaction** is a technology used to amplify one or more DNA sequence, generating large number of copies, (e.g. thousands, millions).

Polyp is a small growth protruding from a mucous membrane.

Progenitor cell is a cell, characterised by the ability of limited division number.

Protein is a molecule built from amino acids, (formed from triplets of nucleotide bases), which governs functionality. Proteins characterise major components, (e.g. hair, skin, etc.).

Proto-oncogene is a normal gene that can become an oncogene due to mutations or increased expression.

Radiotherapy is standard for more advanced cancer stages and aims to destroy malignant cells.

Recurrent (cancer) refers to malignant tumours; initially susceptible to treatment, which reoccurs in an individual after a time-period during which the disease is non-detectable (latent).

Residue is a small quantity of a substance.

Risk factor is a factor such as chemical exposure, tobacco usage, that may increase the probability of disease development.

RNA, or ribonucleic acid, is a single-stranded macromolecule transcribing the genetic information from DNA to proteins.

Sensitive to age-related methylation is a gene feature that indicates that methylation levels of a given gene are influenced by ageing.

Sporadic refers to a tumour that occurs in a population without involving heritable predisposition.

Stage (cancer) is the degree of the cancer development and extension.

Stem cells can differentiate into diverse specialized cell types. Referred to as self-renewable due to ability to produce new stem cells and considered to be the repair system of the body.

Signalling pathway represents a set of inter-related molecules within a cell working to control cell functions, including cell division or apoptosis.

Targeted therapy aims to detect and destroy malignant cells without affecting normal cells. It can be applied for more advanced cancer stages, (e.g. metastasis), and also in combination with chemotherapy.

Tumour refers to abnormal and uncontrolled cell proliferation.

Tumour suppressor gene is a gene that protects a cell from progressing one step along the path to disease; it governs cell apoptosis.

Appendix B

Extended information on CRC

Biological Background

Table B.1: Key genes in CRC development

Gene Symbol	**Gene characteristics**
APC	The molecular changes of the *adenomatous polyposis coli, (APC)*, gene are linked to the earliest beginning of CRC, being associated with FAP and sporadic cases, [Fodde et al., 2001; Grady and Markowitz, 2002; Frank, 2007; Suehiro et al., 2008]. The results presented in Suehiro et al. [2008] show an association between the hypermethylation of APC and mutations of TP53 in colorectal neoplasms.
KRAS	The *v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog, (KRAS)*, gene is the most commonly mutated RAt Sarcoma (RAS) family member in CRC, [Grady and Markowitz, 2002]. Studies have reported that approximately 50% of advanced adenomas with APC mutations also have KRAS mutations, [Grady and Markowitz, 2002]. A high influence is seen also in lung cancer, where KRAS was observed as being mutated in 30% of cases, [Risch and Plass, 2008].
TP53	The *tumour protein p53, (TP53)*, gene is a stress response gene involved in maintaining genomic stability through the control of cell cycle progression, DNA repair and apoptosis. Its inactivation leads to both decrease of tumour-cell death rate and increase of their division rate, [Knudson, 2001]. A significantly high level of changes in this TGS was reported in lung and colorectal cancers, [Risch and Plass, 2008; Suehiro et al., 2008].
	Continued on next page

Gene Symbol	Gene characteristics
RASSF1A	The *Ras-associated domain family member 1, (RASSF1A)*, gene is considered to be one of the most hypermethylated TSGs in various cancer types, such as breast, [Dworkin et al., 2009], lung, [Dammann et al., 2003], ovarian, renal, prostate and colorectal cancers, [Yoon et al., 2001]. Studies show that RASSF1A expression is significantly reduced in tumours, although it has a high level in normal cells, [Dworkin et al., 2009]. Therefore, identification of high RASSF1A methylation level in carcinoma cells could become important in early detection of cancer, [Dammann et al., 2003].
MCC	*Mutated In Colorectal Cancers (MCC)* gene is considered to be involved in several cellular processes, (e.g. proliferation, [Fukuyama et al., 2008], movement of epithelial cells, [Arnaud et al., 2009]), and pathways, (e.g. Wnt, [Fukuyama et al., 2008]), and to act as a TSG. MCC silencing, (due to promoter methylation), has been observed in sporadic forms of CRC, [Kohonen-Corish et al., 2007; Fukuyama et al., 2008].
MLH1	The *mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli), (MLH1)*, gene is found to be mutated or aberrant methylated in the *hereditary nonpolyposis* forms of the CRC, [Ogino et al., 2006b; Niv, 2007]. MLH1 is considered a 'caretaker' gene, since its expressed protein has a role in repairing DNA mismatches that appear normally during DNA replication; abnormal changes in MLH1 expression lead to genomic instability and cell proliferation, [Niv, 2007].
MGMT	Another 'caretaker' gene that encodes DNA repair protein, [Herman and Baylin, 2003], the *O^6-methylguanine-DNA methyltransferase, (MGMT)*, is found to be highly hypermethylated in adenomas, [Ogino et al., 2006b; Suehiro et al., 2008]. Additionally, aberrant methylation of APC followed by MGMT or MLH1 hypermethylation is associated with abnormal G-to-A transition in APC, [Suehiro et al., 2008].

