

The Automated Analysis and Statistical Characterisation of Subcellular Biological Images

Simon W. Li

Linacre College

Trinity 2011

Institute of Biomedical Engineering
Department of Engineering Science
University of Oxford

Supervisors:
Prof. J. Alison Noble
Dr. James G. Wakefield

The Automated Analysis and Statistical Characterisation of Subcellular Biological Images

Simon W. Li, Linacre College, Oxford
Submitted for the degree of Doctor of Philosophy, Engineering Science
Trinity 2011

Biomedical research has been revolutionised by the development of high throughput systems which enable thousands of experiments to be carried out simultaneously, leading to the generation of vast amounts of microscopy data. At the same time smaller scale experiments are increasingly dependent on accurate quantification of relatively small changes since the intricate network of interactions within a cell often counteracts attempts to perturb the function of any single protein or pathway. This has motivated the development of algorithms and software packages to assist with the analysis of biological imaging data.

This thesis develops and applies image analysis tools to multiple biological problems, and also illustrates the whole process required to label live samples so they can be imaged in the first place, thus giving an appreciation of the practical difficulties of biomedical research. We begin with the development of a kymograph tool, which is used for the spatio-temporal subcellular quantification of the chromosomal passenger complex, a complex of proteins involved in regulating mitosis. Following this we start to consider inter-cellular signalling interactions in a study of the IGF2 pathway in clumped cells. A random forest classification framework using cell-based image features is developed, and shown to be able to distinguish between cells with different levels of activation. Importantly the random forest feature importance measures are shown to be useful in helping to identify biologically relevant differences. Finally we apply the random forest classifier to the analysis of immunofluoresence images of Ewing's sarcoma tumour biopsies, which are associated with clinical outcome data, allowing us to develop a framework for integrating imaging data with patient prognosis data. The samples originate from multiple sources, making this a very challenging data set due to a lack of normalisation, but we are still able to obtain clinically useful results, in addition to identifying areas where further research is required.

This thesis is submitted to the Department of Engineering Science, University of Oxford, in partial fulfillment of the requirements for the degree of Doctor of Philosophy. The thesis is entirely my own work and, except where otherwise stated, describes my own research.

Simon W. Li, Linacre College

Acknowledgements

Completing my DPhil thesis would have been impossible without the help and support of many people.

Firstly my supervisors, Professor Alison Noble and Dr. James Wakefield, who guided me throughout my research whilst allowing me the freedom to work on an interdisciplinary project, and my unofficial supervisor Professor Bass Hassan, who was instrumental in setting the final direction of my thesis.

I would also like to thank my friends and colleagues in all departments I have worked with, including Claudia Buehnemann and Haiyue Yu who spent many long hours collecting and processing images, Tommy Duncan and Pete Jones for their help with experiments, Boguslaw Obara for providing a refreshing outlook on the field of bioimaging, Sana Fathima for her help in life, Luke Beaumont for leading the way, and Alan Wainman for his knowledge of the important and not so important things in life.

It has been a privilege to be part of the Life Sciences Interface Doctoral Training Centre, and I am grateful to the EPSRC for funding me throughout my studies. The actual writing in my thesis was mostly done in G&D's in Cowley and The Boston Tea Party in Exeter, so I thank them for providing excellent coffee, food and free wifi.

Last but not least, I thank my parents for supporting me throughout my studies.

Contents

1 Introduction **1**
- 1.1 Purpose of this thesis . . . 2
- 1.2 Thesis outline . . . 2

2 Background **4**
- 2.1 Quantitative biology . . . 4
  - 2.1.1 High throughput screening . . . 5
  - 2.1.2 High content analysis . . . 6
  - 2.1.3 Machine learning and cell biology . . . 6
  - 2.1.4 Visualisation techniques . . . 8
- 2.2 Histology studies . . . 9
  - 2.2.1 Tissue microarrays . . . 9
- 2.3 Segmentation . . . 10
  - 2.3.1 Watershed related methods . . . 11
  - 2.3.2 Snakes and level sets . . . 11
  - 2.3.3 Region based level sets . . . 12
  - 2.3.4 Multiregion segmentation . . . 13
- 2.4 Graph based segmentation . . . 13
- 2.5 Comparison of cell segmentation methods . . . 15
- 2.6 Tracking . . . 16
  - 2.6.1 Contour evolution . . . 16
  - 2.6.2 Segment and associate . . . 16
  - 2.6.3 Bayesian methods . . . 17
- 2.7 Random forests . . . 17
  - 2.7.1 Classification and regression trees . . . 18
  - 2.7.2 Feature relevance . . . 19
  - 2.7.3 Unsupervised clustering . . . 20
  - 2.7.4 Variations of random forests . . . 21
  - 2.7.5 Comparison with other methods . . . 21
- 2.8 Fluorescence imaging . . . 21
  - 2.8.1 Principles of fluorescence microscopy . . . 21
  - 2.8.2 Optical sectioning . . . 23
  - 2.8.3 Green fluorescent protein . . . 23

2.9 Heterogeneity in cell populations . . . 25
2.9.1 Non-genetic variation . . . 25
2.9.2 The consequences of cell heterogeneity . . . 25
2.9.3 Heterogeneity in cancer . . . 26
2.9.4 Inter-cell communication . . . 26
2.9.5 Single cell measurements . . . 27
2.10 Mitosis and cytokinesis . . . 28
2.10.1 The cell cycle . . . 29
2.10.2 Mitosis (M) and cytokinesis . . . 30
2.11 *Drosophila melanogaster* as a model organism . . . 32
2.12 The Chromosomal passenger complex . . . 32
2.12.1 Overview of the complex . . . 33
2.12.2 Aurora B (Ipl1-aurora-like-kinase) . . . 33
2.12.3 Incenp . . . 35
2.12.4 Borealin . . . 36
2.12.5 Deterin/Survivin . . . 37
2.12.6 Putative CPC members . . . 37
2.12.7 Interactions with Polo . . . 37
2.12.8 Central spindle . . . 38

3 Chromosomal passenger complex experiments 39
3.1 Collaborations . . . 39
3.2 Overview . . . 40
3.3 Gene sequences . . . 40
3.4 Entry vectors . . . 41
3.4.1 Polymerase chain reaction . . . 41
3.4.2 TOPO cloning . . . 42
3.4.3 Sequencing . . . 43
3.5 Expression vectors . . . 45
3.5.1 LR reaction . . . 45
3.6 Fly crosses . . . 46
3.7 Biochemistry . . . 48
3.7.1 Preparation of tissues for biochemistry . . . 48
3.7.2 Western blotting . . . 49
3.8 Imaging . . . 50
3.8.1 Embryo fixation . . . 50
3.8.2 Live embryo preparation . . . 51
3.8.3 Microscopy . . . 51

4 Quantitative analysis of disruptions to the chromosomal passenger complex 54
4.1 Initial results . . . 54
4.1.1 GFP expression . . . 54

