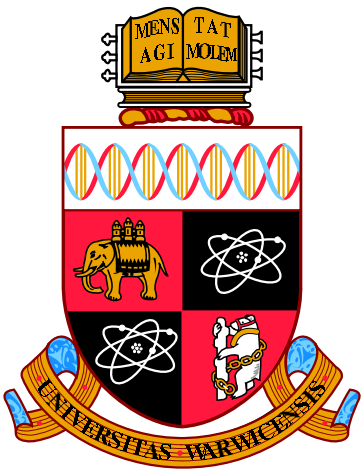


Modelling and Analysis of the Tumour Microenvironment of Colorectal Cancer

by

Violeta Naskova Kovacheva

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Systems Biology Doctoral Training Centre

September 2015

THE UNIVERSITY OF WARWICK

Contents

List of Tables

List of Figures

Acknowledgments

I would like to give special thanks to my supervisor Dr Nasir Rajpoot for his invaluable ideas and guidance. His constant support and encouragement, even from half-way across the world, have helped me complete this work. I would also like to thank my co-supervisor Dr Mike Khan for his input and guidance on the workings of the TIS machine. I am also very grateful to Prof David Epstein for his ideas and kind suggestions throughout my doctorate research.

I am very grateful to Dr David Snead, Dr Ian Cree, Dr Hesham El-Daly and Dr Yee Wah Tsang for helping me understand colon histopathology and providing histology data and invaluable feedback into the development of the colorectal microenvironment model. I am also thankful to my PhD advisory panel Dr Sara Kalvala and Dr Victor Sanchez for their guidance and valuable suggestions on my PhD progress. I would also like to thank Dr Richard Savage for his helpful ideas.

My special thanks to the Systems Biology DTC and the BBSRC for funding my PhD and supporting this project. I am thankful to the administrative staff both within the Systems Biology DTC and the Department of Computer Science for their support throughout these years.

I would like to thank all the members of my lab Dr Adnan Khan, Dr Shan-e-Ahmed Raza, Guannan Li, Nicholas Trahearn, Korsuk Sirinukunwattana, Mike T. Song, Samuel Jefferyes and Najah Alsubaie for their help and support.

I would like to especially thank my family for their unconditional love and support throughout my life. Finally, I would like to thank my boyfriend Archer Sapte for his love, support, understanding and listening to my rants.

Declarations

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. I declare that, except where acknowledged, the material contained in the thesis is my own work, and has not been previously published for obtaining an academic degree.

Violeta N. Kovacheva

September 25, 2015

List of Publications

- V. N. Kovacheva, D. Snead, N. M. Rajpoot, A Model of the Spatial Microenvironment of the Colonic Crypt, In Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on, 172–176, April 2015.
- V. N. Kovacheva, K. Sirinukunwattana, and N. M. Rajpoot. A bayesian framework for cell-level protein network analysis for multivariate proteomics image data. In SPIE Medical Imaging, pages 904110–904110. International Society for Optics and Photonics, 2014.
- V. N. Kovacheva, D. B. A. Epstein, N. M. Rajpoot, Advances in Discovery of Complex Biomarkers for Colorectal Cancer Using Multiplexed Proteomics Imaging, Oncology News, 8(6): 191–193, 2014.
- V. N. Kovacheva, A. M. Khan, M. Khan, D. B. A. Epstein, N. M. Rajpoot, DiSWOP: A Novel Measure for Cell-Level Protein Network Analysis in Localised Proteomics Image Data, Bioinformatics, 30(3): 420–427, 2014.

Abstract

New bioimaging techniques have recently been proposed to visualise the colocation or interaction of several proteins within individual cells, displaying the heterogeneity of neighbouring cells within the same tissue specimen. Such techniques could hold the key to understanding complex biological systems such as the protein interactions involved in cancer. However, there is a need for new algorithmic approaches that analyse the large amounts of multi-tag bioimage data from cancerous and normal tissue specimens in order to begin to infer protein networks and unravel the cellular heterogeneity at a molecular level.

In the first part of the thesis, we propose an approach to analyses cell phenotypes in normal and cancerous colon tissue imaged using the robotically controlled Toponome Imaging System (TIS) microscope. It involves segmenting the DAPI-labelled image into cells and determining the cell phenotypes according to their protein-protein dependence profile. These were analysed using two new measures, Difference in Sums of Weighted cO-dependence/Anti-co-dependence profiles (DiSWOP and DiSWAP) for overall co-expression and anti-co-expression, respectively. This approach enables one to easily identify protein pairs which have significantly higher/lower co-dependence levels in cancerous tissue samples when compared to normal colon tissue. The proposed approach could identify potentially functional protein complexes active in cancer progression and cell differentiation.

Due to the lack of ground truth data for bioimages, the objective evaluation of the methods developed for its analysis can be very challenging. To that end, in the second part of the thesis we propose a model of the healthy and cancerous colonic crypt microenvironments. Our model is designed to generate realistic synthetic fluorescence and histology image data with parameters that allow control over differentiation grade of cancer, crypt morphology, cellularity, cell overlap ratio, image resolution, and objective level. The model learns some of its parameters from real histology image data stained with standard Hematoxylin and Eosin (H&E) dyes in order to generate realistic chromatin texture, nuclei morphology, and crypt architecture. To the best of our knowledge, ours is the first model to simulate image data at subcellular level for healthy and cancerous colon tissue, where the cells are organised to mimic the microenvironment of tissue *in situ* rather than dispersed cells in a cultured environment. The simulated data could be used to validate techniques such as image restoration, cell segmentation, cell phenotyping, crypt segmentation, and

differentiation grading, only to name a few. In addition, developing a detailed model of the tumour microenvironment can aid the understanding of the underpinning laws of tumour heterogeneity.

In the third part of the thesis, we extend the model to include detailed models of protein expression to generate synthetic multi-tag fluorescence data. As a first step, we have developed models for various cell organelles that have been learned from real immunofluorescence data. We then develop models for five proteins associated with microsatellite instability, namely MLH1, PMS2, MSH2, MSH6 and p53. The protein models include subcellular location, which cells express the protein and under what conditions.

Abbreviations

AHC	Agglomerative Hierarchical Clustering
APC	Affinity Propagation Clustering
CMP	combinatorial molecular phenotype
CIN	chromosomal instability
CSC	cancer stem cells
CRA	colorectal adenocarcinoma
CRC	colorectal cancer
DAPI	4',6-diamidino-2-phenylindole
DiSWAP	Difference in Sum of Weighted Anti-co-dependence profiles
DiSWOP	Difference in Sum of Weighted cO-dependence profiles
FFPE	formalin-fixed, paraffin-embedded
GBHC	Gaussian Bayesian Hierarchical Clustering
H	Healthy
H&E	Haematoxylin and Eosin
HPA	Human Protein Atlas
IHC	Immunohistochemistry

IF	Immunofluorescence
IMS	Imaging mass spectrometry
LS	Lynch syndrome
MCEP	molecular co-expression pattern
MD	moderately differentiated
MIC	maximal information coefficient
MMR	mismatch repair
MSI	microsatellite instability
PD	poorly differentiated
PPD	Protein-protein dependence
PPDP	Protein-protein dependence profile
PSF	point spread function
TIS	Toponome Imaging System
TNM	tumor node metastasis
WD	well differentiated

Chapter 1

Introduction

In order to understand biological processes, there is an increasing need to quantitatively characterise phenotypes [1]. This is particularly true if we are to fully understand how cancers form, develop and spread through the body. Cancer refers to a group of diseases involving dynamic changes in the genome resulting in defects in regulatory circuits that govern normal cell proliferation, differentiation and death. It is widely believed that tumour development proceeds in a manner similar to Darwinian evolution, where a succession of genetic changes, each conferring a type of growth advantage, leads to a progressive conversion of normal cells into cancer cells [2]. Despite there being over 100 distinct types of cancer, there are certain features that are shared by most, if not all, tumours. These include self-sufficiency in growth signals, evading growth suppressors, evasion of programmed cell death (apoptosis), enabling replicative immortality, inducing angiogenesis, activating invasion and metastasis, reprogramming of energy metabolism, and evading immune destruction [3, 4]. It is now also understood that normal cells, forming the tumour-associated stroma (connective tissue between glandular or tumour regions), are active participants in tumourigenesis rather than passive bystanders. It has been shown that they are critical for the development and expression of some of the hallmarks of cancer [5]. Hence, in order to understand the biology of tumour, we need to encompass the contributions of the "tumour microenvironment" rather than considering tumours as insular masses [4].

Modern cancer treatment is based on accurate tissue diagnosis of samples obtained from needle biopsy or surgical excision. Morphological interpretation, including shape, texture and spatial context of the various cells in histological sections forms the basis of diagnosis and prognosis for cancer [6]. The histopathologist assigns a histopathological classification as part of the clinical diagnosis by microscopically

analysing routinely stained tissue sections, firstly at low magnification to observe the overall staining pattern, and then, if necessary, at higher magnification. Immunostaining and selected molecular tests are also used to help establish a specific cancer diagnosis. The most widely used system for colorectal cancer (CRC) staging is the tumour - node - metastasis (TNM) classification, which considers the main parameters of local growth, the presence of cancer cells in regional lymph nodes and evidence for distant spread [7]. The TNM classification has been used for over 80 years and has been continually refined with the seventh edition of the guidelines being issued by the Union for International Cancer Control in 2009. However, it is now recognised that the clinical outcome can vary significantly among patients within the same stage [8]. With the increasing understanding of cancer pathology, additional features have been reported to improve the prognostic value. These include lymphatic and vascular invasion, tumour budding and immune response [9, 10, 11]. Similarly, intratumour genetic heterogeneity is now recognised as a fundamental driver of therapeutic resistance in most human cancers [12].

Immunohistochemistry (IHC) studies are also a major source of data on cell phenotypes, protein expression and location. However, performing most of the clinical work using visual examination to assess changes is a difficult, subjective and time-consuming task. It also leaves important questions unanswered, such as what makes the observed structures, and if there are any specific protein interactions associated with the observed morphological phenotypes. With the improvements in high-throughput acquisition technologies like tissue microarrays and automated whole-slide scanners, automated analysis of tissue images is highly desirable, and studies have shown that quantitative software can detect changes in disease states that are missed by visual inspection [13]. Several frameworks have been developed to statistically characterise the histological features of the nuclei and cells [6, 14]. Usually these methods focus on the cell nuclei as the nucleus can hold the key to understanding cell function [15].

In this thesis we focus on CRC which accounts for about 10% of all cancers (after exclusion of non-melanoma skin cancer) and it is the fourth leading cause of cancer death in the world [16]. CRC is the second leading cause of death from malignancy in the industrialised world [17]. Every year, nearly one million people world wide develop CRC, of which 50% die within 5 years [18]. Many Asian countries, including China, Japan, South Korea and Singapore, have experienced an increase of two to four times in the incidence of colorectal cancer during the past few decades. The rising trend in incidence and mortality from colorectal cancer is more striking in affluent than in poorer societies and differs substantially among ethnic groups [19].

1.1 Cancer Heterogeneity

Extensive genetic and phenotypic variation exists not only between tumours (inter-tumour heterogeneity) but also within individual tumours (intratumour heterogeneity). Oncologists are increasingly using molecular characterisation of a sample from a primary or metastatic tumour to guide treatment selection for the patient. Both inter- and intra-tumour heterogeneity have significant implications for the choice of biomarkers to guide clinical decision-making in cancer medicine [20] and can affect the patient outcome.

1.1.1 Intra-tumour Heterogeneity

Cancer is continuously revealed to be ever more complex than previously thought. The remarkable complexity and heterogeneity of cells within an individual tumour has been demonstrated by sequence analyses of cancer cell genomes [21, 22] and metabolomic and proteomic techniques largely based on mass spectrometry [23, 24]. These findings demonstrate the various functionally important cell phenotypes within any given tumour, including cancer cells [25, 26], cancer stem cells (CSCs) [27], stromal cells [28, 29], vascular endothelium [30], and immunocytes [31]. In addition to this, the cancer cell population may be even more heterogeneous than previously anticipated on the basis of clonal evolution, at least in part due to the effects of continuing lineage differentiation [21, 32]. Most of this diversity results from genomic instability which can arise through various routes, such as deregulated DNA replication, defects in chromosome segregation, or mutations in components of the DNA repair pathways. The genomic instability is thought to enhance intercellular heterogeneity, broadening the pool of cells that are subject to selection, and therefore the likelihood of selective expansion of multiple different subclones [33, 34]. Clonal dynamics may lead to the emergence of clinical resistance during disease progression despite the matching of targeted treatment to the mutation. Adding to the complexity are findings demonstrating that epigenetic coding within tumours can be highly heterogeneous and associated with tumour behaviour [35]. These results point to the importance of a side population of stem cell-like cancer cells that may be responsible for malignant behaviour [36, 27].

Intratumour heterogeneity may be so profound that the DNA copy number profiles of single tumour biopsies may more closely resemble those of tumours from different patients than those of adjacent biopsies of the same tumour [37]. All of the above pose serious challenge to conventional 'omics' technologies, which rely on the average expression profiles of genes or proteins in a tissue that has been destroyed

prior to the analyses [23, 38, 39, 24, 22]. The destruction of the tissue means that we can't draw conclusions correlating phenotype, function and morphology. Analysing microscopy images of the intact tissue with multiple protein markers could help address this issue.

1.1.2 Inter-tumour Heterogeneity

Colorectal cancer is a heterogeneous group of diseases which have distinctive genetic and epigenetic background [40]. It arises following one of the three pathways: the microsatellite instability (MSI), the chromosomal instability (CIN) or CpG island methylator phenotype (CIMP) pathways. Figure 1.1 summaries the current understanding of the molecular pathways involved in colorectal tumourigenesis [41]. The CIN pathway is the most common and is characterised by widespread imbalances in chromosome number and loss of heterozygosity (loss of an entire gene). It can result from accumulation of mutations in specific tumour suppressor genes and oncogenes that activate pathways critical for CRC such as chromosomal segregation, telomere stability, and the DNA damage response [42]. On the other hand, epigenetic instability is now believed to be implicated in the pathogenesis of almost one third of colorectal cancers [43]. Colorectal cancers with CIMP are characterised by epigenetic loss of function of tumour suppressor genes without mutations [44, 43]. The MSI pathway is discussed in more detail below.

Microsatellite Instability Pathway

Microsatellites are simple repeat sequences of 1 to 6 base pairs (also known as short tandem repeats) and are particularly prone to replication errors. Defects in one of the four DNA mismatch repair (MMR) genes (MLH1, MSH2, MSH6, PMS2) causes small changes in the number of repeats of microsatellites throughout the genome, hence manifesting MSI. Mismatch repair is a complex process that depends on the MMR proteins and multiple proteins that interact directly with DNA [45].The MSH2 and MSH6 proteins exist as a heterodimer, which forms a sliding clamp on the DNA strand. When MSH2 recognises a DNA base pair mismatch, it recruits the MLH1-PMS2 heterodimer. Repairing the mismatch requires coordinated activity of DNA repair proteins and the precise mechanisms are still under investigation [46, 47].

Around 15% of CRCs are characterised by a high degree of MSI (MSI-high) [46], and of these, about 1 in 5 (3%–5% overall, [48]) are due to Lynch syndrome (LS), previously known as hereditary nonpolyposis colorectal cancer (HNPCC). LS is the most common inherited colorectal cancer syndrome and it predisposes the

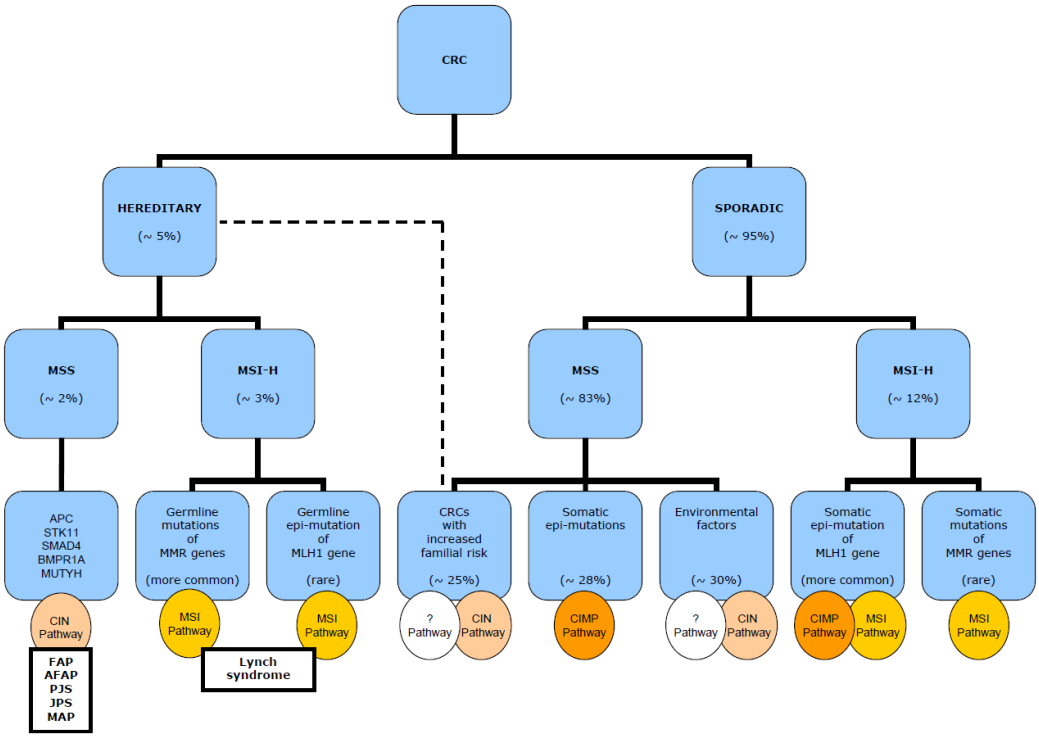


Figure 1.1: Molecular pathways in colorectal tumourigenesis. CRC, colorectal cancer; MSS, microsatellite stable; MSI-H, high level microsatellite instability; FAP, familial adenomatous polyposis; AFAP, attenuated FAP; PJS, Peutz-Jeghers syndrome; JPS, juvenile polyposis syndrome; MAP, MUTYH-associated polyposis; CIN, chromosomal instability pathway; MSI, microsatellite instability; CIMP, CpG island methylator phenotype; ?, pathways yet undefined (Image credit: [41]).

patient to cancers of multiple organ systems, including the gastrointestinal tract. It is important to identify patients with LS as it allows for increased surveillance of the affected individual and of potentially affected family members. Hence, preliminary screening is often done using IHC to detect MSI.

Most inherited MSI-high cancers are caused by epigenetic silencing of the MLH1 gene ($\approx$ 50%) or the MSH2 gene ($\approx$ 40%) [49]. Mutations in MSH6 and PMS2 occur only in about 10% of LS patients [50, 51]. Diagnosing LS is further hindered by findings that there are common missense mutations of MLH1 which may be associated with expression of an abnormal protein with normal IHC results [49]. On the other hand, sporadic MSI is usually caused by epigenetic silencing of MLH1. This typically occurs in CIMP-high tumours [52]. In addition, Samowitz et al. [53] considered the relationship between p53 mutations and MSI in CRCs. The study considered mutation in the p53 gene to be indicated by overexpression (over

50% of tumour cells expressing) of the protein in IHC data. They found that p53 overexpression occurred in 56% of stable tumours and only 20% of unstable tumours.

MSI tumours have a more favourable prognosis and are less prone to lymph node or distant metastasis [54]. This could relate to the high numbers of tumour infiltrating immune cells observed in these kind of tumours. Furthermore, MSI has been associated with a lack of response to fluorouracil-based adjuvant chemotherapy [55], although these findings do not currently influence the patient therapy [56].

Since the 1980s, it has been recognised that cancers arising in different parts of the colon involve different genetic mechanisms [57, 58]. For instance, the Lynch syndrome is most commonly found in the proximal colon (the right side of the colon). In contrast, familial adenomatous polyposis (FAP) tends to show more polyps in the left colon and arises in patients with inherited mutations in the Adenomatous polyposis coli (APC) gene, which has been the centre of the original Fearon-Vogelstein model of colorectal tumourigenesis [59] that forms the basis of the CIN pathway. This suggests that there are epigenetic or environmental factors playing a role in the development of genomic instability.

The main goal of this research is to develop quantitative frameworks for studying tumour heterogeneity. We first propose a framework that can identify cell phenotypes with different protein-protein co-dependence and highlight protein pairs that exhibit different levels of interaction in normal and cancerous tissue. We then develop a model of the colorectal tumour microenvironment after quantitatively studying inter- and intra-tumour heterogeneity. The model attempts to mimic intra-tumour heterogeneity in the synthetic data, allowing us to better understand the underlying principles of tumour heterogeneity and quantitatively evaluate image analysis frameworks developed for histology and fluorescence data.

The rest of this chapter is organised as follows. In Section 1.2 a detailed description of the normal architecture of colon tissue is presented, followed by a description of how the architecture changes as cancer develops and becomes more malignant, and how it is graded in clinical practice. Section 1.3 outlines the motivation for this work. Section 1.4 briefly presents the thesis organisation.

1.2 Tissue Architecture

The tumour microenvironment is a complex, dynamic environment, consisting of cells of various types, soluble factors, signalling molecules and an extracellular matrix that can promote tumour growth and invasion, as well as protect the tumour from the host immune system [60]. The importance of the tumour microenvironment has only

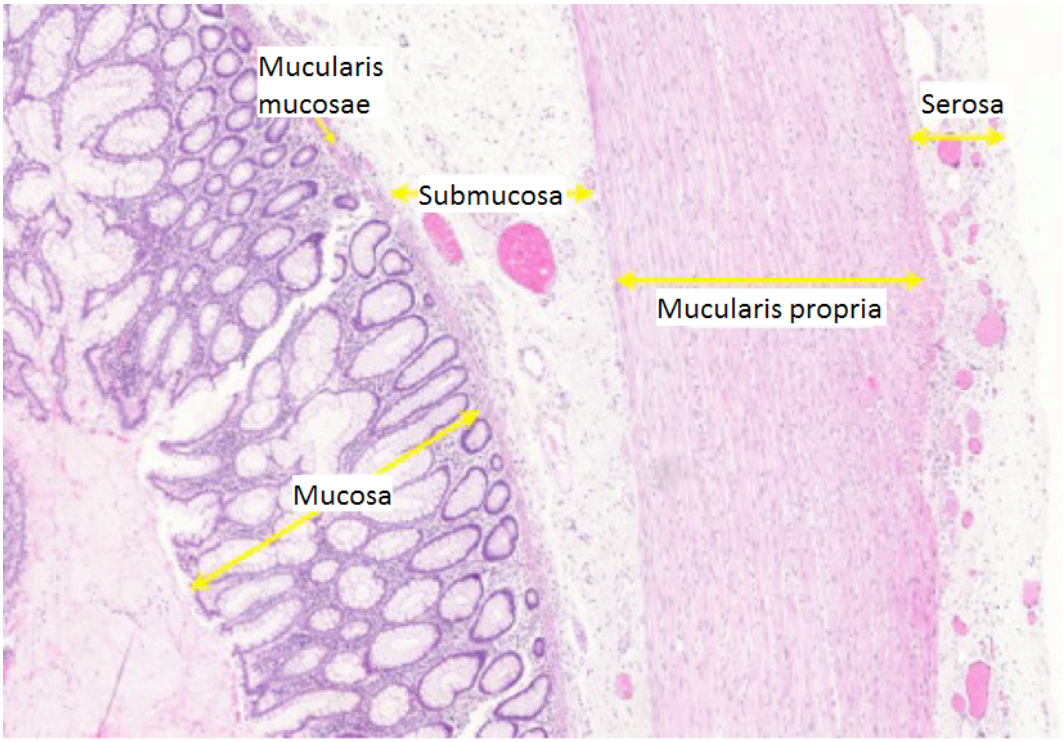


Figure 1.2: The histological structure of healthy colon tissue.

recently begun to be appreciated as the view of cancer diseases has shifted from a cell-autonomous condition, whereby epithelial cells mostly comprise the tumour, to a complex multicellular disease involving both epithelial cells and the surrounding stromal elements.

In this work, we look into understanding the microenvironment of healthy and cancerous colon tissue, as shown in Figure 1.2. The histological structure of the healthy colon consists of four major layers. The innermost layer is the mucosa. This is composed of a single layer of epithelium lining the innermost surface of the colon, crypts of Lieberkuhn, lamina propria, which is the connective tissue beneath the epithelium, and the lamina muscularis mucosae. Beyond the mucosa is the submucosa where blood vessels, nerves and lymph nodes can be found. Further outwards are a layer of smooth muscle called the muscularis propria and the serosa. In this study we focus on the mucosa as this is where tumours usually arise. Such epithelial tumours are known as adenocarcinomas.

The crypts consist mostly of three types of cells: epithelial (absorptive) cells, goblet cells and stem cells (Figure 1.3), and extend down to sit on the muscularis mucosae. Goblet cells predominate in the base of the glands, whereas the luminal surface is almost entirely lined by columnar absorptive cells [61]. The tall columnar

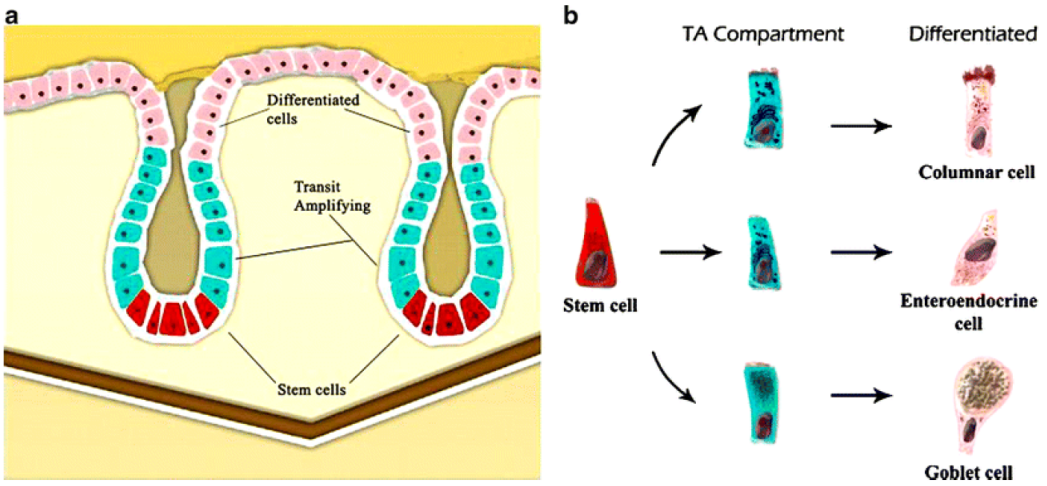


Figure 1.3: Colonic crypt organization. (a) In the epithelial lining of normal colonic mucosa, stem cells (red) are located at the bottom of the crypts. Upon asymmetrical divisions, the daughter cells undergoing differentiation migrate upward to give rise in turns to transit-amplifying (TA) precursors (light blue) and terminally differentiated cells (pink). (b) Cell types in the colon epithelium. Intestinal stem cells generate three epithelial cell types: the absorptive columnar cells, the hormone-producing enteroendocrine cells, and the mucous-producing goblet cells. (Image credit: [62]).

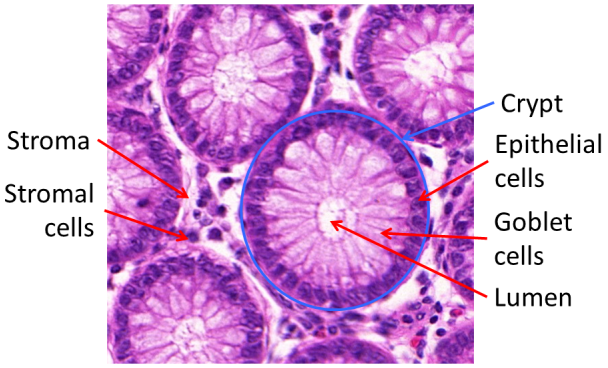


Figure 1.4: The structure of healthy colon tissue demonstrated using H&E markers.

absorptive cells have oval basal nuclei. In contrast, goblet cell nuclei are small and condensed. There are also stem cells at the base of the crypts, which continuously replace the epithelium. Lamina propria (also known as stroma) fills the space between the crypts. This contains some blood vessels, lymphocytes, plasma cells and fibroblasts (Figure 1.4).

As adenocarcinoma develops from normal tissue, the epithelium exhibits increased dysplasia (pre-malignant change in the epithelium with disordered growth and mutation). The epithelial nuclei become larger in size. There are also fewer

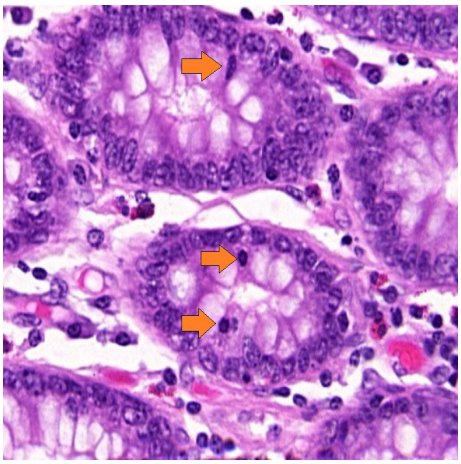
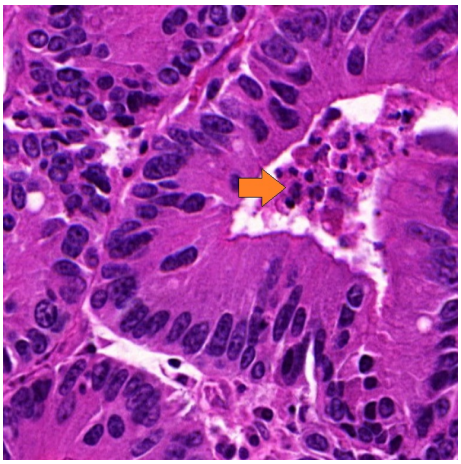
a)  b) 

Figure 1.5: Abnormal colorectal tissue showing mitotic figures (a) and areas of necrosis (dead tissue) within the glands (b). Objects of interest are identified by the orange arrows.

mucus-containing goblet cells, reflecting a lack of normal cellular differentiation. Often one can observe mitotic figures (Figure 1.5 (a)) or areas of necrosis (dead tissue) within the glands (Figure 1.5 (b)).

Histopathological grading of tumours is performed to provide some indication of their aggressiveness, which relates to prognosis and/or choice of treatment. The traditional tumour node metastasis (TNM) classification system of grading also used by the International Union Against Cancer (UICC) distinguishes four grades of differentiation:

G1 : well differentiated

G2 : moderately differentiated

G3 : poorly differentiated

G4 : undifferentiated

The percentage of tumour showing formation of gland-like structures can be used to define the grade. Well differentiated (grade 1) CRA lesions exhibit glandular structures in >95% of the tumour; moderately differentiated (grade 2) adenocarcinoma has 50-95% glands; poorly differentiated (grade 3) adenocarcinoma has 5-50%; and undifferentiated (grade 4) carcinoma has <5%. Grades 3 and 4 are often combined. This will be the case in the discussion and analysis presented here. There are some additional characteristics that can be used to differentiate between the different grades. Well differentiated tumours have well formed but slightly irregular

glands (Figure 1.6 (b)). Nuclei are basally oriented and exhibit slight atypia, which is characterised by variations in the size of the nuclei and visible nucleoli. In moderately differentiated CRAs there is still a glandular configuration, but the glands are irregular and often very crowded (Figure 1.6 (c)). There can be loss of mucin and budding of the crypts (asymmetric crypt division, Figure 1.7). One can also observe loss of nuclear polarity and increased nuclei atypia. On the other hand, in poorly differentiated tumours majority of the tumour (excluding the advancing edge) is sheets of cells without gland formation. Some glands may still be observed, but also single cells or clumps of cancerous cells, which are usually bigger than the stromal cells (Figure 1.6 (d)).

In practice, most colorectal adenocarcinomas ($\sim$70%) are diagnosed as moderately differentiated. Well and poorly differentiated carcinomas account for 10% and 20%, respectively [41]. Tumour grade is generally considered as a stage-independent prognostic variable, and high grade histology is associated with poor patient survival [63, 64, 65].

1.3 Aims of the Thesis

The tumour microenvironment is a complex and dynamic system. It consists of a multitude of components including cells of various types, signalling molecules and the extracellular matrix. It has been shown to play an important part in promoting tumour growth and invasion [60]. In this thesis, we aim to develop methods for analysing and modelling the tumour microenvironment of colorectal carcinoma (CRC) and studying cancer heterogeneity.

1.3.1 Cell-Level Protein Network Analysis

New bioimaging techniques have recently been proposed to visualise the co-location or interaction of several proteins within individual cells, displaying the heterogeneity of neighbouring cells within the same tissue specimen. Such techniques could hold the key to understanding complex biological systems such as the protein interactions involved in cancer. In this thesis, we aim to develop new algorithmic approaches that analyse the large amounts of multi-tag bioimage data from cancerous and normal tissue specimens in order to begin to infer protein networks and unravel the cellular heterogeneity at a molecular level.

As there is now evidence that rearrangement and different protein interactions, rather than up-or down- regulation of proteins could be key to generating new cell functionalities [66], we aim to consider the protein-protein dependence instead

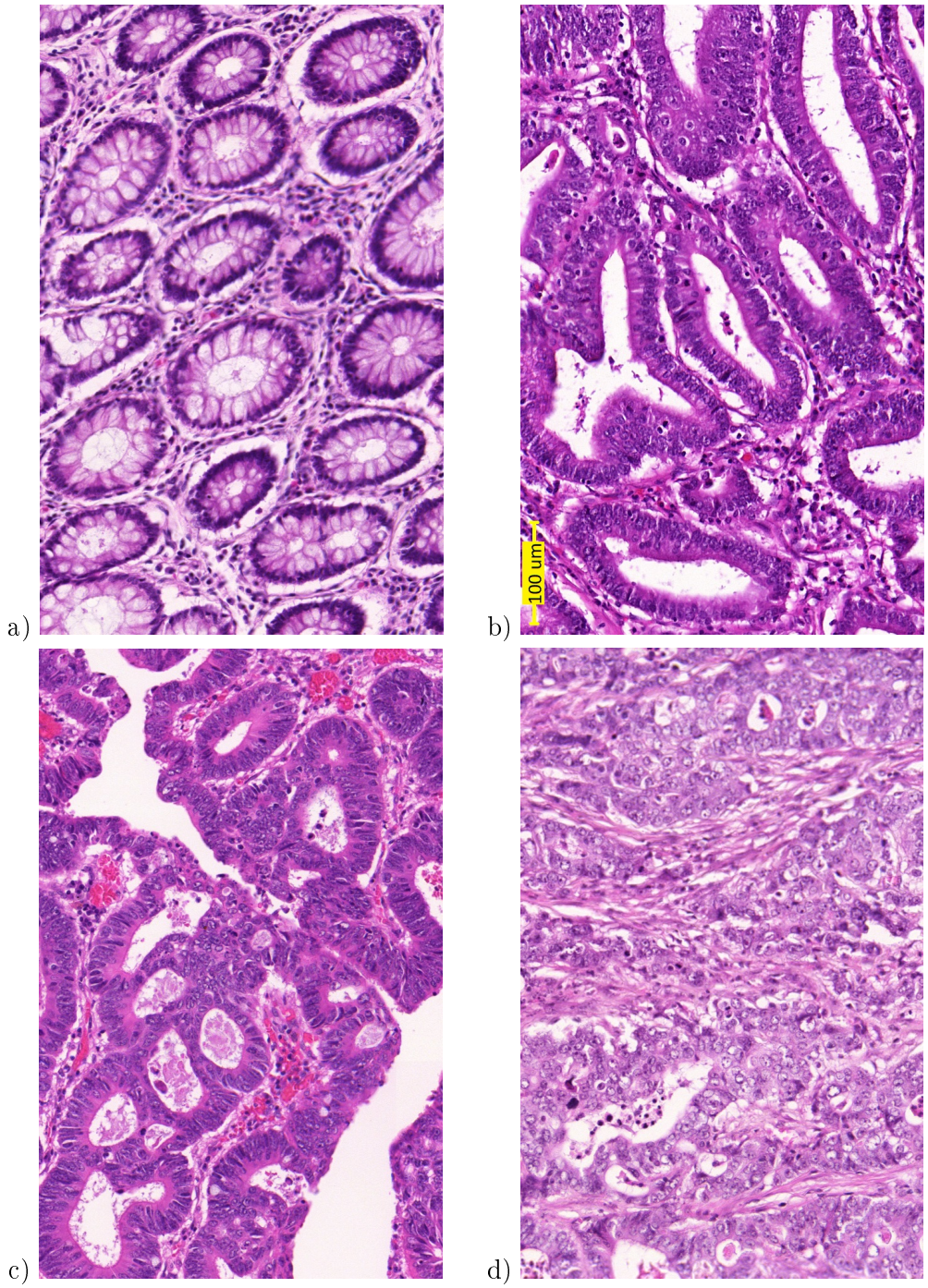


Figure 1.6: Examples of real images for different grades: (a) healthy tissue, (b) well, (c) moderately and (d) poorly differentiated cancerous tissue. Images are at 20× magnification. Size of the scalebar is 100 μm.

Figure 1.7: Example of crypt budding.

of the raw protein expression profiles. Furthermore, we perform the analysis at cell level rather than pixel level. This enables us to gain a better understanding of the heterogeneity within the cancer cell population.

Our aim is to investigate if multi-channel imaging techniques could be used to find new multiplex biomarkers to improve patient diagnosis by distinguishing between different types of samples using protein co-dependence. For this purpose we develop a framework for studying the localised protein networks.

1.3.2 Modelling the Tumour Microenvironment

There have been great advancements in the field of digital pathology and multiplex immunofluorescence (IF). As studies relying on analysis of the digital images produced by these technologies become popular, the validation of such analytical tools gains significance. A common approach for validation is to compare the algorithm's results with expert-labelled data. Nevertheless, the repeatability and accuracy of manual labelling can always be questioned due to human error sources [67] and the process is very time-consuming. We aim to address this problem by developing benchmark synthetic datasets for objectively validating and comparing these methods. In addition, developing a detailed model of the tumour microenvironment can aid our understanding of the underpinning laws of tumour heterogeneity.

Our aim here is to develop a realistic model of the tumour microenvironment by performing detailed quantitative study of real data and incorporating the various learned parameters into the model. In order to achieve this, we study not only how the overall architecture of the colon tissue changes from healthy into increasingly

more malignant cancer, but also take into account the distributions of various cellular phenotypes associated with each stage.

1.3.3 Sub-cellular Protein Expression

Finally, the thesis aims to study the tumour heterogeneity by combining the detailed study of the spatial microenvironment with the study of the molecular microenvironment. This is achieved by developing sub-cellular models for protein expression within the tissue architecture. The models include the sub-cellular location and strength of expression of the proteins, which types of cells express them and under what conditions (i.e. presence of mutation). Developing realistic models is achieved by detailed analysis of high-resolution confocal images of cell cultures, demonstrating the sub-cellular expression patterns, and histology images of CRA samples, allowing for analysis of cell phenotypes.

1.4 Thesis Organisation

Chapter 2 contains a brief review of current literature on multiplex imaging techniques, including the Toponome Imaging System (TIS), data from which has been used within the thesis, and other similar systems, as the analytical frameworks presented within the thesis can be easily generalised to other multiplex imaging data. It also reviews existing literature on generating synthetic fluorescence images and modelling protein expression.

In Chapter 3, we propose a framework to extract cell-level protein networks from multiplex IF data. The framework highlights protein pairs with different co-localisation patterns in healthy and cancerous tissue samples, which could potentially be useful cancer biomarkers.

Chapter 4 proposes a model of tumour heterogeneity capable of simulating IF and histology images of healthy colon tissue and CRCs of different differentiation grades.

The tumour heterogeneity model is then expanded to generate multiplex IF data in Chapter 5. In this chapter we consider methods for simulating protein expression patterns by considering a group of five proteins associated with MSI.

Finally, Chapter 7 concludes the thesis, discusses limitations of the work, possible application and future directions.

Chapter 2

Literature review

In this chapter, we review multiplex fluorescence imaging techniques and the methods developed for analysing such data. The chapter also reviews current methods for generating synthetic fluorescence data and frameworks for simulating protein expression.

2.1 Multiplex Imaging

A cell in a human tissue can be defined as an assembly of thousands of proteins which interact together to define cell functions [68, 69]. In order to understand cellular biology on a systems level, relationships between molecular components must be understood not only at a functional level but also localised in the spatial domain [70]. This is due to the fact that proximity of key proteins provides an indication of the possible existence of functional protein complexes. Furthermore, it is increasingly important to measure not just the average expression of molecules in homogenised tissue but also their spatial distribution while preserving cellular and tissue architectural features [71]. This results from the complexity of tissue samples studied in anatomic pathology. This complexity is made evident by multi-parameter detection methods such as gene/protein expression arrays and flow cytometry. The protein compositions can be decoded by using modern fluorescence imaging techniques. Most fluorescence microscopy techniques are limited to up to ten fluorescent tags which can point to simultaneous localisation of the corresponding biomolecules and protein structures inside the cells of a tissue specimen [72]. In order to more fully understand complex cellular systems, new bioimaging techniques have been recently proposed to visualise the colocation or interaction of several proteins in cells in intact tissue specimen. These include Toponome Imaging System (TIS) [73], MALDI imaging

[74], Raman microscopy [75], multi-spectral imaging methods [76], MxIF [77], and imaging mass cytometry [78, 79]. These techniques are discussed below.

2.1.1 Toponome Imaging System

TIS is an automated high-throughput technique able to co-map up to a hundred different proteins or other tag-recognisable bio-molecules onto the same pixel on a single tissue section [80]. It runs cycles of fluorescence tagging, imaging and soft bleaching *in situ* (Figure 2.1). While co-location does not necessarily imply interaction, it has been consistently found that clusters containing particular proteins are found in specific sub-cellular compartments, hence allowing such a hypothesis to be generated [68]. For instance, the spherical and the exploratory cell states of rhabdomyosarcoma cells had identical average protein profiles. In contrast, when sub-cellular protein clusters were determined, striking differences were found [66]. Hence, rearrangement, rather that up- or down-regulation of proteins is (or can be) key to generating new cell functionalities [80]. This shows the importance of co-dependence of proteins rather than abundance on its own. Also, co-dependence between two proteins is a potential indication for an interaction that is not necessarily direct. The importance of studying protein interactions is further highlighted by evidence that cancer proteins interact with higher number of proteins and tend to play a more central role in proteome networks [81, 82]. TIS has a sub-cellular maximum lateral resolution of 206×206 nm/pixel [68] which allows the determination of sub-cellular protein network architectures. The combination of proteomic information with spatial sub-cellular level topographical data in morphologically intact cells and tissues has been termed 'toponomics' [83, 80].

Biomarkers used in current clinical practice are limited to the simultaneous analysis of only a handful of proteins. They, therefore, fail to assess the true complexity of cancer, and the resulting biomarkers have a low prognostic value [22]. The capabilities of the TIS hold promise for developing a new generation of multiplex biomarkers [84] which could aid the development of personalised medicine. Studying the protein interactions in cancer could uncover previously unknown mechanisms of tumour formation and could identify new potential drug targets in the form of protein interactions.

The standard way to analyse TIS images is to apply a threshold to each image of the stack and so reduce it to binary values, representing a combinatorial molecular phenotype (CMP) [73]. The CMP code consists of either tag present (=1) or tag absent (=0) for each data point [85, 69, 73, 86, 80]. The CMPs are grouped together into 'CMP motifs' according to certain rules. All CMPs within a

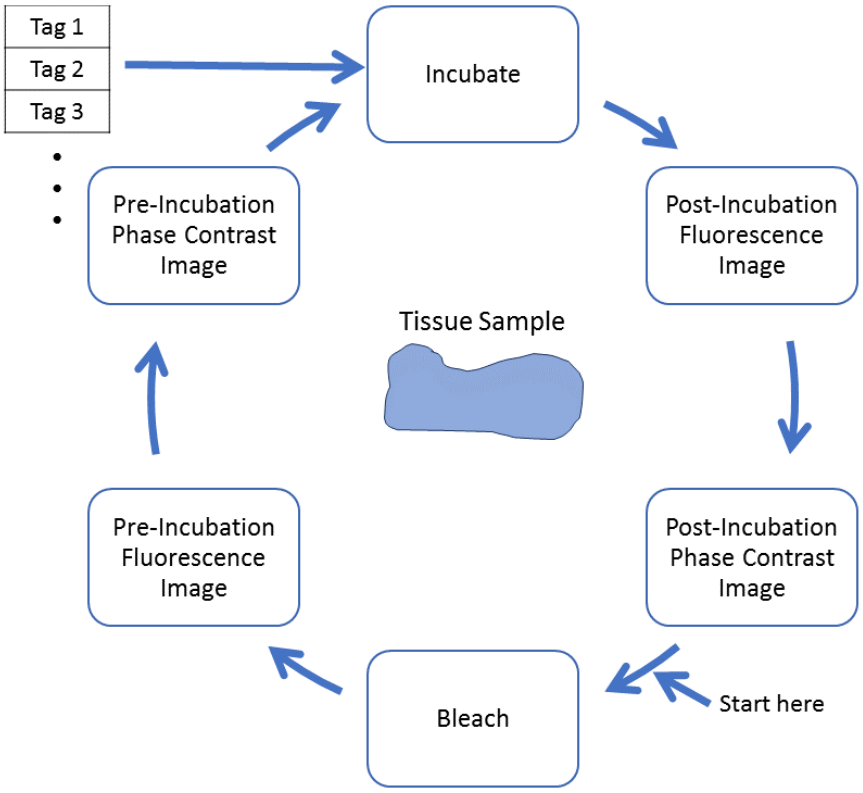


Figure 2.1: Toponome Imaging System (TIS) data acquisition cycle.