Table B.2: Colorectal cancer types

CRC Type	Major characteristics
CIN	CIN is characterised by genetic mutations involving the inactivation of TGSs such as APC, p53, and the activation of oncogenes, (e.g. KRAS), [Grady, 2004].
CIMP	CIMP is characterised by inactivation of different TSGs, (e.g. CDKN2A:p16, IGF2, RUNX3), and DNA repair genes, (e.g. MLH1), caused by CpG islands hypermethylation in their promoters, (reviewed in [Hughes et al., 2012]). In addition, activation of the BRAF proto-oncogene is considered to be an early event in CIMP, [Hughes et al., 2012].
	Continued on next page

CRC Type	Major characteristics
HNPCC	HNPCC is characterised by germline mutations in the DNA mismatch repair genes, e.g. in MLH1, MSH2, or MSH6 genes. In addition, the time-period for developing CRC decreases in comparison with sporadic cases, (i.e. 2-3 years in the former instead of 8-10 years in the later). Moreover, in the former case, the malignant phenotype can be identified in younger individuals, (with average age $\approx$ 45 years), in comparison with the later, (with the average age $\approx$ 63 years), [Cunningham et al., 2010].
FAP	FAP is characterised by large numbers of adenomas in both colon and rectal tissues, which can progress during short time-periods to CRC if are not treated. Initially, the polyps develop as consequence of germline mutations in APC gene, [Half et al., 2009].
MAP	MAP is characterised by recessively inherited mutations in MUTYH gene causing a number of polyps, (lower than in the FAP cases), [Half et al., 2009].
BRAF	BRAF is a proto-oncogene, mutated in several human cancers, including melanoma, ovarian, prostate and CRC, [McCubrey et al., 2006; Cho et al., 2006; Ahmed et al., 2013]. In CRC, BRAF is one of the most commonly affected genes, in addition to e.g. TP53 and KRAS, [Ahmed et al., 2013]. Drugs targeting BRAF mutation have been already approved by FDA, (e.g. *Vemurafenib* , [Bollag et al., 2012]) and are in use for melanoma.

Table B.3: Cancer predisposition based on ageing

Group name	Description
Group 1	refers to children, (aged 0-14 years), and young adults, (15-29 years), and is characterised by very low number of diagnosed cancer cases. An exception to this observation is that of leukaemia, the most common childhood cancer, which has a very low rate in adults, [Cancer Research UK, 2014a]. CRC incidence rate is estimated to be below 5 per 100,000 population UK, (2009-2011), [Cancer Research UK, 2014b].
Group 2	includes adults of 30-49 years and shows a slightly higher overall cancer predisposition in females than in males. This is caused mainly by the high number of breast cancer cases identified in this age range, [Frank, 2007]. CRC incidence is still low, with the rate around 45 for male and 40 for females per 100,000 UK population, (2009 - 2011), [Cancer Research UK, 2014b].
	Continued on next page

Group name	Description
Group 3	is composed of adults aged 50-74 years and includes the highest number of detected cancer cases from all age groups, with more cancer cases diagnosed in males than in females. Some early studies reported that CRC incidence increases markedly after 50 years, [Issa et al., 1994]. However, more recent reports noted that although abnormal changes can appear in the colon even 10-15 years earlier than malignant symptoms, CRC development mostly occurs between the ages 65-75, with screening recommended between ages 50-75 or 40-75 for people with familial colon cancer history, [Scholefield, 2002]. Estimated CRC incidence per 100,000 of the population was reported to be around 360 world-wide in 2012, and around 220 in US, (2009-2011), [SEER, 2013a; Ferlay et al., 2014].
Group 4	is associated with the elderly, (i.e. of age greater than 75 years), and it characterised by relatively high CRC incidence rate. For instance, estimated CRC incidence per 100,000 individuals was reported to be around 200 world-wide in 2012, and around 250 in US, (2009-2011), [SEER, 2013a; Ferlay et al., 2014].