4.1.2 Live imaging . . . 55
4.1.3 Fixed images . . . 56
4.1.4 Micro-injections . . . 60
4.2 Motivation for further analysis . . . 63
4.3 The analysis tools . . . 65
4.3.1 Terminology . . . 65
4.3.2 Detection . . . 66
4.3.3 Segmentation . . . 67
4.3.4 Alignment of nuclei . . . 69
4.3.5 Kymographs . . . 71
4.3.6 Run times . . . 72
4.4 Anti-Borealin microinjections . . . 73
4.5 Coiled-coil microinjections . . . 75
4.5.1 Possible effects of the coiled-coil peptides . . . 76
4.6 Conclusions . . . 77
4.7 Improvements . . . 77

5 Clumped cell analysis **79**
5.1 Introduction . . . 79
5.1.1 Division of work . . . 81
5.2 Biological background to this chapter . . . 81
5.2.1 The IGF signalling pathways . . . 82
5.2.2 Signalling Markers . . . 83
5.3 Description of the data . . . 84
5.3.1 Experiment 1: time dose . . . 85
5.3.2 Experiment 2: Cell clumps . . . 87
5.4 Cell segmentation methods . . . 90
5.4.1 Nuclei segmentation . . . 90
5.4.2 Clumped cell segmentation . . . 92
5.5 Segmentation of clumped cells . . . 93
5.6 Nuclei segmentation (DAPI) . . . 94
5.6.1 Detection . . . 94
5.6.2 Segmentation . . . 95
5.6.3 Clump splitting . . . 97
5.7 Cytoplasm segmentation (phalloidin) . . . 99
5.7.1 External clump boundary . . . 99
5.7.2 Boundary enhancement . . . 100
5.7.3 Multi-region level set segmentation . . . 102
5.7.4 Local Minima Avoidance . . . 104
5.7.5 Random walker segmentation . . . 105
5.7.6 Graph colouring . . . 107
5.7.7 Manual comparison of all methods . . . 109
5.7.8 Run times . . . 116

5.8 Segmentation conclusions . . . 116
5.9 Analysis of cells . . . 118
5.9.1 Feature extraction . . . 118
5.9.2 Image features . . . 120
5.9.3 Random forests classification of cells . . . 121
5.10 Feature importance . . . 124
5.10.1 Time-dose classification . . . 126
5.11 Classification conclusions . . . 128
5.12 Spatial analysis . . . 133
5.12.1 Analysis of spatial interactions . . . 133
5.12.2 Geographic measures of spatial autocorrelation . . . 134
5.12.3 Clump sizes . . . 137
5.12.4 Cell distances . . . 139
5.12.5 Moran's I: Egr1 and Foxo3a . . . 140
5.12.6 Semi-variogram: Egr1 and Foxo3a . . . 141
5.12.7 Spatial analysis conclusions . . . 144

6 Survival analysis of Ewing's sarcoma tissue microarrays 147
6.1 Ewing's sarcoma . . . 147
6.2 Hypothesis . . . 148
6.3 Collaborations . . . 149
6.4 Overview of the datasets . . . 149
6.4.1 Variation within the data . . . 150
6.5 Image features . . . 150
6.5.1 Derived image features . . . 153
6.5.2 Summary features . . . 154
6.6 Survival analysis . . . 155
6.6.1 Survival times and censoring . . . 155
6.6.2 Survival analysis methods . . . 156
6.7 Image normalisation . . . 160
6.7.1 Comparison of feature values in each dataset . . . 160
6.8 Distribution of outcomes . . . 162
6.9 Ki67 biomarker . . . 165
6.9.1 Identification of Ki67 positive cells . . . 168
6.9.2 Manual grouping of images . . . 170
6.9.3 Conclusions . . . 173
6.10 Random forest classification . . . 174
6.10.1 Unsupervised clustering . . . 174
6.10.2 Supervised classification . . . 175
6.10.3 Metastasis classification . . . 177
6.10.4 Feature importance . . . 182
6.10.5 Parameter selection . . . 187
6.10.6 Random survival forests . . . 190

6.11 Conclusions 196

7 Conclusions and future work **198**
7.1 Future work 199
7.1.1 Spatio-temporal subcellular protein localisation 199
7.1.2 Spatially-correlated interactions between cells 200
7.1.3 Machine learning framework for partially normalised data 201

A Segmentation and analysis **202**
A.1 Fast level set algorithm 202
A.2 Solution of the random walker algorithm 203
A.3 Moran's index 205

B Supervised classification **206**
B.1 Confusion matrices 206
B.2 Harrell's concordance index 207
B.3 Feature importance characteristics of random forests 208
B.3.1 Regression: synthetic data 208
B.3.2 Classification: Synthetic data 211

C Protein Alignments **213**
C.1 Aurora B, CG6620 213
C.2 Borealin, CG4454 214
C.3 Survivin, CG12265 214
C.4 Incenp, CG12165 215

D MCF7 experimental work **216**
D.1 MCF7 fixing and staining 216

E TMA data preparation **217**
E.1 TMA experimental protocols 217
E.2 TMA Image acquisition and analysis 218
E.3 Microscope bias field 218
E.4 Feature naming convention 219

References **222**

List of Figures

2.1 The basic principles of fluoresence 22
2.2 Principles behind a confocal microscope 24
2.3 The eukaryotic cell cycle 29
2.4 Fluorescent images of mitosis in a *Drosophila* embryo 30

3.1 Gateway entry and LR reactions for *deterin* 47
3.2 Fly crosses 48
3.3 Microscope slide for live imaging of *Drosophila* embryos. 51

4.1 Western blot showing expression of GFP tagged CPC proteins 55
4.2 Deterin-GFP during mitosis in a *Drosophila* syncytial embryo 57
4.3 Closeup of the localisation of Aurora B-GFP during mitosis 58
4.4 Localisation of Aurora B-GFP, Borealin-GFP and Deterin-GFP 59
4.5 Aurora B-GFP metaphase localisation in fixed embryos 60
4.6 Incenp-GFP localisation in fixed embryos 61
4.7 Anti-Borealin micro-injected into an embryo expressing Aurora B-GFP . . 62
4.8 Predicted coiled-oil domain injected into an Aurora B-GFP embryo 63
4.9 Predicted coiled-coil domain injected into dual labelled embryos 64
4.10 Flowchart of the CPC segmentation procedure. 65
4.11 Nuclei detection, segmentation and alignment 70
4.12 A kymograph created from one nucleus 72
4.13 Aurora B-GFP spindle intensity in anti-Borealin injected embryos 74
4.14 Time differences in anaphase transition in PredCC injected embryos 76

5.1 An overview of the insulin-like growth factor 1 receptor pathway 83
5.2 An example image of a MCF-7 cell stained with three markers 87
5.3 MCF-7 cells under different conditions with Egr1 labelled in red. 87
5.4 MCF-7 cells under different conditions with Foxo3a labelled in red. 88
5.5 MCF-7 cells under different conditions with pAkt labelled in red. 88
5.6 MCF-7 cells under different conditions with pS6 labelled in red. 89
5.7 Overview of the segmentation process. 94
5.8 Main steps of the nuclei segmentation algorithm 96
5.9 Measurements used to find the optimal split between two nuclei. 98
5.10 External cell boundaries tend to fade into the background 100