CMP motif contain at least one or more of the same protein (lead proteins), they never contain certain proteins (absent proteins), and they variably contain additional proteins (wild-card proteins). This suggests a hierarchical organisation, and lead proteins have been found to control protein network topology and (dys)function. Several studies have found that when a lead protein is blocked or down-regulated, the corresponding functional network disassembles [73, 80, 87]. This phenomenon has been observed experimentally in chronic neuropathic pain and in cultured tumour cells [73], mouse models for Amyotrophic Lateral Sclerosis (ALS) [88] and clinical trails for ALS [89]. Another study using TIS studied the immune system in CRC and showed that it induces a tremendous modification of protein expression profiles in the lumina propria [90].

It has been shown that TIS imaging can be used in cancer research for protein network mapping [68]. However, while thresholding is straightforward and can be performed objectively [91], by reducing the image to binary, a lot of potentially important information is lost. Recently, such non-threshold methods have been presented [92, 93]. These algorithms cluster molecular co-expression patterns (MCEPs)

on a pixel level and therefore fail to capture the variation at a cell level. This can be crucial when analysing cancerous samples due to the heterogeneity of cancer cells [22]. Furthermore, these algorithms are based on the raw expression levels, which are intensity dependent and hence may vary between different stacks. A similar approach is used in the Web-based Hyperbolic Image Data Explorer (WHIDE) [94], which allows analysis of the space and colocation using a H2SOM clustering [95]. While this tool is very effective at identifying molecular co-expression patterns, the cellular structure is lost and hence the method is unable to analyse the different cell phenotypes that may be present in the samples. More recently, focus has been shifted towards cell-level analysis. In a study by Khan et al. [96] cells were phenotyped after dimensionality reduction of their raw expression vector using t-Distributed Stochastic Neighbour Embedding (t-SNE) [97].

2.1.2 Other Techniques

While the frameworks presented in Chapter 3 have been developed for the analysis of data obtained by the TIS microscope, they are easily generalisable to other multiplex techniques.

Raman

Raman microscopy [98] can image protein and gene expression directly, i.e. without the need for labelling (Figure 2.2). Several variations of this have been developed and used for cellular imaging. These include resonant Raman scattering [99, 100], coherent anti-Stokes Raman scattering (CARS) [101, 102], and Fourier transform infra-red absorption (FTIR) [103, 104]. However, these techniques use wavelength bands which cannot identify specific proteins. They are insensitive to protein secondary structure and can only detect the number of protein CH_2 and CH_3 groups. Raman spectral images have been used to visualise mitochondrial distribution [105] and to distinguish normal and malignant cultured cell lines in a variety of cancers, including thyroid [106], lymphoma [107], cervical [108] and colorectal [109]. It has also proven effective in grading tissue biopsies from prostate cancer [110, 111]. However, due to its difficulty of identifying specific proteins and the low resolution, Raman microscopy fails to capture the true complexity of cancer and the variations in protein interactions. In addition, Raman imaging causes thermal damage to the tissue [112].

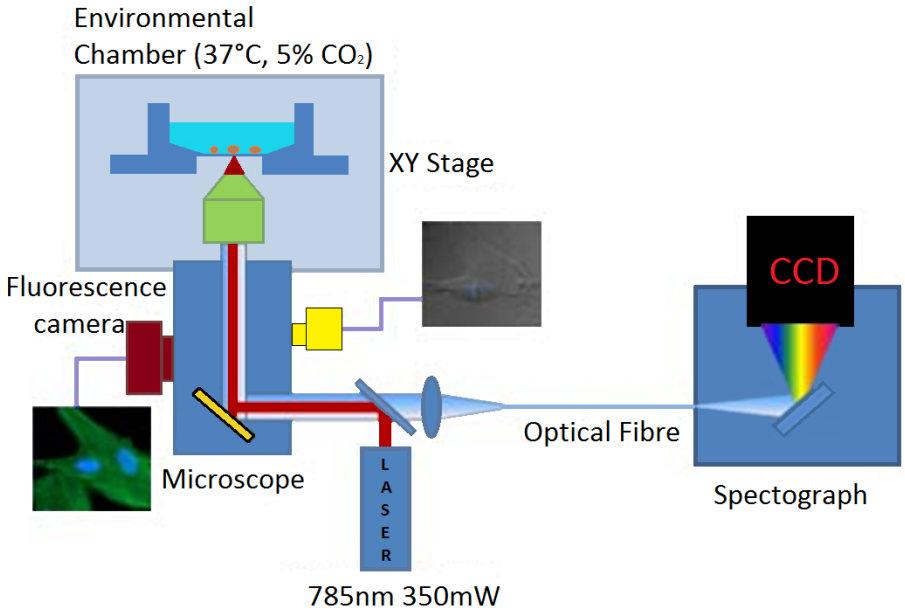


Figure 2.2: Typical diagram of Raman-based apparatus for investigating cells in their natural physiological conditions. The integration of the inverted microscope and environmental enclosure allows time-course CRMS imaging over extended periods of time. The fluorescence imaging enables label-based molecular-specific assays on the same cells (Image credit: [113]).

MALDI

Another emerging technique is matrix-assisted laser desorption/ionization (MALDI) imaging mass spectrometry (IMS) (Figure 2.3) [114]. It can determine the distribution of hundreds of unknown compounds in a single measurement [74]. Key advantages of this technology are that it does not require molecule-specific tags or chemical modifiers to facilitate detection and does not rely on any prior knowledge of the tissue proteome. It can achieve a lateral resolution of approximately 30–50 μm [115]. However, similarly to Raman microscopy, it destroys the sample and it is difficult to identify particular proteins. In addition, it has inherent limitations owing to its requirement for a crystalline chemical matrix [116, 117]. The matrix, combined with the instrument sensitivity, reduces achievable resolution and obscures the signal from elemental reporters. MALDI IMS has been used to investigate several forms of cancers including gliomas [118, 39], breast cancer [119, 120], prostate cancer [121, 122], colon cancer [123] and lung cancer [124]. Several studies have demonstrated the potential of the technology to identify new candidate biomarkers of disease [125, 126, 127, 128, 129]. Simple methods for analysing the data such as hierarchical clustering and principal component analysis have been successfully

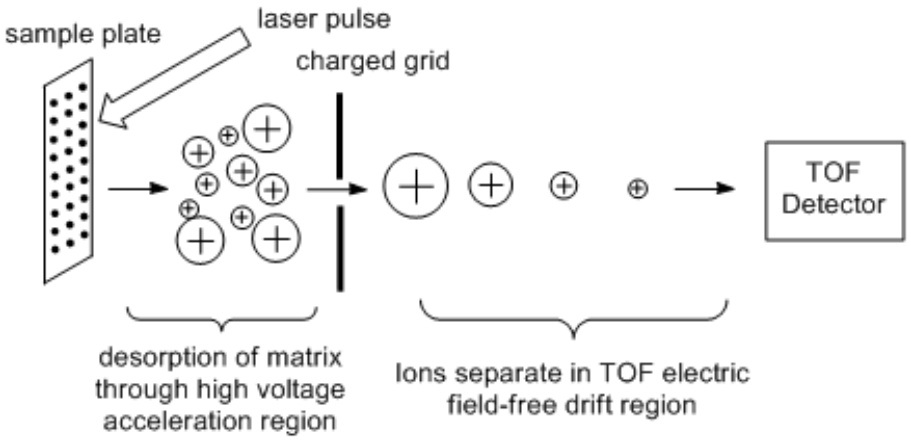


Figure 2.3: MALDI-TOF MS uses laser light in conjunction with a chemical matrix to impart a charge to the sample (ionization) in question and then accelerates the charged ions through a flight tube to the detector, which measures particle counts as a function of time. The time-of-flight (TOF) is directly proportional to the mass of the molecule (Image credit: [131]).

applied for classification of cancerous tissue [130, 128]. While Raman imaging and MALDI IMS provide spectral information, considering the height of selected peaks at each pixel can produce protein expression images similar to those obtained using fluorescent imaging. Hence, the same analytical methods can be applied after feature selection.

Multi-Spectral Imaging

The capabilities of standard fluorescent microscopy have been enhanced by the use of multi-spectral imaging (MSI), which enables one to resolve multiple overlapping fluorophores [132]. MSI enables the analysis of multi-color IHC, and drastically reduces the impact of contrast-robbing tissue autofluorescence common in formalin-fixed, paraffin-embedded (FFPE) tissues [71]. In fact, MSI approaches can result in a 99% reduction in auto- fluorescence and a concomitant reduction in limits of detection and increase in signal to noise ratio [133, 134]. The most common data analysis method applied to microscopy-based MSI data is linear unmixing [135]. The method is a least squares fit, or linear regression, of a number of given spectral shapes (basis functions; often termed a 'spectral library') into a spectrum acquired from the sample [134]. It is worth noting that obtaining correct and quantitative results from this process relies on having accurate examples of the spectral shapes of the fluorophores that will be found in the sample. A number of automated spectral decomposition methods have been developed to find the correct signatures [133, 136, 134]. However,

many prefer to create a set of singly stained spectral control samples from which the signatures can be created. In these cases it is necessary to use computational method to discriminate the spectral signature of the fluorophore from the autofluorescence in the sample [133, 136, 134]. MSI has been used for the multiplexed analysis of proteins [76] and the automated localisation and quantification of proteins [137] in tissue sections. It has also been used in numerous cancer studies including prostate [138, 139], ovarian [140, 141], breast [142, 143], liver [144, 145] and pancreas cancers [146, 147]. However, despite the use of MSI methods, these studies use only a handful of antibody markers due to the increasing complexity of the signal. On the other hand, a proposed method for exciting several fluorophores at the same wavelength and unmixing their emission signals could increase the number of biomolecules considered [76].

MxIF

More recently, the number of fluorescent antibodies that can be imaged has been greatly improved by the introduction of multiplex cyclic technologies. One example is MxIF [77], which uses iterative staining and chemical inactivation of the dyes (Figure 2.4). The study used 61 protein antigens to stain 747 colorectal cancer specimen placed in tissue microarrays. K-median cluster analysis of the data allowed clustering to phenotype segmented cells and studying the tumour heterogeneity. However, the system requires manual bleaching which could potentially damage the tissue. In addition, the conclusions in this study about pathways in colon cancer were drawn by only visual inspection of the phenotypes obtained and without considering any control samples.

Imaging Mass Cytometry

Another multiplex imaging technique is imaging mass cytometry (IMC), which combines IHC and immunocytochemical methods with high-resolution laser ablation with CyTOF mass cytometry [78, 79]. It uses antibodies labelled with rare earth metal to achieve simultaneous imaging of up to 32 proteins with resolution of 1 μm (Figure 2.5). IMC is highly quantitative as there is no sample autofluorescence, there are no matrix effects as found in MALDI, and there is no need for an amplification step such as is often needed in IHC. However, the antibodies used pose a limitation as often antibodies are not available for a given target or in the format needed for mass cytometry, and those that work well in single-plex assays may behave differently in multiplex assays [79]. IMC, combined with spanning-tree progression analysis of

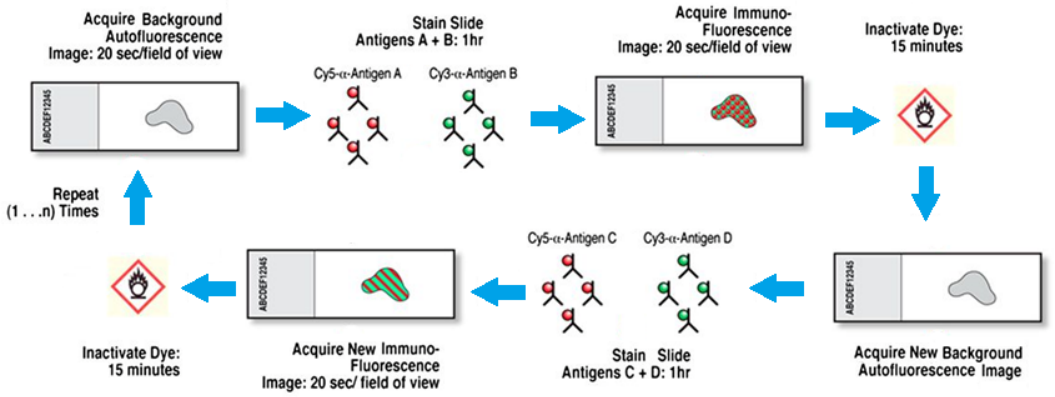


Figure 2.4: Overview of the MxIF technique. Background autofluorescence tissue images are acquired before subsequent application of fluorescent dye-conjugated primary antibodies. Stained images are then acquired, followed by dye inactivation and re-staining with new directly conjugated antibodies. New images are acquired, and the cycle is repeated until all target antigens are exhausted. Times associated with each step are indicated (Image credit: [77]).

density-normalized events (SPADE) analysis [148], has been applied to human breast cancer samples, allowing delineation of cell subpopulations and cell-cell interactions and highlighting tumour heterogeneity [79].

2.2 Synthetic Fluorescence Images

The recent emergence of Digital Pathology is generating massive amounts of digital histopathology image data produced by pathology laboratories embracing the digital slide scanning technologies. Similar trends can be observed with the popularisation of multiplex imaging. By consequence, the demand for development of robust analytical methods for quantitative morphometric analysis of the histopathology as well as multiplex IF image data is on the rise. The acceptance of analytical technologies for such image data depends largely on their ease-of-use and usefulness in terms of accurate quantification. A common approach for validation is to compare the algorithm's results with expert-labelled data. Nevertheless, the repeatability and accuracy of manual analysis can always be questioned due to human-based error sources [67] and the process is very time-consuming. In order to overcome these difficulties, several frameworks for synthetic fluorescent image data generation have been proposed. The simplest of these simulate populations of spots. Grigoryan et al. [149] proposed a toolbox which simulates each spot as a sphere randomly placed in 3D space and overlap between objects is allowed only under certain conditions. Man-

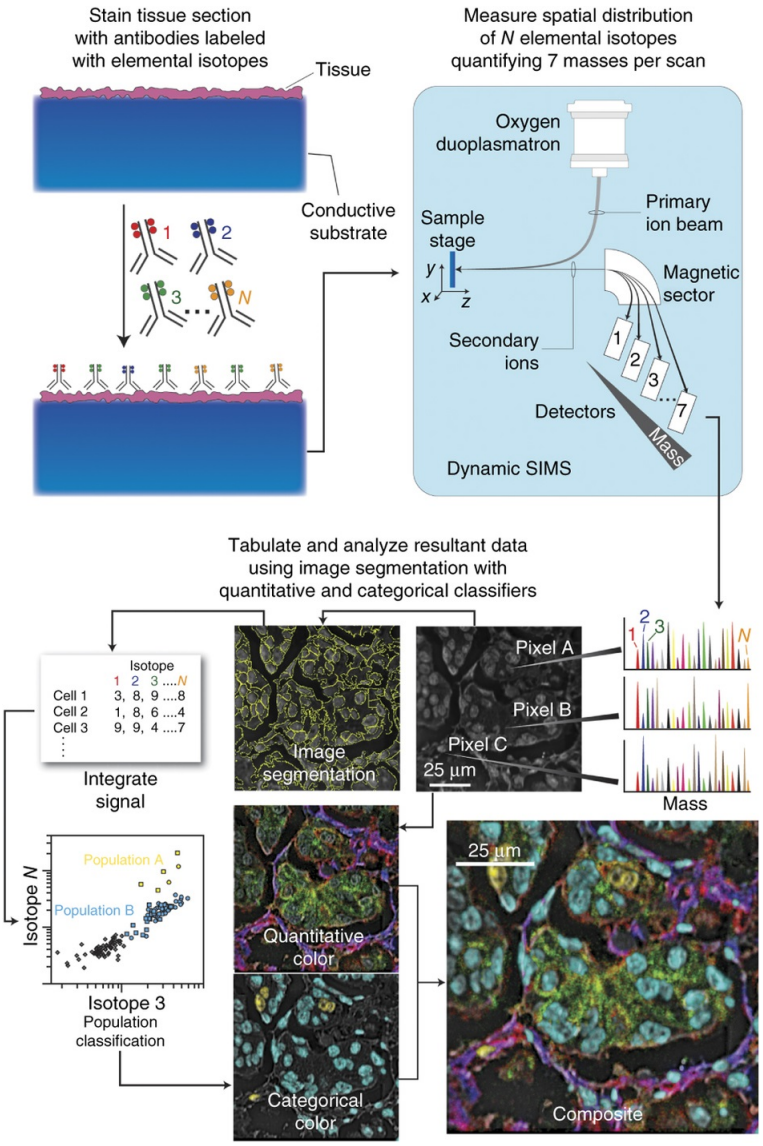


Figure 2.5: Overview of the IMC technique. Biological specimens, such as FFPE tissue or cell suspensions, are immobilised on a conductive substrate. Samples are subsequently stained with antibodies conjugated to unique transition element isotope reporters, dried and loaded under vacuum for MIBI analysis. The sample surface is rasterised with an oxygen primary ion beam that sputters the antibodyspecific isotope reporters native to the sample surface as secondary ions. Metal-conjugated antibodies are quantified via replicate scans of the same field of view, where up to seven metal reporters are measured with each scan. ROIs demarcating nuclear and cytosolic compartments of each cell are integrated, tabulated and categorised. Composite images comprised of pseudocolored categorical features and quantitative three-colour overlays are constructed to summarise multidimensional expression data (Image credit: [78]).

ders et al. [150], on the other hand, used a large grid of Gaussian-like 3D objects to verify their region-growing segmentation algorithm. Lockett et al. [151] used a more complex set of shapes, such as curved spheres, discs, bananas, satellite discs, and dumbbells. Graner and Glazier [152] simulated large cell populations by adopting the statistical large-Q Potts model to simulate the reorganisation of uniformly distributed cell-like objects. This ensured the natural shape and distribution of the cells. More recently, more realistic simulations have been presented. For example, Lehmussola et al. [153] designed a simulator called SIMCEP, which can simulate large 2D cell populations with realistically looking cytoplasm, nuclei and cell organelle. Svoboda et al. [154] generated a model to simulate fully 3D image data of cell nuclei of cell populations, with realistic distribution [155], and later of healthy colon tissue [156]. There has also been considerable advances in modelling time-lapse microscopy of cell populations [157] and evolving chromatin texture [158]. However, these models only include cell nuclei. In addition, the shape of the nuclei in the colon tissue model of [156] is not very realistic and does not reflect the variety of cell phenotypes found in real tissue. Heterogeneous cell populations expressing different protein markers can be simulated using the SimuCell toolbox [159]. The first method for simulating bright-field microscopy was proposed for generating synthetic cervical smears [160].

A different approach has been to use statistical generative models based on distributions of cellular morphology and organisation [161, 162, 163, 164]. These methods use imaging data to learn models that describe the relationship between compartments and the distribution of markers within them and build models based on the underlying population variances within the cell populations. These generative modelling methods have been combined in CellOrganizer, an open source tool currently developed by the Murphy Lab incorporating both parametric and non-parametric models [165]. The parametric models in CellOrganizer can be used to describe cell and nuclear shape, vesicular shape, frequency and location, and microtubule number, length and linearity. These models are limited by their parameterisation of the cell, as, for instance, B-spline models of nuclear shape are restricted to modelling star polygons [161, 162, 164]. On the other hand, the non-parametric large deformation diffeomorphic metric mapping [166] approach in CellOrganizer can be used to model arbitrary polygons or sets of polygons jointly assuming shapes can be properly registered using non-rigid image registration [167, 165]. This approach generates a "shape space" by reducing the dimensions of the pairwise differences in cell shape obtained through non-rigid image registration. The main advantage of the machine learning approach is that it could extract a more precise shape model

from real data, but the model cannot be described in precise mathematical terms. In addition, these generative models are restricted to individual cells in culture and, hence, fail to capture the heterogeneity and organisation of tissue cells.

When simulating microscopy data, it is essential that the method also includes consideration of the degradation of the image during the acquisition process. In optical microscopy, the first stage to consider is the signal transmission where characteristics of the environment can intervene with the signal. The most typical of these is the impulse response of the system, often called the point spread function (PSF) [168]. This drastically affects the final results and is usually simulated by convolving the incoming signal with the given or estimated PSF [154]. However, finding the PSF is commonly simplified by approximating it as a simple Gaussian kernel [151, 169, 150, 170]. Other phenomena that may affect the transmitted signal include uneven illumination [171], chromatic [172] and monochromatic aberrations [173], and reflection or refraction on lens surfaces that could result in artefacts. The second stage corresponds to the device sensors detecting the signal and converting it to a digital representation. The use of sensors introduces Poisson noise [151, 153, 169, 154], which can typically be observed even with the naked eye. In addition to this, the A/D converter and amplification electronics introduce additive white Gaussian noise [151, 170, 154]. If the equipment is not properly cooled, CCD detectors can also introduce fixed-pattern noise and blooming effect.

2.3 Protein Expression Models

Building accurate models for protein expression requires not only the chemical properties of the molecules involved, but also their spatial distributions. This is especially important for proteins because the subcellular location of a protein is so critical to its function that the same protein can have different functions at different locations [174]. In addition, for some proteins such as β-catenin [175] and NF-κB [176], the extent of localisation in the nucleus can be used as a biomarker to predict cancer patient prognosis. A number of studies have been concerned with simulating protein expression within a single cell. Most of these models consider the dynamic behaviour of interacting molecules over time. Simplified models analyse protein-protein interaction networks using homogeneous methods in which chemical species are assumed to be well mixed. Such methods include systems of ordinary differential equations (ODEs) and the Gillespie method [177, 178]. These methods can be extended to include compartmental models in which one can define homogeneous computational "compartments" determining which molecules can interact with each other. The high

efficiency of these methods have made them very popular for modelling systems in which the copy number of each species is large and compartments are expected to be reasonably well mixed.

However, for some proteins the number of molecules found in a cell can be very low and vary greatly between cells [179]. In addition, the heterogeneous nature of cells is critical to their function [180]. As a result, significant efforts have been made to develop spatial models for these biochemical systems. For example, a simple model of an idealised cell demonstrated how the eccentricity of the cell affects plasma membrane signalling [181]. The Virtual Cell project [182] enables the formulation of both compartmental and spatial partial differential equation models, the latter with either idealized or experimentally derived geometries of one, two or three dimensions. Similarly, Monte Carlo Cell (MCell) and Smoldyn [183, 184, 185] use agent-based methods which simulate each molecule individually and evaluate their diffusion and probability of interactions on a per-particle basis for each time step. Although extremely computationally expensive, these methods have very high spatial resolution and are very successful at modelling interactions of small numbers of heterogeneously distributed molecules. However, as these methods are stochastic, they require multiple random initialisations of the simulation in order to determine the expected behaviour of the system, further adding to the computational cost of these simulations.

Nevertheless, majority of cellular modelling continues to be with a homogeneous spread of the molecules despite the development of these spatially resolved simulation tools. This is due in part to the limited realistic geometries available for simulation which are often either hand segmented or manually fabricated, both of which can be very time-consuming. In addition, there is still need for the development of efficient ways to study cellular response using targeted geometries and organisations, as these simulation tools currently require a large amount of training to properly use. Furthermore, while these methods can be useful for studying the dynamics of protein interaction, they do little to simulate the corresponding microscopy image data, which is necessary for validation of image analysis methods such as cell-compartment classification methods [186, 161, 162, 167, 163]. To address this issue, Zhao and Murphy [164] presented a machine learning method to generate realistic cells with labelled nuclei, membranes and a protein expressed in a cell organelle. Parameters for these models were learned from real images of cells in culture. However, these generative models are restricted to individual cells in culture and only one protein of interest at a time. Hence, this method struggles to capture the dynamic interplay between cells.

Chapter Summary

In this chapter, we have reviewed the existing literature on multiplex imaging. The review covered quantitative data mining methods developed for analysis of the TIS imaging data. These include pixel-level analyses both with and without tresholding the intensity values. Due to the general nature of the analysis framework presented in the next chapter, we also briefly reviewed studies that have been performed with other multiplex techniques such as MALDI, Raman, multi-spectral imaging, MxIF and imaging mass cytometry. The chapter also included a review of frameworks for the generation of synthetic image data. Currently, the majority of these methods focus on the generation of homogeneous cell populations in culture. We have also briefly reviewed current methods for simulating protein expression.

Chapter 3

DiSWOP: A Novel Measure for Cell-Level Protein Network Analysis in Localised Proteomics Image Data

In this chapter, we propose a framework for analysing multiplex image data. As discussed in Chapter 2, the standard way of analysing image data obtained using TIS is to threshold it and then cluster CMPs into CMP motifs. While the lead proteins identified using this approach have been shown to be of functional significance, by thresholding the data a lot of potentially important information is lost. On the other hand, if one considers the raw protein expression profiles without tresholding, the data first needs to be normalised in a robust manner. This is due to inter-sample and inter-protein intensity variations that could result from small differences in sample preparation, imaging and antibody concentrations. This could be a very difficult issue to address due to the lack of controls and ground truth data. Instead here we focus on obtaining the protein interaction networks by considering the protein-protein dependence profile (PPDP) of the cells instead of the raw protein expression profiles. In Section 3.3 we present several measures that could be used to calculate the PPDP and demonstrate why some of them and the raw expression profiles fail.

Furthermore, we perform the analysis at cell level rather than pixel level. This minimises noise from unspecific binding of the protein antibodies to the extra-cellular matrix, stroma and lumen. In addition, the pixel size is not of any biological relevance. Hence, clustering of pixels gives large amounts of noisy data of little biological meaning. Our approach phenotypes the cells according to their PPDP.

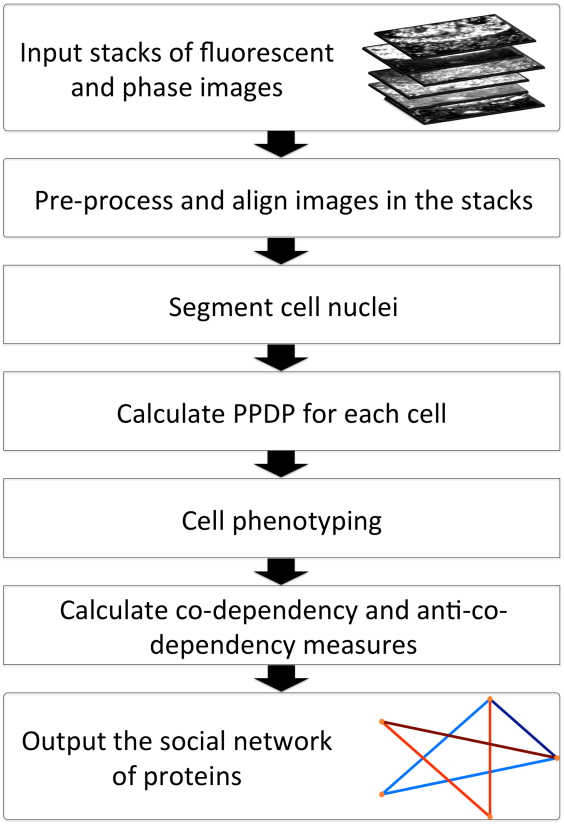


Figure 3.1: Overview of the proposed framework.

This enables us to gain a better understanding of the heterogeneity within the cancer cell population. We can also compare cell phenotypes present in healthy tissue and different cancers.

Lastly, two new measures are proposed to enable us to infer small-scale protein networks. These new measures highlight protein pairs which have very different interaction in cancer and normal tissue. An overview of the approach is presented in Figure 3.1.

3.1 Data

The image data used in this study was acquired using a TIS microscope [73] installed at the University of Warwick. Samples had been surgically removed from colon cancer patients. One sample was taken from the surface of the tumour mass, and another one was selected from apparently healthy colonic mucosa at least 10cm away from the visible margin of the tumour. Two visual fields were manually selected in

each tissue sample, resulting in up to four TIS data sets from each patient. However, many of the visual fields were not available for analysis as they had been identified as poor quality by the biologists performing the experiments. This was due to issue with the microscope during image acquisition. The results presented here were obtained by considering a total of 11 samples – 6 healthy and 5 cancerous. The samples were obtained from 5 patients, with one patient having all four visual fields, one patient missing only a cancer sample, one patient with two normal samples and two patients with a single cancer sample. The data used was obtained from 26 cycles of the TIS machine chosen based on recent findings [68]. However, some of these were excluded from the analysis using the following criteria:

1. Function of the tag not relevant to the study - this way we excluded 2 DAPI channels with different tag concentrations and 5 PBS runs, which were performed to remove autofluorescence.

2. Tag was not registered properly by the RAMTaB algorithm [69] (see below).

3. Invalid expression - images were checked by a pathologist to validate expression of the protein tag in the image. This resulted in all images of the Ki67 tag to be excluded. This protein is expected to be found in increased concentrations only in proliferating cells but this was not the case (Figure 3.2).

This resulted in a library of 12 antibody tags that were used, details of which are shown in Table 3.1. Some of the tags are known tumour markers or cancer stem cell markers. These were CD133, CK19, Cyclin A, Muc2, CEA, CD166, CD36, CD44, CD57, CK20, Cyclin D1 and EpCAM. The stacks also included a DAPI tag used to identify the cell nuclei. A previously presented protocol for sample preparation and image acquisition was used [68].

Table 3.1: List of all antibodies with known information.

Molecule (official symbol)	Location	Known function
CD133 (PROM1)	Expression in luminal membranes of glandular epithelia.	Prominin-1 is a pentaspan transmembrane glycoprotein. The protein localises to membrane protrusions and is often expressed on adult stem cells, where it is thought to function in maintaining stem cell properties by suppressing differentiation.

Table 3.1: Continued

Molecule	Location	Known function
Ck19 (KRT19)	Cytoplasmic and membranous expression in epithelium tissue.	This smallest known acidic cytokeratin is not paired with a basic cytokeratin in epithelial cells. It is specifically expressed in the periderm, the transiently superficial layer that envelopes the developing epidermis.
Cyclin A (CCNA2)	Nuclear and to some extent cytoplasmic staining in proliferative cells.	This cyclin is expressed in all tissues tested. This cyclin binds and activates CDC2 or CDK2 kinases, and thus promotes both cell cycle G1/S and G2/M transitions.
Muc2 (MUC2)	Selective cytoplasmic expression in mucus producing cells of the gastrointestinal tract.	Mucin 2 is secreted and forms an insoluble mucous barrier that protects the gut lumen. The protein polymerises into a gel of which 80% is composed of oligosaccharide side chains by weight.
CEA (CEACAM1)	Membranous expression mostly in epithelium cells.	This is a glycoprotein, with a series of Ig like domains. Its normal tissue distribution includes columnar epithelial cells and goblet cells in colon, mucous cells in stomach, squamous epithelium of tongue, esophagus and cervix and prostate. It is used clinically as a tumor marker for colorectal cancer.
CD166 (ALCAM)	Cytoplasmic and membranous expression mostly in epithelium cells.	Activated leukocyte cell adhesion molecule is a member of a subfamily of immunoglobulin receptors with five immunoglobulin-like domains (VVC2C2C2) in the extracellular domain. This protein binds to T-cell differentiation antigene CD6, and is implicated in the processes of cell adhesion and migration.

Table 3.1: Continued

Molecule	Location	Known function
CD36 (CD36)	No data available	CD36 is a major glycoprotein of the platelet surface and serves as a receptor for thrombospondin in platelets and various cell lines. The protein may have important functions as a cell adhesion molecule. It binds to collagen, thrombospondin, anionic phospholipids and oxidized LDL.
CD44 (CD44)	Mainly localised to the plasma membrane but also to the Golgi apparatus	CD44 is a cell-surface glycoprotein involved in cell-cell interactions, cell adhesion and migration. It participates in a wide variety of cellular functions including lymphocyte activation, recirculation and homing, hematopoiesis, and tumor metastasis.
CD57 (B3GAT1)	Cytoplasmic expression in several cell types.	This glycoprotein is expressed normally in hematopoietic cells (Natural killer cells and CD8 positive Tlymphocytes), neuroectodermal cells, neuro endocrine cells, striated muscles and epithelium of prostate. It is possibly related to cell-cell interaction. Laminin, P-selectin and N-selectin are its natural ligands.
Ck20 (KRT20)	Selective cytoplasmic expression in gastrointestinal epithelium.	The keratins are intermediate filament proteins responsible for the structural integrity of epithelial cells and are subdivided into cytokeratins and hair keratins. This cytokeratin is a major cellular protein of mature enterocytes and goblet cells and is specifically expressed in the gastric and intestinal mucosa. It is known to be overexpressed in CRA as compared to normal colon.

Table 3.1: Continued

Molecule	Location	Known function
Cyclin D1 (CCND1)	Localised to the nucleus but excluded from the nucleoli.	This cyclin forms a complex with and functions as a regulatory subunit of CDK4 or CDK6, whose activity is required for cell cycle G1/S transition. Mutations, amplification and overexpression of this gene, which alters cell cycle progression, are observed frequently in a variety of tumors and may contribute to tumorigenesis.
EpCAM (EpCAM)	Selective expression in the cytoplasm and cell membranes of glandular cells.	Epithelial cell adhesion molecule is expressed on most normal epithelial cells and gastrointestinal carcinomas and functions as a homotypic calcium-independent cell adhesion molecule. The antigen is being used as a target for immunotherapy treatment of human carcinomas.
4,6-diamidino-2-phenylindole (DAPI)	Chromatin	DAPI is an intercalating agent. It is a fluorescent molecule and has been used to stain DNA nucleic acids.

3.2 Pre-processing

Background autofluorescence is digitally subtracted at an early stage. Hence, any remaining fluorescence should be true protein expression. In each of the stacks, the images were aligned using the RAMTaB (Robust Alignment of Multi-Tag Bioimages) algorithm [69]. This is done in order to prevent possible noise resulting from the slight mis-alignment of the multi-tag images obtained using TIS. A measure of confidence in the registration results is given by the standard deviation of shifts computed by different blocks and images were discarded if this exceeded a pre-defined threshold. This method has been shown to achieve sub-pixel accuracy of registering this data [69]. Then, if there are K tags, each having a corresponding image of size $m \times n$,

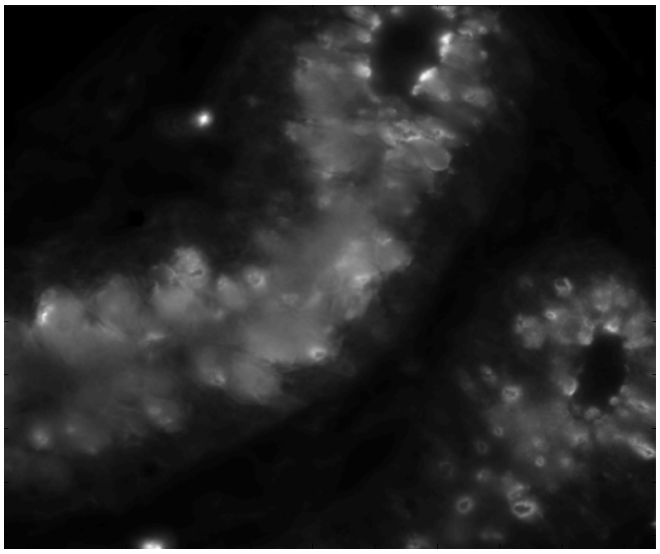


Figure 3.2: Expression of Ki67 in a normal sample

the data can be represented as a $K \times mn$ matrix

$$\mathbf{X} = \left[\begin{array}{c|c|c|c} x_{1,1}^1 & x_{1,2}^1 & \cdots & x_{m,n}^1 \\ x_{1,1}^2 & x_{1,2}^2 & \cdots & x_{m,n}^2 \\ \vdots & \vdots & \ddots & \vdots \\ x_{1,1}^K & x_{1,2}^K & \cdots & x_{m,n}^K \end{array}\right], \tag{3.1}$$

where $x_{i,j}^k$ is the expression level of protein k ($k < K$) at pixel (i, j). In our experiments $K = 12, m = 1027$ and $n = 1056$.

A method was recently proposed to perform cell segmentation of TIS stacks in order to restrict the analysis to cellular areas only [187]. This ensures that signals from stroma and lumen are removed as they can potentially add noise to the subsequent analysis. To follow best practice, one should segment entire cells since some of the proteins observed are located in parts of the cells other than the nucleus, such as the cytoplasm, vesicles or the Golgi apparatus. However, this is challenging in cancerous tissues because of the variable orientation of cells due to disrupted tissue architecture and a tag of the cell membrane was not used in this set of experiments to enable us to precisely identify entire cells. Instead, each image was segmented using a modified form of the graph cut method [188] applied to a DAPI channel [187] (Figure 3.3). Initially, each image is binarised using graph-cut based algorithm to extract the foreground. Next, an initial segmentation is performed by detecting seed points on the foreground of the binarised image by using a multi-scale Lapla-

cian of Gaussian (LoG) filter [189]. The initial segmentation is then refined using a second graph-cut based algorithm. Finally, the nuclei segmentation results obtained using the framework are post-processed by either eliminating very small nuclei or merging them with nearby nuclei, as they usually result from segmentation errors. This final step ensures that analysis is restricted only to clearly distinguishable nuclei. This serves as a rough approximation of the pixels belonging to the cells. This was necessary in order to extract pixel locations of the nuclei and their immediate neighbourhood only, as the DAPI tag stains the DNA. Using only nuclei may reduce the amount of cell available for analysis but is comparatively unambiguous and can be used as a rough approximation of the cells. Segmentation is an issue as gold-standard data is not available and perfect cutting of sections is impossible. Future experiments should include a membrane tag, which would resolve this problem.

Segmentation resulted in a total of 2945 cells being identified. The cell-localised protein expression values for each of the K proteins is collected in a protein expression matrix $\mathbf{X}_c$ of the order $K \times N_c$ for each cell c

$$\mathbf{X}_c = \{\mathbf{x}_{i,j} | \, (i,j) \in \Omega_c\}, \tag{3.2}$$

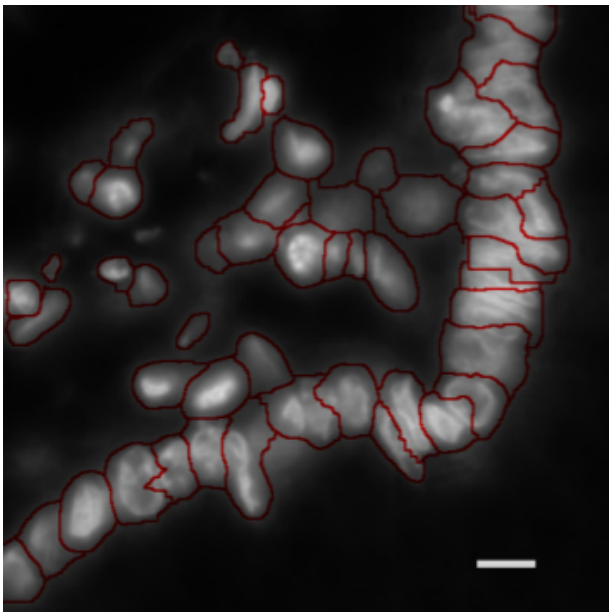
where $\Omega_c = \{(i_1, j_1), (i_2, j_2), ..., (i_{N_c}, j_{N_c})\}$ denotes the set of pixel coordinates in cell c, $N_c = |\Omega_c|$ denotes the number of pixels in each cell c and the vector $\mathbf{x}_{i,j} = [x^1_{i,j}, ..., x^K_{i,j}]$ is the expression levels of each tag at pixel (i,j). In matrix form this is given by

$$\mathbf{X}_c = \left[\begin{array}{c|c|c|c} x^1_{i_1,j_1} & x^1_{i_2,j_2} & \cdots & x^1_{i_{N_c},j_{N_c}} \\ x^2_{i_1,j_1} & x^2_{i_2,j_2} & \cdots & x^2_{i_{N_c},j_{N_c}} \\ \vdots & \vdots & \ddots & \vdots \\ x^K_{i_1,j_1} & x^K_{i_2,j_2} & \cdots & x^K_{i_{N_c},j_{N_c}} \end{array}\right]. \tag{3.3}$$

3.3 Protein-protein dependence profile (PPDP)

The pairwise maximal information coefficient (MIC) [190] for each pair of proteins, localised to an individual cell c, is calculated to obtain the protein-protein dependence profile (PPDP) of the cell. We used this statistic since it has been shown to capture a wide range of associations, both functional and not, and it gives similar scores to equally noisy relationships of different types [190]. The MIC for each pair of proteins, localised to an individual cell c, is calculated by considering the intensities of the two proteins pixel by pixel. The MIC is calculated by exploring all grids on the scatter-plot up to a maximal grid resolution dependent on the grid size,

a)



b)

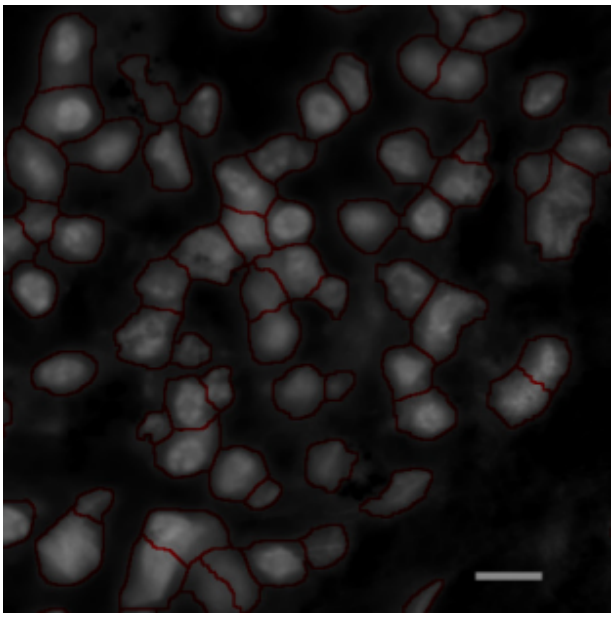


Figure 3.3: Segmentation results on a part of a normal sample (a) and a part of a cancer sample (b). The size of the scale bars is 10 μm.

computing for every pair of integers (k, l) the largest possible mutual information achievable by any k-by-l grid applied to the data (Figure 3.4 (a)). The values found are then normalised as follows: for a grid G, let I_G denote the mutual information of the probability distribution induced on the grid boxes of G, where the probability of a box is proportional to the number of points that fall in the box. The highest nor-

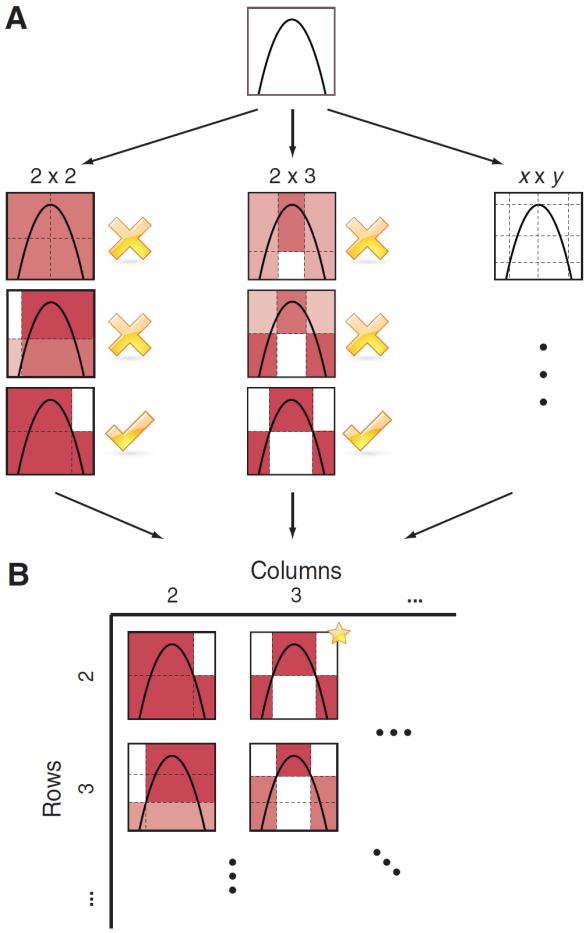


Figure 3.4: Computing MIC (A) For each pair (x,y), the MIC algorithm finds the x-by-y grid with the highest induced MI. (B) The algorithm normalises the MI scores and compiles a matrix that stores, for each resolution, the best grid at that resolution and its normalised score. The normalised scores form the characteristic matrix; MIC corresponds to the maximum of this matrix. In this example, there are many grids that achieve the highest score. The star in (B) marks a sample grid achieving this score. Image credit: [190].

malised mutual information achieved by any k-by-l grid is recorded as the element $m_{k,l}$ of a characteristic matrix $\mathbf{M}$, where

$$m_{k,l} = \frac{max(I_G)}{log(min\{k,l\})}, \tag{3.4}$$

with the maximum being taken over all k-by-l grids G (Figure 3.5 (b)). The normalisation ensures a fair comparison between grids of different sizes and obtains values between 0 and 1. The MIC is the maximum value of $\mathbf{M}$ [190]. As suggested by [190],

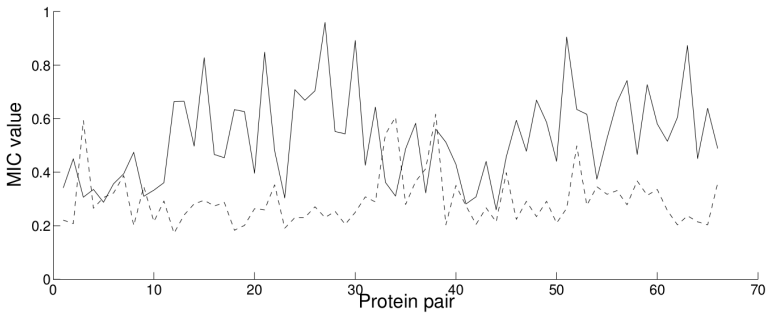


Figure 3.5: Protein-protein dependence profile (PPDP) of two cells from the same specimen.

the maximum size of the grids considered was set to be $kl < N_c^{0.6}$ where N_c is the number of pixels in the cell c. For each cell c, a $K(K-1)/2$-dimensional vector μ_c of pairwise MIC scores is obtained. The vector represents the PPDP of the cell and can be expressed as

$$\mu_c = \left[\mu_c^{1,2}\mu_c^{1,3}\cdots\mu_c^{1,K}\mu_c^{2,3}\mu_c^{2,4}\cdots\mu_c^{2,K}\cdots\mu_c^{K-1,K}\right], \tag{3.5}$$

where $\mu_c^{i,j} \in [0,1]$ is given by the MIC between rows i and j of the matrix $\mathbf{X}_c$. The PPDP for two sample cells from the same tissue specimen is shown in Fig 3.5.