Appendix C

Resources

Table C.1: A list of Resources for Epigenetics Research (Urls accessed on 09/06/2014)

Resource name	**Description and url**.
Epigenesys - Network of Excellence	European community on Epigenetics; url: http://www.epigenesys.eu/en/.
Epigenie - Informally informative coverage of Epigenetics	Resources for Epigenetics research including information on technologies, webminars, meetings and products etc.; url: http://epigenie.com/.
e3: Scientific Networking for Epigenetic Experts	Information on conferences, meetings, resources for Epigenetics researchers; url: http://epiexperts.com/.
Galaxy	A web-based, free framework for biomedical data research, [Goecks et al., 2010]; url: http://galaxyproject.org/.
NCBI Epigenomics	Repository for epigenetic modification data for whole-genome, [Fingerman et al., 2011]; url: http://www.ncbi.nlm.nih.gov/epigenomics.

Table C.2: Databases accessed for genetic and epigenetic mechanisms in cancer and other human diseases (Urls accessed on 09/06/2014)

Resource	**Description/ Database content**
ArrayExpress	Gene expression data from studies using microarray and high throughput sequencing technologies; url: https://www.ebi.ac.uk/arrayexpress/.
	Continued on next page

Resource	Description/ Database content
COSMIC	*Catalogue Of Somatic Mutations In Cancer* database: Somatic mutations observed in human cancers; url: http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/.
DiseaseMeth	Methylation level in human diseases; url: http://202.97.205.78/diseasemeth/.
Ensembl Genome Browser	The genome of different organisms, (including human, mouse); url: http://www.ensembl.org/index.html.
GeneCards	Gene annotation; url: http://www.genecards.org/
HGMD	*Human Gene Mutation Database*: Methylation level and mutation of cancer-related genes; url: http://www.hgmd.org/.
HHMD	*Human Histone Modification Database*: Human histone modification; url: http://202.97.205.78/hhmd/.
Histone Database	Histone sequences; url: http://genome.nhgri.nih.gov/histones/.
HIstome	Information on histone modifications, variants and modifying enzymes in human systems; url: http://www.actrec.gov.in/histome/index.php.
ICGC Data Portal	Repository for output information from the ICGC projects; url: https://dcc.icgc.org/.
KEGG Pathway	Pathways involved in e.g. cellular processes, human diseases; url: http://www.genome.jp/kegg/pathway.html.
MethDB	Cancer methylation; url: http://www.methdb.de/.
MethyCancer	Methylation level of cancer-related genes; url: http://methycancer.psych.ac.cn/.
MPromDb	Information from ChIP-Seq experiments on gene promoters in mammalian biological systems; url: http://mpromdb.wistar.upenn.edu/.
NCBI GEO	Genomic data generated from microarray and NGS technologies; url: http://www.ncbi.nlm.nih.gov/geo/.
NCBI Gene	Gene annotation; url: http://www.ncbi.nlm.nih.gov/gene
PubMeth	Gene methylation level specific to various systems; url: http://www.pubmeth.org/.
StatEpigen	Conditional relationships between genetic and epigenetic events affecting different genes in various cancer types; url: http://statepigen.sci-sym.dcu.ie/.
TCGA Data Portal	The TGCA data sets on exome, methylation, miRNA and clinical information for different malignant phenotypes; url: https://tcga-data.nci.nih.gov/tcga/.