5.11 The ridge filter responses and final potential field . . . 103
5.12 Schematic of a random walker binary segmentation . . . 106
5.13 Four colours may be insufficient to colour an evolving graph . . . 108
5.14 Example cytoplasm segmentation overlaid on the original image . . . 109
5.15 An example image showing the β-catenin staining of the membrane . . . 110
5.16 Comparison of segmentations for different algorithm parameters . . . 113
5.17 Comparison of segmentations for selected parameter values . . . 115
5.18 Probabilities from the random walker segmentation . . . 116
5.19 CDF plot of the ratio of the mean nuclear to cytoplasmic EGR1 levels . . . 123
5.20 CDF plots of the nuclear size for multiple concentrations and times . . . 131
5.21 Example definition of an interaction weight . . . 136
5.22 Moran's I calculated on an artificial network of cells . . . 138
5.23 Semivariogram for an artificial network of cells . . . 138
5.24 Distribution of clump sizes for the Egr1 and Foxo3a stained images . . . 139
5.25 Example of a cell-cell distance measure . . . 140
5.26 Moran's I calculated for mean Egr1 or Foxo3a levels . . . 142
5.27 Moran's I calculated for the area of each nucleus . . . 143
5.28 Semi-variograms of the Egr1 and Foxo3a mean intensity . . . 145

6.1 An example of a TMA confocal image . . . 153
6.2 Histograms of cell features in the Egr1 labelled images . . . 163
6.3 Density plots of the principal components of cell features . . . 164
6.4 Scatter plots of the first two principal components for all images . . . 165
6.5 Scatter plot of the principal components of the patient features . . . 166
6.6 Distributions of the main patient survival outcomes . . . 167
6.7 Kaplan-Meier survival functions for the three datasets . . . 168
6.8 Density plots of the mean CD99 and Ki67 nucleus and cytoplasm levels . . . 171
6.9 Determination of the optimal threshold for defining CD99+ and Ki67+ cells 172
6.10 Unsupervised clustering of all patients . . . 176
6.11 Results of the random forest metastasis classification . . . 179
6.12 The random forest out-of-bag error estimates . . . 181
6.13 MDS plots of the proximity matrix for the metastasis classifiers . . . 182
6.14 Feature importances for the combined dataset metastasis classifier. . . . 184
6.15 Feature importances for the metastasis classifiers . . . 185
6.16 Distributions of three potentially informative features . . . 186
6.17 EFS feature importance in the augmented metastasis classifier . . . 187
6.18 Random forest OOB error rates as a function of the number of trees . . . 188
6.19 Random forest error rates against feature reduction iterations . . . 189
6.20 Error rates for each of the random survival forest classifiers . . . 192
6.21 Feature importances for the RSF using variable hunting . . . 193
6.22 Feature importances for the all cell Düsseldorf classifier. . . . 194
6.23 Kaplan Meier plots of two features on the Düsseldorf dataset . . . 194
6.24 Predicted relative mortality as a function of selected features . . . 195

6.25 Classifier error and feature importance for the augmented dataset 196

A.1 The fast level set representation of an image. 203

B.1 Contribution of the first six synthetic variables 209
B.2 The random forest regression with different numbers of samples 210
B.3 Feature importance with different numbers of samples 212

E.1 Average intensity images and segmentations for each channel 220

List of Tables

3.1 CPC genes in *Drosophila melanogaster* 41
3.2 Primers used to amplify the four genes to be cloned 42
3.3 Primers used for the sequencing reactions 43
3.4 Mutations found in the sequence of the chosen entry clones 44

4.1 Summary of parameters used for detection and segmentation 72

5.1 Mean dice similarity coefficients for each image 112
5.2 Mean Hausdorff distances for each image 114
5.3 Dice similarity coefficients for the random walker segmentation 115
5.4 Run-times for the cytoplasm segmentation. 117
5.5 Classification of whole images of Egr1 or Foxo3a stained cells 124
5.6 Classification of individual Egr1 or Foxo3a stained cells 125
5.7 20 most important features for classifying Egr1 stained cells 126
5.8 20 most important features for classifying Foxo3a stained cells 127
5.9 Automatic time point classification of whole images 129
5.10 Automatic dose classification of whole images 130

6.1 Number of patients with known survival outcome 150
6.2 Hierarchy of principal sources of variation 151
6.3 Numbers of images and cells in each dataset 152
6.4 Number of patients alive/dead at overall and event free survival 166
6.5 Results of the random forest metastasis classification 180
6.6 Confusion matrices for the metastasis classifiers 180
6.7 Significance values from a two-sided Kolmogorov-Smirnov test 183
6.8 Error rates for the random survival forest survival prediction 191

B.1 Example confusion matrices for good and reasonable classifications 207
B.2 The random forest classification with different numbers of samples 211

Chapter 1

Introduction

Robert Hooke first described a cell in his book *Micrographia* [1665], and since then biologists have been pushing the limits of microscopes in an effort to observe the intricate mechanisms that govern how cells divide and interact.

Imaging technologies have come a long way since the invention of the basic light microscope in the 17th century, from the phase contrast microscope invented by Frits Zernike in the early 1930s to the development of fluorescence imaging, the discovery of the green fluorescent protein, and the development of modern optical sectioning techniques such as the confocal microscope which allow 3D images to be acquired in real time.

We are now at a stage where high-throughput technologies are enabling thousands of experiments to be rapidly and automatically performed, including genome-wide screens, long term studies of cell growth and differentiation involving millions of cells, and automatically analysing thousands of potential drug compounds, resulting in databases filled with vast amounts of information. At the same time smaller scale experiments are increasingly dependent on accurate quantification of relatively small changes since the intricate network of interactions within a cell often counteracts attempts to perturb the function of any single protein or pathway.

This has motivated the development of algorithms and software packages to assist with the analysis of biological imaging data. The high variability of biological data means this is not an easy task, and as new discoveries are made and technologies developed the complexity of the data analysis increases.

1.1 Purpose of this thesis

This thesis explores applications of cellular imaging at multiple levels of research, beginning with low level research on the fundamentals of cell division and ending with the investigation of tumour biopsies, and the integration of imaging data with clinical outcomes.

The aim is to apply quantitative analysis techniques to multiple problems in cell biology, with the intention of taking advantage of expert knowledge whilst reducing the biases that are unavoidable in any analysis.

The thesis is not limited to the analysis of imaging data— it also illustrates the full experimental workflow, starting from the molecular biology required to fluorescently tag a protein, enabling an appreciation of the challenges faced by both biologists and image analysts.

1.2 Thesis outline

Chapter 2 sets the background to the work described in this thesis, including an overview of existing applications of imaging, and an introduction to the biology of cells.

The core part of the thesis begins with an investigation of the subcellular localisation of a protein complex involved in regulating mitosis. Chapter 3 describes the research carried out to obtain fruit fly embryos in which each component of a vital protein complex involved in regulating mitosis was tagged with GFP. This includes the molecular biology required to clone the gene and attach the GFP DNA sequence, the biochemistry needed to verify that the tagged protein was being expressed, and the sample preparation and imaging protocols for obtaining live timelapse movies of *Drosophila* embryos as they underwent mitosis.