Other co-dependence measures were also considered for the analysis. Pearson's and Spearman correlations fail to capture non-linear relationships between protein expression profiles, which often occur due to the inhomogeneous structure of the cells. An example of this can be seen in Figure 3.6. In Figure 3.6 (a) we can see that the two proteins are weakly dependent on each other. However, the Pearson's coefficient for this cell was -0.01, whereas the MIC was 0.33. Mutual information and normalised mean expression values were also tested. However, each of these resulted in a batching effect where some phenotypes (See section 3.4) were predominantly located in a single, usually cancerous, sample and the samples were split into a handful of phenotypes (Figure 3.7). This seems biologically unlikely as we expect that there should be some normal cells within the tumour tissue and that cancers share some common types of cells. These findings are consistent with the findings that functionality can be determined by colocation rather than changes in abundance levels [66]. For result comparison Distance Correlation (DC) [191] was also used. The final results obtained were very similar to the ones obtained using MIC (See Table 3.2). However, it has been found that the distance correlation has a strong preference for some types of dependencies and gives different scores at the same noise levels [190]. Therefore, the MIC is preferred due to its robustness to

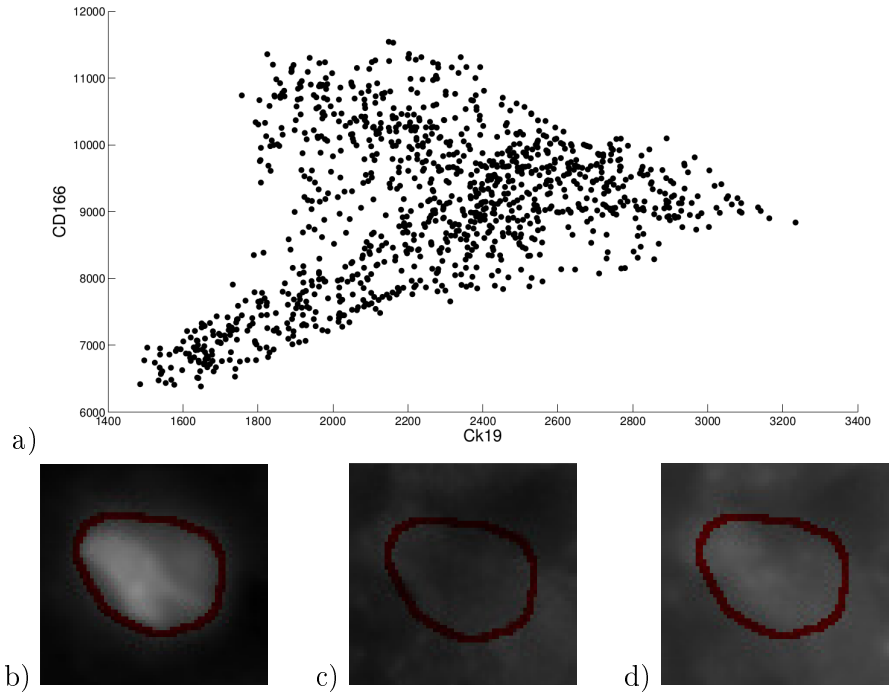


Figure 3.6: An example of non-linear dependence between protein expressions in a cell. Figure (a) shows a scatter plot of the pixel intensities of CK19 and CD166 in a cancer cell. Figures (b) - (d) show the DAPI, CK19 and CD166 expression of the cell outlined in red.

variations in the type of dependence.

3.4 Cell phenotyping based on localised PPDP

We consider and compare the results from three different clustering frameworks. Ideally, the final results of the analysis should be independent of the phenotyping method.

3.4.1 Affinity Propagation Clustering (APC)

The vector μ_c is the PPDP of the cell c and can be used to determine the cell phenotype using a clustering algorithm. Affinity Propagation Clustering (APC) is a clustering method, which takes as input a matrix containing measures of similarity between pairs of data points. Real-valued messages are passed between data points until a high-quality set of exemplars and corresponding set of clusters gradually emerges [192]. We have used a Gaussian Kernel based on the Euclidean distance between the protein co-dependence profiles of cells as an affinity matrix, so for a

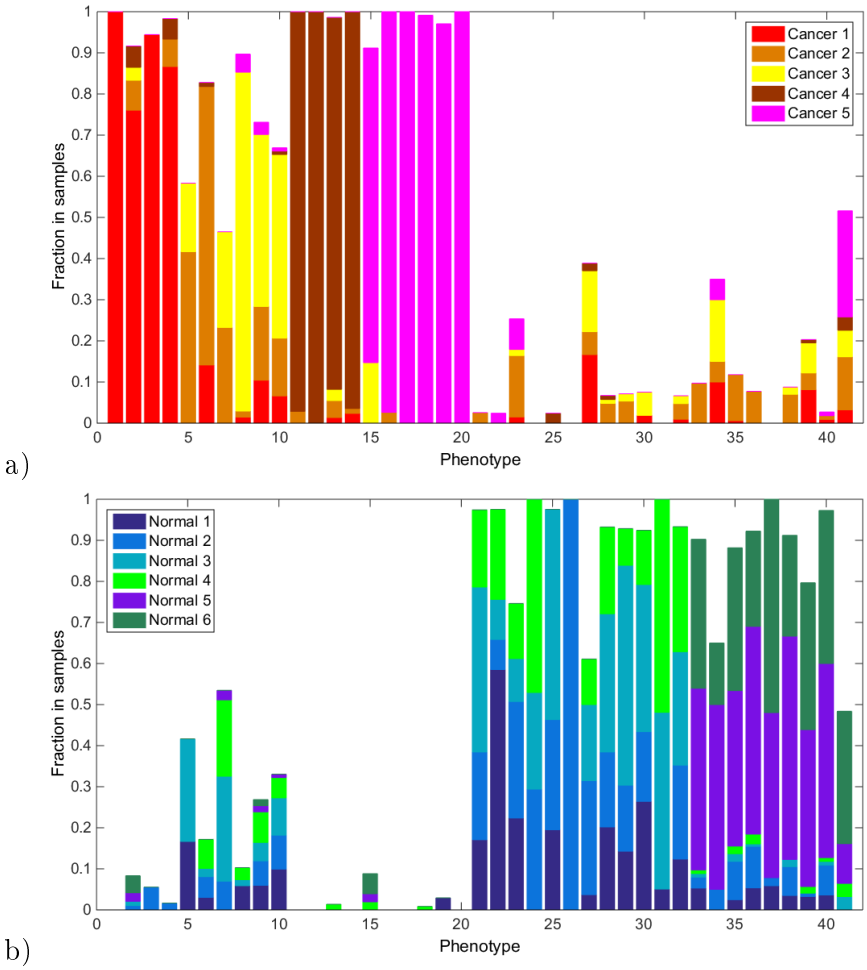


Figure 3.7: Distribution of phenotypes obtained using APC based on the mutual information profile of the cells amongst (a) cancerous samples and (b) normal samples. Each colour corresponds to a different sample. The 41 phenotypes are shown along the x-axis. The y-axis shows proportion of the phenotype located in each sample.

pair of cells a and b with PPDPs μ_a and μ_b, respectively, the (a, b) entry of the similarity matrix (for $a \neq b$) is given by

$$s_{a,b} = exp\left(\frac{-\|\mu_a - \mu_b\|^2}{2\sigma^2}\right), \tag{3.6}$$

where $\sigma = (max_{a,b}\|\mu_a - \mu_b\|)/3$ and $\|\cdot\|$ is the Euclidean distance. All diagonal entries of the matrix are set to equal the minimum value of the matrix. This means that each cell is equally likely to be a cluster centroid and results in a moderate number of clusters. We denote the number of cell phenotypes resulting from this approach by $\hat{C}$, which in this instance was found to be 41. An Agglomerative Hi-

Table 3.2: Top and bottom 10 DiSWOP results from different dependency measures (MIC and DC) and clustering methods (APC, GBHC and AHC). Pairs are shown with decreasing DiSWOP score. All results have been obtained by considering the top 5 PPDP scores for each phenotype.

MIC and AP	MIC and GBHC	MIC and AHC	DC and AP
CEA & EpCAM	CK20 & EpCAM	CEA & EpCAM	CEA & EpCAM
CD133 & EpCAM	CEA & EpCAM	CD133 & CK20	CD133 & Muc2
CEA & CK20	Muc2 & EpCAM	CK19 & CK20	CK19 & CEA
CD133 & Muc2	CD133 & CEA	CK19 & EpCAM	CK19 & EpCAM
CD133 & CEA	Muc2 & CEA	CK19 & CEA	CK19 & CD57
CK19 & EpCAM	CK19 & CEA	Muc2 & EpCAM	CD133 & EpCAM
CK20 & EpCAM	CD133 & Muc2	CD57 & EpCAM	Cyclin A & CD57
CD133 & Cyclin D1	CK19 & CK20	CEA & CK20	CD133 & CEA
Muc2 & EpCAM	CEA & CK20	CD133 & CEA	Muc2 & EpCAM
CD57 & EpCAM	CK19 & EpCAM	Cyclin A & EpCAM	CD57 & EpCAM
CD166 & Cyclin D1	Cyclin A & CD57	CD44 & CK20	CD57 & Cyclin D1
Cyclin A & CK20	CD133 & Cyclin D1	Muc2 & CD44	Muc2 & CD44
CD166 & CD57	CD166 & CD57	Muc2 & CD166	Muc2 & CD36
Muc2 & CD44	CD57 & Cyclin D1	CD57 & Cyclin D1	Muc2 & CD57
Muc2 & CD166	CD166 & CD36	CK19 & CD57	Cyclin A & CK20
CK19 & Cyclin A	Cyclin A & CD166	CD166 & CD36	CD166 & CD36
CD166 & CD36	CD44 & EpCAM	CD166 & Cyclin D1	Cyclin A & CD166
CD36 & Cyclin D1	CD166 & Cyclin D1	CD36 & CD57	CD44 & EpCAM
CD36 & CD57	CD36 & Cyclin D1	CD36 & Cyclin D1	CD36 & Cyclin D1
CD44 & EpCAM	CD36 & CD57	CD44 & EpCAM	CD36 & CD57

erarchical Clustering (AHC) [193] and Gaussian Bayesian Hierarchical Clustering (GBHC) [194, 195] approach with the same number of clusters was also considered. It was encouraging to see that these gave similar results, which are shown in Table 3.2.

3.4.2 Agglomerative Hierarchical Clustering (AHC)

This is a bottom-up clustering method [193], which starts with each of the N PPDPs μ as belonging to a different cluster. At each iteration the algorithm merges two clusters together by aiming to minimise the increase in the variance of clusters [196]. If at each iteration level k^* we have clusters $S_j = \{\mu_{(j,1)}, \ldots, \mu_{(j,n_j)}\}$, where $n_j = |S_j|$ and $j \in \{1, \ldots, N - k^*\}$, clusters at level $k^* + 1$ are obtained by finding clusters u and v such that

$$u, v = \underset{u,v \in \{1,\ldots,N-k^*\}}{\operatorname{argmax}} \frac{n_u n_v \left(\|\bar{S}_u - \bar{S}_v\|_2\right)^2}{n_u + n_v}, \tag{3.7}$$

where $\bar{S}_j$ is the mean vector for S_j. This step reduces the number of clusters by one by merging clusters S_u and S_v. The number of clusters was set to equal that found by AP clustering, so the tree was cut at level $N - \hat{C}$.

3.4.3 Gaussian Bayesian hierarchical clustering (GBHC)

As this method is computationally expensive, we employed a feature selection technique to reduce the number of features considered. The protein pairs that best discriminate between cancer and normal samples were selected using the Wilcoxon rank sum test [197]. For a protein pair, this was done by calculating the p-value that the PPD values of the cancer cells and of the normal cells come from distributions with a different median. Then, out of the 66 protein pairs, the 33 with lowest p-values were selected for clustering to be performed on. This drastically speeds up performance of the algorithm. Cells with similar phenotype are expected to have PPDPs with similar nature. In terms of probability, we can hypothesise that different phenotypes are explained by different probability distribution, and cells with similar phenotype should come from the same distribution. Cell phenotyping can therefore be achieved through GBHC [194], which models data as a mixture of probability distributions.

Let the PPDP of the i^{th} cell be denoted by $x^{(i)} = (x_1^{(i)}, \ldots, x_d^{(i)})$, where $x_j^{(i)}$ is a PPD value of the j^{th} protein pair for the i^{th} cell, and $d = 33$ is the number of protein pairs after feature selection. Without loss of generality, we assume that the whole PPD data have zero mean and unit variance. Let D_k denote a set of PPD data for n_k cells belonging to the k^{th} phenotype. According to the assumptions of GBHC, for each protein pair j, $x_j^{(i)}$ are independent and identically normal distributed with unknown mean μ_j and variance σ_j^2, i.e.

$$x_j^{(i)} \sim \mathcal{N}(x|\mu_j, \sigma_j^2) \quad \forall x^{(i)} \in D_k. \tag{3.8}$$

Furthermore, μ_j and σ_j^2 are assumed to be normal-gamma distributed with hyperparameters λ_0, β_0, and κ_0. The marginal likelihood of D_k based on this hierarchical probabilistic model can be expressed as

$$P(D_k|\lambda_0, \beta_0, \kappa_0) = \prod_{j=1}^{d} \left[\frac{\Gamma(\lambda_{n_k})}{\Gamma(\lambda_0)} \frac{\beta_0^{\lambda_0}}{\beta_{n_k,j}^{\lambda_{n_k}}} \left(\frac{\kappa_0}{\kappa_{n_k}} \right)^{\frac{1}{2}} (2\pi)^{\frac{-n_k}{2}} \right], \tag{3.9}$$

where

$$\lambda_0, \beta_0, \kappa_0 > 0, \tag{3.10}$$

$$\kappa_{n_k} = \kappa_0 + n_k, \tag{3.11}$$

$$\lambda_{n_k} = \lambda_0 + \frac{n_k}{2}, \tag{3.12}$$

$$\bar{x}_j = \frac{1}{n_k} \sum_{i=1}^{n_k} x_j^{(i)}, \tag{3.13}$$

$$\beta_{n_k,j} = \beta_0 + \frac{1}{2} \left[\sum_{i=1}^{n_k} \left(x_j^{(i)} - \bar{x}_j \right)^2 + \frac{\kappa_0 n_k \bar{x}_j^2}{\kappa_{n_k}} \right], \tag{3.14}$$

and $\Gamma(\cdot)$ denotes a gamma function. This likelihood term indicates how likely it is that cells in D_k have the same phenotype, and it will be used as an alternative to a distance-based dissimilarity measure, which is normally used in agglomerative hierarchical clustering methods.

GBHC uses Bayesian model selection to decide which pair of small data sets D_k and D_l is the most probable to belong to the same distribution, and should be merged together to form a larger data set D_m. This is done through Bayes' rule:

$$r_m = \frac{\pi_m P(D_m|\lambda_0, \beta_0, \kappa_0)}{\pi_m P(D_m|\lambda_0, \beta_0, \kappa_0) + (1 - \pi_m) P(D_k|\lambda_0, \beta_0, \kappa_0) P(D_l|\lambda_0, \beta_0, \kappa_0)}, \tag{3.15}$$

in which $P(D_k|\lambda_0, \beta_0, \kappa_0)$ is the marginal likelihood of a cluster D_k as defined in Equation 3.9, $\pi_m = \alpha\Gamma(n_m)/\rho_m$, $\rho_m = \alpha\Gamma(n_m) + \rho_k\rho_l$, we set $\pi_k = 1, \rho_k = \alpha$ for every initial cluster set and α is a concentration parameter related to the expectation of the number of clusters in the data. As we climb up a hierarchical tree, the probability that two clusters being merged come from the same distribution gets lower. Using this information, GBHC does not consider merges with probability less than 0.5 as valid merges. This in turn results in the algorithm automatically giving the final number of clusters, here found to equal 25.

Since there is no ground truth available for the number and distribution of cell phenotypes in these samples, evaluating the accuracy of the clustering methods is challenging. Hence, this clustering method was selected mainly due to its contrasting approach from APC. This allows us best to demonstrate the robustness of the DiSWOP results.

3.5 Protein-protein co-dependence and anti-co-dependence measures

Once the cell phenotype clusters have been obtained, an average PPDP, $\bar{\mu}_S$ is calculated for each cluster S. For a protein pair (i, j) (with $i < j \leq K$) $\bar{\mu}_S^{i,j}$ is given

by

$$\bar{\mu}_S^{i,j} = \frac{\sum_{c \in S} \mu_c^{i,j}}{|S|}. \tag{3.16}$$

Then $\bar{\mu}_S$ is the vector

$$\bar{\mu}_S = \left[\bar{\mu}_S^{1,2} \bar{\mu}_S^{1,3} \cdots \bar{\mu}_S^{1,K} \bar{\mu}_S^{2,3} \bar{\mu}_S^{2,4} \cdots \bar{\mu}_S^{K-1,K} \right]. \tag{3.17}$$

In order to more objectively investigate the protein pairs which have higher dependency and are more frequent in cancer samples, a difference of weighted sums was calculated by considering the top N (here set to equal 5 or 10) dependency scores of the ten most frequent phenotypes in each sample. The measure weights the dependency score with the phenotype probability in the sample, and sums all occurrences of the protein pair in all the cancerous samples and of all the normal samples. The sums are normalised by the number of samples. It then subtracts the score for the normal from the score for the cancer samples, hence giving a positive score if a pair appears more frequently and with higher dependency scores in the cancerous samples. More formally, if $\hat{\mu}_S$ is the vector with the elements of $\bar{\mu}_S$ (lying in $[0, 1]$) sorted in descending order, p_S^r is the probability of phenotype S in sample r, $S_{\alpha,r}$ is the α^{th} most frequent phenotype in sample r, and

$$M_S^{i,j} = \begin{cases} \bar{\mu}_S^{i,j}, & \text{if } \bar{\mu}_S^{i,j} \text{ is one of the first } N \text{ elements of } \hat{\mu}_S \\ 0, & \text{otherwise} \end{cases}, \tag{3.18}$$

then the difference of the sum of frequency-weighted localised protein-protein co-dependence values for a protein pair (i, j), $w_{i,j}$ is given by

$$w_{i,j} = \frac{1}{|\psi|} \sum_{r \in \psi} \sum_{\alpha=1}^{10} p_{S_{\alpha,r}}^r M_{S_{\alpha,r}}^{i,j} - \frac{1}{|\nu|} \sum_{r \in \nu} \sum_{\alpha=1}^{10} p_{S_{\alpha,r}}^r M_{S_{\alpha,r}}^{i,j}. \tag{3.19}$$

where ψ is the set of cancerous samples, ν is the set of normal samples.

A similar quantity of anti-co-dependence has also been considered by looking at the bottom N dependency scores, so we define

$$\hat{M}_S^{i,j} = \begin{cases} \bar{\mu}_S^{i,j}, & \text{if } \bar{\mu}_S^{i,j} \text{ is one of the last } N \text{ elements of } \hat{\mu}_S \\ 0, & \text{otherwise} \end{cases}, \tag{3.20}$$

and use $1 - \hat{M}_S^{i,j}$ instead of $M_S^{i,j}$ to measure anti-co-location of protein pairs, i.e.

$$\hat{w}_{i,j} = \frac{1}{|\psi|} \sum_{r \in \psi} \sum_{\alpha=1}^{10} p^r_{S_{\alpha,r}} \left(1 - \hat{M}^{i,j}_{S_{\alpha,r}}\right) - \frac{1}{|\nu|} \sum_{r \in \nu} \sum_{\alpha=1}^{10} p^r_{S_{\alpha,r}} \left(1 - \hat{M}^{i,j}_{S_{\alpha,r}}\right). \tag{3.21}$$

Hence, we introduce two new measures called Difference in Sum of Weighted cO-dependence/Anti-co-dependence profiles, further referred to as DiSWOP (Equation 3.19) and DiSWAP (Equation 3.21). Large positive values of DiSWOP indicate that the protein pair (i, j) is more co-dependent in cancer samples, while a low negative DiSWOP value means that the protein pair is more co-dependent in the normal samples. Similarly for DiSWAP a large positive value suggests that the protein pair is more anti-co-dependent in cancer and a large negative value that the protein pair is more anti-co-dependent in healthy samples. The DiSWOP and DiSWAP scores are shown in Figures 3.8 and 3.9, respectively. Various combinations of number of phenotypes and dependency scores were also considered. Altering the number of clusters caused very little change to the results as the phenotypes that were added or excluded have very low probability in the samples. On the other hand, increasing the number of dependency scores considerably changed the protein pairs highlighted. However, if more than the top ten scores are included, the average dependency score added to the analysis is below 0.5 and so the proteins are more anti-co-dependent than they are co-dependent. Therefore, these scores should not be included as part of the DiSWOP measure. Further biological validation and analysis of a greater number of samples is needed to determine the optimal number of dependency scores to be considered as part of the dependency measures.

3.6 Results

The results presented in Figures 3.8 and 3.9 suggest that it is in fact the combinations of protein pairs with high dependency scores that identify cancer cells, which is to be expected, considering the complexity of the system. Calculating the DiSWOP and DiSWAP measures identified pairs which are significantly more co-dependent or anti-codependent in cancer samples than in normal tissue. As can be seen in Figure 3.8 and Table 3.2, EpCAM and CEA have very high positive DiSWOP score for all results. It is encouraging to see that most of the protein pairs highlighted are the same when different methods are used for phenotyping. This may be due to the fact

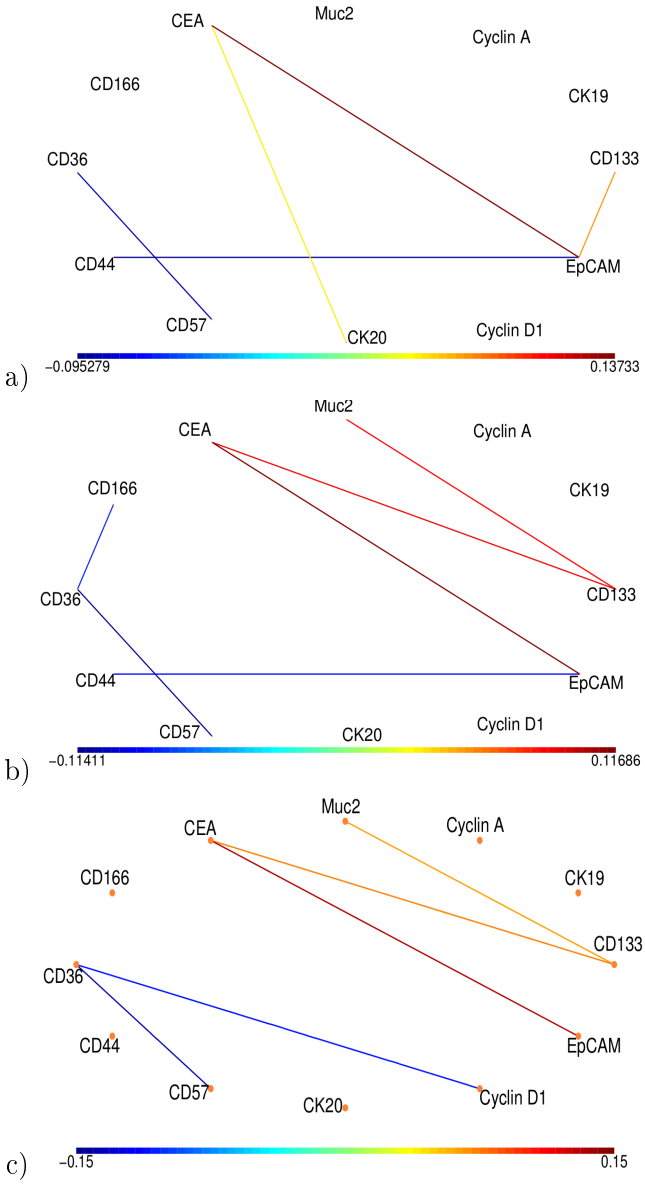


Figure 3.8: The social networks of proteins' colocalisation. Each node represents a protein and each edge colour shows a protein pair with different level of co-expression in the normal and cancer samples. Only edges with the top 10% and the bottom 10% of the DiSWOP values are shown. Figures (a) and (b) show DiSWOP values obtained using APC when considering the top 5 and 10 dependency scores, respectively. Figure (c) show DiSWOP values obtained using GBHC when considering the top 5 dependency scores. Here, a large positive value (shown in red) indicates that the protein pair is more co-dependent in cancer samples, whereas a large negative value (shown in blue) means that the protein pair is more active in normal tissue.

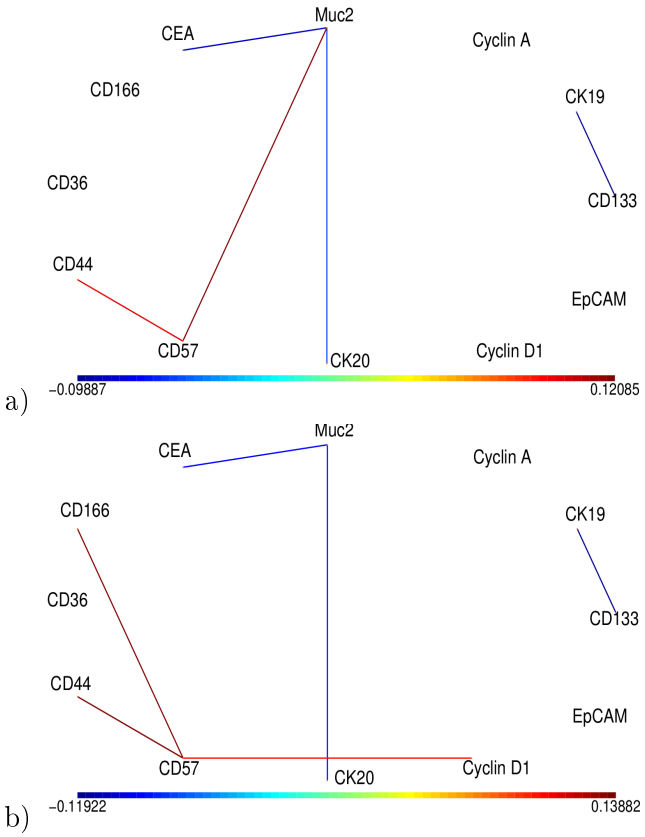


Figure 3.9: The social networks of proteins' anti-colocalisation. Each node represents a protein and each edge colour shows a protein pair with different level of co-expression in the normal and cancer samples. Only edges with the top 10% and the bottom 10% of the DiSWAP values are shown. Figures (a) and (b) show DiSWAP values obtained using APC when considering the top 5 and 10 dependency scores, respectively. In this case, a large positive value (shown in red) indicates that the protein pair is more anti-co-dependent in cancer samples, whereas a large negative value (shown in blue) means that the protein pair is more anti-co-dependent in normal tissue.

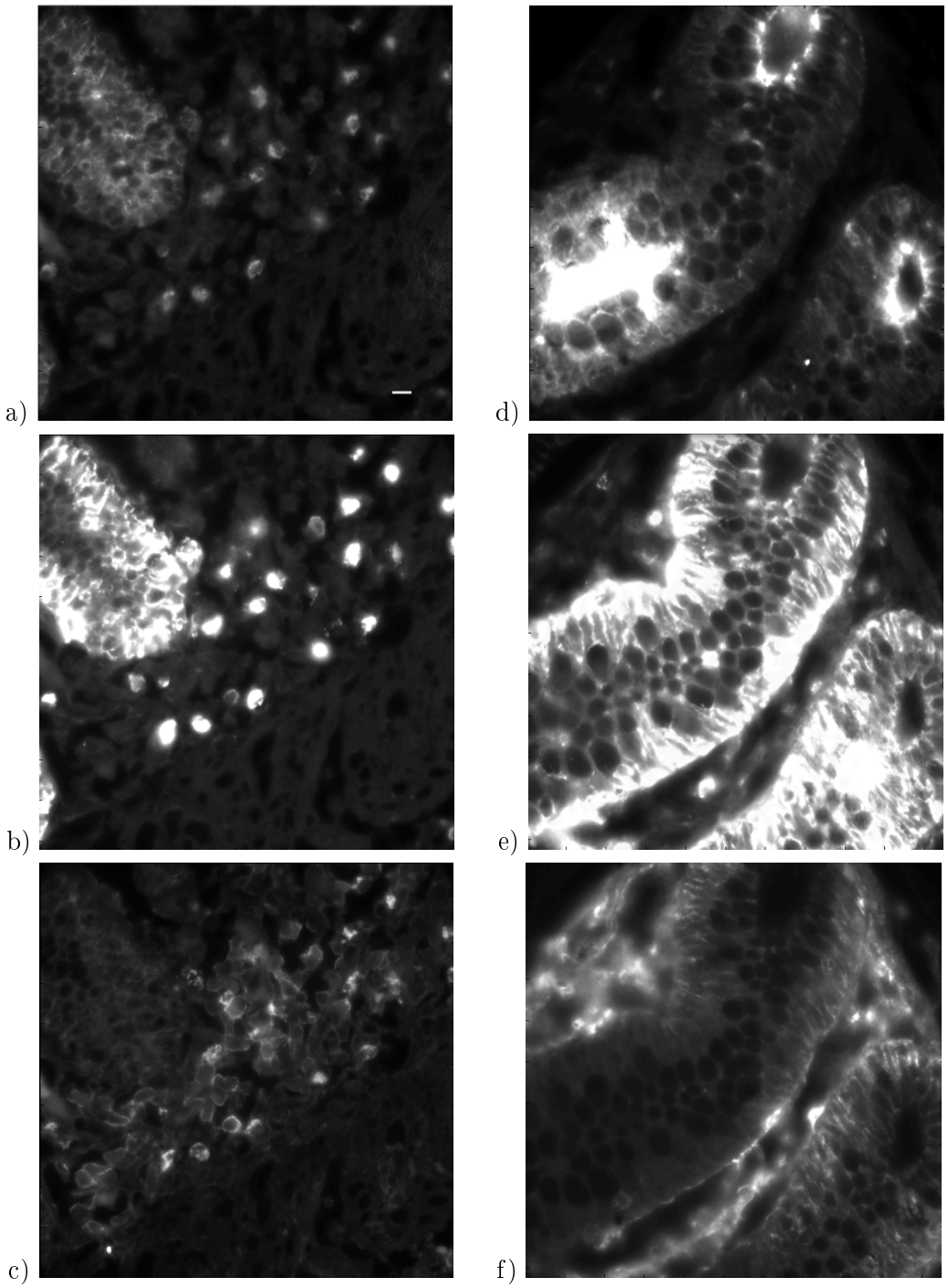


Figure 3.10: Protein expression images. Figures (a) - (c) show CEA, EpCAM and CD44 expression levels, respectively, in a cancer sample. Figures (d) - (f) show CEA, EpCAM and CD44 expression levels, respectively, in a normal sample. The scale bar in (a) is 10 μm

that both proteins are involved in cell adhesion (details of all the proteins considered have been previously presented [68]). On the other hand, the pairs CD36 and CD57, and CD44 and EpCAM were more likely to interact in the normal tissue samples (Figures 3.8). These dependencies can be seen in the data. Figure 3.10 shows the expression levels of CEA, EpCAM and CD44 in a cancer and a normal sample. It is clear that protein expression in Figures 3.10 (a) and (b) illustrate a higher dependence than in Figures 3.10 (d) and (e), whereas the expression patterns in Figures 3.10 (b) and (c) differ more than those in Figures 3.10 (e) and (f). Similar trends can be seen in most of the other samples. Considering the DiSWAP measure also highlights some pairs of proteins such as CD44 and CD57 being more anti-codependent in cancer samples and Ck19 and CD133 in normal samples. It is worth noting that when we compare the results for DiSWOP obtained using APC and GBHC there is very high agreement as to which pairs have high positive or negative DiSWOP values. In order to quantitative evaluate the similarity between the networks we calculate distance measures between the vectors containing the DiSWOP values. The L_1 norm between all the edges in the graphs shown is 0.636 and the mean of the relative absolute difference between the edge weights, as defined by

$$mean\left(\frac{|w^{(1)} - w^{(2)}|}{max(|w^{(1)}|, |w^{(2)}|)}\right) \tag{3.22}$$

is found to be 0.683, where $w^{(1)}$ and $w^{(2)}$ are the weights of the two graphs shown in Figure 3.8 (a) and (c), respectively. The later measure can take values between 0 and 2, with 0 meaning that all the edges are the same, 1 meaning that non of the edges co-occur and 2 showing that all $w^{(1)} = -w^{(2)}$. On the other hand, when all of the edges (thresholded and non-thresholded) are considered, the L_1 norm is 0.89 and the mean of the relative absolute difference between the edge weights is found to be 0.561. In both cases, the maximum absolute difference between the edge weights is 0.0624.

It is important to note that these results were obtained using only 11 samples which, while being a great improvement on previous studies in the toponomics of colon cancer [68, 93], is still insufficient to draw significant biological conclusions.

3.7 Results Significance

In order to further analyse the consistency of the two dependency measures, the analysis was performed on 16 different combinations of 3 cancer and 3 normal samples. The results are shown in Figure 3.11 where it can be seen that the protein pairs with highest and lowest DiSWOP and DiSWAP scores are the same as the ones

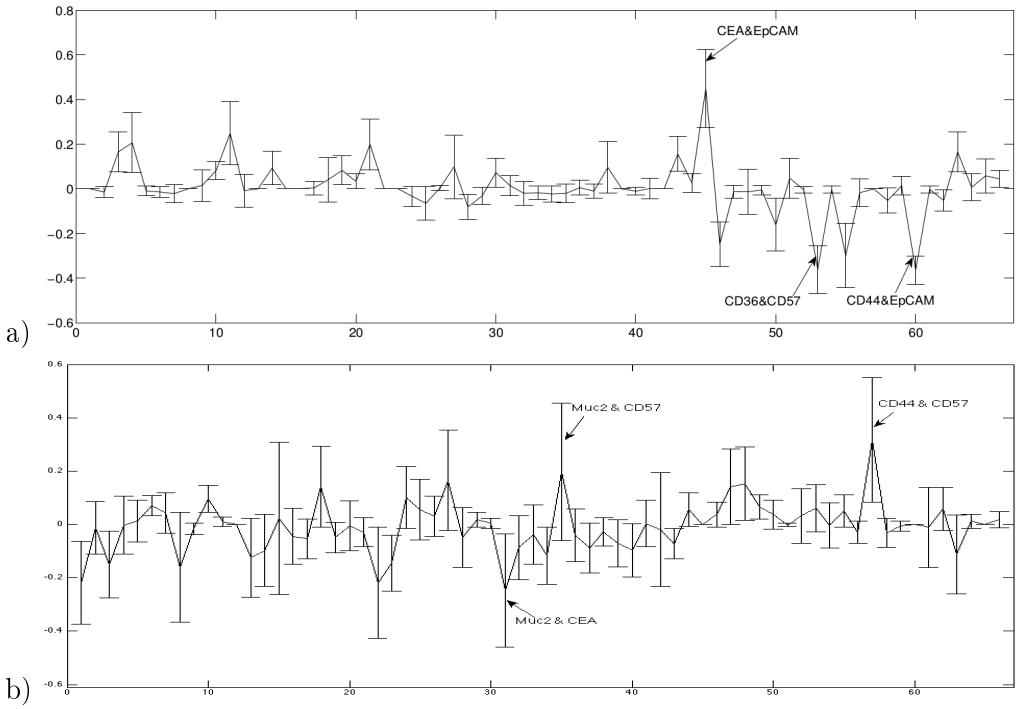


Figure 3.11: Mean (a) DiSWOP and (b) DiSWAP values (using the top 5 dependency scores) obtained using 16 different combinations of 3 cancer and 3 normal samples. The error bars are the size of one standard deviation. Numbers along the x-axis correspond to different protein pairs. Note that the labelled protein pairs are the same as the ones highlighted from the analysis of all 11 samples.

found when all 11 samples were analysed (Figures 3.8 and 3.8). The large standard deviation bars for some of the protein pairs illustrate the inter-sample heterogeneity.

A further experiment was performed in order to assess the significance of the results. For each protein channel and each cell we randomly permuted the pixel intensity values. This should break any real dependence between protein expressions and results in noise. Once the pixel intensities are permuted, the PPDP for each cell is calculated and DiSWOP analysis is performed as above. The experiment was run 11 times and results are shown in Figure 3.12. The scale in Figure 3.12 is different to that in Figure 3.11 due to the normalisation factors in the DiSWOP analysis.

3.8 Protein - Protein Interaction Pathways

In addition, some of these protein pairs have been experimentally found to interact or to be part of a pathway involved in colorectal cancer. For example, several studies have established that CEA and EpCAM interact through the pathway CEA

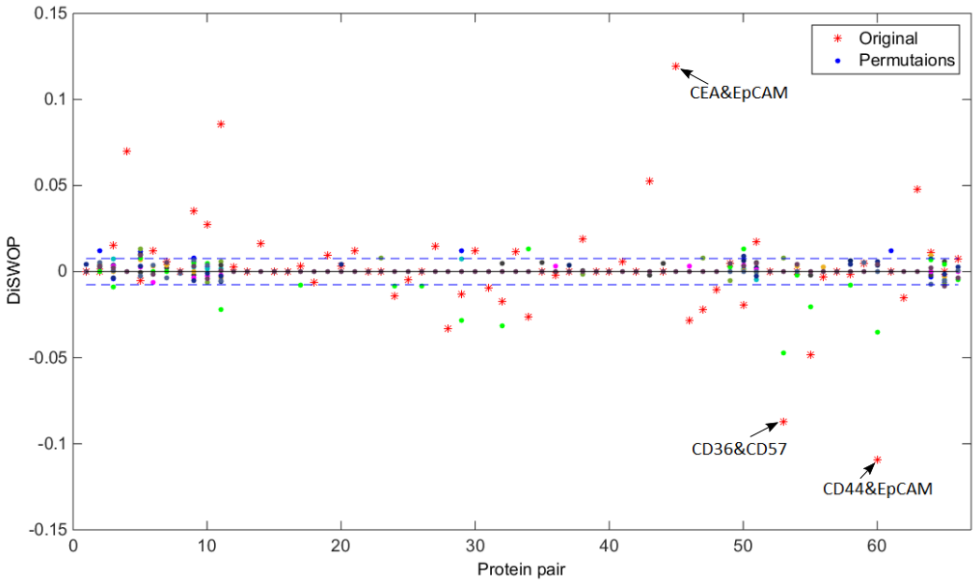


Figure 3.12: Results for DiSWOP from significance experiment using permuted pixel values. Red stars represent the DiSWOP values found in the original data. Circle markers of different colours demonstrate results from different permutation experiments. The dashed lines indicate the standard deviation of all permuted DiSWOP values.

– SOX9 – Claudin7 – EpCAM [198, 199, 200, 201] (Figure 3.13), which plays an important role in determining the morphology of the colon epithelium and promotes colorectal cancer progression [201]. In addition, physical interaction pathways have been established between CD44 and EpCAM, and between CD36 and CD57 (Figure 3.14) [198].

Further analysis of the results have been performed using an interactive tool for localisation of high PPD within the different samples, as shown in Figure 3.15. It enables the user to consider two protein pairs simultaneously and see where their PPD is above manually set thresholds (Figure 3.15). Alternatively, there is the option to see all cells in the samples coloured corresponding to the dependence between a selected protein pair. In this case, the PPDs are binned in intervals of size 0.2 and each cell is displayed in a corresponding colour (Figure 3.16). Screen-shots of the tool has been shown in Figure 3.15, which shows the cells expressing high PPD (above 0.7) of the two pairs CEA and EpCAM, and CD36 and CD57. We can easily see that normal and cancerous samples show differences in the distribution of high PPD for these two protein pairs. This tool confirms the heterogeneity of protein co-localisation both of neighbouring cells within the same tissue specimen and between different cancerous and normal samples. It could help identify complex biomarkers

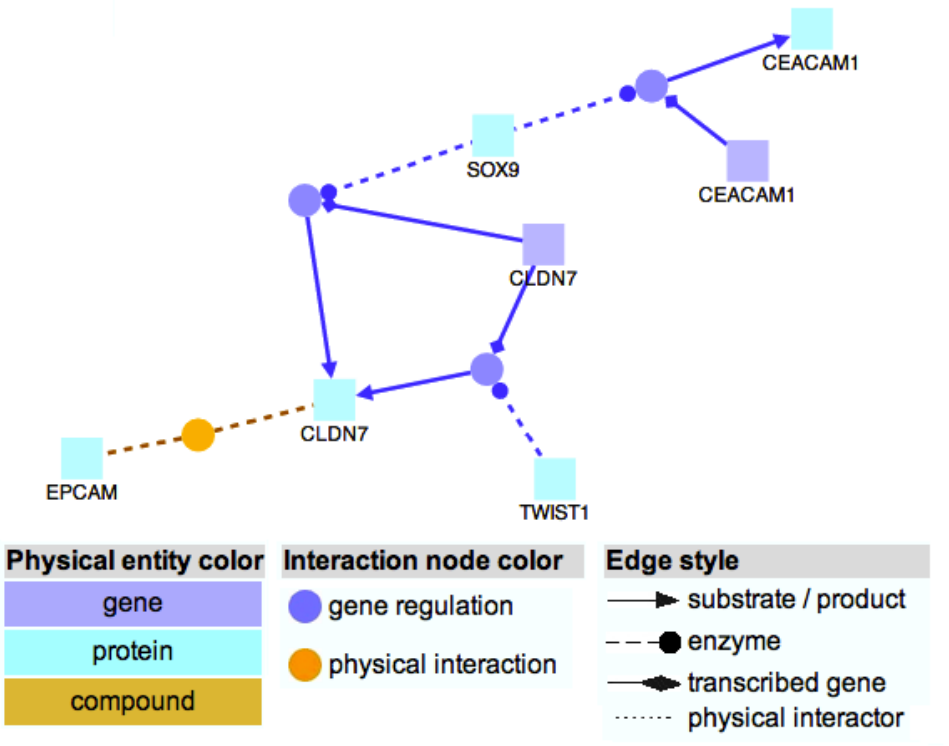


Figure 3.13: CEA and EpCAM interaction pathway [198]. Sox 9 has been found to activate expression of CEA [199] and mediate repression of Claudin-7 by Tcf-4 [200]. Claudin-7 and EpCAM have been found to co-express in colon tissue and possibly be part of a complex [201].

for cancer stem cells or cancer prognosis.

3.9 Discussion

The framework presented here is novel as it clusters the cells found in a sample, rather than the pixels, as in previous methods [92, 93, 94]. Hence this method enables us to consider the heterogeneity of the samples. Using the MIC scores means that the PPDP is considered rather than the raw expression profile. Therefore, the method is independent of the intensity of the images and hence different stacks can be considered simultaneously. Furthermore, it enables the identification of pairs of proteins which are more active in cancer cells than in normal cells and vice versa. The approach has been developed for images obtained using TIS, but it can also be easily used for other multi-variate imaging techniques, such as MALDI imaging [74], Raman microscopy [75] and multi-spectral imaging methods [76].