Table C.3: A list of software developed for generating BN-applications, (software urls accessed 19/06/2014)

Tool type	**Tool name and url**
Free	Bayesian Network tools in Java, url: http://bnj.sourceforge.net/.
	Chordalysis, url: http://sourceforge.net/projects/chordalysis/.
	FDEP, url: http://www.cs.bris.ac.uk/ flach/fdep/.
	GeNIe, url: http://genie.sis.pitt.edu/.
	jBNC, url: http://jbnc.sourceforge.net/.
	PNL: Open Source Probabilistic Networks Library, url: http://sourceforge.net/projects/openpnl/.
	Netice, url: http://www.norsys.com/.
Commercial	Analytica, url: http://www.lumina.com/.
	BayesServer, url: http://www.bayesserver.com/default.aspx.
	BNet, url: https://www.cra.com/commercial-solutions/bnet-builder.asp.
	Flint, url: http://www.lpa.co.uk/fln.htm.
	Precision Tree, url: http://www.palisade.com/precisiontree/.
	WEKA, url: http://www.cs.waikato.ac.nz/ ml/weka/.

Table C.4: Genes included in the E-G Network Model, with information on gene symbol and name from the *GeneCards* database, [Safran et al., 2010]

Gene Symbol	**Gene Name**
APC	Adenomatous Polyposis Coli
BRAF	v-raf murine sarcoma viral oncogene homolog B
CDKN2A:P14	Cyclin-dependent kinase inhibitor 2A
CDKN2A:p16	Cyclin-dependent kinase inhibitor 2
CRABP1	Cellular retinoic acid binding protein 1
GJC1	Gap junction protein, gamma 1, 45kDa
IGF2	Insulin-like growth factor II
KRAS	v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog
MCC	Mutated In Colorectal Cancers
MGMT	O^6-methylguanine-DNA methyltransferase
MLH1	MutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli)
MUTYH	MutY homolog
NORE1	Ras association (RalGDS/AF-6) domain family member 5
NR3C1	Nuclear receptor subfamily 3, group C, member 1 (glucocorticoid receptor)
RASSF1A	Ras-associated domain family member 1
SMAD4	SMAD family member 4
TP53	Tumour Protein p53

Appendix D

Cell Rate Relationships in the LogisticCrypt Component Model

The cell action rate expressions described here are deduced based on biological information on intestinal crypt dynamics and structure indicated in e.g Potten et al. [2003]; Sancho et al. [2003]; Kim et al. [2005]; Canavan et al. [2006]; Frank [2007]; Nicolas et al. [2007]; Humphries and Wright [2008]; Jin et al. [2009]; Khalek et al. [2010]; Umar [2010]; Bock [2012]; Vaiopoulos et al. [2012]; Clevers and Bevins [2013], (Section 2.4).

D.1 Colon Crypt System

1. Rules applied to update cell numbers in colon crypt over time:

   $\text{Stem}(t+1) = \text{Stem}(t) + \text{Stem}(t) \times (\text{Rate}^{Stem}_{SymDiv} - \text{Rate}^{Stem}_{Apoptosis}) \times \text{DDC}_{STEM}(t)$

   $\text{Prog}(t+1) = \text{Prog}(t) + [\text{Stem}(t) \times \text{Rate}^{Stem}_{AsymDiv} + \text{Prog}(t) \times (\text{Rate}^{Prog}_{Div} - \text{Rate}^{Prog}_{FullDiff})] \times \text{DDC}_{PROG}(t)$

   $\text{Diff}(t+1) = \text{Diff}(t) + [\text{Prog}(t) \times \text{Rate}^{Prog}_{FullDiff} - \text{Diff}(t) \times \text{Rate}^{Diff}_{Apoptosis}] \times \text{DDC}_{DIFF}(t)$

   $\text{DDC}_{STEM}(t) \neq 0$; $\text{DDC}_{PROG}(t) \neq 0$; $\text{DDC}_{DIFF}(t) \neq 0$,

   $\text{Stem}(t_0)$, $\text{Prog}(t_0)$, $\text{Diff}(t_0)$ = known values, being equal to respectively, the stem, progenitor and differentiated cell number at initial time, i.e. t_0.