Chapter 4 directly follows on from this, and contains the experimental results and an analysis of the data. This includes visual evidence that one of the tagged proteins, Deterin,

is the *Drosophila* orthologue of human Survivin, something which has previously only been based on sequence homology. A novel use of kymographs for visualising and quantifying spatio-temporal information in mitosis is demonstrated, and includes the application of a levelset algorithm which is used extensively in a subsequent chapter. This work was in collaboration with Dr. James Wakefield's group.

We then move up a level in Chapter 5, with an analysis of the responses of clumps of cultured cells following stimulation with a ligand, IGF2. The study of the heterogeneity of cell responses, and the underlying biological causes, is a relatively new area which will be vital in understanding the mechanisms of cancer and the development of new and personalised treatments, so this chapter demonstrates a methodology for identifying and analysing the variation in cell responses.

This begins with the development of a novel algorithm for segmenting clumps of cells with unclear boundaries, and is followed by the extraction of features and the use of a random forest classifier to test which features are potentially relevant. Finally, preliminary work into the analysis of spatially dependent correlations between cells in a clump is presented. This work was carried out in collaboration with Professor Bass Hassan in the Dunn School of Pathology, University of Oxford.

The collaboration is continued in Chapter 6, where the random forests classification framework developed in Chapter 5 is applied to clinical biopsies from Ewing's sarcoma patients, in the form of tissue microarrays (TMAs). The classification method is further developed to allow for the analysis of patient survival data, with inspiration coming from the field of bioinformatics where genetic screens are correlated to known clinical outcomes.

Finally Chapter 7 summarises the conclusions of the thesis, and suggests future areas of work.

Chapter 2
Background

Bioimage informatics is a relatively new field which is developing due to the desire of biologists and clinicians to obtain quantitative measurements from large volumes of cellular data, and to avoid tedious manual analysis. It encompasses all stages of the analysis pipeline, from pre-processing to remove image artefacts through to segmentation, tracking and feature extraction, and the subsequent analysis of the extracted data.

Applications can vary from quantifying phenotypes in a laboratory experiment through to massive datasets which involve mining thousands of features from millions of samples in order to find targets of interest, perhaps with the integration of information from other databases, and over multiple spatial and temporal scales, linking in with the work done by bioinformaticians and systems biologists [Peng, 2008, Rittscher, 2010].

This chapter will begin by providing an introduction to high throughput and high content imaging in Sections 2.1–2.2. Sections 2.3–2.6 will then introduce some of the image analysis algorithms that will be used later in this thesis, along with random forests in Section 2.7.

A brief introduction to microscopy is given in Section 2.8, and the background to the biological work in this thesis follows in Sections 2.9–2.12.

2.1 Quantitative biology

As researchers start to dig deeper into the underlying mechanisms of cells the phenotypes they discover are becoming less obvious, so accurate measurements over many samples

are required. From an imaging perspective this has led to the development of analysis tools such as ImageJ [Abramoff, 2004] and CellProfiler [Carpenter et al., 2006] to assist biologists in measuring expression levels of proteins, to count cells and to measure lengths of subcellular structures.

In other cases custom tools are written, and the field of mitosis provides many such examples, for instance looking at chromosome segregation [Yau, 2007], spindle length [Goshima and Scholey, 2010], spindle angle [Buttrick et al., 2008], studying microtubule dynamics [Smal et al., 2008] and monitoring cell cycle phase [Padfield et al., 2009], to name but a few.

2.1.1 High throughput screening

The field of *high throughput* science refers to using automated technologies to carry out thousands or millions of experiments simultaneously with minimal need for human intervention. The Human Genome Project was one of the first and is probably the most high profile example to drive the development of automated robotics [IHGSC, 2001], and subsequently the combination of automated microscopes and robotic manipulators with cell incubators has allowed the automated extraction of vast numbers of images.

High throughput screens (HTS) typically involve targeting cells with one of several thousand interfering agents, such as drug compounds or RNAi against particular genes, followed by the identification of cells which show an interesting or abnormal response indicating that a compound or gene should be investigated further [Conrad and Gerlich, 2010, Kittler et al., 2007]. The analysis is often restricted to isolated single cells, as this eases the segmentation procedure [Perlman et al., 2004], but this is gradually changing.

Imaging remains the easiest way to obtain per-cell measurements, but recently *lab-on-a-chip* technologies capable of manipulating the environment of single cells have become available, for instance using micro-fluidics [Wakamoto and Yasuda, 2006]. Here we will

focus on the imaging aspects.

2.1.2 High content analysis

High content analysis is the use of imaging and other technologies to extract as much information as possible from an experiment, and has come about with the development of automated and semi-automated analysis tools. Modern microscopes allow the acquisition of high resolution multi-channel 2D and 3D data with a temporal dimension that ranges from seconds to weeks [Pepperkok and Ellenberg, 2006], and recently the recording of multispectral data has become possible [Levenson and Mansfield, 2006].

There are two main ways in which we can approach high content data:

- Classification of the samples into known predefined categories.
- Data exploration where the outcomes are not well-defined.

The first case encompasses problems such as automatically categorising subcellular protein localisation [Glory and Murphy, 2007] or detecting cell cycle phases [Wang et al., 2008], the latter is a much wider problem, for instance analysing thousands of drug compounds [Loo et al., 2007, Perlman et al., 2004], or the Mitocheck project aimed at finding all genes involved in mitosis [Walter, 2008].

2.1.3 Machine learning and cell biology

The combination of high throughput and high content screens leads to the generation of large high-dimensional datasets, and this poses major challenges when it comes to analysing the data. The complexity of segmentation and feature extraction algorithms may be limited by the available computational resources, but a bigger problem is how the massive database of feature values can be converted into something that is useful and biologically meaningful. This is where machine learning methods come into play.

2.1.3.1 Subpopulations of cells

In the context of cell-based screens, subpopulations refers to groups of cells in different biological states (see Section 2.9). The general idea is that each cell can be approximated as lying in one of a finite number of states, and the task is to see how changing the conditions of the cell changes the proportion of cells in each state.

For instance, principal components analysis (PCA) can be followed by clustering using Gaussian mixture models (GMM) on marker expression levels [Singh et al., 2010] or morphological features [Yin et al., 2008]. Slack et al. [2008] used this method to study the effects of drug treatments on cells, and interestingly showed that different drugs with similar underlying mechanisms could be grouped using this method. In addition, novel phenotypes could be distinguished using a measure of model fit.

2.1.3.2 Temporal data

Temporal data (live cell imaging) is significantly harder to obtain compared to fixed images due to the need to maintain stable conditions in which the cells remain viable, and the requirement for fluorescent labels which do not interfere with the function of the proteins of interest. However, when it is available it can lead to valuable insights into the behaviour and temporal heterogeneity of cells [Muzzey and van Oudenaarden, 2009].