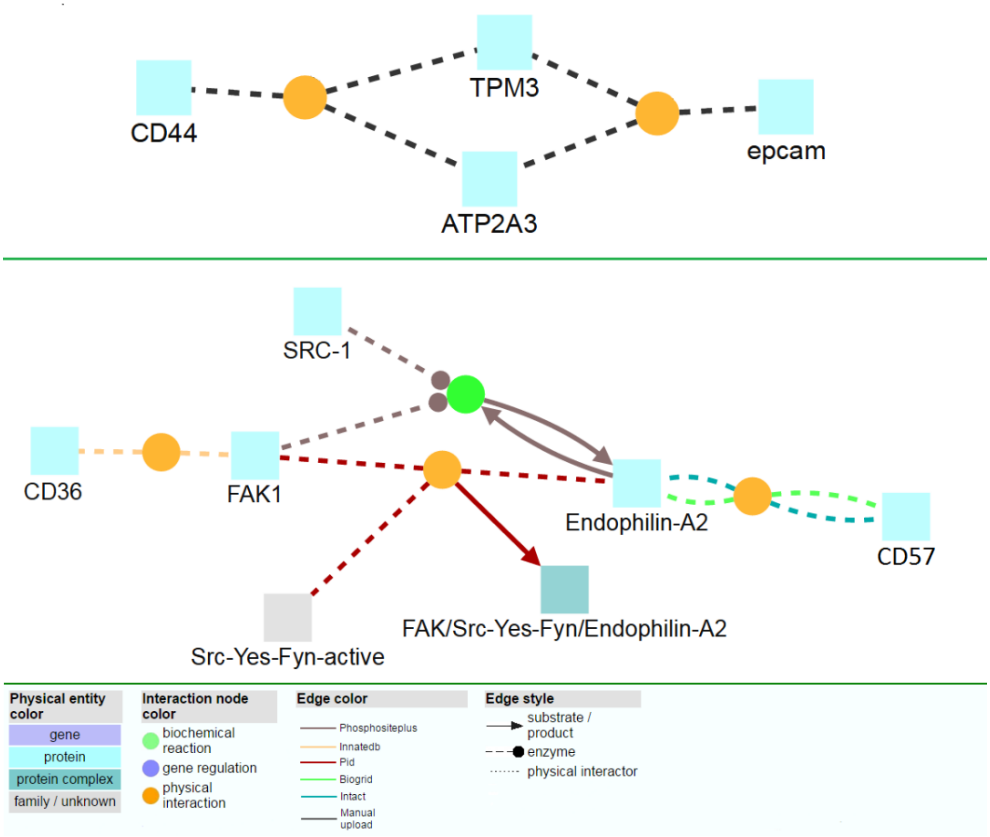
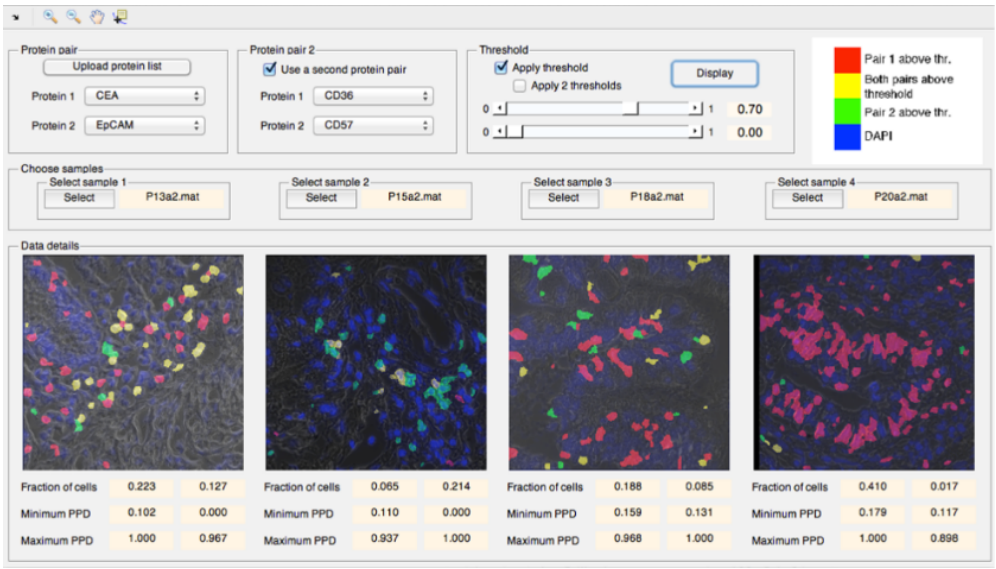


Figure 3.14: CD44 and EpCAM, and CD36 and CD57 interaction pathways [198].

The proteins used were not chosen because links between them were expected to show up in a protein network, but for a different scientific purpose, namely to help identify cell type. For this reason, relatively few links were considered significant, though with a compensating chance that these links were previously unknown. In the future, we will use additional proteins and we expect to find additional links. Previous work on exploring protein networks in colon cancer have used techniques like microarrays which, unfortunately, destroy all anatomical details. The advantage of our approach is that links in the protein network are found by studying individual cells. A disadvantage, however, is that we are restricted to at most 100 proteins, whereas microarrays measure expression of thousands of genes simultaneously.

The proposed measures could prove more useful once a membrane tag is used to help in a more accurate segmentation of cells. Many of the proteins considered are located in parts of the cell other than the nucleus and these interactions are currently not fully taken into account. Furthermore, a study with an extended tag library may reveal more prominent dependencies specific to cancerous tissue.

a)



b)

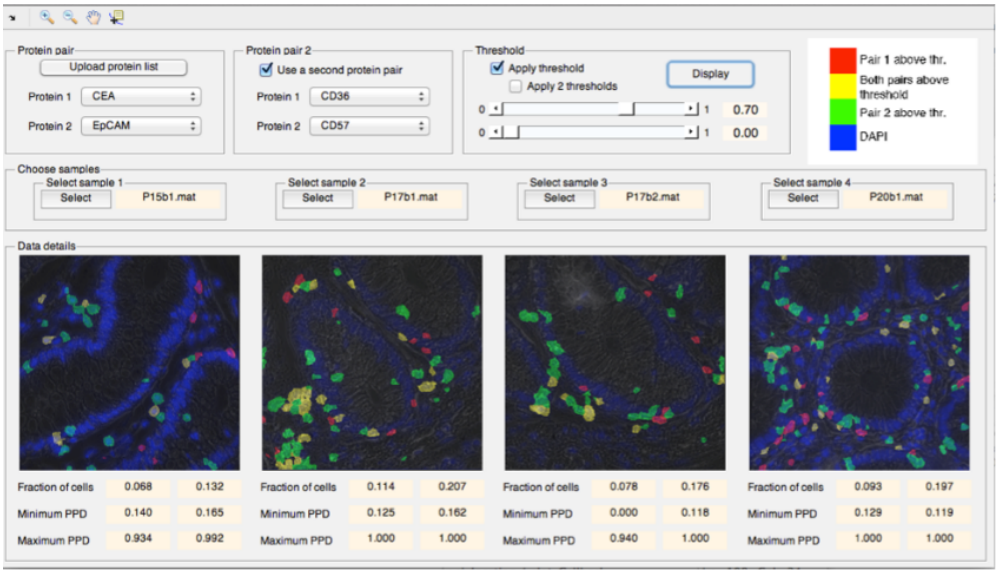


Figure 3.15: Screen-shots of the interactive tool for high PPD localisation. The tool displays the location of PPD above a threshold of 0.7 between CEA and EpCAM (in red) and between CD57 and CD36 (in green). Overlap between the two is shown in yellow and other nucleic regions are shown in blue. We can observe the heterogeneity of protein interactions in four (a) cancerous and (b) normal samples. In both figures colours are overlaid on top of a phase image. Below each sample is information about the fraction of cells above the threshold, the minimum and maximum PPD between each of the two protein pairs.

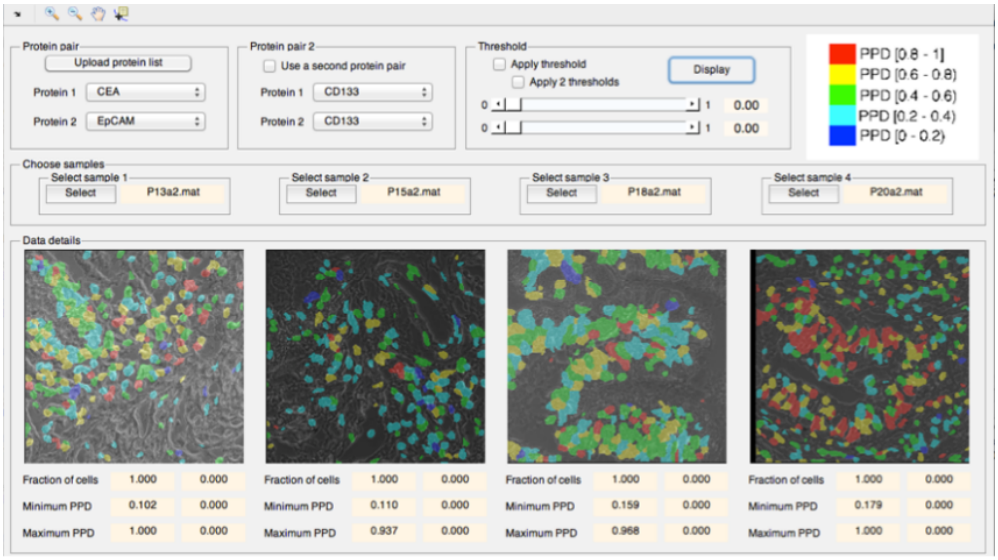


Figure 3.16: Screenshot of the interactive tool for high PPD localisation. It displays the heterogeneity of the distribution of cell phenotypes characterised by the co-localisation of CEA and EpCAM in four cancer samples. Each colour corresponds to an interval of values of PPD as shown in the legend. Colours are overlaid on top of a phase image. Below each sample is information about the fraction of cells above the threshold, the minimum and maximum PPD between each of the two protein pairs.

The binarisation method [73] introduced the ideas of lead and absent proteins in motifs of protein clusters, where a lead protein is one which is present after binarisation in all clusters and an absent protein is one which is not present in any of the clusters. These ideas in a way have been expanded by the DiSWOP and DiSWAP measures, which also identify colocation and anti-co-location, respectively. The quantities introduced here provide a measure of the degree, rather than a simple Yes-No classification, of the co-dependence of proteins. Furthermore, they overcome the fact that these proteins are found in both types of tissue by considering the difference between cancer and normal samples.

Chapter Summary

In this chapter we have introduced a novel method for analysing multi-label image data such as the TIS image data. The main novelties of the algorithm are that it performs cell rather than pixel level analysis of the samples, intensity independence, and phenotyping of cells based on their protein co-expression profile. We have considered several measures of dependence between protein expression and have selected the MIC due to its abilities to capture a wide variety of associations and its robustness. We have also compared three different methods for cell phenotyping and have shown that the results obtained from the new measures of co-dependence and anti-co-dependence, DiSWOP and DiSWAP, are independent of the choice of clustering. Due to the general nature of the framework, the method could be applied to other tissues and/or images obtained from other multivariate imaging techniques.

Applying these over a TIS dataset of eleven samples of cancerous and normal colon tissue, we have found protein pairs that are much more co-dependent or anti-codependent in cancerous than in normal tissue, pointing to the possibility that combinations of protein pairs rather than single proteins will lead to specific markers for cancer. The results presented here are only preliminary and need to be validated using a larger number of samples and subsequently by other biological techniques. While the number of samples considered is insufficient to draw significant biological conclusions, this is the largest study of colon cancer using TIS conducted to date. Furthermore, we have performed several validation checks which give confidence that our novel measures can help identify and quantify important examples of codependence and anti-codependence of protein pairs.

Chapter 4

A Model of Spatial Tumour Heterogeneity in Colorectal Adenocarcinoma Tissue

The validation of quantitative results from analysis of bioimages poses a great challenge. This is due to the lack of ground truth data. One way of solving this problem is to generate realistic synthetic data where the ground truth is generated as part of the model. Hence, in this chapter we develop a model of the colon tissue architecture both for healthy and cancerous cases at different stages. As the architecture is mostly characterised by the differentiation grade, we focus on how this parameter affects the crypt structures and cell phenotypes present in the tissue. As most imaging in clinical pathology labs is done using histology techniques, we consider extending the model from IF to IHC data. This greatly improves the model's usability. An overview of the model is presented in Figure 4.1.

4.1 Materials and Methods

4.1.1 Data acquisition

In order to make the model realistic, H&E slides from colon cancer patients were analysed. The slides were digitally scanned at 40× magnification by Zeiss MIRAX MIDI Slide Scanner. For cell-level analysis, a total of 42 visual fields at 40× magnification were considered. These, including a context at 4× magnification, were graded by three pathologists and the majority vote was taken. The visual fields were categorised as 7 healthy, 4 well-differentiated, 26 moderately differentiated and 5 poorly differentiated samples. Individual nuclei in each image were hand-marked

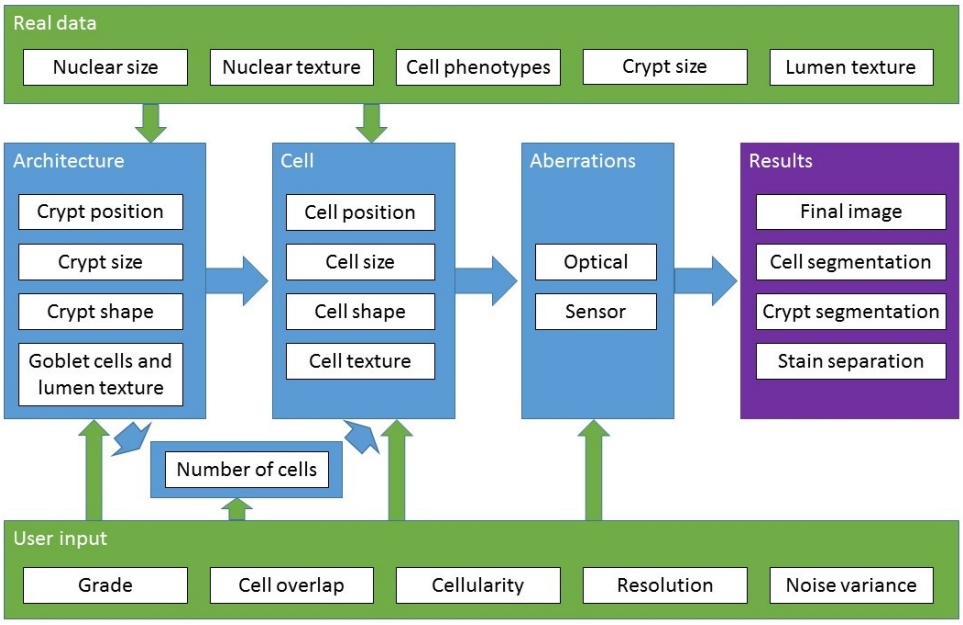


Figure 4.1: Flowchart of the simulation process. Blue boxes indicate parts of the model, green boxes contain model inputs, purple box shows the outputs. The sample grade and crypt sizes from real data input into the architecture generated. The number of cells is determined by the architecture and the user-defined cell overlap and cellularity. Cells are then iteratively generated with input of the cell phenotype distributions and the nuclear sizes and texture found in the real data. Ideal images are then degraded in order to mimic errors in an image acquisition system with parameters of noise variance defined by the user. In addition to the final image, various ground truth data is output.

as epithelial or stromal. A total of 5,826 nuclei were hand-marked for analysis. In addition, 31 visual fields at 20× were selected for analysis of the crypt structures. These were split into 9 healthy and 22 cancerous samples. In these, 480 healthy and 396 cancerous crypts were hand-marked. More cancerous samples were required to obtain a similar number of crypts as cancerous crypts tend to be significantly larger.

4.1.2 Learning from the real data

As whole-cell segmentation is difficult to obtain from H&E slides, we concentrate on studying the nuclear regions. This approach is supported by findings that the nucleus can hold the key to understanding cell function [15]. In order to extract cell information visual fields at 40× magnification were analysed. Size and 13 Haralick texture features were extracted for each nucleus. Affinity Propagation [192] was used to phenotype the nuclei according to the textural features. This clustering

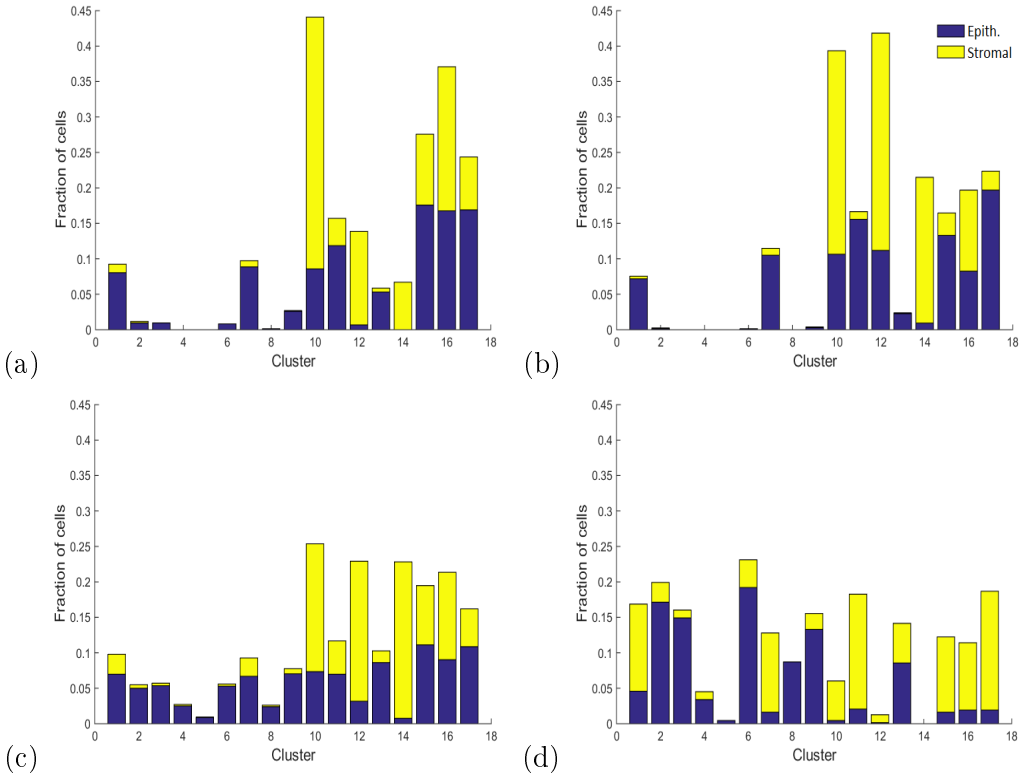


Figure 4.2: Frequency of each type of cell belonging to a phenotype. (a) shows frequencies for healthy epithelial and stromal cells. (b)-(d) show frequencies for well, moderately and poorly differentiated, respectively.

algorithm automatically determines the number of clusters found in the data. The main purpose for clustering the cells is to group together cells with similar texture and increase the texture samples available for texture synthesis. For each of the 17 phenotypes found in this way, mean and standard deviation of the length of the major axis and the ratio between the minor and major axes were obtained (Table 4.1). In addition, we calculated the frequency with which nuclei belonging to each phenotype are found to be epithelial or stromal, and incorporate the phenotype frequency into our model as described in Section 2.4. These frequencies are shown in (Figure 4.2). Some of these phenotypes were found to contain mostly cancerous epithelial cells (Figure 4.3 top row), whereas others consisted of predominantly stromal cells (Figure 4.3 bottom row). The average profiles for size and texture features are shown in Tables 4.2 and 4.3, respectively. In addition, we obtained hand-marked images for crypt texture. One image was used to obtain healthy lumen texture (Figure 4.4 (a)). Seven crypts from different cancer samples were also marked and texture was extracted. Figures 4.4 (c) and (e) show two of these.

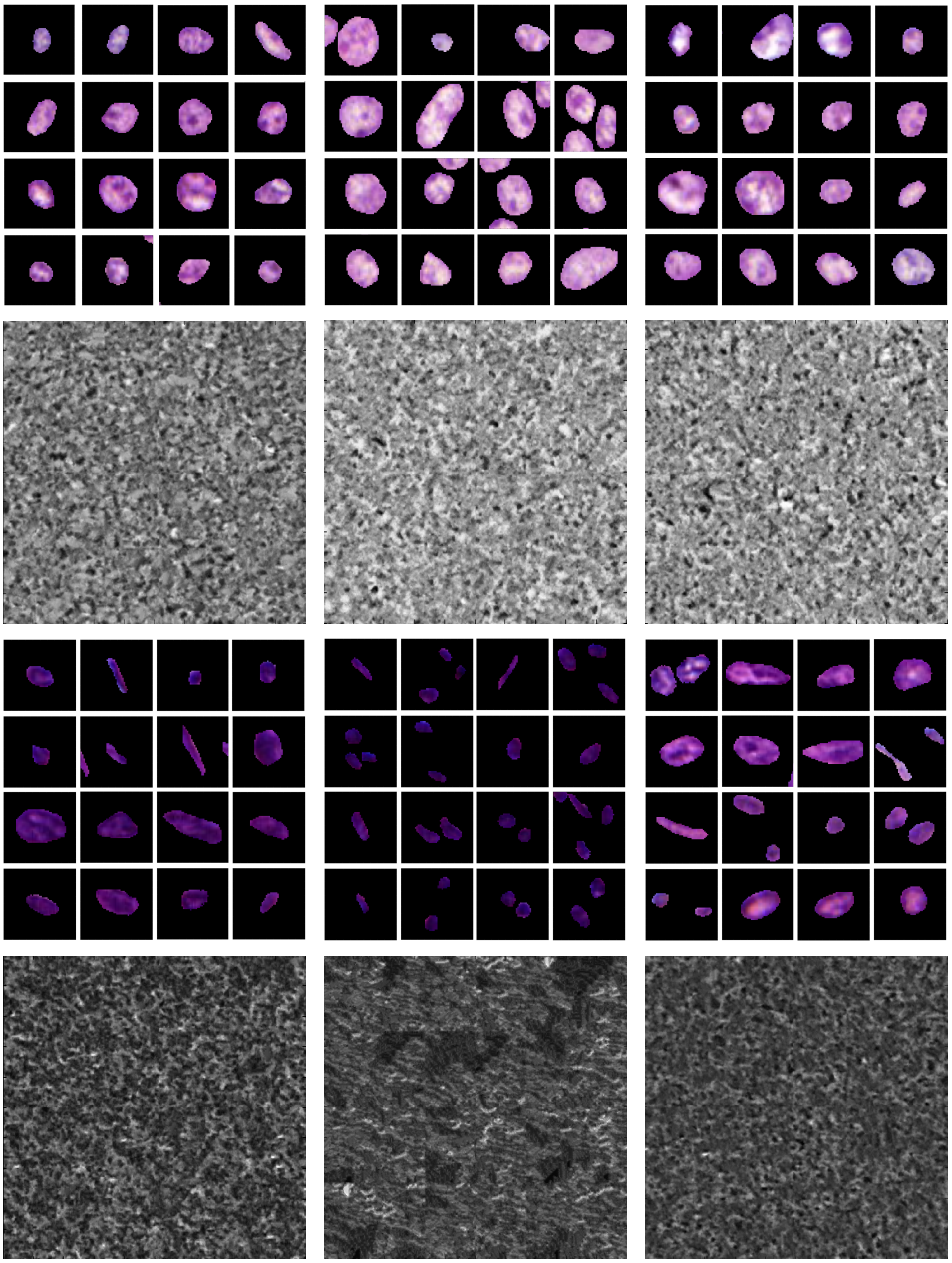


Figure 4.3: Selection of cells belonging to different phenotypes with the corresponding texture images below. The phenotypes shown are numbers 2, 3, 8, 12, 14, and 17 from Figure 4.2. One can easily see that the first row of phenotypes contains mostly tumour and epithelial cells, whereas the second one consists mostly of stromal cells.

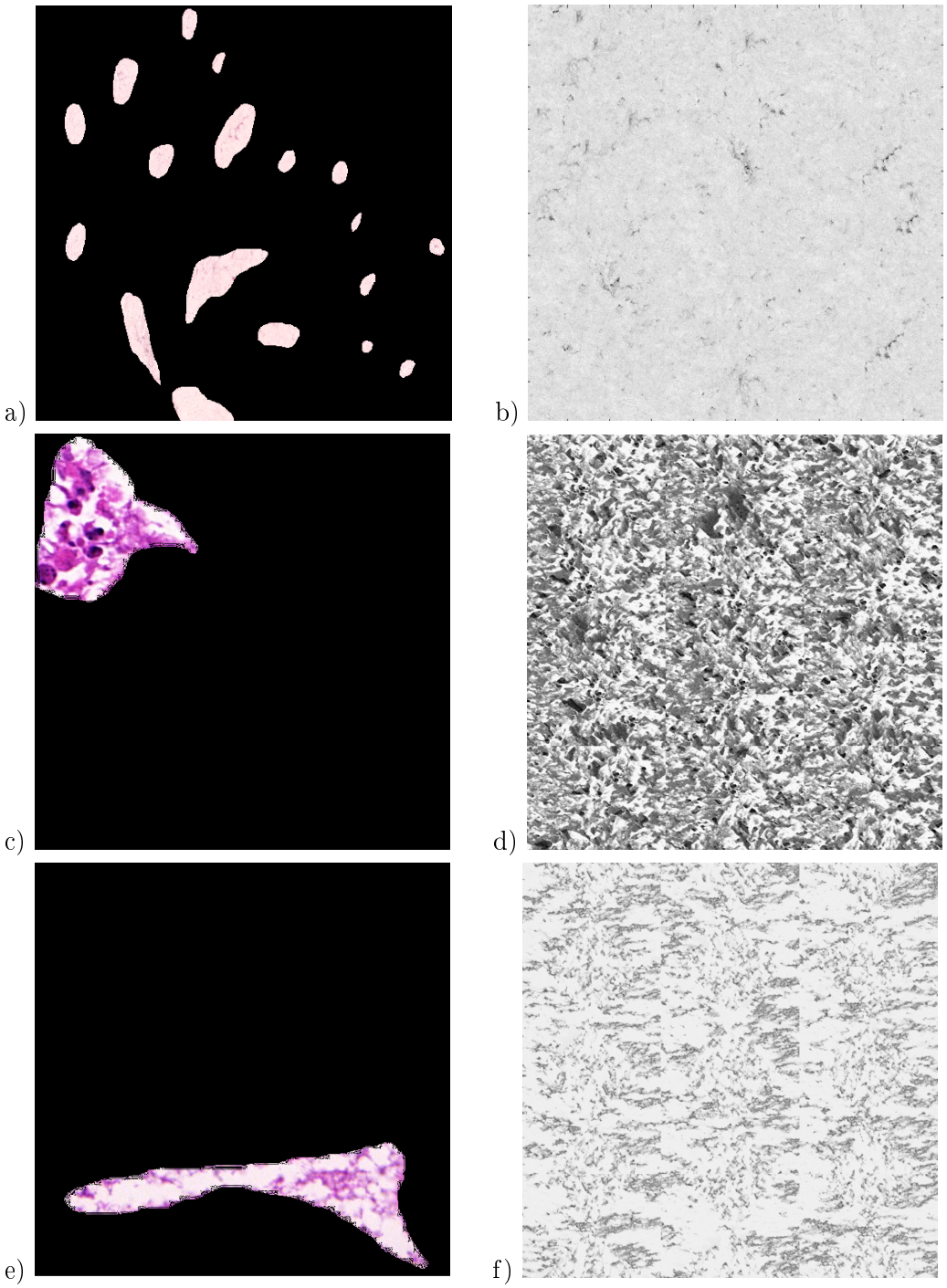


Figure 4.4: Obtaining lumen texture. Figure (a) shows extracted lumen texture from a healthy sample. Figures (c) and (e) show two of the extracted lumen texture from cancer samples. Figures (b), (d) and (f) show the respective generated texture images.

Table 4.1: Main parameters of the model.

Description	Annotation	Source	Typical Values
Image size	$i_h \times i_w$	User-defined	1000×1000
Magnification		User-defined	$40\times$, $20\times$
Size of CCD pixel		User-defined	$11\mu m$
Cancer grade	S	User-defined	$\{0, 1, 2, 3\}$
Cellularity of epithelial cells	ν_e	User-defined	$[0, 1]$
Cellularity of stromal cells	ν_s	User-defined	$[0, 1]$
Cell overlap	L_{max}	User-defined	$[0, 1]$
Variance of point spread function	G	User-defined	1
Variance of the CCD detector noise	σ_G	User-defined	0.00025
Stain matrix		User-defined	
Distribution of nuclei major axis length	μ_l, σ_l	H&E data	
Distribution of nuclei minor axis length	μ_w, σ_w	H&E data	
Distribution of crypt minor axis length	μ_b, α_b, β_b	H&E data	
Distribution of crypt ratio between axes	α_e, β_e	H&E data	
Distribution of cell phenotypes		H&E data	
Approximate cell radius	r	[202]	$6\mu m$
Number of crypts	N_c	Eq. 4.1	
Rotation of crypts	ϕ	Random	$[0, 2\pi]$
Number of cells	N	Section 4.1.3	

Table 4.2: Size feature profiles for nuclear texture phenotypes found in the real data. Sizes are in pixels for 40× images.

Phenotype	Area	Minor axis	Minor/ Major
1	379.52 ±196.62	26.71 ±8.75	0.682 ±0.161
2	420.13 ±175.84	27.68 ±7.25	0.704 ±0.150
3	464.99 ±217.80	28.54 ±7.27	0.729 ±0.147
4	494.08 ±228.90	28.99 ±7.99	0.744 ±0.134
5	577.00 ±291.98	32.23 ±10.68	0.716 ±0.150
6	446.53 ±192.06	28.46 ±7.12	0.709 ±0.152
7	391.79 ±197.94	27.19 ±8.14	0.675 ±0.172
8	504.45 ±225.19	29.48 ±7.10	0.732 ±0.127
9	419.92 ±187.77	27.76 ±7.40	0.702 ±0.150
10	179.70 ±166.93	18.62 ±9.51	0.637 ±0.197
11	337.19 ±188.38	25.54 ±8.81	0.656 ±0.172
12	167.03 ±143.80	20.16 ±10.02	0.562 ±0.217
13	418.80 ±210.43	28.18 ±8.28	0.678 ±0.172
14	103.40 ±56.67	16.97 ±6.83	0.548 ±0.231
15	323.81 ±205.63	25.45 ±9.38	0.626 ±0.175
16	247.69 ±200.77	21.86 ±10.03	0.634 ±0.195
17	356.95 ±205.14	26.17 ±8.82	0.654 ±0.172

Table 4.3: Texture feature profiles for phenotypes found in the real data.

Phenotype	Energy	Contrast	Corr.	SOSV	IDMN	Sum average	Sum variance	Sum entropy	Entropy	Difference variance	Difference entropy	Inf1	Inf2
1	0.172 ±0.055	0.800 ±0.381	0.797 ±0.123	12.54 ±0.74	0.988 ±0.005	6.46 ±0.26	28.53 ±1.46	1.93 ±0.24	2.19 ±0.30	0.800 ±0.381	0.772 ±0.119	-0.470 ±0.084	0.850 ±0.065
2	0.164 ±0.048	1.017 ±0.283	0.825 ±0.071	18.01 ±0.88	0.987 ±0.003	7.75 ±0.29	44.74 ±1.46	1.98 ±0.22	2.20 ±0.27	1.017 ±0.283	0.771 ±0.099	-0.478 ±0.069	0.859 ±0.050
3	0.155 ±0.047	1.177 ±0.328	0.838 ±0.061	22.08 ±1.03	0.985 ±0.004	8.57 ±0.29	56.64 ±2.12	2.06 ±0.21	2.27 ±0.26	1.177 ±0.328	0.784 ±0.090	-0.484 ±0.064	0.869 ±0.049
4	0.154 ±0.046	1.312 ±0.346	0.844 ±0.055	27.13 ±1.12	0.984 ±0.004	9.57 ±0.28	72.93 ±2.70	2.04 ±0.21	2.24 ±0.24	1.312 ±0.346	0.768 ±0.085	-0.488 ±0.058	0.869 ±0.043
5	0.153 ±0.064	1.384 ±0.529	0.847 ±0.074	31.55 ±1.82	0.983 ±0.006	10.38 ±0.40	87.08 ±4.50	2.05 ±0.25	2.25 ±0.27	1.384 ±0.529	0.773 ±0.077	-0.483 ±0.082	0.863 ±0.078
6	0.159 ±0.048	1.096 ±0.270	0.833 ±0.055	19.80 ±0.82	0.986 ±0.003	8.12 ±0.25	49.95 ±1.62	2.02 ±0.21	2.24 ±0.27	1.096 ±0.270	0.780 ±0.101	-0.482 ±0.063	0.866 ±0.043
7	0.191 ±0.072	0.652 ±0.268	0.801 ±0.118	10.88 ±0.69	0.991 ±0.004	6.04 ±0.24	24.36 ±1.02	1.85 ±0.26	2.08 ±0.32	0.652 ±0.268	0.728 ±0.108	-0.477 ±0.081	0.844 ±0.077
8	0.143 ±0.039	1.191 ±0.308	0.848 ±0.052	24.73 ±1.01	0.985 ±0.004	9.11 ±0.26	64.43 ±2.36	2.10 ±0.20	2.32 ±0.24	1.193 ±0.308	0.784 ±0.084	-0.490 ±0.059	0.877 ±0.043
9	0.171 ±0.055	0.961 ±0.276	0.822 ±0.070	16.26 ±0.86	0.987 ±0.003	7.35 ±0.27	39.61 ±1.54	1.96 ±0.24	2.19 ±0.29	0.961 ±0.276	0.768 ±0.103	-0.479 ±0.068	0.857 ±0.053

Table 4.3: Continued

Phenotype	Energy	Contrast	Corr.	SOSV	IDMN	Sum average	Sum variance	Sum entropy	Entropy	Difference variance	Difference entropy	Inf1	Inf2
10	0.333 ±0.106	0.375 ±0.186	0.554 ±0.203	4.03 ±0.48	0.994 ±0.003	3.79 ±0.21	8.04 ±0.83	1.25 ±0.24	1.53 ±0.29	0.375 ±0.186	0.613 ±0.128	-0.346 ±0.130	0.665 ±0.142
11	0.220 ±0.093	0.650 ±0.291	0.765 ±0.139	9.46 ±0.69	0.991 ±0.004	5.65 ±0.23	21.00 ±0.96	1.73 ±0.29	1.97 ±0.34	0.650 ±0.291	0.717 ±0.119	-0.458 ±0.094	0.816 ±0.092
12	0.337 ±0.073	0.302 ±0.143	0.574 ±0.156	2.91 ±0.35	0.996 ±0.002	3.22 ±0.21	5.40 ±0.70	1.16 ±0.15	1.38 ±0.21	0.302 ±0.143	0.571 ±0.126	-0.310 ±0.128	0.608 ±0.127
13	0.179 ±0.066	0.852 ±0.298	0.820 ±0.089	14.30 ±0.92	0.988 ±0.004	6.88 ±0.32	33.94 ±1.60	1.93 ±0.26	2.16 ±0.32	0.852 ±0.298	0.757 ±0.109	-0.478 ±0.079	0.852 ±0.066
14	0.503 ±0.167	0.212 ±0.098	0.459 ±0.175	1.81 ±0.40	0.997 ±0.002	2.54 ±0.26	3.49 ±0.39	0.86 ±0.26	1.01 ±0.31	0.212 ±0.098	0.483 ±0.132	-0.209 ±0.116	0.434 ±0.158
15	0.231 ±0.068	0.518 ±0.262	0.745 ±0.131	6.91 ±0.57	0.993 ±0.004	4.85 ±0.24	14.40 ±1.02	1.61 ±0.23	1.89 ±0.27	0.518 ±0.262	0.675 ±0.122	-0.454 ±0.097	0.807 ±0.078
16	0.251 ±0.098	0.490 ±0.256	0.635 ±0.206	5.41 ±0.58	0.993 ±0.004	4.34 ±0.25	10.86 ±0.95	1.46 ±0.25	1.77 ±0.28	0.490 ±0.256	0.671 ±0.133	-0.404 ±0.122	0.751 ±0.124
17	0.226 ±0.085	0.576 ±0.375	0.766 ±0.138	8.22 ±0.64	0.992 ±0.005	5.27 ±0.24	17.68 ±1.03	1.6920 ±0.27	1.95 ±0.31	0.576 ±0.375	0.694 ±0.115	-0.460 ±0.095	0.816 ±0.091

In addition to this, visual fields at 20× magnification were selected for the analysis of crypt shapes and sizes. We calculated the distributions of the minor axis and the ratio between minor and major axes for each group. These were modelled as Gamma functions and the parameters were incorporated into the model. The fit of the Gamma distributions is shown in Figure 4.5.

4.1.3 Tissue structure

In this section we describe how the tissue microenvironment in CRA is modelled. We begin by explaining the overall organisation in terms of the crypts and stroma. We then describe how individual cells are modelled.

Crypts

Given an image resolution and magnification level, we assume the appropriate radius, r, of the cells to be $6\mu m$ [202], while a suitable value for the radius of the crypts corresponds to the mean length on the minor axis, μ_b, found from the H&E images (Section 4.1.2). The generated image depends on the differentiation grade, S, of the CRA, which can take the values of $[0, 1, 2, 3]$, corresponding to healthy tissue, well differentiated, moderately differentiated and poorly differentiated cancers, respectively. The number of crypts and cells to be simulated in the image are determined using their rough sizes. The number of crypts, N_c in an $i_h \times i_w$ image is determined as follows:

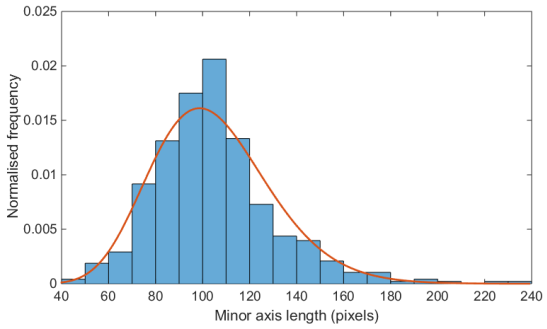
$$N_c = f_c \lfloor i_h/(2\mu_b) \rfloor \lfloor i_w/(2\mu_b) \rfloor. \tag{4.1}$$

where f_c is the fraction of the sample covered in crypts and is given by

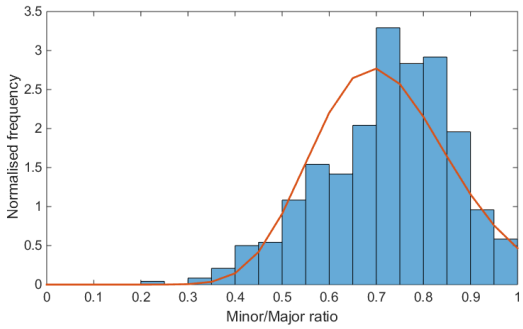
$$f_c = \begin{cases} 1, & \text{if } S = 0, 1 \\ U(0.5, 0.95)|U(1, 1.3), & \text{if } S = 2 \\ U(0, 0.5), & \text{if } S = 3, \end{cases} \tag{4.2}$$

where $U(x_1, x_2)$ is a number uniformly drawn from the range $[x_1, x_2]$. There are two cases if $S = 2$ selected at random with equal probability, corresponding to fewer crypts than normal tissue or overcrowding (with overlapping) of crypts. Both of these phenomena can be observed in moderately differentiated CRA. The value ranges for f_c were determined from pathology guidelines and discussions with pathologists. To create colon tissue structure (Figure 1.4) crypts are simulated as elliptical structures. For each crypt, the minor axis b is sampled from the Gamma distribution $\Gamma(\alpha_b, \beta_b)$,

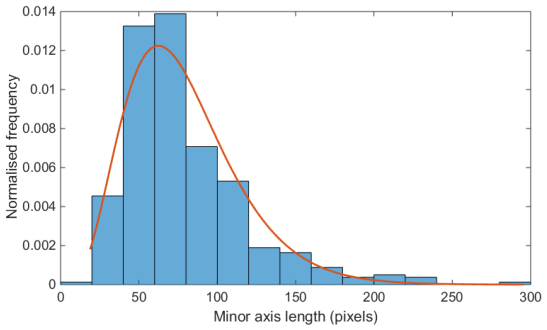
a)



b)



c)



d)

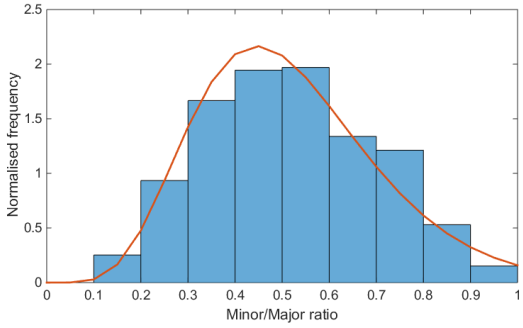


Figure 4.5: Distribution of crypt shape parameters extracted from the real data. Figures (a) and (b) show the minor axis length and the ratio between the minor and major axes, respectively, for healthy crypts. Figures (c) and (d) show the minor axis length and the ratio between the minor and major axes, respectively, for cancerous crypts. Frequencies are normalised so that sum of areas of bars equals 1.

where α_b and β_b are the parameters for the distribution of the minor axis estimated from the H&E images. To determine the length of the major axis, a, we use the ratio between the minor and major axes, $e = b/a$. Then a is given by $b/(\Gamma(\alpha_e, \beta_e))$, where α_e and β_e are the parameters for the distribution of the ratio (Table 4.1). The degree of rotation of the major axis, ϕ, of the crypts is chosen at random. The crypt outline is then computed as follows,

$$R(\theta) = \frac{ab\sqrt{2}}{\sqrt{(b^2 - a^2)\cos(2\theta - 2\phi) + a^2 + b^2}} + u, \tag{4.3}$$

where $R(\theta)$ is the polar radius, $\theta \in [0, 2\pi]$ is the polar angle and $u = (S^2 + 1)U(-0.06, 0.1)$ is a degree of deformation of the crypts, a function of the grade S.

Then, the crypt centres, $\mathbf{c} = (x_c, y_c)$, are selected so that the crypts don't overlap for healthy or well differentiated samples. For tissues of grades 2 and 3, at most 2 ellipses can overlap to a certain extent. In these cases, one crypt would be modelled by several overlapping deformed ellipses. This generates the "gland within gland" phenomenon and more complex glandular structures often observed in higher grade cancers. In order to speed up the selection of the crypt centre, we only consider a sample of points in a randomly placed grid structure with distance between vertices of $0.6b$. The epithelial cells are placed at a random location along or close to the crypt edge according to the equation

$$\begin{aligned} x &= x_0 + rGu_x \\ y &= y_0 + rGu_y, \end{aligned} \tag{4.4}$$

where (x_0, y_0) is a randomly selected point on the outline of the crypt, and u_x and u_y are random scaling factors with $u_{x,y} = U(-0.75, 0.25)/3$. It is difficult to extract the exact value of this parameter from real data, so the range was chosen with the aim to maximise the visual similarity between the real and synthesised images. Hence, in healthy tissue the epithelial cells are attached to the crypt boundary and the structure becomes increasingly distorted for higher differentiation grades. Once the cells are placed, they are rotated so they point towards the crypt centre and, if $G < 2$, their nuclei are displaced closer to the edge of the crypt. The stromal cells are placed uniformly in the space outside the crypts. All stromal cells are rotated in a direction given by $\phi + U(-\pi/6, \pi/6)$ (Table 4.1), to reflect the structure of the stromal tissue that can be observed in histology images.

Number of cells

The maximum amount of cell overlap is controlled by a parameter L_{max}. The relative amount of overlap, L_{ij}, that is caused on the region of pixels R_i defined by one simulated cell by the region of pixels R_j of another cell is measured by

$$L_{ij} = \frac{|R_i \cap R_j|}{|R_i|}, i \neq j \tag{4.5}$$

where $|\cdot|$ is the cardinality of a set. With this definition setting $L_{max} = 1$ doesn't pose any restrictions on overlap, whereas $L_{max} = 0$ doesn't allow any overlap. Overlap can be controlled either on the cytoplasm or nuclei regions. When a cell is placed randomly, if it overlaps with an already placed cell to an extent that is greater than L_{max}, a new set of coordinates is chosen.

In addition to this, in poorly differentiated samples, we place clusters of cancer cells in the stroma. Stromal cells are placed within a cluster with probability of 50%. A cluster is a region of size $10r \times 10r$ and cells placed in it have value of maximum overlap equal to $min(2L_{max}, 0.8)$.

Once the number and size of crypts has been determined and the crypts have been placed, we calculate the number of cells, N that will be placed in the image. Firstly, an estimate of the area of a stromal cell, A is calculated:

$$A = \pi[(1.7 - 0.7L_{max})r]^2. \tag{4.6}$$

Here the multiplication factor of r accounts for the effect of overlap and doesn't go below 1 as stromal cells are generally sparse. The area covered by stroma, A_s is found by counting the pixels outside the outlines of the crypts. Then the number of stromal cells is given by $N_s = \nu_s A_s / A$, where $\nu_s \in [0, 1]$ is a user-defined parameter for the cellularity (density) of stromal cells.

Similarly, the number of epithelial cells is determined by

$$N_e = \frac{\nu_e P}{2(1.25 - L_{max})r}. \tag{4.7}$$

where P is the sum of the perimeters of the crypts in the image, $\nu_e \in [0, 1]$ is a user-defined parameter for the cellularity of epithelial cells, and the factor in the denominator accounts for the effects of overlap. The overlap factor is smaller than the one for stromal cells because epithelial cells are more tightly packed. Then the final number of cells is given by $N = N_s + N_e$.

Lumen and goblet cells

When a sample is being generated, the inside of the crypts is filled with lumen texture. In order to generate the lumen, we employed the non-parametric model presented by Efros and Leung [203] which generates texture from a given source image. In this framework, the value of a pixel is determined by finding all patches in the source image that resemble the filled part of the neighbourhood of the pixel in question. One of these patches is selected at random and the value of the centre pixel is assigned to the pixel to be filled. We model the gray-scale texture of hand-marked lumen regions from the H&E training images (Section 4.1.2) in order to generate a large texture image corresponding to each crypt texture (Figure 4.4). Seven textures were generated for cancer crypts and one for normal lumen texture. When a crypt is being synthesised, a random part of a texture image is selected and used as the texture. For healthy samples, the normal lumen texture is used. When a cancer sample is being generated, a texture image is selected at random for each crypt.

In healthy samples once the lumen texture is placed, we generate the goblet cells structure. This is done using Voronoi diagrams [204]. The crucial step when generating a Voronoi diagram is to select the centres of gravity for the regions. The observed structure of the goblet cells depends on the angle at which the crypt is sliced through (Figure 4.6). Alternatively, we can consider the ratio e between the minor and major axes of the crypt as a surrogate indicator of the structure observed. If $e \approx 1$, (i.e., a round crypt) we get a single ring of goblet cells (Figure 4.6 (a)). The number of goblet cells in this ring for a particular crypt is given by $\gamma = a/r$. However, if $e < 1$, we define $\kappa \approx 1/e, \kappa \in \mathbb{N}$, and we get additional $2\kappa(\kappa - 1)$ goblet cells around each end of the major axis of the crypt. To determine their location, we take even angular increments from the centre of the ellipse and place the points on the outer ring a distance from the crypt boundary equal to the cell radius r. The additional points are placed along the 2κ angles closest to the major axis a distance $2i, i = 2, ..., \kappa$ from the boundary (Figure 4.7). A centre of gravity for the Voronoi diagram is also added at the centre of the crypt. A small amount of variation is allowed for the location of each point and the Voronoi diagram is generated. To make the boundaries more realistic, they are dilated and the corners at each Voronoi vertex are rounded using dilation. Some texture [205] is added to the boundaries, they are convolved with a Gaussian and added to the final image.