2. Relationships showing no variations on cell number over time:

   Stem(t+1) - Stem(t) = 0

   Prog(t+1) - Prog(t) = 0

   Diff(t+1) - Diff(t) = 0

3. Calculation of relationship between parameters:

   Stem(t+1) - Stem(t) = 0 $\Rightarrow$

   Stem(t) + Stem(t) $\times$ (Rate$^{Stem}_{SymDiv}$ - Rate$^{Stem}_{Apoptosis}$) $\times$ DDC$_{STEM}$(t) - Stem(t) = 0 $\Rightarrow$

   (Rate$^{Stem}_{SymDiv}$ - Rate$^{Stem}_{Apoptosis}$) $\times$ DDC$_{STEM}$(t) = 0 $\Rightarrow$

   Rate$^{Stem}_{SymDiv}$ - Rate$^{Stem}_{Apoptosis}$ = 0, (given DDC$_{STEM}$(t) $\neq$ 0) $\Rightarrow$

   Rate$^{Stem}_{Apoptosis}$ = Rate$^{Stem}_{SymDiv}$

   Prog(t+1) - Prog(t) = 0 $\Rightarrow$

   Prog(t) + [Stem(t) $\times$ Rate$^{Stem}_{AsymDiv}$ + Prog(t) $\times$ (Rate$^{Prog}_{Div}$ - Rate$^{Prog}_{FullDiff}$)] $\times$ DDC$_{PROG}$(t)
   Prog(t) = 0 $\Rightarrow$

   [Stem(t) $\times$ Rate$^{Stem}_{AsymDiv}$ + Prog(t) $\times$ (Rate$^{Prog}_{Div}$ - Rate$^{Prog}_{FullDiff}$)] $\times$ DDC$_{PROG}$(t) =
   0, (given DDC$_{PROG}$(t) $\neq$ 0) $\Rightarrow$

   Stem(t) $\times$ Rate$^{Stem}_{AsymDiv}$ + Prog(t) $\times$ (Rate$^{Prog}_{Div}$ Rate$^{Prog}_{FullDiff}$) = 0 $\Rightarrow$

   Stem(t) $\times$ Rate$^{Stem}_{AsymDiv}$ + Prog(t) $\times$ Rate$^{Prog}_{Div}$ = Prog(t) $\times$ Rate$^{Prog}_{FullDiff}$

   At initial time, i.e. t_0, $\Rightarrow$

   Rate$^{Prog}_{FullDiff}$ = $\frac{Stem(t_0)}{Prog(t_0)}$ $\times$ Rate$^{Stem}_{AsymDiv}$ + Rate$^{Prog}_{Div}$

   Diff(t+1) - Diff(t) = 0 $\Rightarrow$

   Diff(t) + [Prog(t) $\times$ Rate$^{Prog}_{FullDiff}$ - Diff(t) $\times$ Rate$^{Diff}_{Apoptosis}$] $\times$ DDC$_{DIFF}$(t) - Diff(t)
   = 0 $\Rightarrow$

[Prog(t) $\times$ Rate$^{Prog}_{FullDiff}$ - Diff(t) $\times$ Rate$^{Diff}_{Apoptosis}$] $\times$ DDC$_{DIFF}$(t) = 0 (given DDC$_{DIFF}$(t) $\neq$ 0) $\Rightarrow$

Prog(t) $\times$ Rate$^{Prog}_{FullDiff}$ - Diff(t) $\times$ Rate$^{Diff}_{Apoptosis}$ = 0

At initial time, i.e. t_0, $\Rightarrow$

Rate$^{Diff}_{Apoptosis}$ = $\frac{Prog(t_0)}{Diff(t_0)}$ $\times$ Rate$^{Prog}_{FullDiff}$ $\Rightarrow$

Rate$^{Diff}_{Apoptosis}$ = $\frac{Prog(t_0)}{Diff(t_0)}$ $\times$ [$\frac{Stem(t_0)}{Prog(t_0)}$ $\times$ Rate$^{Stem}_{AsymDiv}$ + Rate$^{Prog}_{Div}$]

D.2 Small Intestine Crypt System

1. Rules applied to update cell numbers in small intestine crypt over time:

   Stem(t+1) = Stem(t) + Stem(t) $\times$ (Rate$^{Stem}_{SymDiv}$ - Rate$^{Stem}_{Apoptosis}$) $\times$ DDC$_{STEM}$(t);

   Diff(t+1) = Diff(t) + [Prog(t) $\times$ Rate$^{Prog}_{FullDiff}$ - Diff(t) $\times$ Rate$^{Diff}_{Apoptosis}$] $\times$ DDC$_{DIFF}$(t)

   Prog(t+1) = Prog(t) + [Stem(t) $\times$ Rate$^{Stem}_{AsymDiv}$ + Prog(t) $\times$ (Rate$^{Prog}_{Div}$ - Rate$^{Prog}_{FullDiff}$ - Rate$^{Prog}_{Paneth}$)] $\times$ DDC$_{PROG}$(t)