For example, in a study by Cohen et al. [2008] intracellular levels of multiple tagged proteins were shown to increase, decrease, or increase and then decrease in different cells of the same line in response to a drug. The analysis of dynamic features, such as cell motility, can only be done with live imaging [Veronika et al., 2009].

2.1.3.3 Marker optimisation

A recent use of machine learning is to optimise experiments by reducing the number of markers needed to distinguish between phenotypes. Loo et al. [2009] used a classifier to identify the least important marker, allowing it to be replaced by a more informative one. Peng et al. [2010] return to the problem of classifying subcellular protein localisation and show it is possible to unmix different patterns from a limited number of markers.

2.1.4 Visualisation techniques

A field closely related to machine learning is that of data visualisation. The previously mentioned examples of Slack et al. [2008] and Singh et al. [2010] used multidimensional scaling (MDS) [Borg and Groenen, 1997] to visualise the differences between the effects of drugs.

MDS can be thought of as a non-parametric analogue of PCA, but using a pairwise distance/dissimilarity matrix between all samples, for instance calculated between the feature vectors of each cell or sample. Samples which show a similar response should cluster together in the MDS plot, whereas those which are dissimilar from all others should appear isolated. There are several choices of distance measures, such as the Komogorov-Smirnoff, Kullback-Liebler, and Bhattacharya distances [Shutin and Zlobinskaya, 2010]. Closely related to MDS are Kohonen networks (self organising maps), used by Caie et al. [2010] to predict the mode of action of novel compounds using a reference set.

Loo et al. [2007] used the novel approach of screening several hundred drug compunds by building separate support vector machines (SVMs) [Vapnik, 1995] for each control vs. compound comparison. The normal vector of the hyperplane was used to define a characteristic profile for each compound, which could then be compared with other doses and compounds.

Interactive data exploration tools can also be useful. Hamilton and Teasdale [2008] created a tool for visualising and comparing subcellular protein localisation. CellProfiler Analyst [Jones et al., 2008] and CellClassifier [Rämö et al., 2009] both require a small number of representative cells to be manually selected, and will automatically find similar cells.

2.2 Histology studies

The previous sections mostly considered *in-vitro* studies, involving the monitoring and perturbation of cultured cells. Recently work on automating the analysis of histology samples has gained pace, see Gurcan et al. [2009] for a review of this field.

This research can be divided into two main areas

- Automated classification or grading of tumour biopsies.
- Screening for biomarkers related to cancer grade, and survival.

The former is motivated by the need to standardise the grading of tumours which at present is done visually by a pathologist, and is subject to inconsistencies as well as being inefficient [Longacre et al., 2006]. Examples include the use of textural features with a SVM [Doyle et al., 2008, Masood and Rajpoot, 2009].

The latter is partly motivated by the success of bioinformatic screens where microarrays have been used to produce gene signatures of tumours which are linked to patient survival, and the desire to obtain *in-situ* information on gene expression [van de Vijver et al., 2002, Chibon et al., 2010].

2.2.1 Tissue microarrays

On the experimental side tissue microarrays (TMAs) are increasing the feasibility of analysing hundreds of biopsies under increasingly standardised conditions [Kononen et al.,

1998]. A TMA is formed of tens or hundreds of small cylindrical tissue cores, each less than 2 mm in diameter, and may include replicates from the same biopsy as well as multiple subjects. These cores are embedded in a grid pattern in a paraffin block which is then thinly sliced and laid out onto microscope slides.

Each slide therefore contains a grid of tissue samples, which means any experimental protocols should be standardised across the whole slide, in contrast to the small errors which would invariably creep in if each sample was prepared separately.

Systems for automating the acquisition of images from across a TMA have been developed [Chen et al., 2004, Liu et al., 2002] but in general most of the work is still often carried out manually. Few automated tools are available for the analysis of those images, and normalisation of images remains a big challenge [Camp et al., 2002, Decaestecker et al., 2009].

2.3 Segmentation

The simplest methods of segmentation involve global or local thresholding to define the background, for instance using Otsu's method [Otsu, 1979], or setting a cutpoint in the intensity histogram [Camp et al., 2002]. Fitting a multi-component probabilistic distribution to the intensity distribution can also give good results, as can combinations of morphological operations and filtering with for example a Laplacian kernel [Perlman et al., 2004].

These are very fast methods which are ideal for high throughput screens where most images are of good quality, and the markers used give a very clear labelling of each cell. They tend to fail when presented with images with poor signal-to-noise, or when cells clump together.

2.3.1 Watershed related methods

In order to obtain a more robust segmentation it is necessary to incorporate the spatial relationship between pixels, for instance by using region growing or watershed methods.

Watershed methods are fast, and have some capability of handling multiple touching objects, though in practice they suffer from over-segmentation so a marker controlled version, for instance using a nuclei segmentation, can be used instead [Malpica, 1997]. Alternatively merging operations can be used to join fragments of nuclei or cells [Chen et al., 2006].

A variant of watershed methods are those based on the generalised Voronoi diagram, which can be combined with image gradients to obtain a segmentation [Jones et al., 2005].

2.3.2 Snakes and level sets

A potential issue with watershed methods is that in general there is no way to incorporate long range spatial information. Energy minimisation methods which use information from across the whole of the segmented region are one solution.

Active contours or snakes introduced by Kass et al. [1988] represent a boundary as a spline whose energy is given by an internal regularisation term which controls the length and curvature, and an external data term, such as the image gradient.

Level sets are an active contour implementation that avoids the need to parameterise the snake, by representing a boundary implicitly as the zero level set of a higher dimensional function whose evolution is governed by a partial differential equation, see Sethian [1997] for an intuitive explanation.

Malladi et al. [1995] and Caselles et al. [1997] show how to derive the update equations for a level set attracted to edges in an image— the essential problem is to derive a speed function that causes the level set to stop at a suitably strong edge.

The evolution of the higher dimensional function ϕ is given by:

$$\frac{\partial \phi}{\partial t} = g(u)(c + \kappa)\nabla\phi$$
$$\kappa = div\left(\frac{\nabla\phi}{\|\nabla\phi\|}\right)$$

where c is a balloon force in the form of a positive constant which allows the level set to conform to non-convex boundaries and κ is the Euclidean curvature, which forms part of the regularisation term. $g(u)$ is an idealised edge function of the image u which takes a value of 0 on strong edges.

In practice $g(u)$ is never 0, so alternative formulations with a stopping term are used in practice [Caselles et al., 1997].

2.3.3 Region based level sets

Chan and Vese [2001] derived a new version of active contours using a region based approach which should perform better in images where there is no distinct edge, but it is possible to distinguish regions based on their different intensities.

A piecewise constant intensity segmentation model is assumed:

$$E_{ext} = \int_{inside(C)} \|u(x,y) - c_1\|^2 \, dxdy + \int_{outside(C)} \|u(x,y) - c_2\|^2 \, dxdy$$

where c_1 and c_2 are the mean values of u inside and outside the curve C respectively.