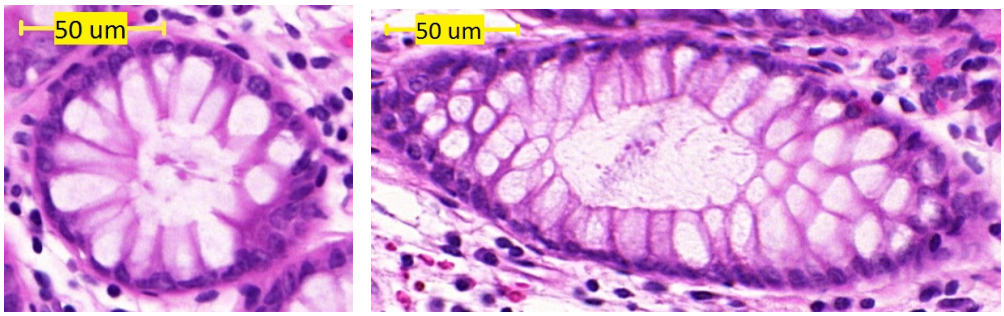


Figure 4.6: Different goblet cell structures. A roughly circular crypt is shown on the left ($\kappa = 1$) and a more elliptical ($\kappa = 3$) on the right. Scalebars are $50\mu m$.

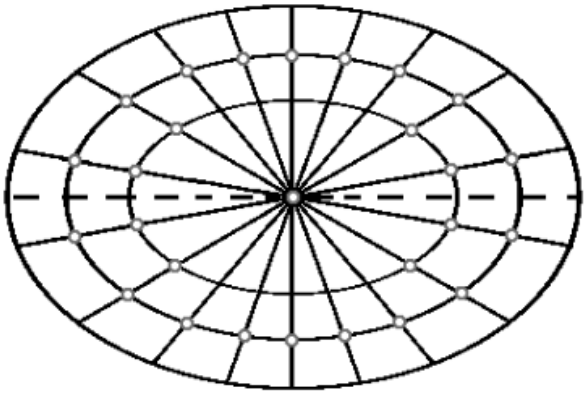


Figure 4.7: An illustration of the initial locations for the centres of gravity (grey circles) for Voronoi diagram in a crypt with $\kappa = 2$. Dashed line gives the major axis.

4.1.4 Single cell

Each of the N cells is constructed separately. Before a cell is synthesised, it is assigned to one of the phenotypes found in the real data with probability equal to the probability of the phenotypes in real H&E tissues of the same grade. We then generate images for the cell cytoplasm and nucleus.

Shape

Two types of shapes are included in our model. First, the cytoplasm for stromal cells and cell nuclei are generated using a parametric model proposed in [153]. In this case, the shapes are initialised as a circle parametrised by $(x(\theta), y(\theta))$, where $\theta \in [0, 2\pi]$ is the polar angle. The angle θ is sampled at $k(k = 10)$ equidistant points to generate a regular polygon (Figure 4.8 (a)). Then a random polygon is created

by randomising the spatial locations of the vertices as follows:

$$\begin{aligned} x_i(\theta_i) &= [U(-\alpha,\alpha) + cos(\theta_i + U(-\beta,\beta))], \\ y_i(\theta_i) &= [U(-\alpha,\alpha) + sin(\theta_i + U(-\beta,\beta))] \end{aligned} \tag{4.8}$$

for $i = 1, ..., k$, where α controls the randomness of the circle radius and β controls the randomness of the angle of sampling. The value for α is dependent on the cancer grade by $\alpha = 0.1(G+1)$, whereas the value of β has been set to 0.05. Taking $k = 10$ is a good compromise between taking too few points and not allowing sufficient control over the shape (Figure 4.8 (e, f)), and taking too many points and obtaining complicated shapes unrealistic for cells in a tissue environment (Figure 4.8 (g, h)). Then we obtain the means, μ_l and μ_w, and standard deviations, σ_l and σ_w, for the nuclei major and minor axes, respectively, from the H&E data phenotypes. These are used to obtain the sizes for the modelled nuclei as

$$\begin{aligned} \mu_l^n &= N(\mu_l, \sigma_l), \\ \mu_w^n &= N(\mu_w, \sigma_w). \end{aligned} \tag{4.9}$$

Then, the size of the modelled cell cytoplasm is chosen to be

$$\begin{aligned} \mu_l^c &= U(1.5, 2.2)\mu_l^n, \\ \mu_w^c &= U(1.5, 2.2)\mu_w^n \end{aligned} \tag{4.10}$$

The lack of a membrane marker makes it difficult to obtain exact cell size estimates but the interval 1.5–2.2 gives a good approximation of observation from real data (Figure 1.6). Normal stromal cells are assigned with equal probability to be either fibroblasts or lymphocytes. For cancer samples, the stromal cells are assigned to be cancerous with probability $1 - 0.2S$. In order to ensure realistic appearance of the stromal cells, the fibroblast cell sizes are rescaled as

$$\begin{aligned} \hat{\mu}_w^n &= 0.8\mu_w^n, \\ \hat{\mu}_l^c &= 1.8\mu_l^c, \\ \hat{\mu}_w^c &= 0.5\mu_w^c \end{aligned} \tag{4.11}$$

and for lymphocytes as

$$\begin{aligned} \hat{\mu}_l^n &= 0.8\mu_l^n, \\ \hat{\mu}_w^n &= 0.8\mu_w^n, \\ \hat{\mu}_l^c &= 0.7\mu_l^c, \\ \hat{\mu}_w^c &= 0.7\mu_w^c. \end{aligned} \tag{4.12}$$

This generates fibroblast cells with thin nuclei and long and thin cytoplasm,

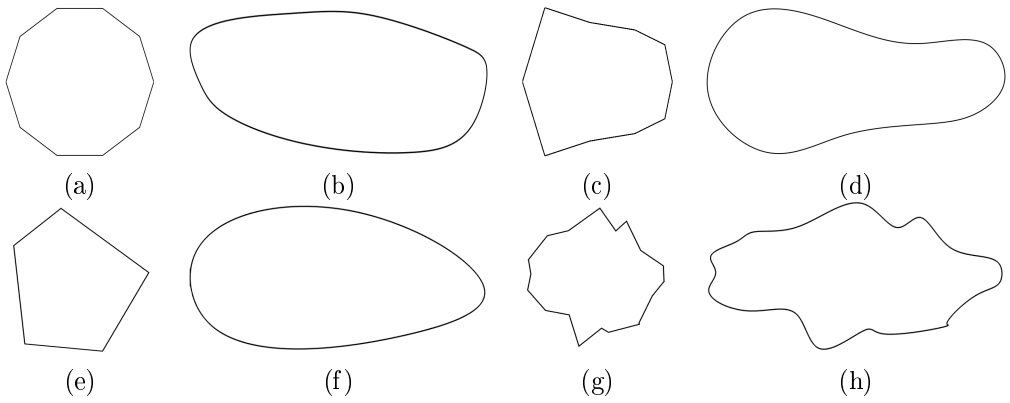


Figure 4.8: Examples of cell nuclei and cytoplasm shapes. Figures (a, c) show polygons without any randomness for $k = 10$ for the (a) stromal and (c) epithelial cells. Figures (b, d) show the corresponding shapes with dislocated vertices after spline interpolation. Figures (d) and (e) show randomised polygons initialised as circles for $k = 5$ and $k = 20$, respectively. Figures (f) and (g) show the corresponding shapes after spline interpolation. Here $\alpha = 0.2, \beta = 0.05$, $\mu_l = 2\mu_w$ and the major axis is shown in the horizontal direction.

and lymphocytes that are smaller than epithelial cells. The cytoplasm of epithelial cells is generated starting from the polygon shown in Figure 4.8 (c). The set of original coordinates $\{(x_i, y_i), i = 1, ..., k\}$ is then randomised and scaled

$$\begin{aligned} \hat{x}_i &= \mu_l^c N\left(x_i, \alpha/2\right), \\ \hat{y}_i &= \mu_w^c N\left(y_i, \alpha/2\right). \end{aligned} \tag{4.13}$$

The polygons are scaled with the respective value as

$$\begin{aligned} \hat{x}_i(\theta_i) &= x_i(\theta_i)\mu_l^{n/c}, \\ \hat{y}_i(\theta_i) &= y_i(\theta_i)\mu_w^{n/c}. \end{aligned} \tag{4.14}$$

where $\mu^{n/c}$ refers to either μ^n or μ^c. Finally, the vertices are interpolated using cubic splines (Figure 4.8 (b) and (d)).

Texture

Texture for the cytoplasm is generated using a well-known procedural model [205] for texture synthesis. In location (x, y), the texture t is given by a weighted sum of n octaves of basic noise function η_{xy}. This is defined as

$$t(x,y) = B + \sum_{i=0}^{n-1} p^i \eta_{xy}(2^i) \tag{4.15}$$

where scaling of the noise functions is controlled by the persistence parameter p, and the bias is given by B. As the nuclei chromatin texture is an important factor when grading the CRA, a more sophisticated method was adopted for synthesising it. In particular, we used the non-parametric model presented by Efros and Leung [203]. The model is applied to the grey-scale texture of all the nuclei found to belong to the phenotypes (Section 4.1.2) in order to generate a large texture image (Figure 4.3). When a nucleus of a particular phenotype is being synthesised, a random part of the corresponding texture image is selected and used as the texture. The sampling is done with replacement, and hence, although unlikely, two nuclei could have the same texture. Although this texture synthesis method produces more realistic results, it is very computationally expensive and so its use has been limited within the model. Texture images and sample of cells belonging to the corresponding phenotype for several of the phenotypes found in the real data have been shown in Figure 4.3. The same method is also used to generate the lumen textures shown in Figure 4.4.

4.1.5 Measurement error

The final step of the simulation degrades the ideal images constructed in the previous sections. This resembles the degradation caused by the real measurement system. For histology images, convolution with a 2D Gaussian, G, is used to simulate the leaking of photons between neighbouring pixels. We also add zero mean Gaussian noise, N_G with variance σ_G to approximate the CCD detector noise (Table 4.1). Hence, the simulated image degraded by the acquisition system, $\hat{I}$, obtained from an ideal image I is given by:

$$\hat{I} = I * G + N_G, \tag{4.16}$$

where $*$ denotes the convolution operator. For fluorescence images there is a larger number of degradation effects that need to be simulated. Firstly, uneven illumination, I_s, is simulated by adding a second degree parabolic polynomial with parameters as established by Svoboda et al. [206]. The centre of the simulated illumination source can be input. The energy of the illumination source is controlled by a parameter E_s. The auto-fluorescence effect, I_a with energy E_a, is simulated as a spatially slowly changing random texture Perlin [205]. In addition, convolution with a 2D Gaussian, G, is used to simulate the point spread function. Finally, we add zero mean Gaussian noise, N_G with variance σ_G to approximate the CCD detector noise.

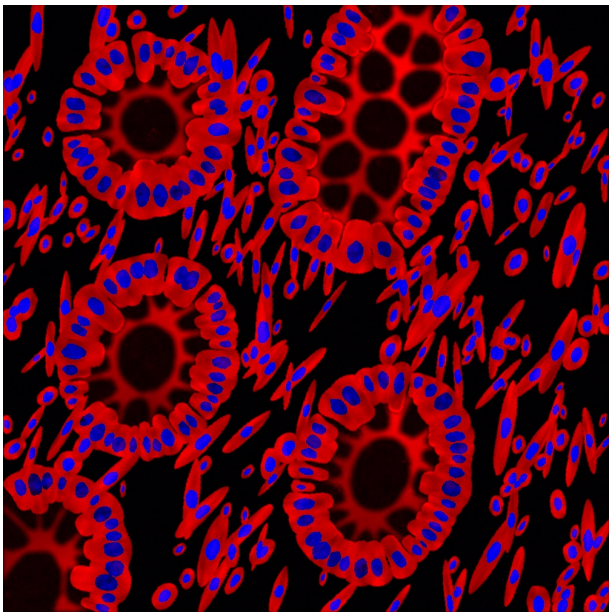


Figure 4.9: Example of a synthesised fluorescence image of a healthy sample at magnification 40×. The cytoplasm is shown in red and nuclei in blue. Here $L_{max} = 0.6$ and the cellularity $\nu_s = \nu_e = 1$.

Hence, the simulated image degraded by the acquisition system, $\hat{I}$, obtained from an ideal image I is given by:

$$\hat{I} = [(I + E_s I_s + E_a I_a) * G] + N_G. \tag{4.17}$$

The resulting fluorescence image is shown in Figure 4.9.

4.1.6 Histology Simulation

The generated cytoplasm and nuclei channels are converted into H & E stains (Figure 4.10) using the colour deconvolotion matrix suggested by Ruifrok and Johnston [207] (unless otherwise stated). By simulating IHC stains, the usability of the model is expanded to verification of a wide range of methods for analysis of H & E images. As one can choose the stain vector used to generate the images, the model can be easily utilised to validate stain normalisation methods. In addition, H & E images are easily assessed by pathologists who routinely deal with histology slides.

4.2 Discussion and Validation

THeCoT models the tumour heterogeneity in colorectal tissue. An example of the resulting images for a healthy sample is shown in Figure 4.10 (a). There are several user defined parameters which allow control over the appearance of the imaged tissue. Figures 4.10 (a) and (b) illustrate how changing the parameters for overlap and cellularity affects the resulting images. Depending on the purpose for image synthesis, one may require to have fewer, easily separable cells (Figure 4.10 (b)), or more crowded and overlapping cells (Figure 4.10 (a)). Manipulating these parameters could be very important when testing cell segmentation algorithms, for instance. The results from cell counting experiments, similar to the ones in [153], using ImageJ [208] and CellProfiler (CP) [209] are shown in Table 4.4. Cell counting was done on a total of 20 simulated samples, 10 healthy and 10 moderately differentiated cancerous images at 40× magnification, and cellularities $\nu_s = \nu_e = 1$. It was performed both on the non-overlapping nuclei regions and on the cytoplasmic regions where overlap of 0.4 was allowed. In CP segmentation was performed by first using an Otsu thresholding with an adaptive threshold. When performing nuclei segmentation minimising the weighted variance gave the best results. However, for segmenting the overlapping cytoplasms, minimising the entropy gave better results and these are reported in Table 4.4. Objects outside the diameter range [8, 50] pixels for nuclei and [8, 100] pixels for cytoplasm were considered mis-segmented and hence were discarded. In ImageJ, two different approaches of segmentation were adopted. Firstly, cells were counted using the ITCN (Image-based tool for counting nuclei) Plugin for ImageJ developed by Thomas Kuo and Jiyun Byun at the Center for Bio-image Informatics at UC Santa Barbara [210]. Its algorithm assumes nuclei to be blob-like structures with roughly convex local intensity distributions whose iso-level contour is approximately ellipsoidal; nuclei are fitted by an inverted Laplacian of Gaussian filter [210]. Images were inverted before using ITCN. Cell detection was performed by detecting dark peaks with the following parameters: cell width = 22, minimum distance = 4, threshold = 1. This method was unable to segment the cytoplasmic images due to their more complex shapes. Hence, a second method for segmentation was tested where the images were first thresholded manually and then watershed was used to attempt to segment regions further. We can see that CellProfiler performed significantly better on the healthy than the cancerous images due to the more consistent nuclei sizes. Similar behaviour was observed for ImageJ using both segmentation algorithms, with cell counting results closer to the ground truth for the healthy images. However, we can see from Figure 4.11 that, in fact

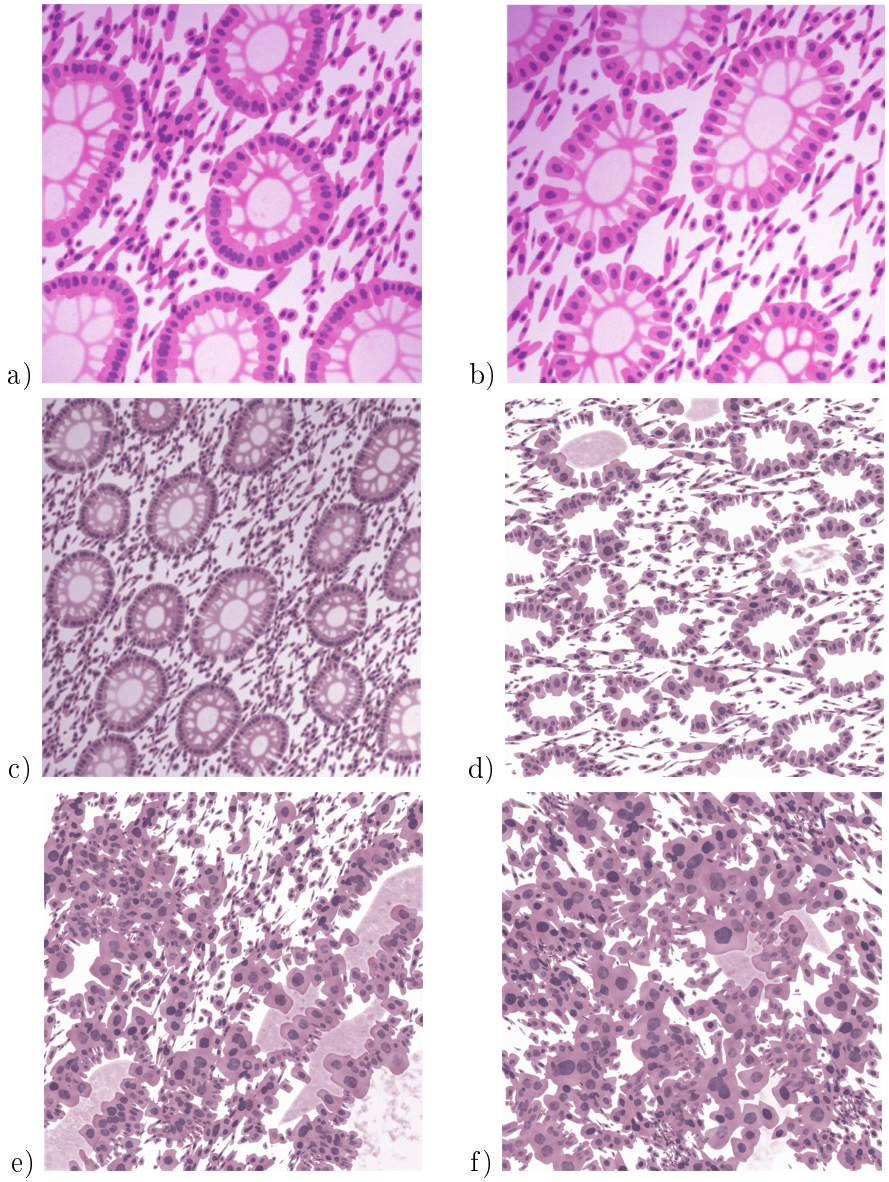


Figure 4.10: Examples of synthesised images demonstrating the effects of different parameter values. Figures (a)-(c) show examples of simulated images of healthy colon tissue. The images are 1000×1000 pixels, at magnification (a, b) $40\times$ and (c) $20\times$. In figure (b) the overlap $L_{max} = 0.2$ and the cellularity $\nu_s = \nu_e = 1$. All other figures have $L_{max} = 0.6$ and the cellularity $\nu_s = \nu_e = 1$. Figures (d)-(f) show various differentiation grades. The images are 1000×1000 pixels, at magnification $20\times$. The figures show (d) well differentiated, (e) moderately differentiated, and (f) poorly differentiated cancers. Figures (a) and (b) were generated using the stain vector proposed by Ruifrok and Johnston [207], whereas the rest of the Figures were generated using a different stain vector.

ITCN tended to over-segment larger nuclei while missing smaller ones (Figure 4.11 (b)). On the other hand, watershed under-segmented cells but picked up regions of the goblet cells cytoplasmic architecture (Figure 4.11 (d)). This is confirmed further by the large under-segmentation of the cancerous images. It is important to note that above algorithms may perform better with further tuning of their parameters. This study only aimed to demonstrate how such algorithms could be compared based on performance on the synthetic data generated by THeCoT.

Figures 4.10 (c) - (f) show how the tissue structure changes as the differentiation grade is increased. When the user specifies the cancer grade, there is a number of parameters integrated as part of the model that also change. These include the size, shape and appearance of the crypts, whether nuclei are basally orientated, and the frequency of cell phenotypes (Table 4.1). It is worth noting that in the model we assume that Eosin is highly specific to marking the cytoplasm. While in reality this is not necessarily the case, the lack of a membrane marker in the ground truth data makes it difficult to separate and model the non-specific binding. We plan to address this issue and model the extracellular matrix in the future.

To assess how realistic the appearance of the images generated by the model is, we asked three pathologists to grade them. They were presented with images for the four grades, at magnifications of 40× and 20× and with overlap of 0.2 and 0.6 (total of 16 images). They consistently rated the number of crypts, epithelial and stromal cells as realistic, suggesting that this is a suitable range for the overlap parameter. Grades for the appearance of the tissue are shown in Table 4.5. The average grade given was 4.28 out of 5. The pathologists on the whole graded the stromal cells as being less realistic because, while one could tell they are stromal cells, one couldn't determine what type of stromal cells they are. This is something that will be addressed in future developments of the model.

The most distinguishing characteristic of the colon microenvironment is the crypt structure. We evaluate this by comparing the overall distributions of morphological features of the synthesised crypts with those calculated from the hand-marked histology images. We have found excellent agreement between the distribution of the minor axis length and the ratio between the minor and major axes and the Gamma distributions estimated from the real data. The results are shown in Figure 4.12. In order to evaluate the overall appearance of tissue, we utilised a thresholded probability map method proposed by Sirinukunwattana et al. [194]. We generated a database of 15 images for each grade (60 in total). The H&E images were generated using a stain vector of a real image used to train the crypt segmentation method. The stain vector was determined using the method proposed by Trahearn et al. [211].

Table 4.4: Cell counting results for ImageJ(IJ) and CellProfiler (CP) with counting based on non-overlapping nuclei or cytoplasm regions with $L_{max} = 0.4$. Mean ± standard deviation are shown normalized by the ground truth. A value over 1 shows over-segmentation, whereas a value under 1 demonstrates under-segmentation.

Image type	CP nuclei	CP cytoplasm	IJ nuclei ITCN	IJ nuclei	IJ cytoplasm
Mean All	1.007 ± 0.014	0.919 ± 0.149	0.952 ± 0.036	1.094 ± 0.041	0.945 ± 0.283
Mean Healthy	1.014 ± 0.011	1.046 ± 0.084	0.976 ± 0.022	1.062 ± 0.023	1.139 ± 0.291
Mean Cancer	1.001 ± 0.015	0.792 ± 0.071	0.929 ± 0.031	1.125 ± 0.029	0.751 ± 0.021

Table 4.5: Average evaluation of the appearance of synthetic images by 3 pathologists. Healthy (H), well differentiated (WD), moderately differentiated (MD), and poorly differentiated (PD) images were evaluated at magnifications 20× and 40×. (1 = Not realistic at all, 5 = Very realistic, '-' means feature is not relevant).

	H 40×	H 20×	WD 40×	WD 20×	MD 40×	MD 20×	PD 40×	PD 20×
Architecture	5	5	5	4	4	4	5	5
Crypt shape	5	5	5	5	5	5	4.5	4.5
Lumen	5	5	5	5	5	5	-	-
Goblet cells	4	4	-	-	-	-	-	-
Epithelial cells	4	4	4	4	4	4	4	4
Stromal cells	3	3	3	3	3	3	3	4

The results for the Dice coefficient on both pixel-level and object-level are shown in Table 4.6. Most of the results are comparable with results for real data [194]. The method performs worse for high grade cancerous samples when trained and tested on different datasets. This is likely to be due to the fact that the segmentation framework relies heavily on the texture within and outside the cancerous crypts. The model currently does not include a model for the extra-cellular matrix which generates the texture between the stromal cells. In addition, the model may need a wider variety of textures available for inside the cancer crypts.

A further set of 20 samples (10 healthy and 10 moderately differentiated) were simulated at 40× with an average of 360 cells per sample. In order to check that the synthesis of nuclei texture has produced satisfying results, we analysed

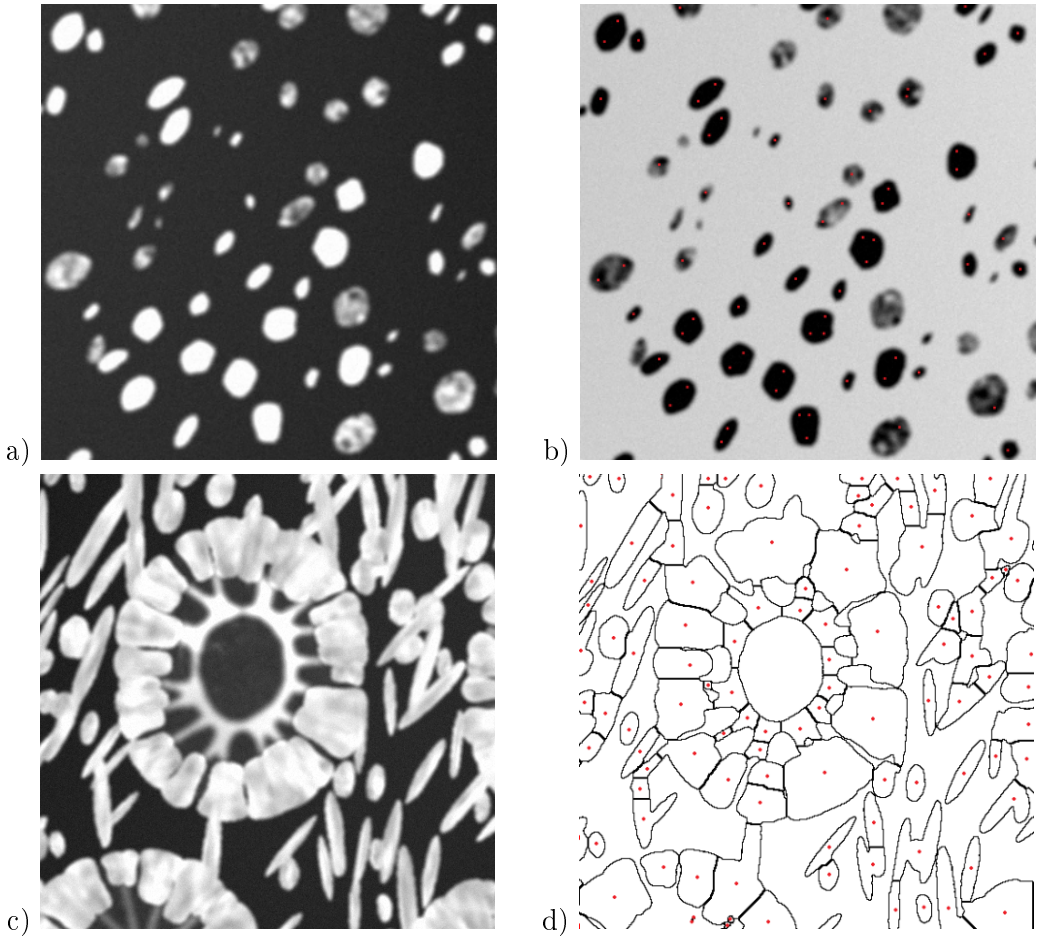
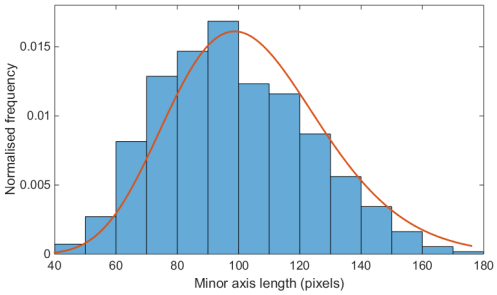
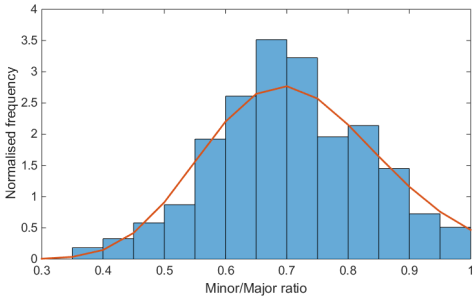


Figure 4.11: Examples of segmentation results using ImageJ. Figures (a) and (b) show original data for a cancerous image and results from segmentation using the ITCN plugin. Red dots mark centres of detected regions. Figures (c) and (d) show original data for a healthy image and results from segmentation using thresholding and watershed segmentation. Figure (d) shows the borders of the segmented regions with a red cross identifying the proposed segmented cells.

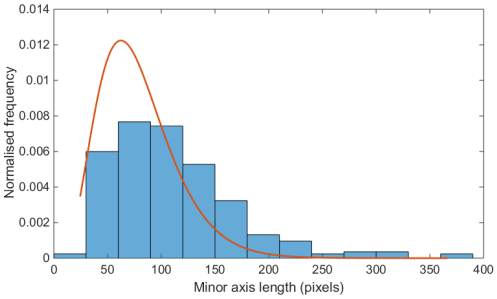
a)



b)



c)



d)

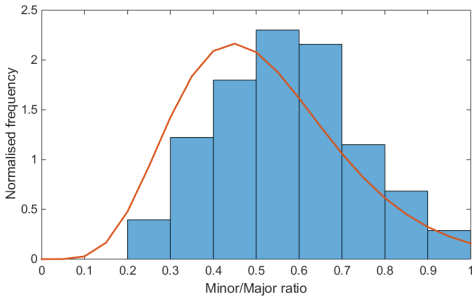


Figure 4.12: Distribution of parameters extracted from synthetic data. Figures (a) and (b) show the minor axis length and the ratio between the minor and major axes, respectively, for healthy crypts. Figures (c) and (d) show the minor axis length and the ratio between the minor and major axes, respectively, for cancerous crypts. Frequencies are normalised so that sum of areas of bars equals 1. The probability distribution functions shown are the ones estimated for the real data.

Table 4.6: Pixel-level and object-level dice coefficient for crypt segmentation of synthetic images of various grades at 20× magnification. Crypts were segmented using a thresholded probability map method [194]. Results are shown when the method was trained and tested on the synthetic and on real data. The reported figures are the average ± standard deviation.

Training data	Test	Grade	Dice-Pixel	Dice-Object
Synthetic	Synthetic	Healthy	0.96 ± 0.003	0.91 ± 0.03
		Well	0.94 ± 0.005	0.90 ± 0.03
		Moderately	0.91 ± 0.02	0.90 ± 0.03
		Poorly	0.65 ± 0.15	0.65 ± 0.13
Real	Synthetic	Healthy	0.87 ± 0.01	0.85 ± 0.02
		Well	0.89 ± 0.01	0.84 ± 0.03
		Moderately	0.88 ± 0.11	0.52 ± 0.11
		Poorly	0.59 ± 0.16	0.36 ± 0.11
Synthetic	Real	Benign	0.69 ± 0.11	0.53 ± 0.13
		Moderately	0.58 ± 0.16	0.43 ± 0.13
		Poorly	0.60 ± 0.17	0.44 ± 0.17

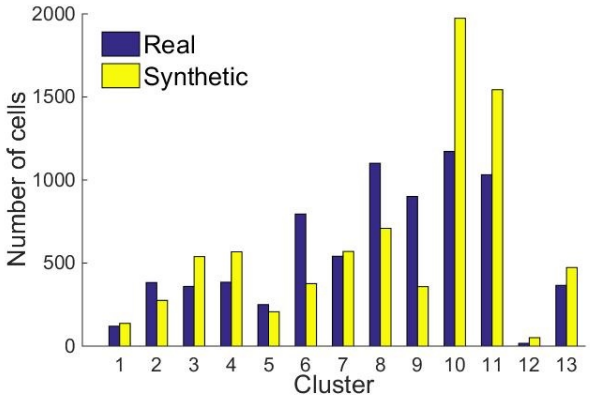


Figure 4.13: Clustering results of real and synthetic nuclei texture.

the nuclei of the 20 synthetic images described above and the hand-marked nuclei from real H&E images. The Haralick features of all the nuclei were calculated and these were phenotyped using Affinity Propagation. The clustering was not able to distinguish between the real and synthetic nuclei and all nuclei phenotypes contained a combination of the two (Figure 4.13). This demonstrates the suitability of the framework adopted for chromatin texture synthesis. In addition, we can see that the distribution of the phenotypes of the real and synthetic nuclear textures are nearly equal.

Chapter Summary

In this chapter we present a model for tumour heterogeneity in colorectal tissue (THeCoT). It is capable of simulating healthy and cancerous colonic crypt architecture and generating both immunofluorescence and histology image data. THeCoT has several user-defined parameters, which allow control over the tissue appearance. We have demonstrated how changing these allows to validate image analysis algorithms such as cell segmentation frameworks. The model also has a number of incorporated parameters learned from real hand-marked H&E images. These help the synthesis of realistic cell phenotypes, chromatin and lumen texture, nuclei morphology, and crypt architecture. Integrated parameters depend heavily on the cancer grade, and control aspects such as the size, shape and appearance of the crypts, whether nuclei are basally orientated and the frequency of cell phenotypes present in the sample.

We have evaluated the main aspects of the model by assessing various features of the generated images. Semi-quantitative evaluation was performed by pathologists who rated the appearance of a range of features. The majority of these were rated as

very realistic. We have also performed validation analysis on the nuclear texture and found that a clustering algorithm could not distinguish between real and synthetic texture. In addition the clustering found very similar distribution of the texture phenotypes in the real and synthetic images.

The synthetic images generated by the model would enable quantitative comparison of different image analysis frameworks, including image restoration, cell and crypt segmentation, cancer grading and stain separation.

Chapter 5

Modelling Protein Expression

As mentioned in Section 1.1, it is important to identify patients with MSI, as this could guide treatment and help diagnose Lynch syndrome. This motivated us to expand the THeCoT model (Chapter 4) to simulate the expression of relevant proteins. We have considered the four MMR proteins (MLH1, PMS2, MSH2, MSH6). Mutations in genes producing these proteins are the cause for MSI. In addition, we consider P53 which has been found to be also associated with the condition. These five proteins have varied sub-cellular expression patterns (Table 5.1) and provide an interesting case study demonstrating how any protein expression could be included within the model.

In order to model the expression of proteins, we first need to have models for the cell organelles where the proteins of interest are expressed. These are detailed in Table 5.1. For this purpose, we use real high-resolution fluorescence data to learn features of the organelles that can then be incorporated into the model. The fluorescence images of cultured cells are utilised instead of the IHC images of CRA as the later do not provide high-enough resolution to consider the sub-cellular protein expression patterns. Once we have realistic models for the cell organelles, we then develop models for the proteins based on where they are expressed and under what conditions. Details of this process are given below.

5.1 Data

We have utilised high resolution immunofluorescence images of cultured cells from the Human Protein Atlas (HPA, http://proteinatlas.org) [212] for learning parameters for our model. In order to model the proteins that we are interested in, we need to develop models for the nucleoli, golgi apparatus and the vesicles. For each organelle,

Table 5.1: Details of the subcellular location of proteins.

Protein	**Subcellular location**
MLH1	Nucleoli, weak expression in the nucleus and cytoplasm
PMS2	Nucleus but not nucleoli, weak expression in cytoplasm
MSH2	Nucleus but not nucleoli, vesicles
MSH6	Mainly in the nucleus but not nucleoli. In addition localised to the cytoplasm, golgi apparatus & vesicles.
P53	Nucleus but not nucleoli

we have used proteins known to be highly specific to that organelle. To avoid bias that could be introduced by the binding of the protein, we have used 2 proteins for each cell organelle, as detailed in Table 5.2. For each cell organelle, we consider a total of 10 images split nearly evenly between the two proteins.

Cell organelle	**Protein tags**
Nucleoli	MLH1 & RRP1B
Golgi	GOLGA2 & GORASP2
Vesicles	ABCD3 & PECR

Table 5.2: Proteins tags used for modelling cell organelles.

5.2 Learning from Real Data

In order to be able to learn from the real immunofluorescence data we needed to be able to reliably segment the individual cells, nuclei and cell organelles. Cell and nuclei segmentation was performed using a seeded watershed segmentation method proposed earlier [213, 186]. The procedure involves thresholding the DAPI image with the threshold being determined as the intensity of the most common pixel. Next, the binary nuclear image is eroded and small objects and objects touching the boundary of the image are removed. Objects with areas outside a specified range are considered erroneous seeds and are also removed. For cell segmentation the endoplasmic reticulum (ER) channel was used to determine background seeds from large areas of pixels with zero intensity. The seeds, along with an inverted ER image, are then used in the seeded watershed algorithm, and resulting regions corresponding to background or erroneous seeds are removed. For nuclear segmentation, the nuclear channel was used to determine both foreground and background seeds. Examples of the results are shown in Figure 5.1.

For the purpose of segmenting the cell nucleoli, a single channel showing a

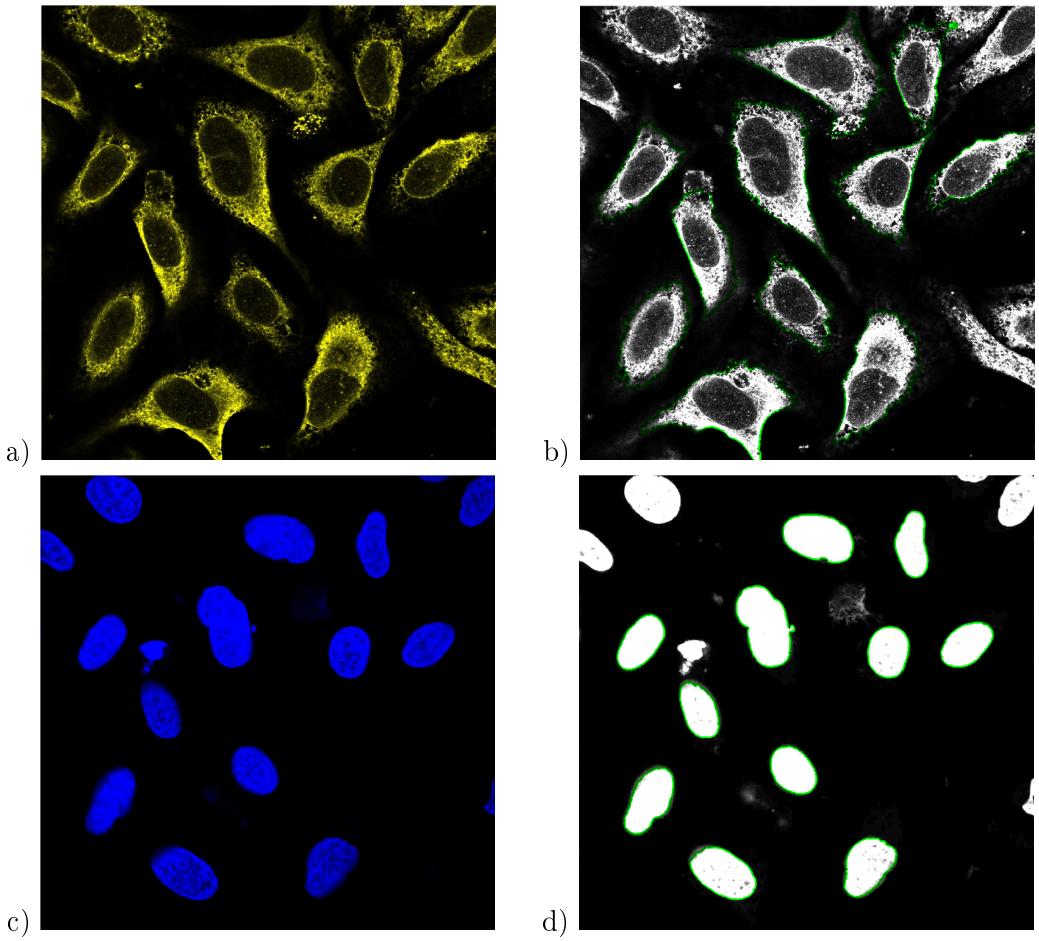


Figure 5.1: Examples of cell and nuclear segmentation. Figures (a) and (c) show the original channels. Figures (b) and (d) show the segmentation borders in green.

relevant antibody was thresholded to remove background noise and used to obtain both background and foreground seeds. The same segmentation as above was then followed. The results are shown in Figure 5.2

When segmenting the vesicles and golgi apparatus, this method didn't perform satisfactorily due to the small size of the objects and the high level of noise in the images (Figures 5.3 (d) and 5.4 (d)). For this reason we have instead used an adaptive thresholding method which uses an adaptive mean filter to highlight image features and then Otsu threshold to segment the image (Figures 5.3 and 5.4). We can see that the method performs very well even at high levels of noise (Figure 5.4 (f)). On the other hand, this method tends to over-segment the nucleoli and detect noise outside the nucleus as possible nucleoli detections (Figure 5.2 (d)). The segmentation procedure resulted in 484 nucleoli from 471 cells, 3433 golgi objects

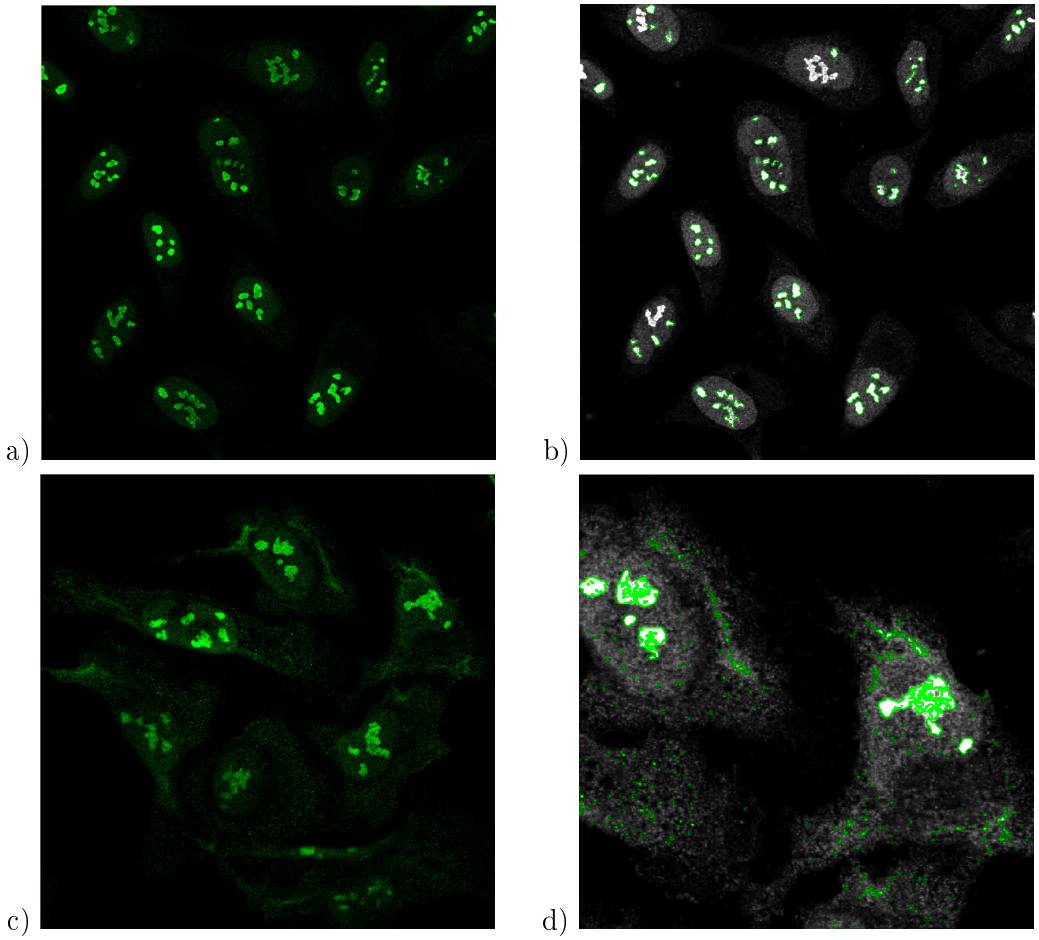


Figure 5.2: Examples of nucleoli segmentation. Figures (a) and (c) show the original channels for MLH1 and RRP1B images, respectively. Figure (b) shows segmentation results from the seeded watershed segmentation method with borders shown in green. Figure (d) shows segmentation results from the adaptive thresholding method.