   Paneth(t+1) = Paneth(t) + [Prog(t) $\times$ Rate$^{Prog}_{Paneth}$ - Paneth(t) $\times$ Rate$^{Paneth}_{Apoptosis}$] $\times$ DDC$_{PANETH}$(t)

   DDC$_{STEM}$(t) $\neq$ 0; DDC$_{PROG}$(t) $\neq$ 0; DDC$_{DIFF}$(t) $\neq$ 0, DDC$_{PANETH}$ $\neq$ 0;

   Stem(t_0), Prog(t_0), Paneth(t_0) = known values, being equal respectively to, the stem, progenitor and Paneth cell number at initial time, i.e. t_0;

   Diff(t_0) = the number of fully-differentiated cells, (excepting Paneth cells that are included in 'Paneth' group), at initial time, (t_0).

2. Relationships showing no variations on cell number over time:

   Stem(t+1) - Stem(t) = 0

   Prog(t+1) - Prog(t) = 0

   Diff(t+1) - Diff(t) = 0

   Paneth(t+1) - Paneth(t) = 0

3. Calculation of relationship between parameters:

Stem(t+1) - Stem(t) = 0 $\Rightarrow$

$\text{Rate}^{Stem}_{Apoptosis} = \text{Rate}^{Stem}_{SymDiv}$ (same as for the colon crypt)

Prog(t+1) - Prog(t) = 0 $\Rightarrow$

$\text{Prog}(t) + [\text{Stem}(t) \times \text{Rate}^{Stem}_{AsymDiv} + \text{Prog}(t) \times (\text{Rate}^{Prog}_{Div} - \text{Rate}^{Prog}_{FullDiff} - \text{Rate}^{Prog}_{Paneth})]$ $\times \text{DDC}_{PROG}(t) - \text{Prog}(t) = 0$ (given $\text{DDC}_{PROG}(t) \neq 0$) $\Rightarrow$

$\text{Stem}(t) \times \text{Rate}^{Stem}_{AsymDiv} + \text{Prog}(t) \times (\text{Rate}^{Prog}_{Div} - \text{Rate}^{Prog}_{FullDiff} - \text{Rate}^{Prog}_{Paneth}) = 0 \Rightarrow$

$\text{Stem}(t) \times \text{Rate}^{Stem}_{AsymDiv} + \text{Prog}(t) \times (\text{Rate}^{Prog}_{Div} - \text{Rate}^{Prog}_{Paneth}) = \text{Prog}(t) \times \text{Rate}^{Prog}_{FullDiff}$

$\Rightarrow$

At initial time, i.e. t_0, $\Rightarrow$

$\text{Rate}^{Prog}_{FullDiff} = \frac{Stem(t_0)}{Prog(t_0)} \times \text{Rate}^{Stem}_{AsymDiv} + \text{Rate}^{Prog}_{Div} - \text{Rate}^{Prog}_{Paneth}$

Diff(t+1) - Diff(t) = 0 $\Rightarrow$

$\text{Diff}(t) + [\text{Prog}(t) \times \text{Rate}^{Prog}_{FullDiff} - \text{Diff}(t) \times \text{Rate}^{Diff}_{Apoptosis}] \times \text{DDC}_{DIFF}(t)$ Diff(t) = 0 $\Rightarrow$

$\text{Rate}^{Diff}_{Apoptosis} = \frac{Prog(t_0)}{Diff(t_0)} \times \text{Rate}^{Prog}_{FullDiff}$ (same as for colon crypt)

At initial time, i.e. t_0, $\Rightarrow$

$\text{Rate}^{Diff}_{Apoptosis} = \frac{Prog(t_0)}{Diff(t_0)} \times [\frac{Stem(t_0)}{Prog(t_0)} \times \text{Rate}^{Stem}_{AsymDiv} + \text{Rate}^{Prog}_{Div} - \text{Rate}^{Prog}_{Paneth}]$

Paneth(t+1) - Paneth(t) = 0 $\Rightarrow$

$\text{Paneth}(t) + [\text{Prog}(t) \times \text{Rate}^{Prog}_{Paneth} - \text{Paneth}(t) \times \text{Rate}^{Paneth}_{Apoptosis}] \times \text{DDC}_{PANETH}(t) -$ Paneth(t) = 0 $\Rightarrow$

$\text{Prog}(t) \times \text{Rate}^{Prog}_{Paneth} - \text{Paneth}(t) \times \text{Rate}^{Paneth}_{Apoptosis} = 0$ (given $\text{DDC}_{PANETH} \neq 0$).