A wide range of modifications have been made to these basic level set formulations, including shape models [Leventon et al., 2002], the introduction of efficient narrowband methods, and more numerically stable formulations [Li, 2005].

2.3.4 Multiregion segmentation

The ability of a level set to seemlessly handle topological changes can be a disadvantage in some circumstances, such as when segmenting multiple separate objects which may touch. This is a particular problem with cells, where it is desirable to handle splitting events but avoid merging two cells if they come into contact.

The standard level set formulation only handles two phases (labelled 0 and 1, or background and foreground), but a number of extensions have been proposed. Vese [2002] use a vector level set in which each of the N regions is represented as the composition of multiple binary level sets, and a total of $log_2 N$ level sets are required. However, it relies on some of the regions disappearing when the number of regions is not a power of two. Mansouri et al. [2006] avoid this problem by reformulating the definition of regions to enforce an exact partitioning of the image by defining N nested levelsets for N regions, at the expense of increasing computational requirements. Yezzi et al. [2002] also require N level sets, but evolve the level sets using a vector value statistic, which necessitates an $N-1$ dimensional image to be created.

An obvious extension when dealing with a large number of objects is to use the fact that an object will typically only have a small number of neighbours, for instance using colouring approaches to allow multiple objects to be represented using the same phase [Nath et al., 2006a], or a topological constraint can be imposed on the level set [Han et al., 2003, Yu et al., 2008].

2.4 Graph based segmentation

Recently, several segmentation algorithms based on a graph representation of the image have been developed.

A graph representation is typically formed by creating a node at each pixel location $\mathbf{x}$ in an image, and creating edges between all pairs of neighbouring nodes, where the neighbourhood is typically 4-connected or 8-connected in 2D. In mathematical terms, the graph $G = (V, E)$ consists of a set of vertices (nodes) $v \in V$ linked by edges $e \in E \subset V \times V$. An edge linking nodes v_i and v_j is denoted e_{ij}, with weight w_{ij}. Only undirected graphs will be considered, so $e_{ij} = e_{ji}$.

The graph cuts algorithm casts the image segmentation problem as one of finding a maximum flow/minimum cut in the graph [Shi and Malik, 2000], where the cost of the cut is based on the sum of the weights of the edges that are cut. It can also be used for supervised or interactive segmentation by defining seeds [Boykov and Jolly, 2001]. In contrast the original random walker algorithm developed by Grady [2006] relies on seeds being manually defined within the image. It encompasses aspects of region growing but also includes resistance to boundary leakage, where the segmentation may leak through a small gap in a boundary. Both of these are suggested by Couprie et al. [2010] to be instances of a more general energy optimisation algorithm, which they term the *power watershed* method.

2.5 Comparison of cell segmentation methods

Voronoi methods	Level sets	Graph based
Local minimum	Variational formulation allows a global minimum to be achieved	Finds global minimum
Can be extremely fast	Solution of PDE is slow	Can be fast, but may have a large memory requirement
	Numerical solution of PDEs is a well researched area	Efficient graph minimisation algorithms exist
	Narrowband methods can be fast	Graph reduction/pruning algorithms can reduce complexity
Uses local information only	Global priors such as shape models are easily incorporated	Research into global priors is ongoing
Requires markers for initialisation	Some formulations are sensitive to initialisation	Methods requiring seed points are sensitive to initialisation
Handles multiple regions	Multi-region segmentation requires an explicit interaction or topological term	Many methods handle multiple regions directly
Evolution can be visualised	Iterative solution allows evolutions to be visualised to aid understanding	Formulated as a matrix inversion
Fixed topology	Topology can change	Ability to change topology depends on method
Cannot be used unsupervised	Unsupervised segmentation possible	Unsupervised segmentation possible
Pixel level segmentation	PDE solution allows subpixel segmentation	Pixel level segmentation

2.6 Tracking

There are several considerations to make when attempting to track biological objects, including the number of objects (single or multiple), appearance changes, splitting/merging, occlusion, presence of other distracting objects or noise, field of view, and temporal resolution. The problem of target tracking has been widely explored in many fields, with many of the original algorithms being developed for radar or sonar tracking by the defence industries. See Meijering [2006] for a simple introduction to biological tracking applications.

The algorithms can be loosely split into three main groups: contour evolution, segment/associate, and Bayesian methods.

2.6.1 Contour evolution

Contour evolution is ideal when the temporal resolution of the data is high enough that there is significant overlap between the foreground of each object in consecutive frames, and fits in well with active contour algorithms [Shi, 2005, Yau, 2007, Zimmer et al., 2002].

A variant of this is to convert a temporal dimension into a spatial one by concatenating all images along the additional dimension, so a 2D+T segmentation and tracking problem becomes a 3D segmentation problem [Padfield et al., 2009]. These easily handle splitting (mitosis) events but can fail when cells come into close contact.

2.6.2 Segment and associate

The two step approach of detecting objects in each frame independently, followed by an association step to link corresponding objects together, is one of the oldest algorithms for tracking multiple objects [Reid, 1979, Fortmann et al., 1983].

In the simplest implementation a greedy algorithm is used in which each object is

associated with the nearest neighbour in the next frame, though in practice this can lead to multiple conflicting associations, so a cost-minimisation procedure, for instance using graph theory methods, can be used instead [Chen et al., 2006]. The definition of nearest neighbour can incorporate features other than physical distance, such as appearance information or a motion model [Sbalzarini and Koumoutsakos, 2005]. Modification of this approach handles merging/splitting and appearance/disappearance [Padfield et al., 2010, Matov et al., 2011].

2.6.3 Bayesian methods

The single target Kalman filter [Kalman, 2002] is one of the oldest examples of a Bayesian tracker, incorporating a model of the target's dynamics (which may incorporate morphological changes) to predict the state of the target, allowing noisy measurements to be unweighted and false targets to be rejected.

The particle filter [Isard, 1998] is better suited for non-linear (non-Gaussian) models, enabling multiple hypotheses to be handled, at the expense of significantly greater computational complexity. Biological applications usually require the filter to be modified to handle multiple interacting objects [MacCormick, 2000, Khan et al., 2005, Smal et al., 2008].

2.7 Random forests

Random forests [Breiman, 2001] is a machine learning algorithm that is increasingly being used by the biomedical community for a variety of tasks, including disease classification, identification of relevant genes in bioinformatics screens, and image segmentation.

Its primary strengths include a resistance to over-fitting, the relatively few parameters which need to be tuned, and good performance comparable with those of other learning algorithms such as support vector machines (SVMs) [Vapnik, 1995]. For a comprehensive

overview of machine learning see Bishop [2006]. A quick introduction to random forests can be found in Liaw and Wiener [2002], and a more in-depth tutorial in Strobl et al. [2009].

2.7.1 Classification and regression trees

The idea of random forests was born out of classification and regression trees (CART) [Breiman et al., 1984], single decision trees which are trained to classify data by recursively splitting the data into two partitions by optimising some criteria, typically involving a measure of the information gain (or entropy reduction) which results from partitioning the data. One of the core ideas behind CART which allows it to handle high-dimensional data is that many response functions are locally dependent on just a few of the covariates, so the recursive partitioning allows these subspaces to be found [Friedman, 1991].