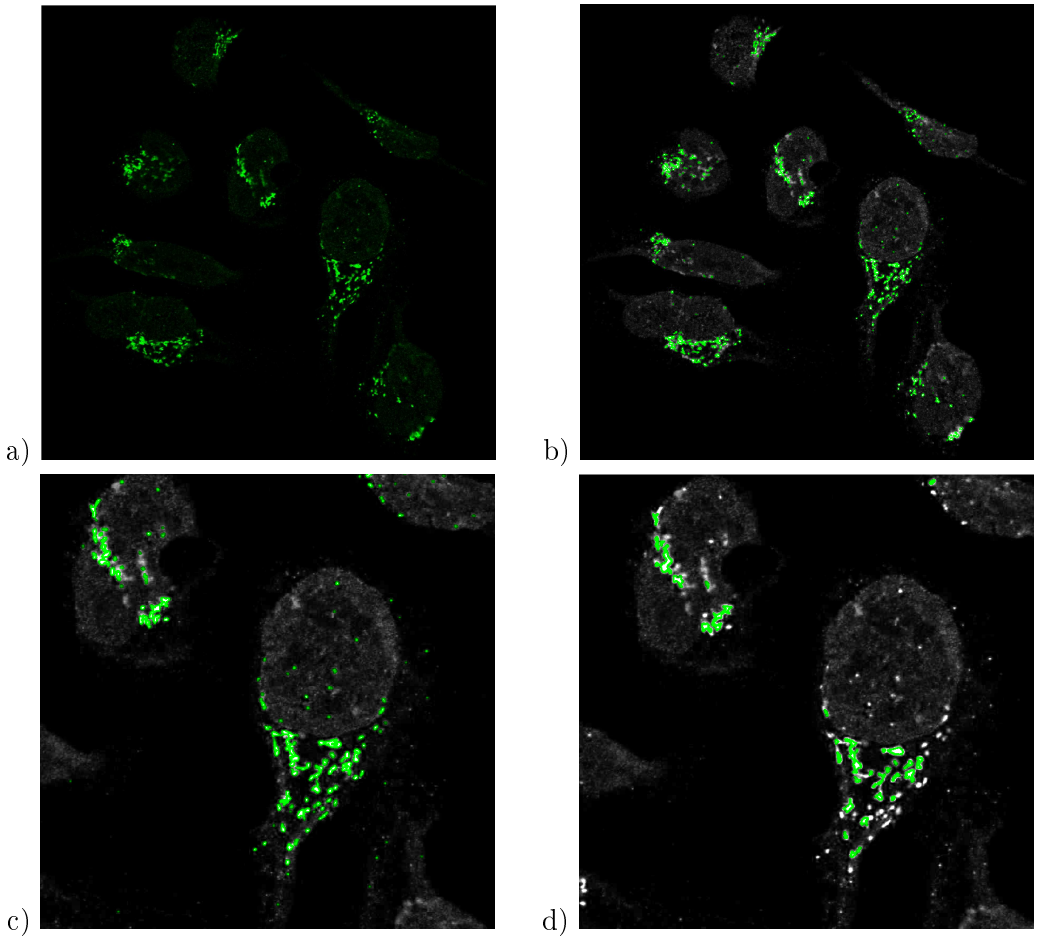


Figure 5.3: Examples of golgi segmentation. Figure (a) shows the original channel for a GOLGA2 image. Figures (b) and (c) show segmentation results from the adaptive thresholding method with borders shown in green. Figure (c) shows a zoomed in section of Figure (b). Figure (d) shows segmentation results from the seeded watershed segmentation method.

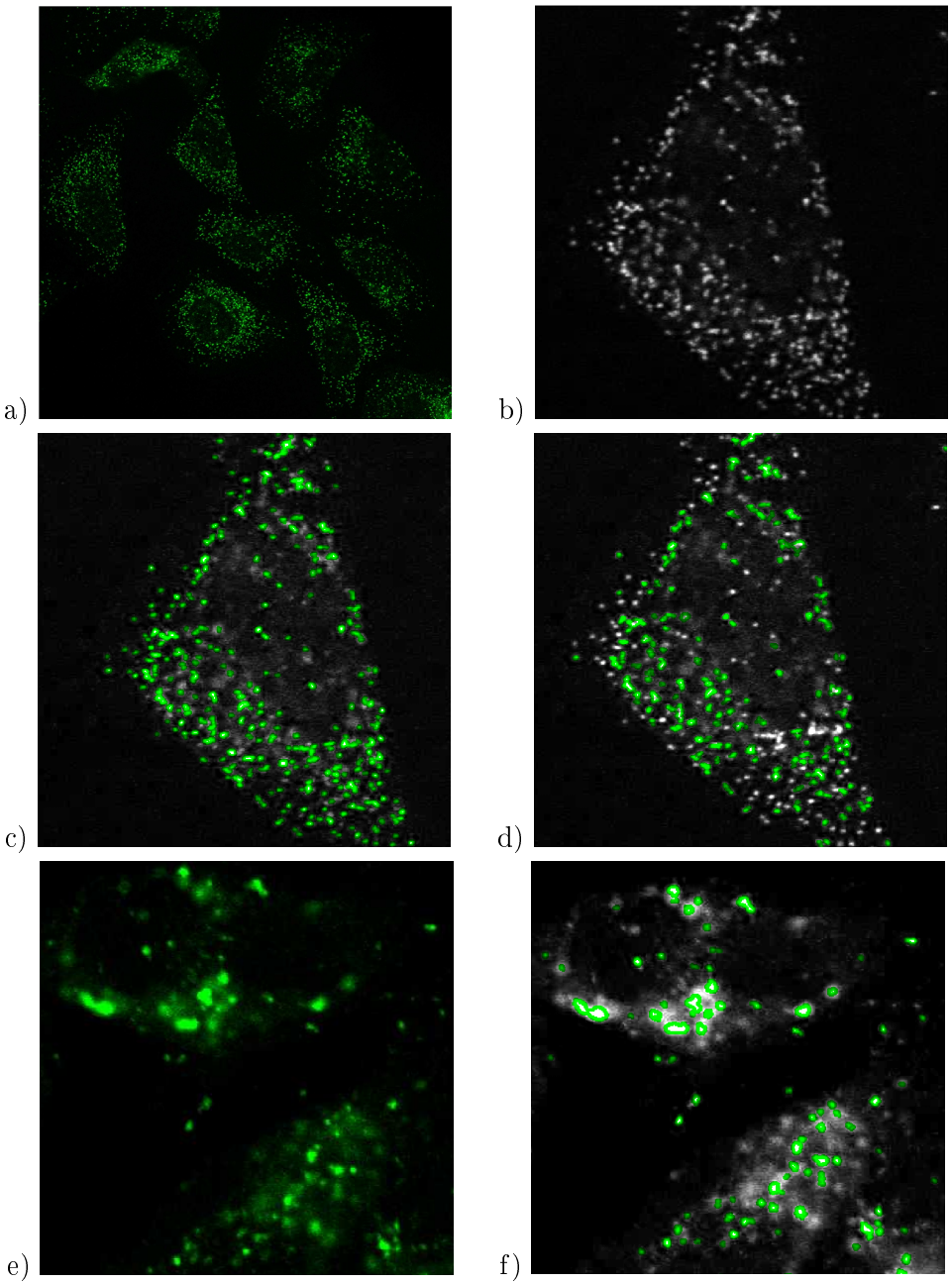


Figure 5.4: Examples of vesicles segmentation. Figure (a) shows the original channel for an ABCD3 image. Figures (b) and (e) show enlarged sections of the original channels for ABCD3 and PSAP images, respectively. Figures (c) and (f) show segmentation results from the adaptive thresholding method with borders shown in green. Figure (d) shows segmentation results from the seeded watershed segmentation method.

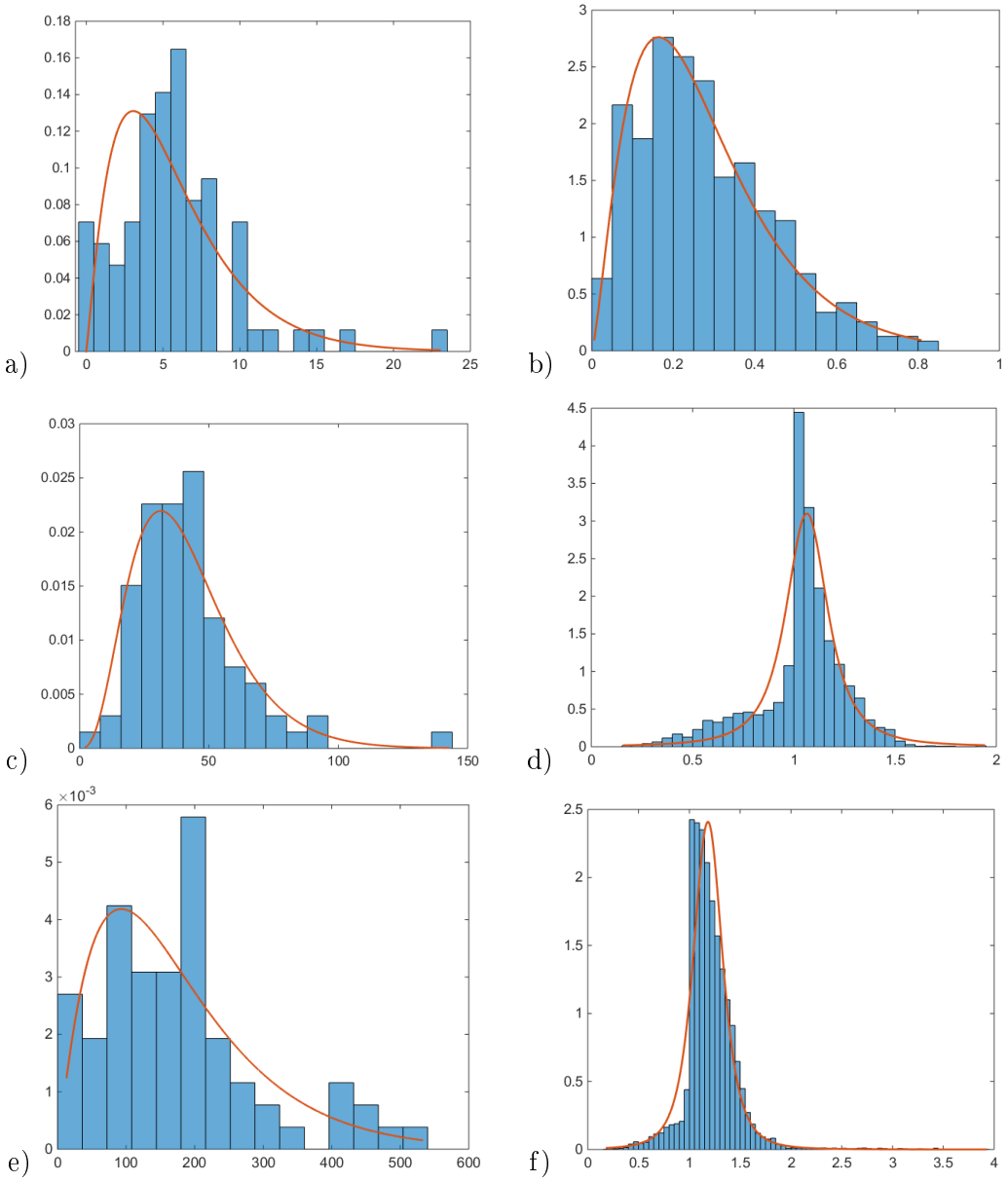


Figure 5.5: Estimated probability distribution functions for the number (left column) and position (right column) of the (a, b) nucleoli, (c, d) golgi and (e ,f) vesicles.

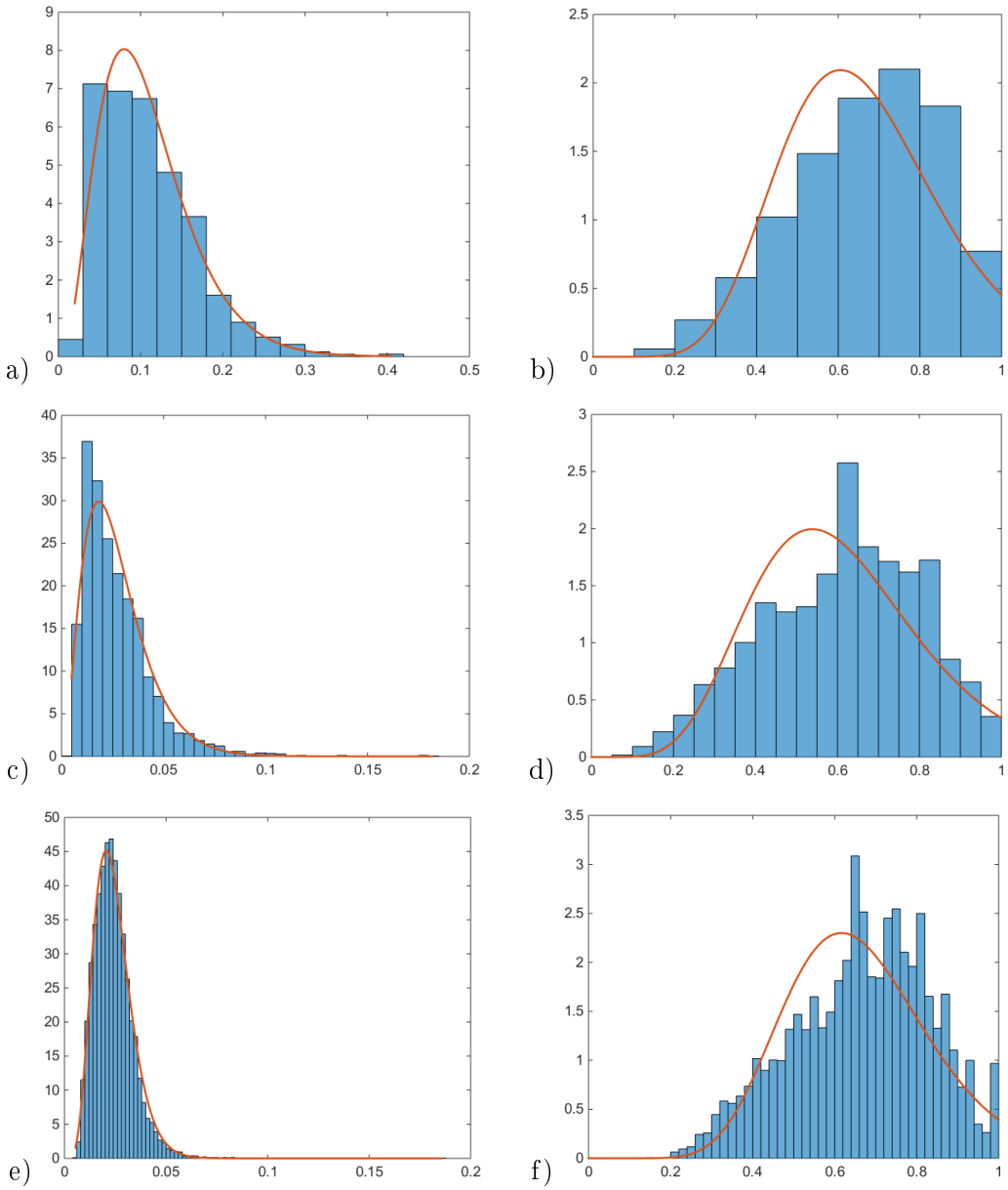


Figure 5.6: Estimated probability distribution functions for the ratios between the minor axes of the organelles and the nucleus of the corresponding cell (left column) and between the minor and major axes of the organelles. Figures show the ratios for (a, b) nucleoli, (c, d) golgi and (e, f) vesicles.

from 83 cells and 12,764 vesicles from 72 cells being identified.

Once we have all objects segmented, we can extract features representing the cell organelles to be incorporated into the model. We extract several features describing the organelles and their distribution within the cell. For each of them, we then estimate a probability distribution function (PDF) which is incorporated into the model. Firstly, we obtain the numbers of organelles within each segmented cell. These are modelled using a Gamma PDF and the results are shown in Figure 5.5 (left column). We then consider the size and shape of the organelles. As the real data available is for different types of cultured cells, instead of estimating the size of the cell organelles directly, we consider the ratio between the minor axes of the cell organelle and the corresponding nucleus. We assume that the cell nucleus approximately holds its shape when the cell is in a tissue and in cell culture. The distributions of this ratio and the estimated Gamma PDFs for each cell organelle are shown in Figure 5.6 (left column). Considering this ratio gives better generalisability to cells in a tissue and at different magnifications. To estimate the shape of the organelles, we consider the ratio between the minor and major axes of the segmented objects. The distributions of this feature and the estimated Gamma PDFs for each cell organelle are shown in Figure 5.6 (right column). The last feature considered describes the position of the organelle within the cell. We considered the line from the centre of the cell nucleus going through the centre of the organelle of interest. Let the distance between the centre of the nucleus and the point where the line crosses the nuclear membrane be given by N. Let the distance between the centres of the nucleus and the organelle be given by O, and the distance between the points where the line crosses the nuclear and plasma membranes be given by C (as shown in Figure 5.7). Then, the distance feature is given by

$$D = 1 - \frac{N - O}{N + C}. \tag{5.1}$$

Consequently, the minimum value of $D = 1 - N/(N + C)$ means the organelle is located at the centre of the nucleus and as $D \to 1$ the cell compartment is located closer to the nuclear membrane but within the nuclear boundary. A value of $D > 1$ describes an organelle that is outside the nuclear boundary and the distance from it is given proportionate to the distance between the centre of the nucleus and the plasma membrane. The distributions of this feature and the estimated PDFs for each cell organelle are shown in Figure 5.5 (right column). The distribution of the nucleoli position was well estimated by a Gamma PDF. On the other hand, most of the vesicles and golgi objects were found close to the nuclear membrane and so a t

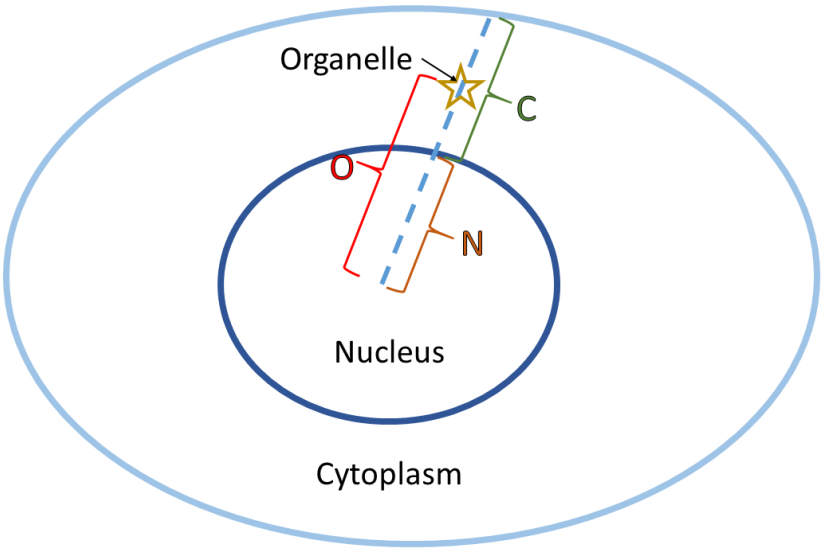


Figure 5.7: Diagram for calculating the position feature. The star marks the position of the organelle of interest

Location-Scale distribution gave a better fit.

5.3 Modelling Cell Organelles

For modelling the different cell compartments, we use the deformed circle model presented in Section 4.1.4. When we are generating cell organelles of a particular type, we draw parameter values from the relevant PDFs. However, we also impose certain restrictions on the parameter values based on the size of the cell in consideration. For each cell, first we choose the number of organelles to be created. Nonetheless, we only place a new cell organelle if that type of organelles are not taking up more than 12% or 18% of the cell area for golgi and vesicles, respectively, and 20% of the nuclear area for nucleoli. These constraints were set up to address the fact that other parameter values are drawn independently and so may result in unrealistic examples where a large number of organelles with relatively great size are generated. The values were set based on observations from the real data where golgi and vesicles took up to 4% and 6% of the cell area, and the nucleoli took up to 19.3% of the nucleus. The first two values were scaled up as the cytoplasm of cells in a tissue has more compact shape and so the 2D projection of it would give a much smaller area. On the other hand, we don't expect the nucleus to significantly change shape and so the threshold was held nearly the same. For each cell organelle to be placed, we choose the length of the minor axis by drawing a value for the ratio between the

nuclear minor axis and that of the organelle. A minimum length of 1 pixel is set. To determine the length of the major axis, we draw a value from the PDF estimated for the ratio between the minor and major organelle axes. Finally, we need to estimate the position of the organelle. For this, we draw a value from the PDF of the distance feature and select the direction from the nuclear centre at random. Using Equation 5.1 we can estimate the distance from the nuclear centre. The resulting organelles are shown in Figure 5.8.

5.4 Modelling Protein Expression

With a view to include the protein channels into the model, three user-defined parameters were introduced per protein. These define whether or not the protein has been imaged, whether there is a mutation in the gene and what fraction of the epithelial cells express the protein. Five proteins were included in the model, namely MLH1, PMS2, MSH2, MSH6 and P53. Details of each are given below. In addition, the user could choose to produce samples that are representative of the population. In this case, the model would include an MMR protein mutation with a 15% probability. If a mutation occurs, it has a probability of 50% of being in the MLH1 gene with , 40% in MSH2, 7% in MSH6 and 3% in PMS2 [49]. In cases without mutation, P53 has 50% probability of being overexpressed in epithelial cells, whereas in MSI cases it is overexpressed in only 20% of the cases Samowitz et al. [53].

5.4.1 Modelling the MLH1 expression

The subcellular expression for MLH1 was modelled at described in Table 5.1 and shown from confocal fluorescence images of cultured cells in Figure 5.9 (a), namely the protein has a strong expression in the nucleoli and weak expression in the rest of the nucleus. We can see that this also agrees with what is observed when the cells are in a tissue (Figure 5.9 (b)). If the user specifies a mutation in the MLH1 gene, the protein is not expressed in the epithelial cells. Otherwise, the user can specify what fractions of the epithelial cells are expressing the protein. It is worth noting that, in practice, even if only a small fraction of epithelial cells express the MMR proteins, the sample is graded as positively stained. Most stromal cells would always express the MMR proteins and, in the clinic, this serves pathologists as a positive control that the tissue has been stained. Within the model, all stromal cell would always express MLH1. Example of the protein expression images generated is shown in Figure 5.9 (c, d).

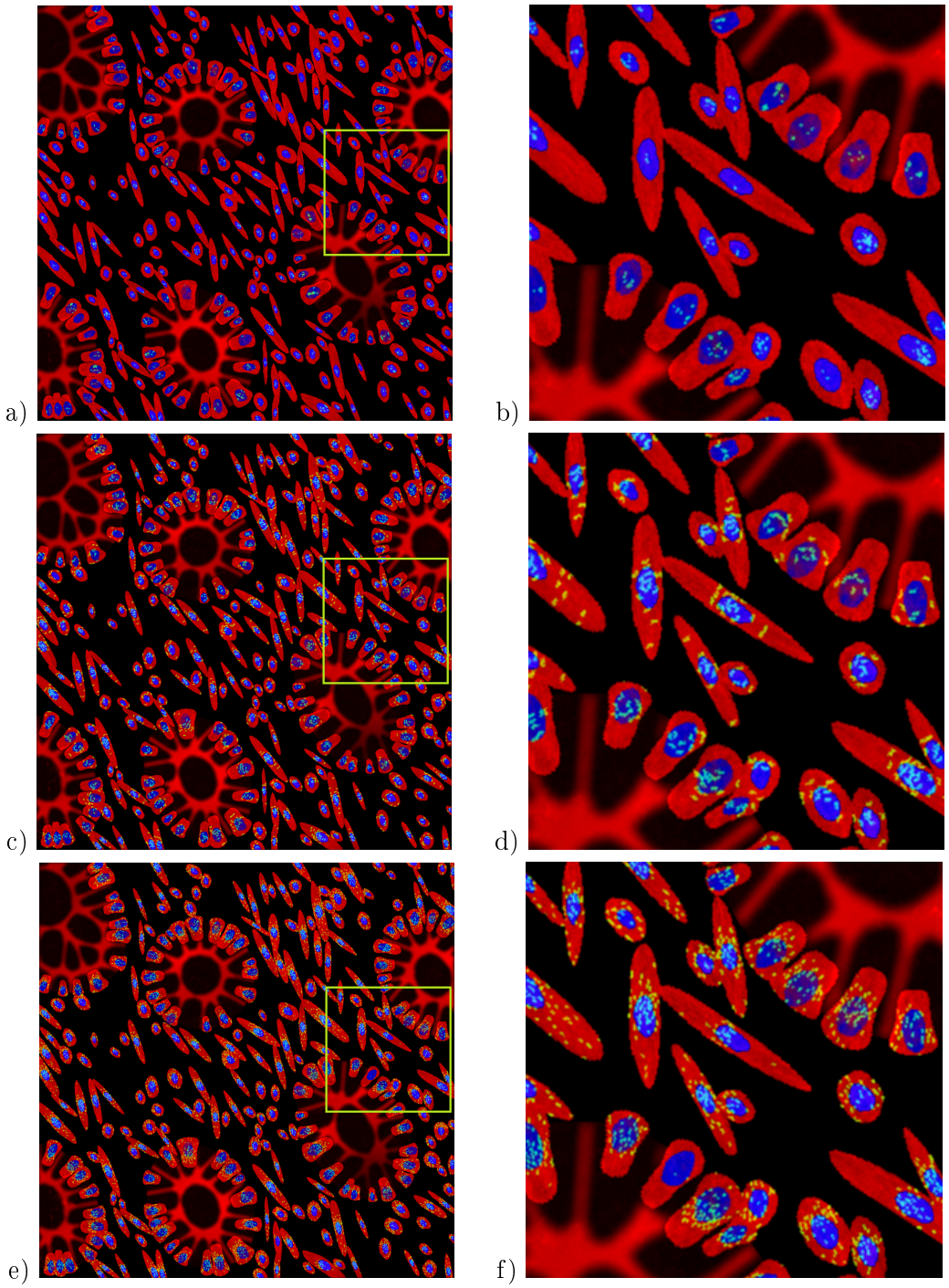


Figure 5.8: Examples of generated cell organelles. In all images the cytoplasm is shown in red, nuclei in blue and the green channel shows (a, b) the nucleoli, (c, d) the golgi and (e, f) the vesicles. Figures (b, d, f) show close-up sections of Figures (a, c, e), respectively, with the section identified by the green square.

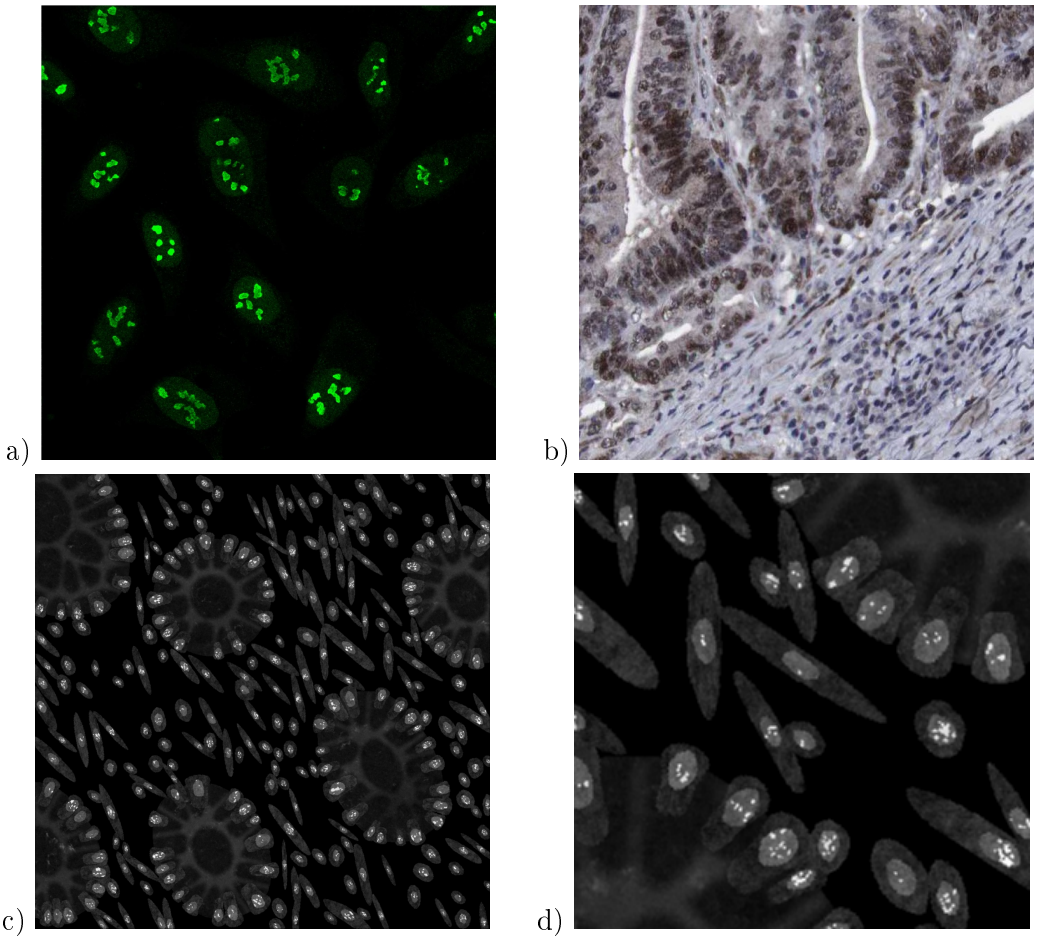


Figure 5.9: Modelling MLH1. Figure (a) shows the subcellular location in cultured cells imaged using a confocal fluorescence microscope. Figure (b) shows MLH1 expression in a histology image of CRA. Images (a, b) are from the HPA. Figures (c, d) show examples of synthetic images for MLH1 with (d) a scaled up sections from (c). Images are from the same sample as shown in Figure 5.8. In this simulation all cells are expressing the protein.

Table 5.3: Effects of mutations in the MMR genes on protein expression in epithelial cells.

Defective gene	Imaging results
MLH1	Loss of MLH1, PMS2
PMS2	Isolated Loss of PMS2
MSH2	Loss of MSH2, MSH6
MSH6	Isolated Loss of MSH6

5.4.2 Modelling the PMS2 expression

The subcellular expression for PMS2 was modelled at described in Table 5.1 and shown from confocal fluorescence images of cultured cells in Figure 5.10 (a), namely the protein has a strong expression in the nucleus excluding the nucleoli and weak expression in the cytoplasm. We can see that this also agrees with what is observed when the cells are in a tissue (Figure 5.10 (b)). If the user specifies a mutation in the PMS2 gene, the protein is not expressed in the epithelial cells. In addition, the same limited expression would occur if there is a mutation in the MLH1 gene as the two are binding partners (Table 5.3). Otherwise, the user can specify what fractions of the epithelial cells are expressing the protein and these are taken to be a subset of the epithelial cells expressing MLH1. As above, all stromal cell would always express PMS2. Example of the protein expression images generated are shown in Figure 5.10 (c, d).

5.4.3 Modelling the MSH2 expression

The subcellular expression for MSH2 was modelled as described in Table 5.1 and shown from confocal fluorescence images of cultured cells in Figure 5.11 (a), namely the protein has a strong expression in the nucleus and weak expression in the nucleoli. We can see that this also agrees with what is observed when the cells are in a tissue (Figure 5.11 (b)). To generate a realistic texture for this protein we use the chromatin texture used for the nuclear channel of the THeCoT model (Chapter4.1.4). If the user specifies a mutation in the MSH2 gene, the protein is not expressed in the epithelial cells. Otherwise, the user can specify what fractions of the epithelial cells are expressing the protein. All stromal cells would always express the molecule. Example of the protein expression images generated are shown in Figure 5.11 (c, d).

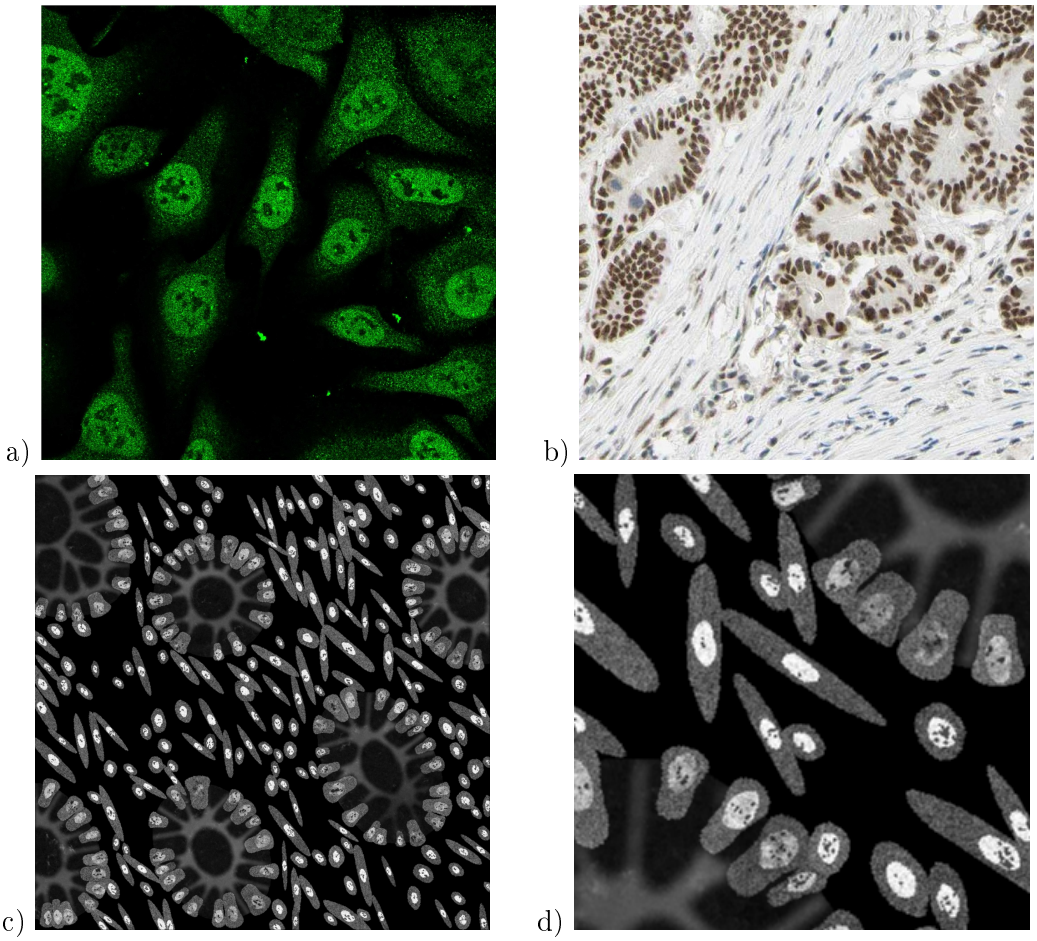


Figure 5.10: Modelling PMS2. Figure (a) shows the subcellular location in cultured cells imaged using a confocal fluorescence microscope. Figure (b) shows PMS2 expression in a histology image of CRA. Images (a, b) are from the HPA. Figures (c, d) show examples of synthetic images for PMS2 with (d) a scaled up sections from (c). Images are from the same sample as shown in Figure 5.8.

5.4.4 Modelling the MSH6 expression

The subcellular expression for MSH6 was modelled as described in Table 5.1 and shown from confocal fluorescence images of cultured cells in Figure 5.12 (a), namely the protein has a strong expression in the nucleus excluding the nucleoli, the vesicles and golgi apparatus, and weak expression in the cytoplasm. We can see that this also agrees with what is observed when the cells are in a tissue (Figure 5.12 (b)). If the user specifies a mutation in the MSH6 gene, the protein is not expressed in the epithelial cells. In addition, the same limited expression would occur if there is a mutation in the MSH2 gene as the two are binding partners (Table 5.3). Otherwise,

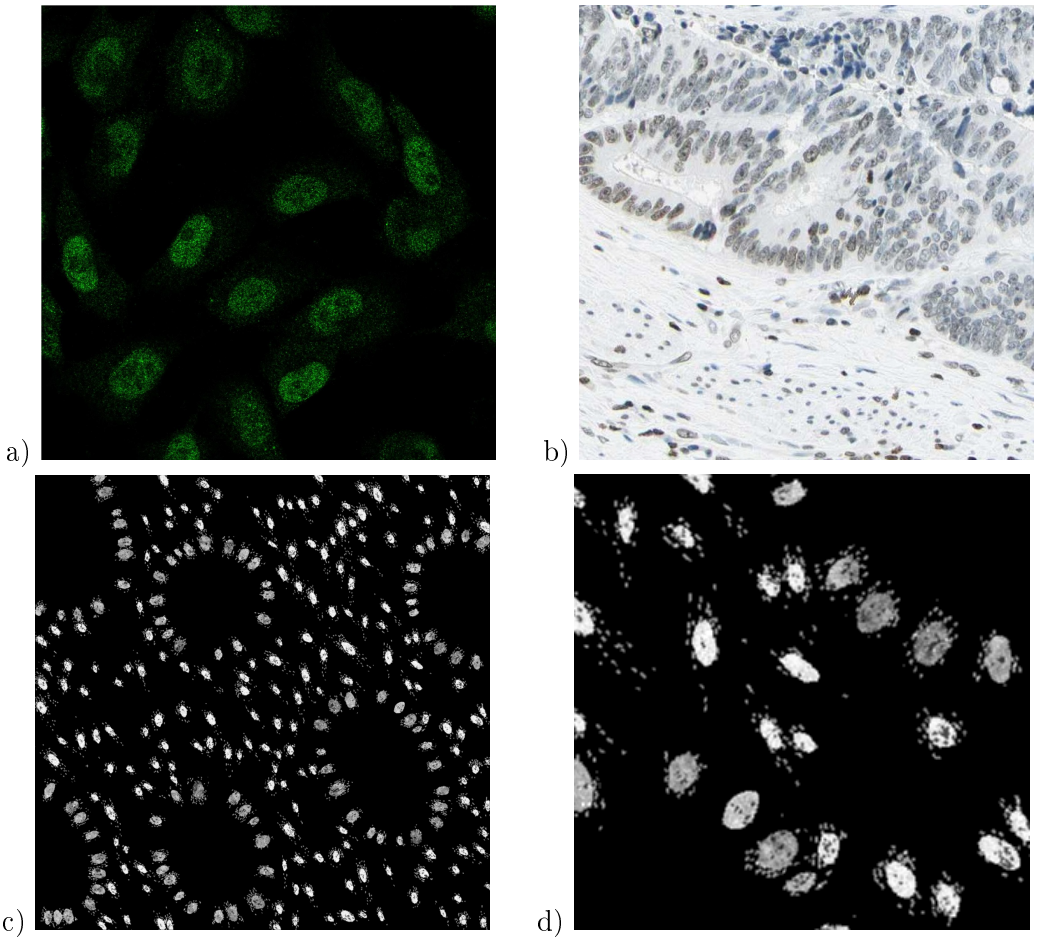


Figure 5.11: Modelling MSH2. Figure (a) shows the subcellular location in cultured cells imaged using a confocal fluorescence microscope. Figure (b) shows MSH2 expression in a histology image of CRA. Images (a, b) are from the HPA. Figures (c, d) show examples of synthetic images for MSH2 with (d) a scaled up sections from (c). Images are from the same sample as shown in Figure 5.8.

the user can specify what fractions of the epithelial cells are expressing the protein and these are taken to be a subset of the epithelial cells expressing MSH2. As above, all stromal cells would always express MSH6. Example of the protein expression images generated are shown in Figure 5.12 (c, d).

5.4.5 Modelling the P53 expression

The subcellular expression for P53 was modelled as described in Table 5.1 and shown from confocal fluorescence images of cultured cells in Figure 5.13 (a), namely the protein has a strong expression in the nucleus excluding the nucleoli. We can see

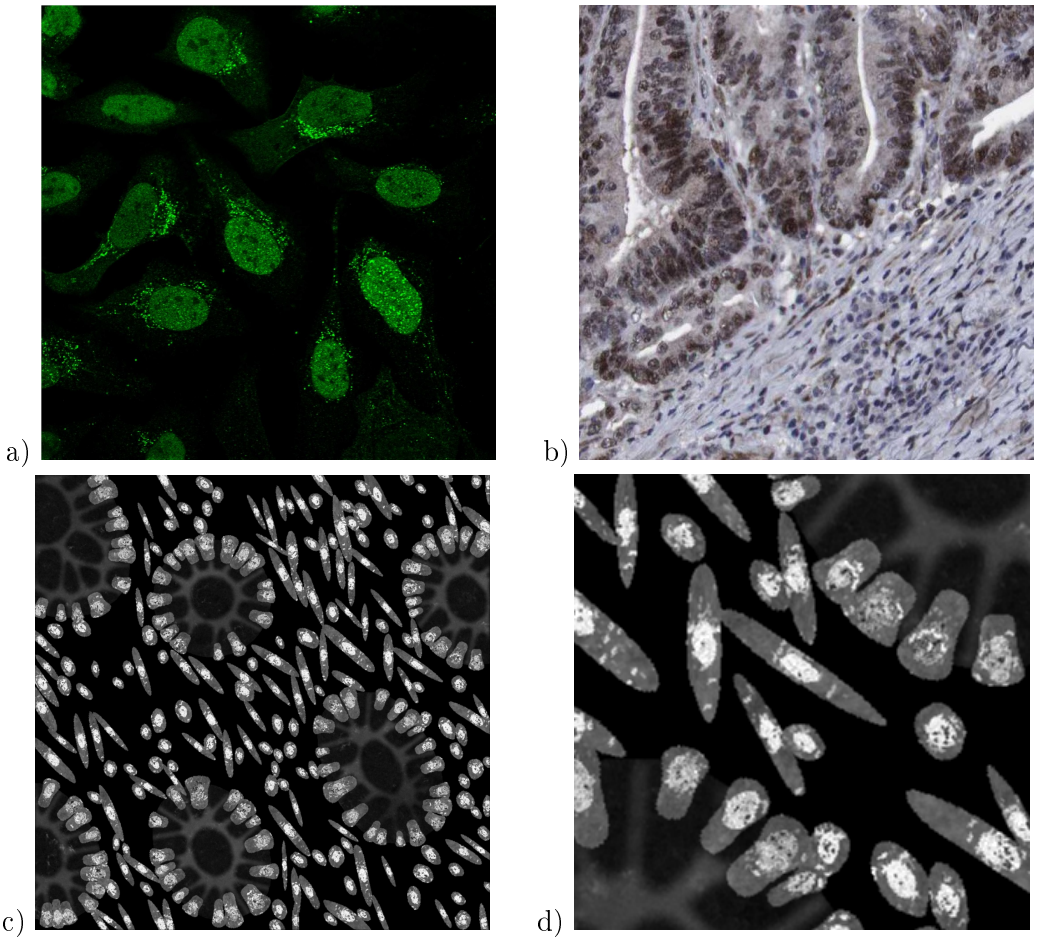


Figure 5.12: Modelling MSH6. Figure (a) shows the subcellular location in cultured cells imaged using a confocal fluorescence microscope. Figure (b) shows MSH6 expression in a histology image of CRA. Images (a, b) are from the HPA. Figures (c, d) show examples of synthetic images for MSH6 with (d) a scaled up sections from (c). Images are from the same sample as shown in Figure 5.8.

that this also agrees with what is observed when the cells are in a tissue (Figure 5.13 (b)). Unlike the MMR genes, P53 is not expressed in the stromal cells. Hence, to avoid a blank image in the stack, the model assumes that there is some expression of the protein in the epithelial cells. The user can specify what fractions of the epithelial cells are expressing the protein. Example of the protein expression images generated are shown in Figure 5.13 (c, d).

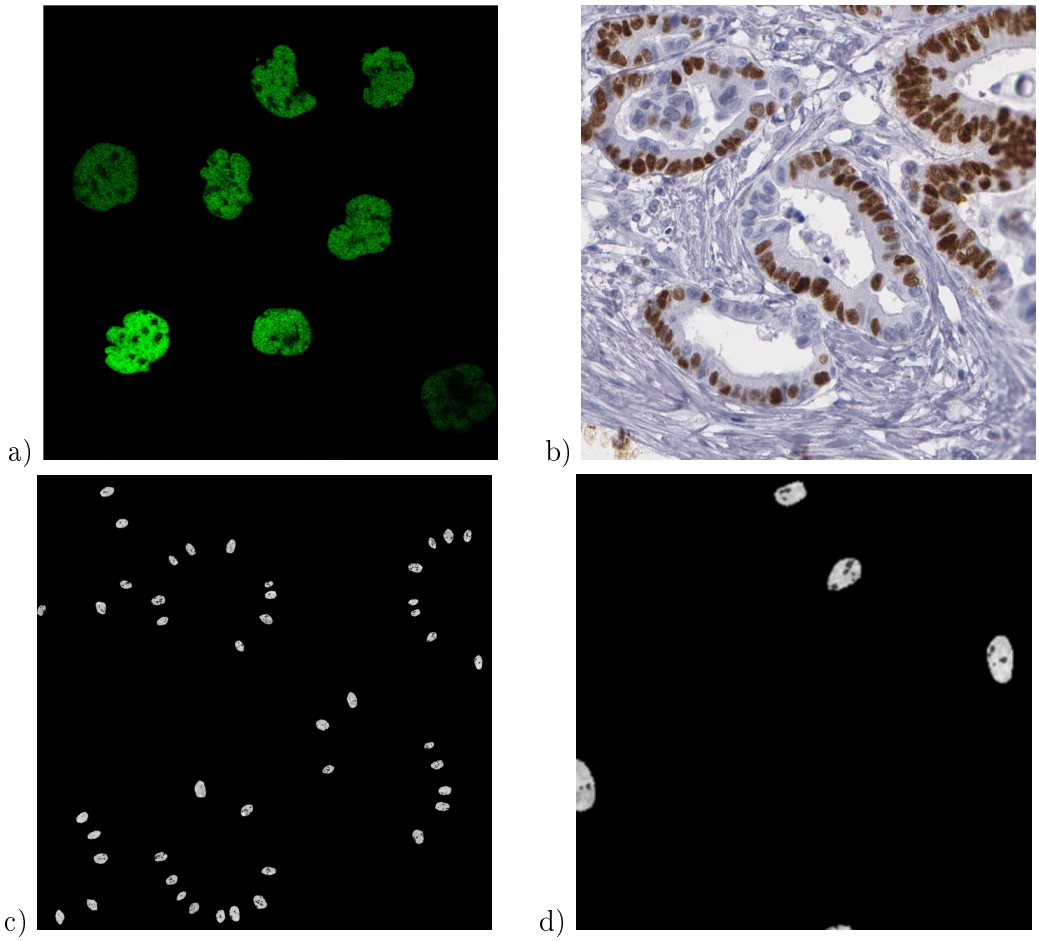


Figure 5.13: Modelling P53. Figure (a) shows the subcellular location in cultured cells imaged using a confocal fluorescence microscope. Figure (b) shows P53 expression in a histology image of CRA. Images (a, b) are from the HPA. Figures (c, d) show examples of synthetic images for P53 with (d) a scaled up sections from (c). Images are from the same sample as shown in Figure 5.8.