At initial time, i.e. t_0, $\Rightarrow$

$\text{Rate}^{Paneth}_{Apoptosis} = \frac{Prog(t_0)}{Paneth(t_0)} \times \text{Rate}^{Prog}_{Paneth}$

Appendix E

List of publications

Journal

- Roznovăţ, I. A. and Ruskin, H. J., A Computational Model for Genetic and Epigenetic Signals in Colon Cancer, *Interdisciplinary Sciences: Computational Life Sciences*, 5, pp. 175 - 186, 2013. DOI: 10.1007/s12539-013-0172-y.

Conference Proceedings

- Roznovăţ, I. A. and Ruskin, H. J., Methylation Inhibitors and Carcinogens in an Agent-based Model for Colon Crypt Dynamics during Cancer Development, In *Proceedings of the UKSim/AMSS 7th European Modelling Symposium (EMS2013)*, Manchester, UK, 20th-22nd November 2013, pp. 144 - 149. Published by IEEE Computer Society - Conference Publishing Service (CPS). DOI: 10.1109/EMS.2013.27.

- Roznovăţ, I. A. and Ruskin, H.J., A Computational Model for Genetic and Epigenetic Signals in Colon Cancer, In *Proceedings of the 2012 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW)*, Philadelphia, USA, 4th-7th October 2012, pp. 188 - 195. doi 10.1109/BIBMW.2012.6470302.

- Roznovăţ, I. A. and Ruskin, H.J., Modelling the Genetic and Epigenetic Signals in

Colon Cancer using a Bayesian Network, In *Proceedings of the 12th European Conference on Complex Systems (ECCS'12)*, Springer Proceedings in Complexity, Brussels, Belgium, 3rd-7th September 2012, pp. 1059-1062. doi 10.1007/978-3-319-00395-5_126.

Book of abstracts

- Szczesna, K., Sandoval, J., Modhukur, V., Kull, M., Rajashekar, B., Perrin, D., Barat, A., Raghavan, K., Roznovăţ, I. A., Huertas, D., Vilo, J. and Ruskin, H. J., Complexity of Interdependent Epigenetic Signals in Cancer Initiation, Book of abstracts for the *Complexity-NET projects: Interdisciplinary Challenges for Complexity Science* Workshop, European Conference on Complex Systems (ECCS'13), Barcelona, Spain, 16th - 20th of September 2013, pp. 16 - 23.

Magazines

- Roznovăţ, I. A. and Ruskin, H. J., Interdependencies of Genetic and Epigenetic Events in a Computational Model for Colon Cancer Dynamics, *ERCIM News*, 95, pp. 42 - 43, October 2013.

Abstracts

Roznovăţ, I. A. and Ruskin, H. J. (2013). *A Computational Model for Genetic and Epigenetic Signals in Colon Cancer*

Cancer, a class of diseases, characterized by abnormal cell growth, has one of the highest overall death rates world-wide. Its development has been linked to aberrant genetic and epigenetic events, affecting the regulation of key genes that control cellular mechanisms. However, a major issue in cancer research is the lack of precise information on tumour pathways, so that delineation of these and the processes underlying disease proliferation

is an important area of investigation. A computational approach to modelling malignant system events can help to improve understanding likely 'triggers', i.e. initiating abnormal micro-molecular signals that occur during cancer development. Here, we introduce a network-based model for genetic and epigenetic events observed at different stages of colon cancer, with a focus on the gene relationships and tumour pathways. Additionally, we describe a case study on tumour progression recorded for two gene networks on colon cancer, *carcinoma in situ*. Our results to date showed that tumour progression rate is higher for a small, closely-associated network of genes than for a larger, less-connected set; thus, disease development depends on assessment of network properties. The current work aims to provide improved insight on the way in which aberrant modifications characterize cancer initiation and progression. The framework dynamics are described in terms of interdependencies between three main layers: genetic and epigenetic events, gene relationships and cancer stage levels.