As with many machine learning methods CART is prone to over-fitting and poor generalisation, though this is partly mitigated by limiting the tree depth.

2.7.1.1 Randomisation

The random forest algorithm uses multiple sources of randomisation to reduce overfitting.

If the training procedure for a tree is sensitive to the dataset then averaging over multiple datasets can improve the accuracy of the tree. Since only a single dataset is usually available bootstrap aggregating, or *bagging*, can be used to simulate the effect of replicate datasets [Breiman, 1996] by sampling with replacement. The sampling and training procedure is repeated multiple times, producing a *forest* of trees.

Since sampling with replacement is used some samples will be left out— these samples are referred to as out-of-bag (OOB) samples, and since they are not used in the training of the associated tree they effectively form a test set for that tree.

The second use of randomisation is in the selection of features used for splitting nodes.

If the feature space is large then it may be impractical to explore all features when building each tree, so instead a small number of candidate features are chosen randomly and multiple thresholds tested to see how informative each feature is [Amit and Geman, 1997]. An additional benefit is that each tree has a different view of the data, causing the training procedure to be more unstable and thus maximising the benefits of the aggregation step.

2.7.2 Feature relevance

There are two main reasons for wanting to select a small set of features from a high dimensional space. The most common case is to build a more efficient learner, both in terms of the required computational resources, and also because a simpler model should be less prone to over-fitting and will therefore generalise better.

The second case which is of increasing interest is to be able to identify a small number of interesting features from a very large set, for instance in the field of bioinformatics where the aim is to identify a small number of genes that are related to a disease, from a screen of several thousand genes [Díaz-Uriarte and Alvarez de Andrés, 2006]. It is important to note that the set of features selected for a learner is not necessarily the same as the set of all relevant features, particularly when several correlated features are present [Kohavi, 1997].

2.7.2.1 Random forest feature importance

Breiman [2001] suggests calculating the importances of feature X_i using a variable permutation test on the OOB data. Feature X_i is randomly permuted thus breaking any association with the response variable Y, and the increase in OOB misclassification rate compared with using the intact feature is the importance measure. A variant of this is to use random node assignment, in which OOB samples are passed down a tree normally but whenever a node is split on feature X_i the left or right branches are chosen at random

[Ishwaran et al., 2008].

Another suggested method is to look at the total Gini scores Breiman [2001] which are found by adding up the decrease in Gini derived from the information gain at each node when splitting on a variable. Although this is computationally less intensive than the permutation tests it can be biased when predictors of different types or scales are used [Strobl et al., 2007].

Recently Ishwaran [2010] has proposed using the depth of a variable in a tree as a measure of importance, motivated by the expectation that important features are more likely to appear near the root of a tree with less important features appearing beneath them. This measure has the benefits of being dimensionless, independent of the error measure used, and amenable to theoretical analysis.

2.7.3 Unsupervised clustering

Samples which appear in the same terminal leaf of a tree must have similar values of the features used at the node splits, so by examining all leaves of all trees a proximity matrix P can be built up in which element $P(i, j)$ is incremented every time sample i appears in the same leaf as sample j. This matrix is converted into a dissimilarity matrix which can then be visualised using multidimensional scaling (MDS) [Shi and Horvath, 2006].

If the resulting clusters are obvious in the dimensionally reduced MDS plots they can be defined manually, or using the k-medoids algorithm [Kaufman and Rousseeuw, 2005]. k-medoids is similar to the k-means algorithm except that it minimises the sum of pairwise dissimilarities instead of the Euclidean distance. This approach was successfully used by Seligson et al. [2005] to pick out a previously unknown association between histone modification patterns and patient prognosis.

2.7.4 Variations of random forests

Several variants of random forests have been proposed in the biomedical field, such as RF++ for classifying correlated (non-independent) data [Karpievitch et al., 2009], conditional random forests to avoid feature importance bias [Strobl et al., 2007], and random survival forests to handle survival data [Ishwaran et al., 2008].

2.7.5 Comparison with other methods

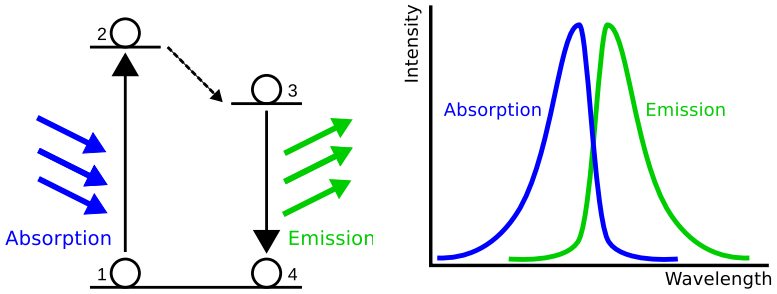
The performance of random forests is comparable with support vector machines (SVMs) [Díaz-Uriarte and Alvarez de Andrés, 2006, Statnikov et al., 2008]. Advantages over SVMs include direct handling of multi-class problems and generally good performance using the default parameters. Disadvantages include the relatively slow prediction time, and unnecessary complexity when a decision plane is not aligned with an orthogonal axis [Menze et al., 2009].

2.8 Fluorescence imaging

Typical light microscopes such as brightfield, phase-contrast and differential interference contrast (DIC) allow detailed structural and histological information to be viewed in biological samples. Fluorescence microscopy extends this by allowing individual components at the molecular level to be labelled with different markers, enabling biologists to tease apart the behaviour and interactions of complex assemblies of proteins within a cell.

2.8.1 Principles of fluorescence microscopy

Fluorescence microscopy takes advantage of fluorophores, molecular labels containing electrons that are excited by light in a certain band of wavelengths, and which return to their



(a) Fluoresence emission (b) Absorption and emission spectra

Figure 2.1: The basic principles of fluoresence. (1) A fluorophore absorbs light of a high enough energy (e.g. GFP maximally absorbs light with wavelength 488 nm, coloured blue), causing its electrons to momentarily enter an excited higher energy state (2). Some of this energy is lost as vibrational energy, known as the Stokes shift (3), and the remainder is emitted in the form of a photon of a lower wavelength than the incident photon (4), in this case green.

ground state within a few nanoseconds by emitting photons of a longer wavelength [Lichtman and Conchello, 2005], as illustrated in Figure 2.1. A suitable choice of filters allows the excitation and emission wavelengths to be separated.

The spatial resolution of light microscopes is inherently limited by the diffraction pattern (called the point spread function, or PSF) formed by photons at the image plane of the microscope, which sets the minimum resolvable distance in the imaging plane to around $0.61\frac{\lambda}{\mathrm{NA}}$ where NA is the numerical aperture of the objective lens and λ the wavelength of the emitted light. For a typical objective with a NA of 1.4 and a green fluorophore emitting light at 510 nm the resolution is around 220 nm.

Although standard microscopes can easily achieve the maximum spatial resolution in the focal (x-y) plane the PSF also extends into the z-axis parallel to the path taken by light through the sample, leading to blurring of the observed image in all but the thinnest of specimens. This is further exacerbated by scattering in thick specimens.