5.5 Discussion and Validation

This chapter has extended the THeCoT model to include models for protein expression. We have focussed on five proteins associated with MSI. These are commonly screened for in clinical practice and developing the protein expression models could aid the development of frameworks for automatic grading. The user could choose to have a sample that is generated with the probability of mutation representative of the general population. In this case, they also need to specify which of the five proteins they wish to be included in the resulting images. Alternatively, they can specify where the mutation occurs. The model takes into account dependencies of

binding pairs of the MMR proteins, and hence, if a mutation occurs in MLH1 or MSH2, its binding partner would also have inhibited expression in epithelial cells. Each protein subcellular expression pattern mimics the behaviour observed in real high-resolution IF data. In this way, we can capture protein co-localisation patterns. In addition, developing realistic protein expression models could potentially aid the discovery of yet unknown protein interactions.

In order to assess the quality of the protein models, we begin by assessing how well the cell organelles have been modelled. We first consider how accurately organelle features that have been used as input to the model have been generated within the synthesised data. The distributions of the numbers of organelles per cell and their position are shown in Figure 5.14. We can see that the distributions of the numbers of organelles are reasonably good approximations of the real PDFs. For the number of golgi, we can see that there are a small number of cells with a very high number of golgi organelles. However, a similar, although smaller peak in the histogram can be observed in the real data (Figure 5.5 (c)). On the other hand, we can see a wider distributions for the position parameter of the synthesised golgi and vesicles. This is due to the fact that when the position of these organelles is being calculated, the method assumes that the nucleus is in the centre of the cell, rather than displaced towards the base of the cell. Hence, the problem does not occur in stromal cells and high-grade cancer samples. On the other hand, the distributions for the ratio between the minor axes of the synthesised organelles and the nucleus of the corresponding cell as shown in Figure 5.15 (left column) and between the minor and major axes of the synthesised organelles in Figure 5.15 (right column) show very good agreement with the PDFs estimated from the real data. We have also considered features that have not been explicitly learned from the real data. Figure 5.16 shows the distributions of the solidity for real and synthesised organelles and we can observe very good agreement between the two. In Figure 5.17 we consider the area taken up by the organelles as a fraction of the total area of the cell, for the golgi and vesicles, or of the nucleus, for the nucleoli. Although the area of the organelles is not specified explicitly within the model, we observe good agreement between the real and synthesised distributions.

5.5.1 Protein Network Analysis

In Chapter 3, we introduced a framework for analysing multiplex IF data, such as the one simulated by the model described above. In this section, we apply the DiSWOP framework to a set of simulated images. For this purpose, we generated 10 healthy and 10 moderately differentiated cancerous samples at 40× magnification. From

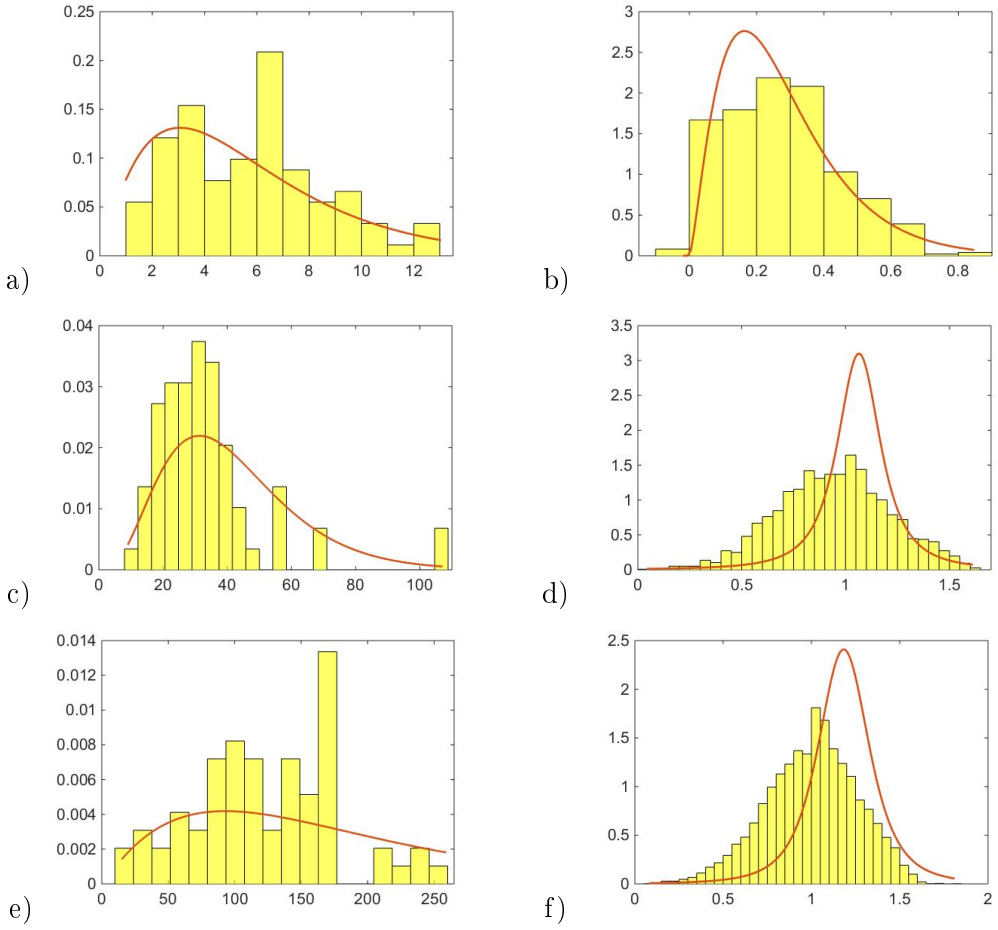


Figure 5.14: Probability distribution functions for the synthesised number (left column) and position (right column) of the (a, b) nucleoli, (c, d) golgi and (e ,f) vesicles. The probability distribution functions shown are the ones estimated for the real data, shown in Figure 5.5.

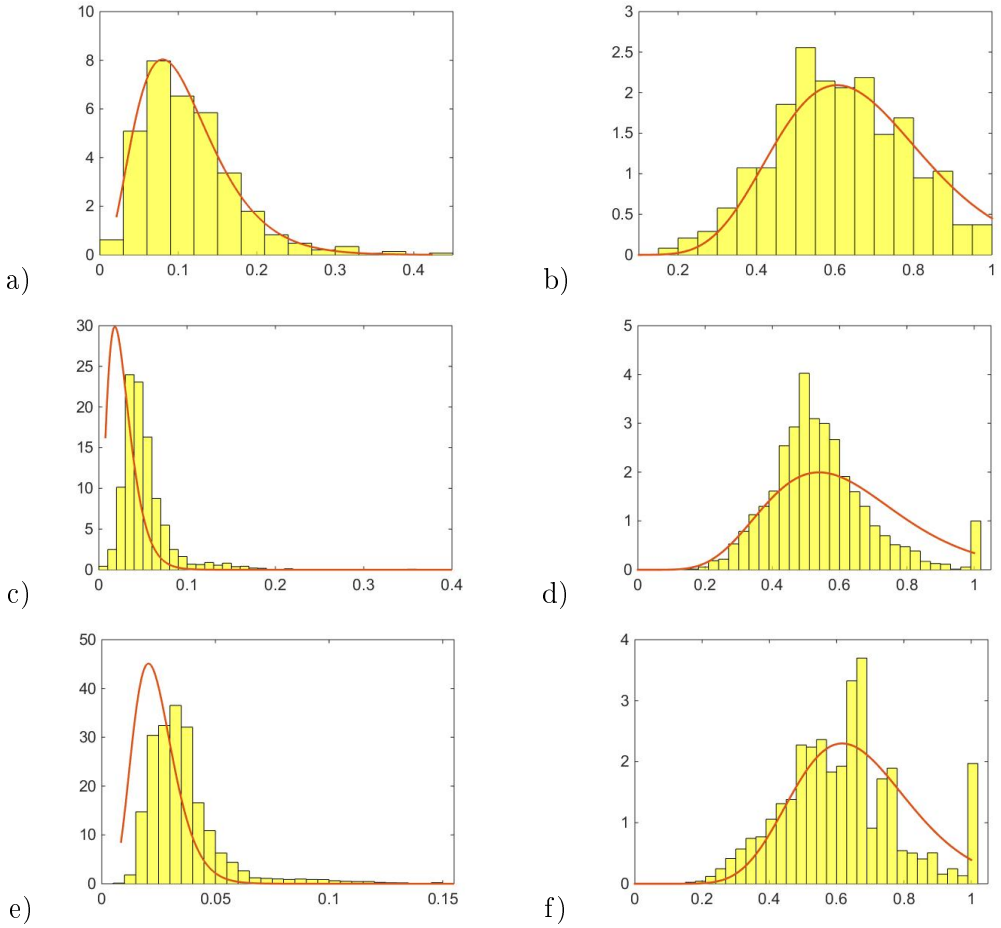


Figure 5.15: Probability distribution functions for the ratios between the minor axes of the synthesised organelles and the nucleus of the corresponding cell (left column) and between the minor and major axes of the synthesised organelles (right column). Figures show the ratios for (a, b) nucleoli, (c, d) golgi and (e, f) vesicles. The probability distribution functions shown are the ones estimated for the real data, shown in Figure 5.6.

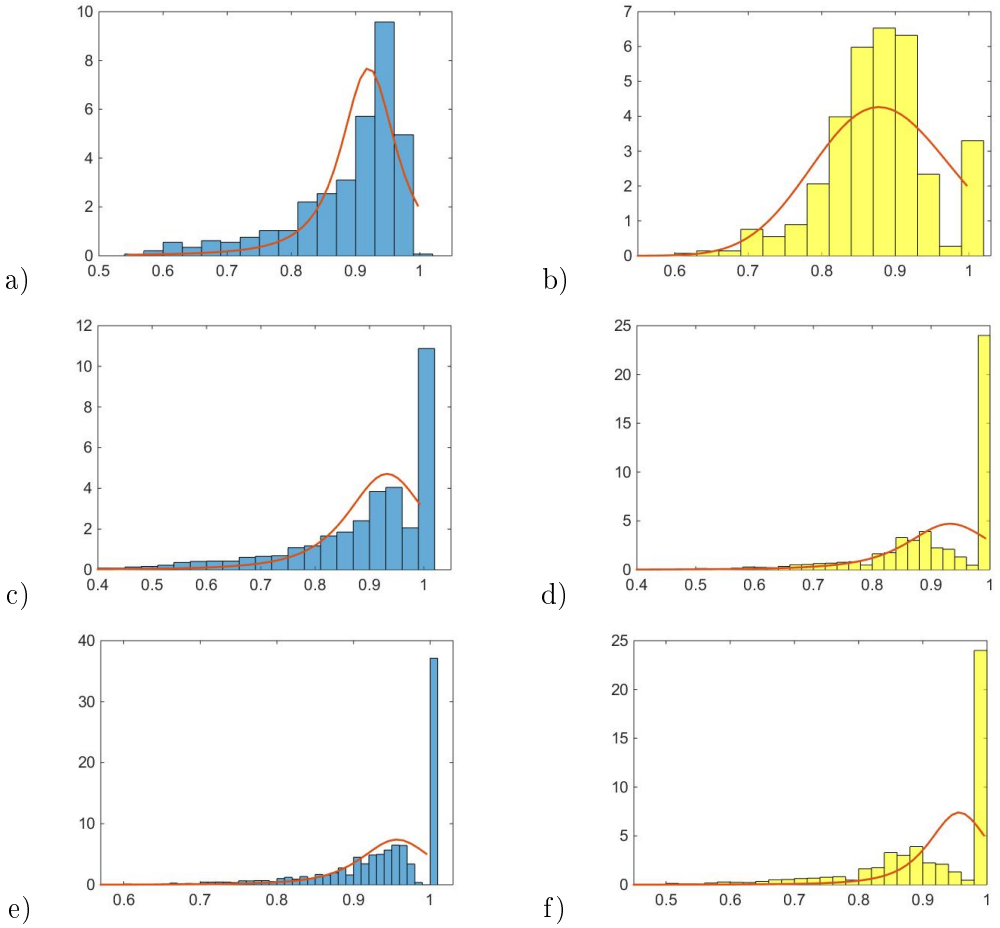


Figure 5.16: Probability distribution functions for the real (left column) and synthesised (right column) solidity of the (a, b) nucleoli, (c, d) golgi and (e ,f) vesicles. The probability distribution functions shown are the ones estimated for the real data.

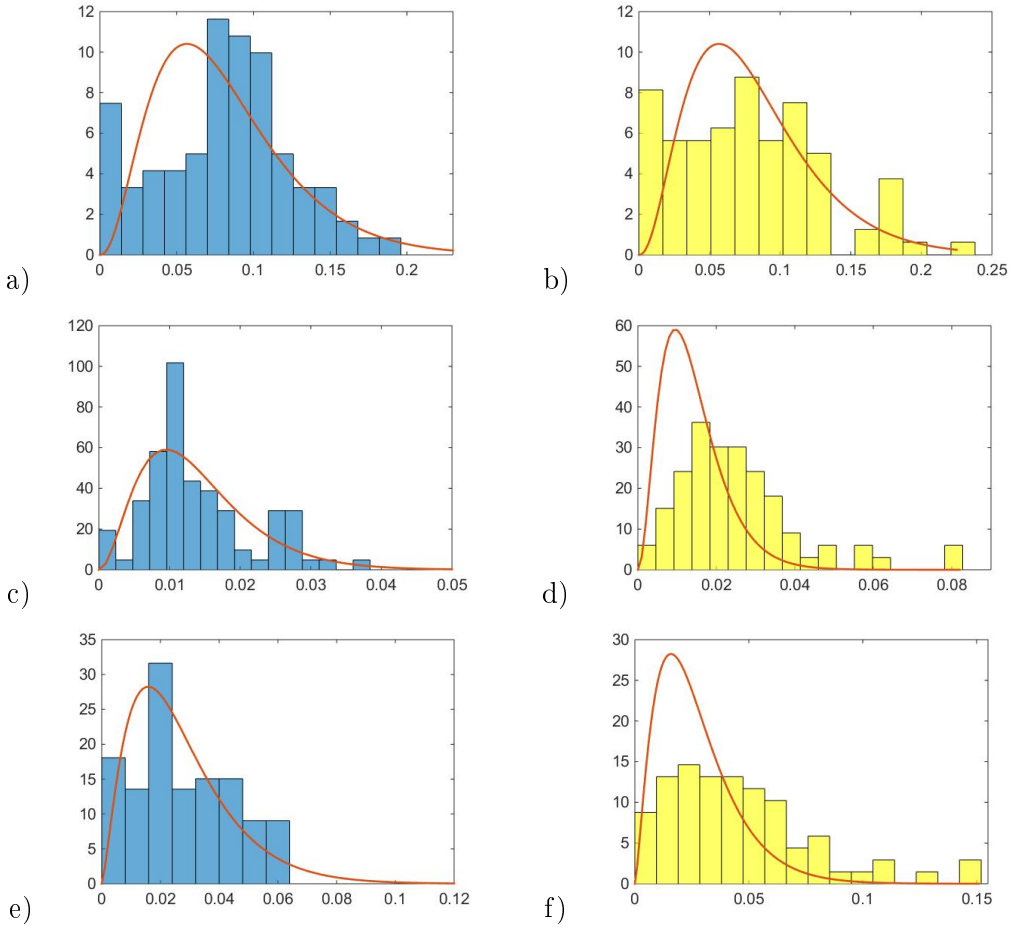


Figure 5.17: Probability distribution functions for the cell area fraction taken up by the real (left column) and synthesised (right column) organelles. Figures (a, b) show the fraction of nuclear area taken up by the nucleoli. The fraction of cytoplasmic area taken up by (c, d) golgi and (e ,f) vesicles is also considered. The probability distribution functions shown are the ones estimated for the real data.

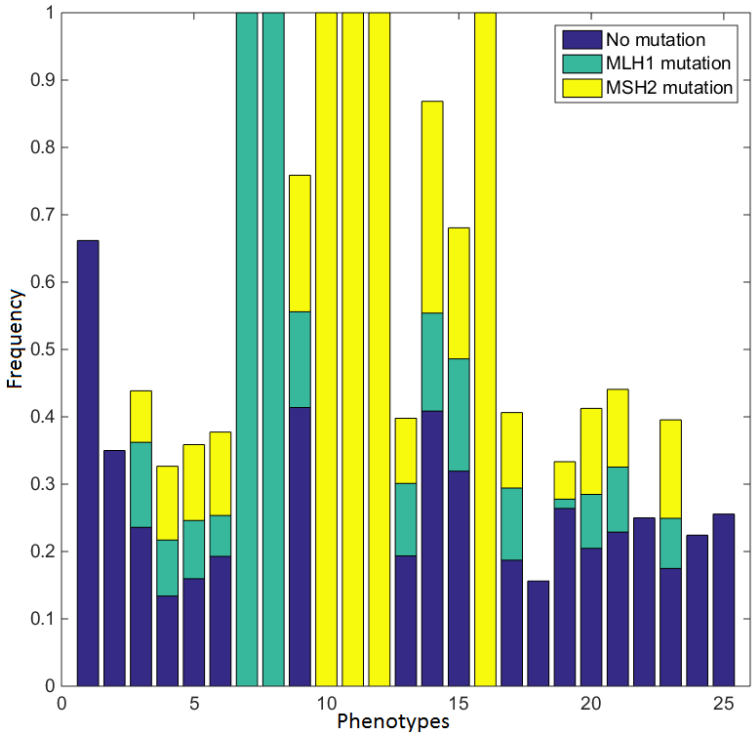


Figure 5.18: Distribution of phenotypes within cancer samples simulated at 40× magnification. Phenotypes are shown along the x-axis and the fraction of the phenotype that is found within each type of cancer samples is shown along the y-axis. Cancer samples without mutation are shown in blue, samples with MLH1 mutation are shown in teal and yellow shows the samples with MSH2 mutation.

the 10 cancerous samples, 4 had no mutation, 3 had a mutation in the MLH1 gene and 3 had a mutation in the MSH2 gene. The same dataset was also simulated at 20× magnification to investigate the dependence of the DiSWOP measure on the magnification scale.

For each of the cells, we calculate the PPDP using the MIC. The protein pairs are shown in Table 5.4. The cells are phenotyped using Affinity Propagation according to their PPDP. The distribution of the phenotypes within the cancerous samples simulated at 40× magnification is shown in Figure 5.18. We can see that phenotypes 7 and 8 that can be found only in samples with MLH1 mutation. Their PPDPs are highlighted in red in Figure 5.19. From Figure 5.19, we can see that phenotype 7 exhibits non-zero dependence only between MSH2 and MSH6, whereas phenotype 8 also has non-zero dependencies between these two proteins and P53. This can also be observed from the real data. We can see in Figure 5.20 that the two phenotypes include all of the epithelial cells, with phenotype 8 including all

	PMS2	MSH2	MSH6	P53
MLH1	1	2	3	4
PMS2		5	6	7
MSH2			8	9
MSH6				10

Table 5.4: Protein pair numbering.

epithelial cells expressing P53. On the other hand, phenotypes 10, 11, 12 and 16 are found only in samples with MSH2 mutation. Phenotypes 10, and 16 (marked in blue in Figure 5.19) show non-zero dependencies between MLH1, PMS2 and P53, splitting the epithelial cells expressing P53 in two phenotypes. These are shown in Figure 5.21. This demonstrates that the clustering is able to detect meaningful cell phenotypes, although the real phenotypes could be split into two or more phenotypes found by the algorithm.

Once we have obtained the phenotypes, we calculate the DiSWOP measure. We consider the top 3 protein pairs in each phenotype. This was chosen because, when the values in the PPDPs were ordered by size, the mean values only for the top 3 protein pairs were above 0.5. The DiSWOP results for the simulated samples at 40× and 20× magnification are shown in Figure 5.22. We can see that nearly the same results are obtained, demonstrating that the measure is independent of the magnification scale and size of the cells. Figure 5.22 also shows that DiSWOP is able to detect that the dependences between MLH1, PMS2 and MSH2 are stronger in the healthy samples, suggesting that they are broken in at least some of the cancer samples. However, it's difficult to interpret the results further as within the cancer samples there are a number of non-MSI samples and cells which have the same protein expressions as the healthy samples. To further analyse the simulated data, we considered dividing the cancer samples into three sets depending on the presence of a mutation. We re-run the analysis framework when considering non-MSI samples versus MSI samples with both mutations (Figure 5.23 (a)), and versus each mutation separately (Figure 5.23 (b) and (c)). When samples with both mutations are considered, the results are very similar to those seen in Figure 5.22. This is due to the fact that the mutations cause all of the protein pair interactions to be broken down in some of the samples. However, the negative values again clearly indicate the lack of co-localisation of the MMR proteins. On the other hand, if we consider non-MSI samples versus samples with MLH1 mutation (Figure 5.23 (b)), we can see that, as expected, the interactions of MLH1 and PMS2 are weaker in the MSI sample while MSH6 shows stronger interactions with MSH2, P53 and MLH1. The latter

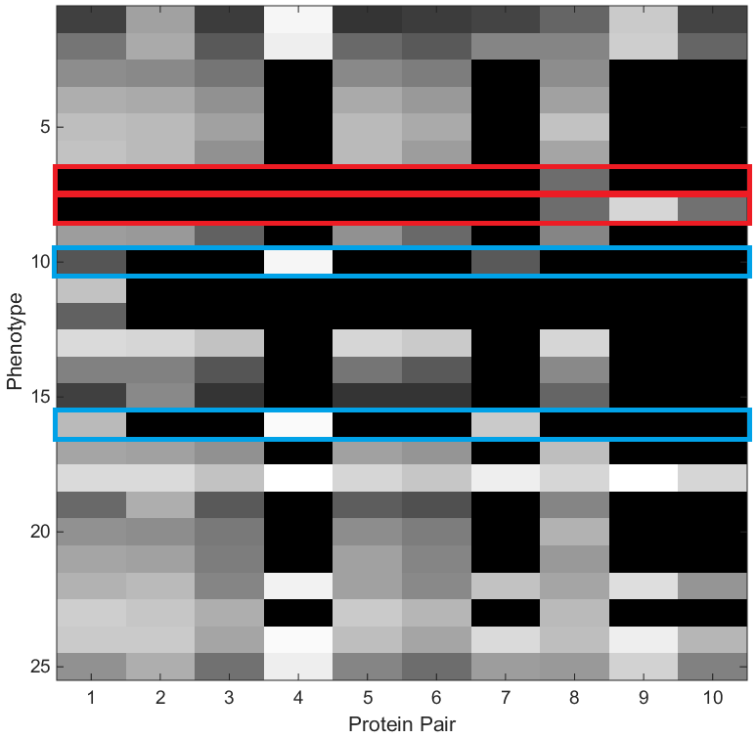


Figure 5.19: Average protein-protein dependence profiles (PPDPs) for the phenotypes found within healthy and cancerous samples simulated at 40× magnification. Phenotypes found only in samples with MLH1 mutation are highlighted in red. Some of the phenotypes found only in samples with MSH2 mutation are highlighted in blue. The numbering of the phenotypes is the same as in Figure 5.18. The numbering of the protein pairs is shown in Table 5.4. Black indicates PPD value of 0, and white shows a PPD value of 1.

interaction is likely to occur only in the stromal cells which express all proteins. Lastly, we compared non-MSI and MSH2 mutated samples (Figure 5.23 (c)). As would be expected, we observe stronger interactions of MSH2 with other proteins in the non-MSI samples. The mutated samples are characterised by increased co-localisation of P53 and PMS2.

With this set of proteins, it would be easier to simply consider the raw protein expression values. This is because there is no evidence to suggest that the expression patterns of these proteins within the cells change as a result of cancer and this has been reflected in the model. Hence, this experiment aims to demonstrate only how the DiSWOP framework could be used to analyse the synthesised data. However, DiSWOP would provide a significantly greater advantage if the simulated proteins changed their subcellular expression patterns. Detecting such changes were the aim of the study presented in Chapter 3. Proteins that exhibit such changes in locali-

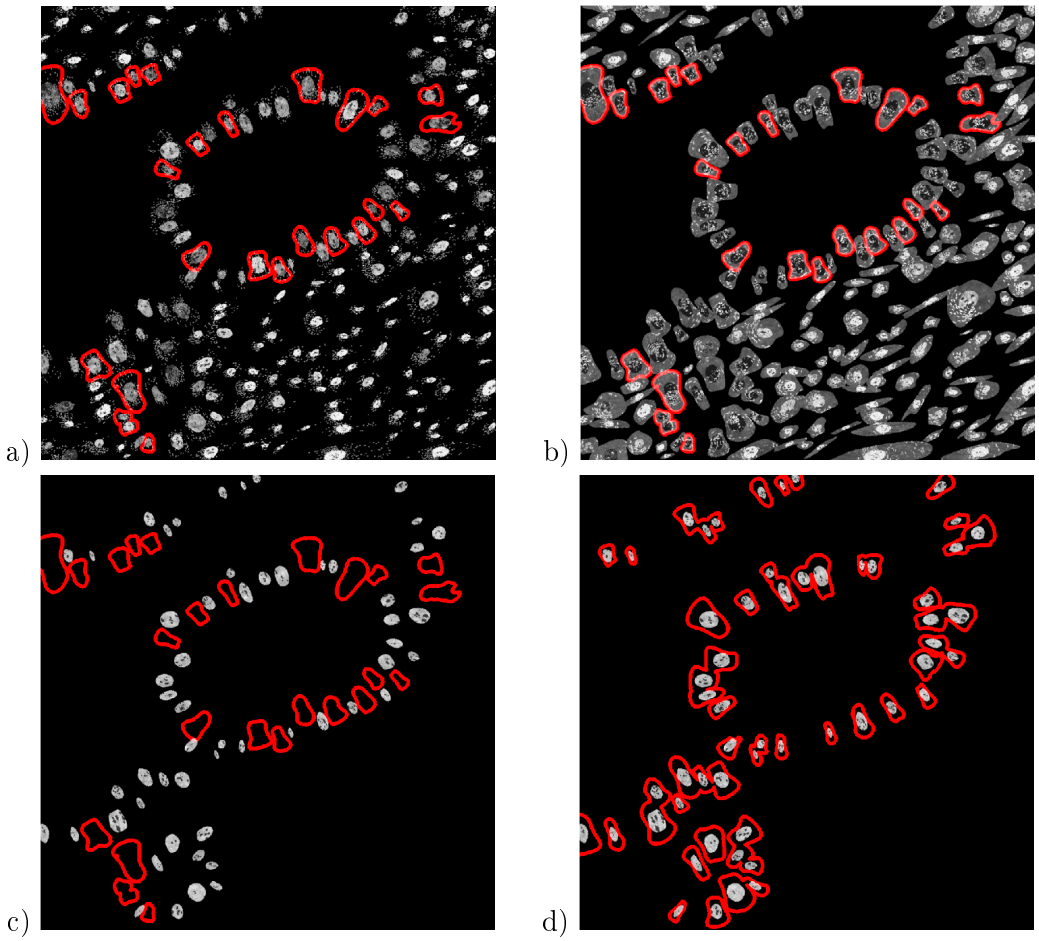


Figure 5.20: Simulated protein expression in cell phenotypes found only in MLH1 mutated samples. The images show the expression for (a) MSH2, (b) MSH6 and (c, d) P53. The red outlines indicate the cells belonging to phenotypes (a - c) 7 and (d) 8.

sation could be easily modelled using the framework presented above. These could be proteins with known response to cancer or one could generate random changes in localisation in order to test hypotheses.

Chapter Summary

In this chapter we have developed models for the protein expression patterns for five proteins associated with MSI, namely MLH1, PMS2, MSH2, MSH6 and P53. The models have been developed as an extension of the THeCoT model presented in the previous chapter. These five proteins have been chosen as a case study to illustrate how the model could be developed to generate synthetic multiplex IF data. Further

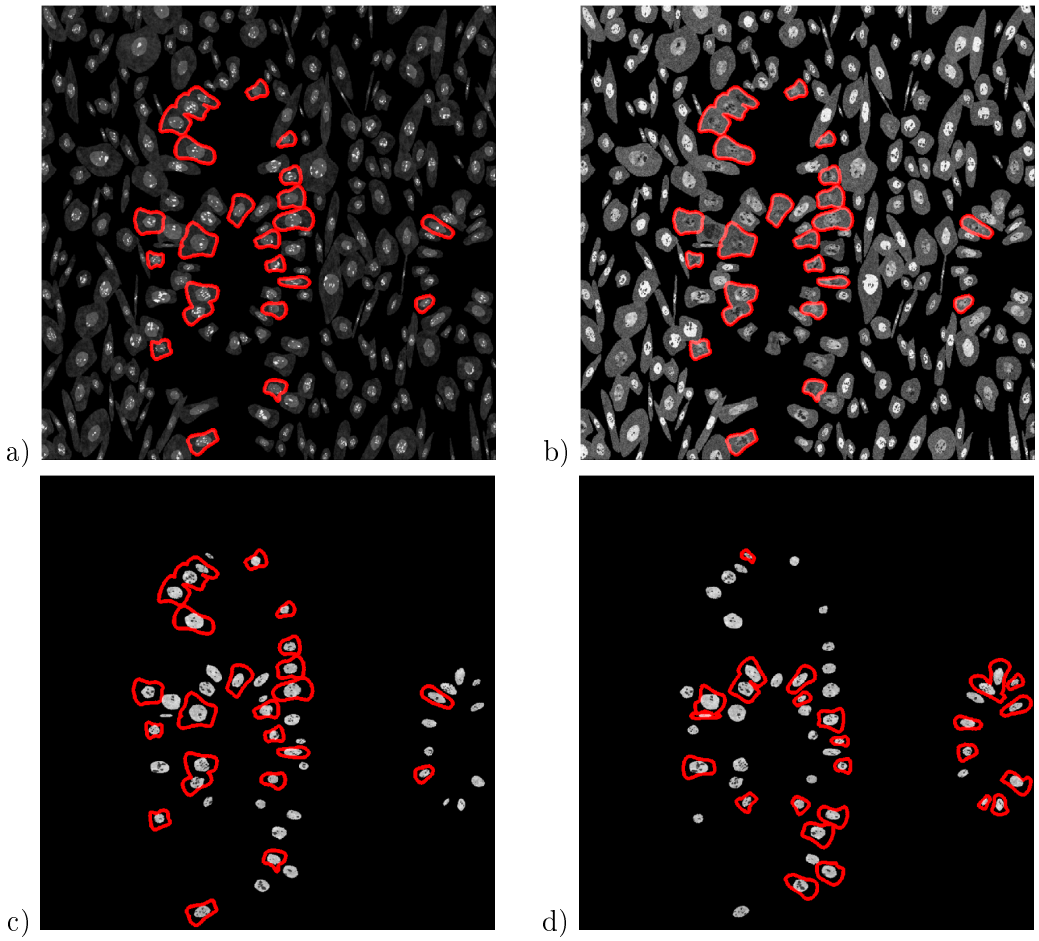


Figure 5.21: Simulated protein expression in cell phenotypes found only in MSH2 mutated samples. The images show the expression for (a) MLH1, (b) PMS2 and (c, d) P53. The red outlines indicate the cells belonging to phenotypes (a - c) 10 and (d) 16.

proteins could be included within the model in a similar manner to enable the study of a larger set of proteins of interest and their interactions.

In order to develop realistic subcellular localisation of the proteins, relevant cell organelles have been modelled. The statistics for these have been obtained from real IF data obtained from the HPA. Comparison between the distribution of various features obtained from the real and synthetic organelles has shown very good agreement. This has included both features that have been used as part of the model input and ones that have not been explicitly considered.

Finally, we presented a study of how the DiSWOP framework presented in Chapter 3 could be used to analyse the synthetic data. This kind of analysis would be invaluable in detecting changes in subcellular expression patterns resulting from

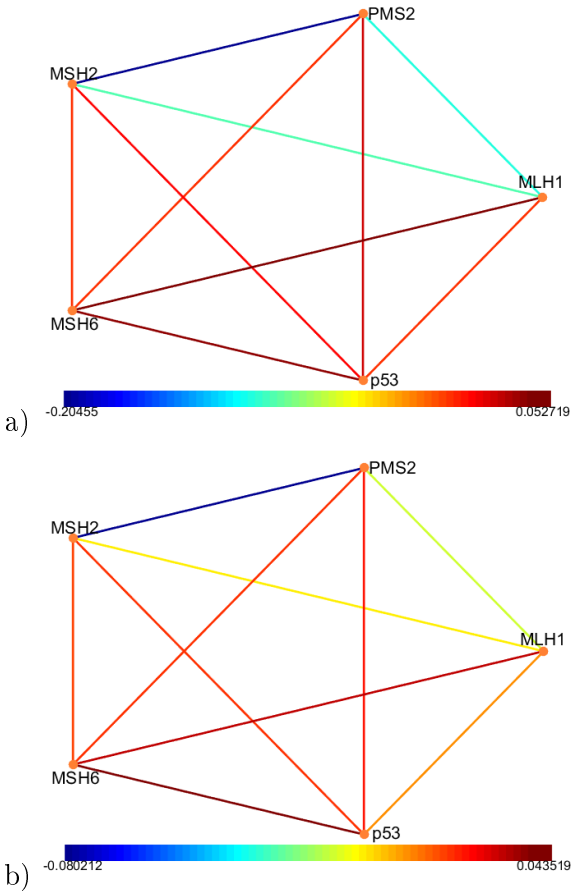


Figure 5.22: DiSWOP results for the simulated samples at (a) 40× and (b) 20× magnification. Each node represents a protein and each edge colour shows a protein pair with different level of co-expression in the normal and cancer samples. Here, a large positive value (shown in red) indicates that the protein pair is more co-dependent in cancer samples, whereas a large negative value (shown in blue) means that the protein pair is more active in normal tissue.

the development of cancer.

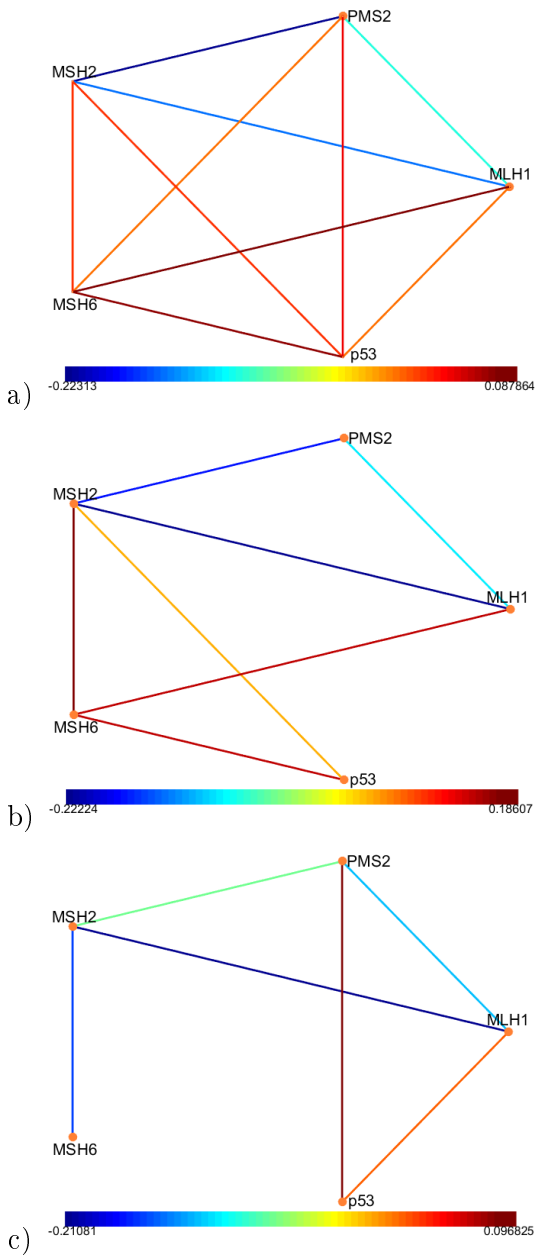


Figure 5.23: DiSWOP results for comparing MSI and non-MSI sets of the simulated cancer samples at 40× magnification. Different results are shown when comparing non-MSI samples to (a) both mutations, (b) MLH1 mutation only, and (c) MSH2 mutation only. Each node represents a protein and each edge colour shows a protein pair with different level of co-expression in the normal and cancer samples. Here, a large positive value (shown in red) indicates that the protein pair is more co-dependent in the mutated samples, whereas a large negative value (shown in blue) means that the protein pair is more active in non-MSI tissue.

Chapter 6

Conclusions and Future Directions

This thesis proposed different methods for studying the tumour microenvironment in colorectal cancer. This was done by analysing multiplex immunofluorescence images and by generating synthetic multiplex immunofluorescence and histology images. This chapter summarises and concludes the work presented in this thesis and discusses some future directions.

In Chapter 1, we have introduced the reader to the heterogeneity present within and between CRC tumours. Recent developments in cell-level analysis techniques have revealed great variation in cell phenotypes present within a tumour. The diversity within the cancer cell population may lead to the emergence of clinical resistance during disease progression. Inter-tumour heterogeneity has been widely studied and there are a number of cancer therapies which address certain mutations present in a sub-population of cancer patients. In CRC, it is crucial to detect cases exhibiting MSI, as some of these could be due to the inherited Lynch syndrome. Identifying patients with LS has implications on the care and monitoring of the patient and potentially affected family members. In addition, a detailed description of the normal architecture of colon tissue is presented. This is followed by a description of how the architecture changes as cancer develops and becomes more malignant, and how it is graded in clinical practice.

In Chapter 2, we review the existing literature on multiplex imaging. The review covered the approaches taken so far to extract meaningful quantitative results from the TIS imaging data. These include pixel-level analyses both with and without thresholding the intensity values. We also briefly reviewed studies that have been performed with other multiplex techniques such as MALDI, Raman, multi-spectral imaging, MxIF and imaging mass cytometry. The review of these techniques has been included, despite the fact that the frameworks developed within this thesis have been

designed for TIS image data, as they can easily be applicable for other multiplex imaging techniques. The same chapter also included a review of frameworks for the generation of synthetic image data. Currently, the majority of these methods focus on the generation of homogeneous cell populations in culture. It also briefly reviewed current methods for simulating protein expression.

We have introduced a novel method for analysing multiplex image data such as the TIS image data in Chapter 3. It is different from previously presented methods in that it considers the samples at cell rather than at pixel level, it is intensity independent, and it allows phenotyping of cells based on their protein co-expression profile. We have presented two new measures of co-dependence and anti-co-dependence, namely DiSWOP and DiSWAP. Applying these over a TIS dataset of eleven samples of cancerous and normal colon tissue, we have found combinations of protein pairs that are much more co-dependent or anti-codependent in cancerous than in normal tissue, pointing to the possibility that combinations of protein pairs rather than single proteins will lead to specific markers for cancer. The resulting protein pair interactions have been validated by consulting with literature on protein interaction pathways established through experimental techniques. In addition, we have investigated where in the data cell phenotypes of interest are located, enabling us to study the inter- and intra- tumour heterogeneity present in our samples. The results presented here are only preliminary and need to be validated using a larger number of samples and subsequently by other biological techniques. While the number of samples considered is insufficient to draw significant biological conclusions, this is the largest study of colon cancer using TIS conducted to date. A study with a larger number of proteins could also aid the optimisation of the number of proteins and cell phenotypes to be considered when calculating the DiSWOP measure. Due to the general nature of the framework, the method could be applied to other tissues and/or images obtained from other multiplex imaging techniques. The method can also be used on other high-throughput techniques that measure localised expression of DNA or RNA, as long as individual cells can be identified or approximated.

Modern high-throughput imaging methods have raised the need for automated analytical frameworks. However, validation of such methods has been challenging since ground truth information in cell biological research is often missing, and verification using manual methods introduces variable results. Hence, simulation is a valuable tool when trying to develop, validate, and compare analytical methods.

In Chapter 4, we presented a model for simulating healthy and cancerous colonic crypt architecture. The proposed model for tumour heterogeneity in colorectal tissue (THeCoT) has several parameters, which allow control over the tissue

appearance. Detailed analysis of hand-marked H&E images has enabled us to make the model realistic by learning parameters to generate realistic cell phenotypes, chromatin and lumen texture, nuclei morphology, and crypt architecture. To the best of our knowledge, ours is the first model to simulate histology image data at subcellular level, where the cells have several compartments and are organised to mimic the microenvironment of tissue *in situ* rather than dispersed cells in a cultured environment. Majority of features of the histology images produced by the model have been rated as being very realistic by the pathologists. The feature rated as least realistic was the appearance of the stromal cells. Dividing stromal cells in the real H&E images according to their functional phenotypes and analysing them separately to obtain their characteristic features to input into the model could address this problem in future developments of the model. We have also shown an example of how a crypt segmentation method can be used on the synthetic data. When the method was trained on real data, it performed worse when trying to segment the high-grade cancer crypts. This is likely to be due to the fact that currently the model does not include model for the extracellular matrix. Including this would generate a more realistic texture outside the cancer glands. We have also demonstrated that phenotyping of the cells on the basis of their textural characteristics showed consistency in the results based on real and synthetic nuclei. The synthesised data could be used to validate techniques such as image restoration, cell and crypt segmentation, stain normalisation, and cancer grading. An interesting application of the synthesised data would be to pre-train convolutional neural networks. These methods are attracting more and more interest in the field of digital pathology but require large amounts of ground truth data for training. The model could be used to generate a significant number of images to pre-train the convolutional networks, which can later be fine-tuned on a more limited set of manually annotated real data. In future, the model could also be extended to include 3D colonic tumour microenvironment taking into account 3D spatial arrangements of all tissue constituents *in situ*. This will allow the model to simulate events due to sectioning commonly observed in histopathology images, such as apparent overlap between cells or cells without nuclei. In addition, extending the model to 3D will allow the simulation of serial sections. Furthermore, the model could be extended to a larger visual field which can contain a wide variety of histopathological features that have been recently shown to be associated with poor prognosis or metastasis. These include immune infiltration, tumour budding, density of macro- and micro-vascular blood vessels, and functional state of lymphatics. While so far we have focused on developing a model only for the spatial microenvironment in CRA, we are also developing a method for generating such a

model. In particular, the same approach can be used to develop models for other epithelial tumours such as prostate, pancreatic or esophageal cancers.

Lastly, in Chapter 5 we extend the THeCoT model to simulate multiplexed IF data. We investigate how to realisticly simulate the expression of five proteins associated with MSI, namely MLH1, PMS2, MSH2, MSH6 and p53. Following the same method, further proteins of interest could be easily added to the model to increase its usability. This could aid the study of toponomics. In order to simulate the subcellular location of the proteins, we developed models for the cell nucleoli, golgi and vesicles, using parameters obtained from real fluorescence data of cells in culture. Comparison between the distribution of various features obtained from the real and synthetic organelles has shown very good agreement. This has included both features that have been used as part of the model input and ones that have not been explicitly considered. The addition of further proteins of interest may require more of the cell organelles to be modelled, such as the cytoskeleton and the endoplasmic reticulum. It would be difficult to represent these using the deformed circle model, so a different approach may need to be developed. Finally, we presented a study of how the DiSWOP framework presented in Chapter 3 could be used to analyse the synthetic data. Using the framework to compare the protein co-localisation in MSI versus non-MSI samples was able to detect the presence of mutations. While for the set of proteins considered in this study, it could also be achieved by consideration of the raw expression values, this kind of analysis would be invaluable in detecting changes in sub-cellular expression patterns resulting from the development of cancer. Proteins that exhibit such changes in localisation could be easily modelled using the framework presented in order to test various hypotheses.

Bibliography

[1] D Houle. Numbering the hairs on our heads: the shared challenge and promise of phenomics. *Proceedings of the National Academy of Sciences*, 107(suppl 1): 1793–1799, 2010.

[2] PC Nowell. The clonal evolution of tumor cell populations. *Science*, 194(4260): 23–28, 1976.

[3] D Hanahan and RA Weinberg. The hallmarks of cancer. *cell*, 100(1):57–70, 2000.

[4] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *cell*, 144(5):646–674, 2011.

[5] NA Bhowmick, EG Neilson, and HL Moses. Stromal fibroblasts in cancer initiation and progression. *Nature*, 432(7015):332–337, 2004.

[6] AH Beck, AR Sangoi, S Leung, RJ Marinelli, TO Nielsen, MJ van de Vijver, RB West, M van de Rijn, and D Koller. Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Science translational medicine*, 3(108):108ra113–108ra113, 2011.

[7] LH Sobin and ID Fleming. Tnm classification of malignant tumors. *Cancer*, 80(9):1803–1804, 1997.

[8] ID Nagtegaal, P Quirke, and HJ Schmoll. Has the new tnm classification for colorectal cancer improved care? *Nature Reviews Clinical Oncology*, 9(2): 119–123, 2012.

[9] J Galon, B Mlecnik, G Bindea, HK Angell, A Berger, C Lagorce, A Lugli, I Zlobec, A Hartmann, C Bifulco, et al. Towards the introduction of the immunoscorein the classification of malignant tumours. *The Journal of pathology*, 232(2):199–209, 2014.

[10] LM Wang, D Kevans, H Mulcahy, J O'Sullivan, D Fennelly, J Hyland, D O'Donoghue, and K Sheahan. Tumor budding is a strong and reproducible prognostic marker in t3n0 colorectal cancer. *The American journal of surgical pathology*, 33(1):134–141, 2009.

[11] I Zlobec, K Baker, P Minoo, JR Jass, L Terracciano, and A Lugli. Node-negative colorectal cancer at high risk of distant metastasis identified by combined analysis of lymph node status, vascular invasion, and raf-1 kinase inhibitor protein expression. *Clinical cancer research*, 14(1):143–148, 2008.