Roznovăţ, I. A. and Ruskin, H. J. (2013). ***Methylation Inhibitors and Carcinogens in an Agent-based Model for Colon Crypt Dynamics during Cancer Development***

Cancer, (uncontrolled cell proliferation), has demonstrably high impact on human life: complex feelings and lifestyle changes, caused by malignancy, affect not only patients, but also family and friends. Cancer origin has been linked to genetic and epigenetic abnormalities that target stem cells. Here, we introduce an agent-based model for colon crypt dynamics, where the agents represent three cell types: *stem*, *progenitor* and *differentiated* cells. Additionally, we describe two test cases developed to analyse the influence of epigenetic inhibitors and carcinogens in tumour development. The focus is on the way in which DNA methylation patterns can be affected by de-regulations in cellular mechanisms in the colon crypt during cancer initiation and progression. Results to date have shown significant differences in average DNA methylation level for crypts, where epigenetic inhibitors have been temporarily or permanently present, in comparison to those with no such inhibitors.

Considerable variation in methylation level has also been observed in systems affected by carcinogens, compared to those unaffected.

Roznovăţ, I. A. and Ruskin, H. J. (2012). ***A Computational Model for Genetic and Epigenetic Signals in Colon Cancer***

`http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6470302`

Cancer, a class of diseases, which are characterized by abnormal cell growth, has one of the highest overall death rates world-wide. Its development has been linked to genetic and epigenetic events that aberrantly affect the regulation of key genes that control cellular mechanisms. However, a distinct problem in cancer research is the lack of precise information on tumor pathways, so the delineation of these and the processes underlying disease proliferation is an important area of investigation. This has motivated the effort to build computational models to help understand the molecular changes that lead to one or more types of malignancy. The current work aims to develop a prototype network-based model to describe cancer initiation and progression, with a focus on three main layers: micro-molecular events, gene interactions and cancer stages.

Roznovăţ, I. A. and Ruskin, H. J. (2012). ***Modelling the Genetic and Epigenetic Signals in Colon Cancer using a Bayesian Network***

`http://link.springer.com/chapter/10.1007%2F978-3-319-00395-5_126`

Cancer, the unregulated growth of cells, has been a major area of focus of research for years due to its impact on human health. Cancer development can be traced back to aberrant modifications in genetic and epigenetic mechanisms within the body over time. Given time and cost implications of human genome experimentation, computational modeling is increasingly being employed to improve understanding of mechanisms which determine cancer initiation and progression. Here, we introduce a network-based model for genetic

and epigenetic signals in colorectal cancer, with the focus on the gene level and tumor pathways. The current framework also considers the influence of ageing for micromolecular events in cancer development.

Szczesna, K., Sandoval, J., Modhukur, V., Kull, M., Rajashekar, B., Perrin, D., Barat, A., Raghavan, K., Roznovăţ, I. A., Huertas, D., Vilo, J. and Ruskin, H. J. (2013). ***Complexity of Interdependent Epigenetic Signals in Cancer Initiation***

`http://www.nwo.nl/en/research-and-results/programmes/complexity/meetings/eccs+meeting+2013`

The aim of the Complexity of Interdependent Epigenetic Signals in Cancer Initiation network is to determine epigenetic signals implicated in cancer initiation in one of the best-characterized mouse carcinogenesis systems, the multistage skin cancer progression model. This has been achieved through efforts on three interconnected layers (a) wet-lab experiments, (b) bioinformatics assisted analysis, (c) computer-based modeling. We have produced DNA methylation, gene expression, and histone modification data profiles from four different cell lines that mimics graduated stages of skin carcinogenesis. The results that we have generated after integration and computational models have been implemented will be presented in the meeting.

Roznovăţ, I. A. and Ruskin, H. J. (2013). ***Interdependencies of Genetic and Epigenetic Events in a Computational Model for Colon Cancer Dynamics***

`http://ercim-news.ercim.eu/en95/ri/interdependencies-of-genetic-and-epigenetic-events-in-a-computational-model-for-colon-cancer-dynamics`

The aim of our current work is to investigate the interdependencies of genetic and epigenetic mechanisms leading to aberrations in cancer initiation and progression. The objectives are to develop a computational model for colon cancer dynamics, linking microscopic

effects to macroscopic outcomes, and to analyse the impact of different risk factors on malignant tumour development.