[12] R Fisher, L Pusztai, and C Swanton. Cancer heterogeneity: implications for targeted therapeutics. *British journal of cancer*, 108(3):479–485, 2013.

[13] M Guillaud, K Adler-Storthz, A Malpica, G Staerkel, J Matisic, D Van Niekirk, D Cox, N Poulin, M Follen, and C MacAulay. Subvisual chromatin changes in cervical epithelium measured by texture image analysis and correlated with hpv. *Gynecologic oncology*, 99(3):S16–S23, 2005.

[14] DA Gutman, J Cobb, D Somanna, Y Park, F Wang, T Kurc, JH Saltz, DJ Brat, LAD Cooper, and J Kong. Cancer digital slide archive: an informatics resource to support integrated in silico analysis of tcga pathology data. *Journal of the American Medical Informatics Association*, 20(6):1091–1098, 2013.

[15] D Zink, AH Fischer, and JA Nickerson. Nuclear structure in cancer cells. *Nature reviews cancer*, 4(9):677–687, 2004.

[16] ML Bisgaard. Young age colorectal cancer and identification of hereditary non-polyposis colorectal cancer cohorts. *British Journal of Surgery*, 94(9): 1055–1056, 2007.

[17] MM Reza, JA Blasco, E Andradas, R Cantero, and J Mayol. Systematic review of laparoscopic versus open surgery for colorectal cancer. *British journal of surgery*, 93(8):921–928, 2006.

[18] K Söreide, EAM Janssen, H Söiland, H Körner, and JPA Baak. Microsatellite instability in colorectal cancer. *British journal of surgery*, 93(4):395–406, 2006.

[19] JJY Sung, JYW Lau, KL Goh, and WK Leung. Increasing incidence of colorectal cancer in asia: implications for screening. *The lancet oncology*, 6(11): 871–876, 2005.

[20] Philippe L Bedard, Aaron R Hansen, Mark J Ratain, and Lillian L Siu. Tumour heterogeneity in the clinic. *Nature*, 501(7467):355–364, 2013.

[21] P Dalerba, T Kalisky, D Sahoo, PS Rajendran, ME Rothenberg, AA Leyrat, S Sim, J Okamoto, DM Johnston, D Qian, et al. Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nature biotechnology*, 29 (12):1120–1127, 2011.

[22] EA Vucic, KL Thu, K Robison, LA Rybaczyk, R Chari, CE Alvarez, and WL Lam. Translating cancer omics to improved outcomes. *Genome research*, 22(2):188–195, 2012.

[23] J Casado-Vela, A Cebrián, MTG del Pulgar, and JC Lacal. Approaches for the study of cancer: towards the integration of genomics, proteomics and metabolomics. *Clinical and Translational Oncology*, 13(9):617–628, 2011.

[24] Q Tian, ND Price, and L Hood. Systems cancer medicine: towards realization of predictive, preventive, personalized and participatory (p4) medicine. *Journal of internal medicine*, 271(2):111–121, 2012.

[25] C Box, SJ Rogers, M Mendiola, and SA Eccles. Tumour-microenvironmental interactions: paths to progression and targets for treatment. *Seminars in cancer biology*, 20(3):128–138, 2010.

[26] N Makrilia, A Kollias, L Manolopoulos, and K Syrigos. Cell adhesion molecules: role and clinical significance in cancer. *Cancer investigation*, 27 (10):1023–1037, 2009.

[27] L Vermeulen, FdS Melo, DJ Richel, and JP Medema. The developing cancer stem-cell model: clinical challenges and opportunities. *The lancet oncology*, 13 (2):e83–e89, 2012.

[28] K Pietras and A Östman. Hallmarks of cancer: interactions with the tumor stroma. *Experimental cell research*, 316(8):1324–1331, 2010.

[29] T Udagawa and M Wood. Tumor–stromal cell interactions and opportunities for therapeutic intervention. *Current opinion in pharmacology*, 10(4):369–374, 2010.

[30] SM Weis and DA Cheresh. Tumor angiogenesis: molecular pathways and therapeutic targets. *Nature medicine*, 17(11):1359–1370, 2011.

[31] SL Shiao, AP Ganesan, HS Rugo, and LM Coussens. Immune microenvironments in solid tumors: new targets for therapy. *Genes & development*, 25(24): 2559–2572, 2011.

[32] Q Shi, L Qin, W Wei, F Geng, R Fan, YS Shin, D Guo, L Hood, PS Mischel, and JR Heath. Single-cell proteomic chip for profiling intracellular signaling pathways in single tumor cells. *Proceedings of the National Academy of Sciences*, 109(2):419–424, 2012.

[33] DP Cahill, KW Kinzler, B Vogelstein, and C Lengauer. Genetic instability and darwinian selection in tumours. *Trends in Genetics*, 15(12):M57–M60, 1999.

[34] RA Burrell, N McGranahan, J Bartek, and C Swanton. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*, 501(7467): 338–345, 2013.

[35] S Dedeurwaerder, D Fumagalli, and F Fuks. Unravelling the epigenomic dimension of breast cancers. *Current opinion in oncology*, 23(6):559–565, 2011.

[36] A Pietras. Cancer stem cells in tumor heterogeneity. *Adv Cancer Res*, 112: 255–281, 2011.

[37] P Martinez, NJ Birkbak, M Gerlinger, N McGranahan, RA Burrell, AJ Rowan, T Joshi, R Fisher, J Larkin, Z Szallasi, et al. Parallel evolution of tumour subclones mimics diversity between tumours. *The Journal of pathology*, 230 (4):356–364, 2013.

[38] JH Malone and B Oliver. Microarrays, deep sequencing and the true measure of the transcriptome. *BMC biology*, 9(1):34, 2011.

[39] M Stoeckli, P Chaurand, DE Hallahan, and RM Caprioli. Imaging mass spectrometry: a new technology for the analysis of protein expression in mammalian tissues. *Nature medicine*, 7(4):493–496, 2001.

[40] S Ogino and A Goel. Molecular classification and correlates in colorectal cancer. *The Journal of Molecular Diagnostics*, 10(1):13–27, 2008.

[41] M Fleming, S Ravula, SF Tatishchev, and HL Wang. Colorectal carcinoma: Pathologic aspects. *Journal of gastrointestinal oncology*, 3(3):153–173, 2012.

[42] MS Pino and DC Chung. The chromosomal instability pathway in colon cancer. *Gastroenterology*, 138(6):2059–2072, 2010.

[43] DC Snover. Update on the serrated pathway to colorectal carcinoma. *Human pathology*, 42(1):1–10, 2011.

[44] JR Jass. Classification of colorectal cancer based on correlation of clinical, morphological and molecular features. *Histopathology*, 50(1):113–130, 2007.

[45] RS Lahue, KG Au, and P Modrich. Dna mismatch correction in a defined system. *Science(Washington)*, 245(4914):160–164, 1989.

[46] Katherine B Geiersbach and Wade S Samowitz. Microsatellite instability and colorectal cancer. *Archives of pathology & laboratory medicine*, 135(10):1269, 2011.

[47] Anna Pluciennik and Paul Modrich. Protein roadblocks and helix discontinuities are barriers to the initiation of mismatch repair. *Proceedings of the National Academy of Sciences*, 104(31):12709–12713, 2007.

[48] A de la Chapelle and H Hampel. Clinical relevance of microsatellite instability in colorectal cancer. *Journal of Clinical Oncology*, 28(20):3380–3387, 2010.

[49] P Peltomäki. Lynch syndrome genes. *Familial cancer*, 4(3):227–232, 2005.

[50] YMC Hendriks, A Wagner, H Morreau, F Menko, A Stormorken, F Quehenberger, L Sandkuijl, P Møller, M Genuardi, H Van Houwelingen, et al. Cancer risk in hereditary nonpolyposis colorectal cancer due to msh6 mutations: impact on counseling and surveillance. *Gastroenterology*, 127(1):17–25, 2004.

[51] NC Nicolaides, N Papadopoulos, B Liu, YF Weit, KC Carter, SM Ruben, CA Rosen, WA Haseltine, RD Fleischmann, CM Fraser, et al. Mutations of two p/ws homologues in hereditary nonpolyposis colon cancer. *Nature*, 371 (6492):75–80, 1994.

[52] Wade S Samowitz, Hans Albertsen, Jennifer Herrick, Theodore R Levin, Carol Sweeney, Maureen A Murtaugh, Roger K Wolff, and Martha L Slattery. Evaluation of a large, population-based sample supports a cpg island methylator phenotype in colon cancer. *Gastroenterology*, 129(3):837–845, 2005.

[53] WS Samowitz, JA Holden, K Curtin, SL Edwards, AR Walker, HA Lin, MA Robertson, MF Nichols, KM Gruenthal, BJ Lynch, et al. Inverse relationship between microsatellite instability and k-ras and p53 gene alterations in colon cancer. *The American journal of pathology*, 158(4):1517–1524, 2001.

[54] L Laghi and A Malesci. Microsatellite instability and therapeutic consequences in colorectal cancer. *Digestive diseases*, 30(3):304–309, 2012.

[55] CM Ribic, DJ Sargent, MJ Moore, SN Thibodeau, AJ French, RM Goldberg, SR Hamilton, P Laurent-Puig, R Gryfe, LE Shepherd, et al. Tumor microsatellite-instability status as a predictor of benefit from fluorouracil-based adjuvant chemotherapy for colon cancer. *New England Journal of Medicine*, 349(3):247–257, 2003.

[56] GY Locker, S Hamilton, J Harris, JM Jessup, N Kemeny, JS Macdonald, MR Somerfield, DF Hayes, and RC Bast. Asco 2006 update of recommendations for the use of tumor markers in gastrointestinal cancer. *Journal of Clinical Oncology*, 24(33):5313–5327, 2006.

[57] RW Beart, LJ Melton III, M Maruta, MB Dockerty, HB Frydenberg, and WM O'Fallon. Trends in right and left-sided colon cancer. *Diseases of the Colon & Rectum*, 26(6):393–398, 1983.

[58] JA Bufill. Colorectal cancer: evidence for distinct genetic categories based on proximal or distal tumor location. *Annals of internal medicine*, 113(10): 779–788, 1990.

[59] ER Fearon and B Vogelstein. A genetic model for colorectal tumorigenesis. *Cell*, 61(5):759–767, 1990.

[60] MA Swartz, N Iida, EW Roberts, S Sangaletti, MH Wong, FE Yull, LM Coussens, and YA DeClerck. Tumor microenvironment complexity: emerging roles in cancer therapy. *Cancer research*, 72(10):2473–2480, 2012.

[61] B Young, P Woodford, and G O'Dowd. *Wheater's functional histology: a text and colour atlas*. Elsevier Health Sciences, 2013.

[62] Lucia Ricci-Vitiani, Eros Fabrizi, Elisabetta Palio, and Ruggero De Maria. Colon cancer stem cells. *Journal of Molecular Medicine*, 87(11):1097–1104, 2009.

[63] WK Blenkinsopp, S Stewart-Brown, L Blesovsky, G Kearney, and LP Fielding. Histopathology reporting in large bowel cancer. *Journal of clinical pathology*, 34(5):509–513, 1981.

[64] CC Compton. Pathology report in colon cancer: what is prognostically important? *Digestive Diseases*, 17(2):67–79, 1999.

[65] JR Jass, WS Atkin, Jet Cuzick, HJR Bussey, BC Morson, JMA Northover, and IP Todd. The grading of rectal cancer: historical perspectives and a multivariate analysis of 447 cases. *Histopathology*, 10(5):437–459, 1986.

[66] W Schubert. On the origin of cell functions encoded in the toponome. *Journal of biotechnology*, 149(4):252–259, 2010.

[67] D Webb, MA Hamilton, GJ Harkin, S Lawrence, AK Camper, and Z Lewandowski. Assessing technician effects when extracting quantities from microscope images. *Journal of microbiological methods*, 53(1):97–106, 2003.

[68] S Bhattacharya, G Mathew, E Ruban, DBA Epstein, A Krusche, R Hillert, W Schubert, and M Khan. Toponome imaging system: in situ protein network mapping in normal and cancerous colon from the same patient reveals more than five-thousand cancer specific protein clusters and their subcellular annotation by using a three symbol code. *Journal of proteome research*, 9(12): 6112–6125, 2010.

[69] SeA Raza, A Humayun, S Abouna, TW Nattkemper, DBA Epstein, M Khan, and NM Rajpoot. Ramtab: robust alignment of multi-tag bioimages. *PLoS ONE*, 7:e30894, 2012.

[70] SG Megason and SE Fraser. Imaging in systems biology. *Cell*, 130(5):784–795, 2007.

[71] JR Mansfield. Multispectral imaging a review of its technical aspects and applications in anatomic pathology. *Veterinary Pathology Online*, 51(1):185–210, 2014.

[72] RF Murphy. Putting proteins on the map. *Nature biotechnology*, 24(10):1223–1224, 2006.

[73] W Schubert, B Bonnekoh, AJ Pommer, L Philipsen, R Böckelmann, Y Malykh, H Gollnick, M Friedenberger, M Bode, and AWM Dress. Analyzing proteome topology and function by automated multidimensional fluorescence microscopy. *Nature biotechnology*, 24(10):1270–1278, 2006.

[74] DS Cornett, ML Reyzer, P Chaurand, and RM Caprioli. Maldi imaging mass spectrometry: molecular snapshots of biochemical systems. *Nature methods*, 4 (10):828–833, 2007.

[75] HJ van Manen, YM Kraan, D Roos, and C Otto. Single-cell raman and fluorescence microscopy reveal the association of lipid bodies with phagosomes in leukocytes. *Proceedings of the National Academy of Sciences of the United States of America*, 102(29):10159–10164, 2005.

[76] E Barash, S Dinn, C Sevinsky, and F Ginty. Multiplexed analysis of proteins in tissue using multispectral fluorescence imaging. *Medical Imaging, IEEE Transactions on*, 29(8):1457–1462, 2010.

[77] MJ Gerdes, CJ Sevinsky, A Sood, S Adak, MO Bello, A Bordwell, A Can, A Corwin, S Dinn, RJ Filkins, et al. Highly multiplexed single-cell analysis of formalin-fixed, paraffin-embedded cancer tissue. *Proceedings of the National Academy of Sciences*, 110(29):11982–11987, 2013.

[78] M Angelo, SC Bendall, R Finck, MB Hale, C Hitzman, AD Borowsky, RM Levenson, JB Lowe, SD Liu, S Zhao, et al. Multiplexed ion beam imaging of human breast tumors. *Nature medicine*, 20:436–442, 2014.

[79] C Giesen, HAO Wang, D Schapiro, N Zivanovic, A Jacobs, B Hattendorf, PJ Schüffler, D Grolimund, JM Buhmann, S Brandt, et al. Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nature methods*, 11(4):417–422, 2014.

[80] W Schubert, A Gieseler, A Krusche, P Serocka, and R Hillert. Next-generation biomarkers based on 100-parameter functional super-resolution microscopy tis. *New biotechnology*, 29(5):599–610, 2012.

[81] PF Jonsson and PA Bates. Global topological features of cancer proteins in the human interactome. *Bioinformatics*, 22(18):2291–2297, 2006.

[82] J Sun and Z Zhao. A comparative study of cancer proteins in the human protein-protein interaction network. *BMC genomics*, 11(Suppl 3):S5, 2010.

[83] W Schubert. Topological proteomics, toponomics, melk-technology. In *Proteomics of Microorganisms*, pages 189–209. Springer, 2003.

[84] RG Evans, B Naidu, NM Rajpoot, DBA Epstein, and M Khan. Toponome imaging system: multiplex biomarkers in oncology. *Trends in molecular medicine*, 18(12):723–731, 2012.

[85] M Friedenberger, M Bode, A Krusche, and W Schubert. Fluorescence detection of protein clusters in individual cells and tissue sections by using toponome

imaging system: sample preparation and measuring procedures. *Nature protocols*, 2(9):2285–2294, 2007.

[86] W Schubert, A Gieseler, A Krusche, and R Hillert. Toponome mapping in prostate cancer: detection of 2000 cell surface protein clusters in a single tissue section and cell type specific annotation by using a three symbol code. *Journal of proteome research*, 8(6):2696–2707, 2009.

[87] W Schubert. Systematic, spatial imaging of large multimolecular assemblies and the emerging principles of supramolecular order in biological systems. *Journal of Molecular Recognition*, 27(1):3–18, 2014.

[88] HA Mohamed, DR Mosier, LL Zou, L Siklós, ME Alexianu, JI Engelhardt, DR Beers, WD Le, and SH Appel. Immunoglobulin fcγ receptor promotes immunoglobulin uptake, immunoglobulin-mediated calcium increase, and neurotransmitter release in motor neurons. *Journal of neuroscience research*, 69 (1):110–116, 2002.

[89] RG Miller, R Zhang, G Block, J Katz, R Barohn, E Kasarskis, D Forshew, V Gopalakrishnan, and MS McGrath. Np001 regulation of macrophage activation markers in als: A phase i clinical and biomarker study. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, 15(7-8):601–609, 2014.

[90] U Berndt, L Philipsen, S Bartsch, B Wiedenmann, DC Baumgart, M Hämmerle, and A Sturm. Systematic high-content proteomic analysis reveals substantial immunologic changes in colorectal cancer. *Cancer research*, 68(3): 880–888, 2008.

[91] A Barysenka, AWM Dress, and W Schubert. An information theoretic thresholding method for detecting protein colocalizations in stacks of fluorescence images. *Journal of biotechnology*, 149(3):127–131, 2010.

[92] D Langenkämper, J Kölling, M Khan, NM Rajpoot, and TW Nattkemper. Towards protein network analysis using tis imaging and exploratory data analysis. In *Workshop on Computational Systems Biology (WCSB)*, 2011.

[93] A Humayun, SeA Raza, C Waddington, S Abouna, M Khan, and NM. Rajpoot. A novel framework for molecular co-expression pattern analysis in multi-channel toponome fluorescence images. *MIAAB 2011 (Proceedings of the 2011 Microscopic Image Analysis with Applications in Biology)*, pages 109–112, 2011.

[94] J Kölling, D Langenkämper, S Abouna, M Khan, and TW Nattkemper. Whidea web tool for visual data mining colocation patterns in multivariate bioimages. *Bioinformatics*, 28(8):1143–1150, 2012.

[95] J Ontrup and H Ritter. Large-scale data exploration with the hierarchically growing hyperbolic som. *Neural networks*, 19(6):751–761, 2006.

[96] AM Khan, SeA Raza, M Khan, and NM Rajpoot. Cell phenotyping in multi-tag fluorescent bioimages. *Neurocomputing*, 134:254–261, 2014.

[97] L Van der Maaten and G Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(2579-2605):85, 2008.

[98] N Uzunbajakava, A Lenferink, Y Kraan, E Volokhina, G Vrensen, J Greve, and C Otto. Nonresonant confocal raman imaging of dna and protein distribution in apoptotic cells. *Biophysical journal*, 84(6):3968–3981, 2003.

[99] GJ Puppels, FFM De Mul, C Otto, J Greve, M Robert-Nicoud, DJ Arndt-Jovin, and TM Jovin. Studying single living cells and chromosomes by confocal raman microspectroscopy. *Nature*, 347:301–303, 1990.

[100] BR Wood, B Tait, and D McNaughton. Micro-raman characterisation of the r to t state transition of haemoglobin within a single living erythrocyte. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1539(1):58–70, 2001.

[101] J Cheng, A Volkmer, LD Book, and XS Xie. Multiplex coherent anti-stokes raman scattering microspectroscopy and study of lipid vesicles. *The Journal of Physical Chemistry B*, 106(34):8493–8498, 2002.

[102] A Zumbusch, GR Holtom, and XS Xie. Three-dimensional vibrational imaging by coherent anti-stokes raman scattering. *Physical Review Letters*, 82(20): 4142–4145, 1999.

[103] M Diem, L Chiriboga, P Lasch, and A Pacifico. Ir spectra and ir spectral maps of individual normal and cancerous cells. *Biopolymers*, 67(4-5):349–353, 2002.

[104] P Lasch, A Pacifico, and M Diem. Spatially resolved ir microspectroscopy of single cells. *Biopolymers*, 67(4-5):335–338, 2002.

[105] C Matthäus, T Chernenko, JA Newmark, CM Warner, and M Diem. Label-free detection of mitochondrial distribution in cells by nonresonant raman microspectroscopy. *Biophysical journal*, 93(2):668–673, 2007.

[106] AT Harris, A Rennie, H Waqar-Uddin, SR Wheatley, SK Ghosh, DP Martin-Hirsch, SE Fisher, AS High, J Kirkham, and T Upile. Raman spectroscopy in head and neck cancer. *Head & neck oncology*, 2(1):26, 2010.

[107] JW Chan, DS Taylor, T Zwerdling, SM Lane, K Ihara, and T Huser. Micro-raman spectroscopy detects individual neoplastic and normal hematopoietic cells. *Biophysical journal*, 90(2):648–656, 2006.

[108] PRT Jess, DDW Smith, M Mazilu, K Dholakia, AC Riches, and CS Herrington. Early detection of cervical neoplasia by raman spectroscopy. *International journal of cancer*, 121(12):2723–2728, 2007.

[109] F Zheng, Y Qin, and K Chen. Sensitivity map of laser tweezers raman spectroscopy for single-cell analysis of colorectal cancer. *Journal of biomedical optics*, 12(3):034002, 2007.

[110] P Crow, N Stone, CA Kendall, JS Uff, JAM Farmer, H Barr, and MPJ Wright. The use of raman spectroscopy to identify and grade prostatic adenocarcinoma in vitro. *British journal of cancer*, 89(1):106–108, 2003.

[111] N Stone, C Kendall, N Shepherd, P Crow, and H Barr. Near-infrared raman spectroscopy for the classification of epithelial pre-cancers and cancers. *Journal of Raman spectroscopy*, 33(7):564–573, 2002.

[112] RE Kast, SC Tucker, K Killian, M Trexler, KV Honn, and GW Auner. Emerging technology: applications of raman spectroscopy for prostate cancer. *Cancer and Metastasis Reviews*, 33(2-3):673–693, 2014.

[113] A Zoladek, FC Pascut, P Patel, and I Notingher. Non-invasive time-course imaging of apoptotic cells by confocal raman micro-spectroscopy. *Journal of Raman Spectroscopy*, 42(3):251–258, 2011.

[114] P Chaurand, SA Schwartz, and RM Caprioli. Imaging mass spectrometry: a new tool to investigate the spatial organization of peptides and proteins in mammalian tissue sections. *Current opinion in chemical biology*, 6(5):676–681, 2002.

[115] A Walch, S Rauser, SO Deininger, and H Höfler. Maldi imaging mass spectrometry for direct tissue analysis: a new frontier for molecular histology. *Histochemistry and cell biology*, 130(3):421–434, 2008.

[116] RM Caprioli, TB Farmer, and J Gile. Molecular imaging of biological samples: localization of peptides and proteins using maldi-tof ms. *Analytical chemistry*, 69(23):4751–4760, 1997.

[117] EH Seeley and RM Caprioli. Imaging mass spectrometry: towards clinical diagnostics. *Proteomics-Clinical Applications*, 2(10-11):1435–1443, 2008.

[118] P Chaurand, SA Schwartz, D Billheimer, BJ Xu, A Crecelius, and RM Caprioli. Integrating histology and imaging mass spectrometry. *Analytical chemistry*, 76 (4):1145–1155, 2004.

[119] DE Palmer-Toy, DA Sarracino, D Sgroi, R LeVangie, and PE Leopold. Direct acquisition of matrix-assisted laser desorption/ionization time-of-flight mass spectra from laser capture microdissected tissues. *Clinical chemistry*, 46(9): 1513–1516, 2000.

[120] BJ Xu, RM Caprioli, ME Sanders, and RA Jensen. Direct analysis of laser capture microdissected cells by maldi mass spectrometry. *Journal of the American Society for Mass Spectrometry*, 13(11):1292–1297, 2002.

[121] N Masumori, TZ Thomas, P Chaurand, T Case, M Paul, S Kasper, RM Caprioli, T Tsukamoto, SB Shappell, and RJ Matusik. A probasin-large t antigen transgenic mouse line develops prostate adenocarcinoma and neuroendocrine carcinoma with metastatic potential. *Cancer Research*, 61(5):2239–2249, 2001.

[122] K Schwamborn, RC Krieg, M Reska, G Jakse, R Knuechel, and A Wellmann. Identifying prostate carcinoma by maldi-imaging. *International journal of molecular medicine*, 20(2):155–159, 2007.

[123] P Chaurand, BB DaGue, RS Pearsall, DW Threadgill, and RM Caprioli. Profiling proteins from azoxymethane-induced colon tumors at the molecular level by matrix-assisted laser desorption/ionization mass spectrometry. *Proteomics*, 1(10):1320–1326, 2001.

[124] K Yanagisawa, Y Shyr, BJ Xu, PP Massion, PH Larsen, BC White, JR Roberts, M Edgerton, A Gonzalez, S Nadaf, et al. Proteomic patterns of tumour subsets in non-small-cell lung cancer. *The Lancet*, 362(9382):433–439, 2003.

[125] LH Cazares, D Troyer, S Mendrinos, RA Lance, JO Nyalwidhe, HA Beydoun, MA Clements, RR Drake, and OJ Semmes. Imaging mass spectrometry of

a specific fragment of mitogen-activated protein kinase/extracellular signal-regulated kinase kinase kinase 2 discriminates cancer from uninvolved prostate tissue. *Clinical Cancer Research*, 15(17):5541–5551, 2009.

[126] R Lemaire, SA Menguellet, J Stauber, V Marchaudon, JP Lucot, P Collinet, MO Farine, D Vinatier, R Day, P Ducoroy, et al. Specific maldi imaging and profiling for biomarker hunting and validation: fragment of the 11s proteasome activator complex, reg alpha fragment, is a new potential ovary cancer biomarker. *Journal of proteome research*, 6(11):4127–4134, 2007.

[127] K Schwamborn. Imaging mass spectrometry in biomarker discovery and validation. *Journal of proteomics*, 75(16):4990–4998, 2012.

[128] A Thomas, NH Patterson, MM Marcinkiewicz, A Lazaris, P Metrakos, and P Chaurand. Histology-driven data mining of lipid signatures from multiple imaging mass spectrometry analyses: application to human colorectal cancer liver metastasis biopsies. *Analytical chemistry*, 85(5):2860–2866, 2013.

[129] SM Willems, A van Remoortere, R van Zeijl, A M Deelder, LA McDonnell, and PCW Hogendoorn. Imaging mass spectrometry of myxoid sarcomas identifies proteins and lipids specific to tumour type and grade, and reveals biochemical intratumour heterogeneity. *The Journal of pathology*, 222(4):400–409, 2010.

[130] SO Deininger, MP Ebert, A Fütterer, M Gerhard, and C Röcken. Maldi imaging combined with hierarchical clustering as a new tool for the interpretation of complex human cancers. *Journal of proteome research*, 7(12):5230–5236, 2008.

[131] Atdbio. Analytical methods, 2014. URL `http://www.atdbio.com/content/8/Analytical-methods`.

[132] RM Levenson and JM Beechem. Practical multispectral fluorescence imaging. *Practical Fluorescence Imaging*, pages 35–62, 2004.

[133] MB Bouchard, SA MacLaurin, PJ Dwyer, J Mansfield, R Levenson, and T Krucker. Technical considerations in longitudinal multispectral small animal molecular imaging. *Journal of biomedical optics*, 12(5):051601, 2007.

[134] JR Mansfield, C Hoyt, and RM Levenson. Visualization of microscopy-based spectral imaging data from multi-label tissue sections. *Current Protocols in Molecular Biology*, 14:14–19, 2008.

[135] ME Dickinson, G Bearman, S Tille, R Lansford, and SE Fraser. Multi-spectral imaging and linear unmixing add a whole new dimension to laser scanning fluorescence microscopy. *Biotechniques*, 31(6):1272–1274, 2002.

[136] JR Mansfield, KW Gossage, CC Hoyt, and RM Levenson. Autofluorescence removal, multiplexing, and automated analysis methods for in-vivo fluorescence imaging. *Journal of biomedical optics*, 10(4):041207–041207, 2005.

[137] M Teverovskiy, Y Vengrenyuk, A Tabesh, M Sapir, S Fogarasi, HY Pang, FM Khan, S Hamann, P Capodieci, M Clayton, et al. Automated localization and quantification of protein multiplexes via multispectral fluorescence imaging. In *Biomedical Imaging: From Nano to Macro, 2008. ISBI 2008. 5th IEEE International Symposium on*, pages 300–303. IEEE, 2008.

[138] WK Hendrickson, R Flavin, JL Kasperzyk, M Fiorentino, F Fang, R Lis, C Fiore, KL Penney, J Ma, PW Kantoff, et al. Vitamin d receptor protein expression in tumor tissue and prostate cancer progression. *Journal of clinical oncology*, 29(17):2378–2385, 2011.

[139] MI Toma, K Friedrich, W Meyer, M Fröhner, S Schneider, M Wirth, and GB Baretton. Correlation of centrosomal aberrations with cell differentiation and dna ploidy in prostate cancer. *Anal Quant Cytol Histol*, 32(1):1–10, 2010.

[140] A Kennedy, H Dong, D Chen, and WT Chen. Elevation of seprase expression and promotion of an invasive phenotype by collagenous matrices in ovarian tumor cells. *International Journal of Cancer*, 124(1):27–35, 2009.

[141] KJ Kimball, AA Rivera, KR Zinn, M Icyuz, V Saini, J Li, ZB Zhu, GP Siegal, JT Douglas, DT Curiel, et al. Novel infectivity-enhanced oncolytic adenovirus with a capsid-incorporated dual-imaging moiety for monitoring virotherapy in ovarian cancer. *Molecular imaging*, 8(5):264, 2009.

[142] JM Krüger, M Thomas, R Korn, G Dietmann, C Rutz, G Brockhoff, K Specht, M Hasmann, and F Feuerhake. Detection of truncated her2 forms in formalin-fixed, paraffin-embedded breast cancer tissue captures heterogeneity and is not affected by her2-targeted therapy. *The American journal of pathology*, 183(2): 336–343, 2013.

[143] D Taylor, CL Pearce, L Hovanessian-Larsen, S Downey, DV Spicer, S Bartow, MC Pike, AH Wu, and D Hawes. Progesterone and estrogen receptors in pregnant and premenopausal non-pregnant normal human breast. *Breast cancer research and treatment*, 118(1):161–168, 2009.

[144] TK Rhee, JY Young, AC Larson, GK Haines, KT Sato, R Salem, MF Mulcahy, LM Kulik, T Paunesku, GE Woloschak, et al. Effect of transcatheter arterial embolization on levels of hypoxia-inducible factor-1α in rabbit vx2 liver tumors. *Journal of Vascular and Interventional Radiology*, 18(5):639–645, 2007.

[145] S Virmani, TK Rhee, RK Ryu, KT Sato, RJ Lewandowski, MF Mulcahy, LM Kulik, B Szolc-Kowalska, GE Woloschak, GY Yang, et al. Comparison of hypoxia-inducible factor-1α expression before and after transcatheter arterial embolization in rabbit vx2 liver tumors. *Journal of Vascular and Interventional Radiology*, 19(10):1483–1489, 2008.

[146] RJ Lewandowski, AC Eifler, DJ Bentrem, JC Chung, D Wang, GE Woloschak, GY Yang, R Ryu, R Salem, AC Larson, et al. Functional magnetic resonance imaging in an animal model of pancreatic cancer. *World journal of gastroenterology: WJG*, 16(26):3292–3298, 2010.

[147] WW Tseng, D Winer, JA Kenkel, O Choi, AH Shain, JR Pollack, R French, AM Lowy, and EG Engleman. Development of an orthotopic model of invasive pancreatic cancer in an immunocompetent murine host. *Clinical Cancer Research*, 16(14):3684–3695, 2010.

[148] P Qiu, EF Simonds, SC Bendall, KD Gibbs Jr, RV Bruggner, MD Linderman, K Sachs, GP Nolan, and SK Plevritis. Extracting a cellular hierarchy from high-dimensional cytometry data with spade. *Nature biotechnology*, 29(10): 886–891, 2011.

[149] AM Grigoryan, G Hostetter, O Kallioniemi, and ER Dougherty. Simulation toolbox for 3d-fish spot-counting algorithms. *Real-Time Imaging*, 8(3):203–212, 2002.

[150] EMM Manders, R Hoebe, J Strackee, AM Vossepoel, and JA Aten. Largest contour segmentation: a tool for the localization of spots in confocal images. *Cytometry*, 23(1):15–21, 1996.

[151] SJ Lockett, D Sudar, CT Thompson, D Pinkel, and JW Gray. Efficient, interactive, and three-dimensional segmentation of cell nuclei in thick tissue sections. *Cytometry*, 31(4):275–286, 1998.

[152] F Graner and JA Glazier. Simulation of biological cell sorting using a two-dimensional extended potts model. *Physical review letters*, 69(13):2013, 1992.

[153] A Lehmussola, P Ruusuvuori, J Selinummi, H Huttunen, and O Yli-Harja. Computational framework for simulating fluorescence microscope images with cell populations. *Medical Imaging, IEEE Transactions on*, 26(7):1010–1016, 2007.

[154] D Svoboda, M Kozubek, and S Stejskal. Generation of digital phantoms of cell nuclei and simulation of image formation in 3d image cytometry. *Cytometry part A*, 75(6):494–509, 2009.

[155] D Svoboda and V Ulman. Towards a realistic distribution of cells in synthetically generated 3d cell populations. In *Image Analysis and Processing–ICIAP 2013*, pages 429–438. Springer, 2013.

[156] D Svoboda, O Homola, and S Stejskal. Generation of 3d digital phantoms of colon tissue. *Image Analysis and Recognition*, 6754:31–39, 2011.

[157] D Svoboda and V Ulman. Generation of synthetic image datasets for time-lapse fluorescence microscopy. In *Image Analysis and Recognition*, volume 7325, pages 473–482. Springer, 2012.

[158] D Svoboda, V Ulman, and I Peterlík. On proper simulation of chromatin structure in static images as well as in time-lapse sequences in fluorescence microscopy. In *Proceedings of 2015 IEEE International Symposium on Biomedical Imaging*, pages 712–716. Engineering in Medicine and Biology Society, 2015. ISBN 978-1-4799-2375-5.

[159] S Rajaram, B Pavie, NEF Hac, SJ Altschuler, and LF Wu. Simucell: a flexible framework for creating synthetic microscopy images. *Nature methods*, 9(7): 634–635, 2012.

[160] P Malm, A Brun, and E Bengtsson. Papsynth: simulated bright-field images of cervical smears. In *Biomedical Imaging: From Nano to Macro, 2010 IEEE International Symposium on*, pages 117–120. IEEE, 2010.

[161] T Peng, W Wang, GK Rohde, and RF Murphy. Instance-based generative biological shape modeling. In *Biomedical Imaging: From Nano to Macro, 2009. ISBI'09. IEEE International Symposium on*, pages 690–693. IEEE, 2009.

[162] T Peng and RF Murphy. Image-derived, three-dimensional generative models of cellular organization. *Cytometry Part A*, 79(5):383–391, 2011.

[163] A Shariff, RF Murphy, and GK Rohde. A generative model of microtubule distributions, and indirect estimation of its parameters from fluorescence microscopy images. *Cytometry Part A*, 77(5):457–466, 2010.

[164] T Zhao and RF Murphy. Automated learning of generative models for subcellular location: building blocks for systems biology. *Cytometry Part A*, 71(12): 978–990, 2007.

[165] TE Buck, J Li, GK Rohde, and RF Murphy. Toward the virtual cell: Automated approaches to building models of subcellular organization learned from microscopy images. *Bioessays*, 34(9):791–799, 2012.

[166] A Khan, E Aylward, P Barta, M Miller, and MF Beg. Semi-automated basal ganglia segmentation using large deformation diffeomorphic metric mapping. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2005*, pages 238–245. Springer, 2005.

[167] GK Rohde, AJS Ribeiro, KN Dahl, and RF Murphy. Deformation-based nuclear morphometry: Capturing nuclear shape variation in hela cells. *Cytometry Part A*, 73(4):341–350, 2008.

[168] WK Pratt. Superposition and convolution. *Digital Image Processing: PIKS Scientific Inside, Fourth Edition*, pages 165–187, 1991.

[169] Y Ma, M Kamber, and AC Evans. 3d simulation of pet brain images using segmented mri data and positron tomograph characteristics. *Computerized medical imaging and graphics*, 17(4):365–371, 1993.

[170] COD Solorzano, EG Rodriguez, A Jones, D Pinkel, JW Gray, D Sudar, and SJ Lockett. Segmentation of confocal microscope images of cell nuclei in thick tissue sections. *Journal of Microscopy*, 193(3):212–226, 1999.

[171] David Svoboda, Marek Kašík, Martin Maška, Jan Hubenỳ, Stanislav Stejskal, and Michal Zimmermann. On simulating 3d fluorescent microscope images. In *Computer analysis of images and patterns*, pages 309–316. Springer, 2007.

[172] M Kozubek and P Matula. An efficient algorithm for measurement and correction of chromatic aberrations in fluorescence microscopy. *Journal of Microscopy*, 200(3):206–217, 2000.

[173] M Born and E Wolf. *Principles of optics: electromagnetic theory of propagation, interference and diffraction of light*. Cambridge university press, 1999.

[174] M Faust and M Montenarh. Subcellular localization of protein kinase ck2. *Cell and tissue research*, 301(3):329–340, 2000.

[175] GG Chung, E Provost, EP Kielhorn, LA Charette, BL Smith, and DL Rimm. Tissue microarray analysis of β-catenin in colorectal cancer shows nuclear phospho-β-catenin is associated with a better prognosis. *Clinical Cancer Research*, 7(12):4013–4020, 2001.

[176] L Lessard, PI Karakiewicz, P Bellon-Gagnon, M Alam-Fahmy, HA Ismail, AM Mes-Masson, and F Saad. Nuclear localization of nuclear factor-κb p65 in primary prostate tumors is highly predictive of pelvic lymph node metastases. *Clinical cancer research*, 12(19):5741–5745, 2006.

[177] ML Blinov, JR Faeder, B Goldstein, and WS Hlavacek. Bionetgen: software for rule-based modeling of signal transduction based on the interactions of molecular domains. *Bioinformatics*, 20(17):3289–3291, 2004.

[178] DT Gillespie. Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry*, 81(25):2340–2361, 1977.

[179] L Cai, N Friedman, and XS Xie. Stochastic protein expression in individual cells at the single molecule level. *Nature*, 440(7082):358–362, 2006.

[180] O Feinerman, J Veiga, JR Dorfman, RN Germain, and G Altan-Bonnet. Variability and robustness in t cell activation from regulated heterogeneity in protein levels. *Science*, 321(5892):1081–1084, 2008.

[181] P Rangamani, A Lipshtat, EU Azeloglu, RC Calizo, M Hu, S Ghassemi, J Hone, S Scarlata, SR Neves, and R Iyengar. Decoding information in cell shape. *Cell*, 154(6):1356–1369, 2013.

[182] LM Loew and JC Schaff. The virtual cell: a software environment for computational cell biology. *TRENDS in Biotechnology*, 19(10):401–406, 2001.

[183] SS Andrews, NaJ Addy, R Brent, and AP Arkin. Detailed simulations of cell biology with smoldyn 2.1. *PLoS Comput Biol*, 6(3):e1000705, 2010.

[184] SS Andrews. Spatial and stochastic cellular modeling with the smoldyn simulator. In *Bacterial Molecular Networks*, pages 519–542. Springer, 2012.

[185] RA Kerr, TM Bartol, B Kaminsky, M Dittrich, JCJ Chang, SB Baden, TJ Sejnowski, and JR Stiles. Fast monte carlo simulation methods for biological

reaction-diffusion systems in solution and on surfaces. *SIAM Journal on Scientific Computing*, 30(6):3126–3149, 2008.

[186] J Li, JY Newberg, M Uhlén, E Lundberg, and RF Murphy. Automated analysis and reannotation of subcellular locations in confocal images from the human protein atlas. *PLoS ONE*, 7(11):e50514, 2012.

[187] AM Khan, A Humayun, SeA Raza, M Khan, and NM Rajpoot. A novel paradigm for mining cell phenotypes in multi-tag bioimages using a locality preserving nonlinear embedding. In *Neural Information Processing*, pages 575–583. Springer, 2012.

[188] Y Al-Kofahi, W Lassoued, W Lee, and B Roysam. Improved automatic detection and segmentation of cell nuclei in histopathology images. *Biomedical Engineering, IEEE Transactions on*, 57(4):841–852, 2010.

[189] A Huertas and G Medioni. Detection of intensity changes with subpixel accuracy using laplacian-gaussian masks. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 8(5):651–664, 1986.

[190] DN Reshef, YA Reshef, HK Finucane, SR Grossman, G McVean, PJ Turnbaugh, ES Lander, M Mitzenmacher, and PC Sabeti. Detecting novel associations in large data sets. *science*, 334(6062):1518–1524, 2011.

[191] GJ Székely and ML Rizzo. Brownian distance covariance. *The annals of applied statistics*, 3(4):1236–1265, 2009.

[192] BJ Frey and D Dueck. Clustering by passing messages between data points. *science*, 315(5814):972–976, 2007.

[193] AK Jain, MN Murty, and PJ Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.

[194] K Sirinukunwattana, RS Savage, MF Bari, D Snead, and NM Rajpoot. Bayesian hierarchical clustering for studying cancer gene expression data with unknown statistics. *PLoS ONE*, 8(10):e75748, 2013.

[195] VN Kovacheva, K Sirinukunwattana, and NM Rajpoot. A bayesian framework for cell-level protein network analysis for multivariate proteomics image data. In *SPIE Medical Imaging*, pages 904110–904110. International Society for Optics and Photonics, 2014.

[196] JH Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

[197] F Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83, 1945.

[198] A Kamburov, C Wierling, H Lehrach, and R Herwig. ConsensuspathdbŮa database for integrating human functional interaction networks. *Nucleic Acids Research*, 37(suppl 1):D623–D628, 2009.

[199] H Zalzali, C Naudin, P Bastide, C Quittau-Prevostel, C Yaghi, F Poulat, Ph Jay, and Ph Blache. Ceacam1, a sox9 direct transcriptional target identified in the colon epithelium. *Oncogene*, 27(56):7131–7138, 2008.

[200] C Darido, M Buchert, J Pannequin, P Bastide, H Zalzali, T Mantamadiotis, JF Bourgaux, V Garambois, P Jay, P Blache, et al. Defective claudin-7 regulation by tcf-4 and sox-9 disrupts the polarity and increases the tumorigenicity of colorectal cancer cells. *Cancer Research*, 68(11):4258–4268, 2008.

[201] S Kuhn, M Koch, T Nübel, M Ladwein, D Antolovic, P Klingbeil, D Hildebrand, G Moldenhauer, L Langbein, WW Franke, et al. A complex of epcam, claudin-7, cd44 variant isoforms, and tetraspanins promotes colorectal cancer progression. *Molecular Cancer Research*, 5(6):553–567, 2007.

[202] D Drasdo and M Loeffler. Individual-based models to growth and folding in one-layered tissues: intestinal crypts and early development. *Nonlinear Analysis: Theory, Methods & Applications*, 47(1):245–256, 2001.

[203] A Efros and TK Leung. Texture synthesis by non-parametric sampling. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 1033–1038. IEEE, 1999.

[204] F Aurenhammer. Voronoi diagramsa survey of a fundamental geometric data structure. *ACM Computing Surveys (CSUR)*, 23(3):345–405, 1991.

[205] K Perlin. An image synthesizer. *ACM Siggraph Computer Graphics*, 19(3): 287–296, 1985.

[206] D Svoboda, V Ulman, L Matyska, M Maska, J Bella, and S Stejskal. On proper simulation of phenomena influencing image formation in fluorescence microscopy. In *Image Processing (ICIP), 2014 IEEE International Conference on*, pages 3944–3948. IEEE, 2014.

[207] AC Ruifrok and DA Johnston. Quantification of histochemical staining by color deconvolution. *Analytical and quantitative cytology and histology/the International Academy of Cytology [and] American Society of Cytology*, 23(4): 291–299, 2001.

[208] MD Abràmoff, PJ Magalhães, and SJ Ram. Image processing with imagej. *Biophotonics international*, 11(7):36–43, 2004.

[209] AE Carpenter, TR Jones, MR Lamprecht, C Clarke, IH Kang, O Friman, DA Guertin, JH Chang, RA Lindquist, J Moffat, et al. Cellprofiler: image analysis software for identifying and quantifying cell phenotypes. *Genome biology*, 7(10):R100, 2006.

[210] J Byun, MR Verardo, B Sumengen, GP Lewis, BS Manjunath, and SK Fisher. Automated tool for the detection of cell nuclei in digital microscopic images: application to retinal images. *Mol Vis*, 12:949–960, 2006.

[211] N Trahearn, D Snead, I Cree, and NM Rajpoot. Multi-class stain separation using independent component analysis. In *SPIE Medical Imaging*, pages 94200J–94200J. International Society for Optics and Photonics, 2015.

[212] L Barbe, E Lundberg, P Oksvold, A Stenius, E Lewin, E Björling, A Asplund, F Pontén, H Brismar, M Uhlén, and H Andersson-Svahn. Toward a confocal subcellular atlas of the human proteome. *Molecular & cellular proteomics*, 7 (3):499–508, 2008.

[213] JY Newberg, J Li, A Rao, F Pontén, M Uhlén, E Lundberg, and RF Murphy. Automated analysis of human protein atlas immunofluorescence images. In *Biomedical Imaging: From Nano to Macro, 2009. ISBI'09. IEEE International Symposium on*, pages 1023–1026. IEEE, 2009.