2.8.2 Optical sectioning

Optical sectioning is the imaging of a single plane within a sample, either physically as in confocal and two-photon microscopy, or mathematically by using deconvolution to invert the PSF [McNally et al., 1999]. Recently several designs of microscopes combining both techniques have been developed, such as light sheet microscopy, optical projection tomography, and structured illumination.

2.8.2.1 Confocal microscope

The confocal microscope is one of the oldest and most widely used methods of optical sectioning [Conchello and Lichtman, 2005]. The key concept in a confocal microscope is the inclusion of a pinhole in front of the light source which focuses the light to a point before it passes through the objective, and another *confocal* pinhole in front of the detector (Figure 2.2). Any light emitted or scattered from above or below the focal plane in the sample will be blocked at the confocal pinhole.

The x-y resolution is determined by the size of the illuminated region, with the simplest solution being to scan a highly focussed laser beam across the sample and to collect the emitted light with a photomultiplier tube (PMT), as in the laser scanning confocal microscope (LSCM). This can be quite slow, so more advanced designs use higher sensitivity detectors and illuminate multiple points concurrently, such as in the spinning disc confocal microscope.

2.8.3 Green fluorescent protein

A special mention should be given to the green fluorescent protein (GFP), which marked a seminal point in the evolution of biological imaging tools [Tsien, 1998]. GFP is a small (27 kDa) protein which behaves as a natural fluorophore— when illuminated by blue light

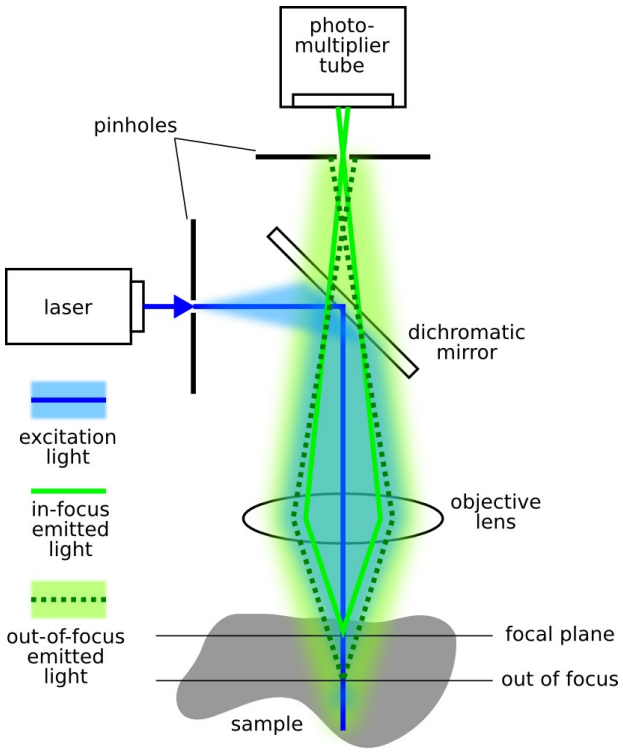


Figure 2.2: Principles behind a confocal microscope. The excitation light beam is focussed using a pinhole and the objective. Fluorescence emitted from an in-focus point in the sample is focussed through the confocal pinhole by the objective and is detected by the photomultiplier tube. Out-of-focus fluorescence is blocked by the confocal pinhole.

it emits visible green light. Following its discovery and the identification of its DNA code several improvements have been made to the original fluorescent protein, including increased fluorescence efficiency, other colours, photoactivatable and photoconvertible variants [Lippincott-Schwartz and Patterson, 2003].

Almost any protein (and therefore any biological structure) can be tagged with a fluorescent marker using molecular cloning and genetic engineering techniques, with the main advantage being the ability to image live samples without any invasive procedures. A potential drawback is that the marker may interfere with endogenous function.

2.9 Heterogeneity in cell populations

A simplistic view of life is that genetically identical organisms in an identical environment will behave identically. In practice there is increasing experimental and theoretical evidence that this is not the case, including at the level of individual cells.

2.9.1 Non-genetic variation

The absolute number of molecules of a protein in a cell is very small, so the effects of stochastic interactions between molecules must be considered, which can lead to a small change having an unexpectedly large effect. This manifests itself in the form of non-genetic variation, or biological noise, which can be roughly categorised into two types.

Extrinsic noise is caused by factors which affect overall gene expression levels in a cell, but vary between different cells, for instance the numbers of transcription and translation complexes. Intrinsic noise is due to stochasticity which is a consequence of the inherent randomness in transcription and translation processes. A good review of both is given by Raj and van Oudenaarden [2008].

This variation in expression can also have a temporal aspect, with protein levels fluctuating regularly or irregularly over long or short timescales.

2.9.2 The consequences of cell heterogeneity

In many systems noise is regarded as a nuisance which gets in the way of perfect measurements. However, as Altschuler and Wu [2010] point out it can have many beneficial biological effects. For instance if a population of cells comes under severe environmental stress a small hidden subpopulation of cells may be the ones that survive.

Continuous fluctuations are not always a good thing though, and Brock et al. [2009]

suggest that gene regulatory networks can be seen as moving between stable "attractor states", explaining how cells can retain some measure of stability.

2.9.3 Heterogeneity in cancer

Unfortunately tumours are also characterised by their very high heterogeneity, effectively consisting of tens or hundreds of subpopulations, all of which may have different responses. Cells at the extreme ranges of behaviour may have a temporary advantage which allows them to proliferate faster than their neighbours, and therefore to have more opportunity to evolve resistance to, for example, an anti-cancer drug.

Cancer treatment regimes often involve repeated treatments on the assumption that cells surviving the first treatment will have changed their state by the time of the second. Niepel et al. [2009] suggests that a log-normal distribution is a good fit for the expression of a protein in a clonal population, that the long tails mean that even repeated treatments will often fail to kill those cells in the "tail", and that in addition each drug treatment may itself cause an increase in heterogeneity.

It is even possible for cancer cells to show contradictory behaviour. Cohen et al. [2008] screened 1000 proteins in cancer cells following treatment with camptothecin, an topoisomerase-1 poison. In some cases there appeared to be two subpopulations of cells for each protein, for instance one in which a protein decreased whereas in the other protein levels increased, and in some cases this was seen to be correlated with cell survival following drug treatment.

2.9.4 Inter-cell communication

Having established that heterogeneity is present in cells, and that it can have significant biological consequences amongst cells, the next question to ask is how might a difference

in behaviour of one cell affect the behaviour of its neighbours?

2.9.4.1 Community effect

The community effect was named by Gurdon [1988], following his discovery that in *Xenopus* embryos cells which would be expected to differentiate into muscle cells failed to do so when cultured as single cells or as a monolayer, but did differentiate as expected when formed into a 3D aggregate, suggesting that the behaviour of one cell reinforced that of its neighbours, and vice-versa. Similarly Clayton et al. [1985] showed that hepatocytes from disaggregated liver cells cultured in a monolayer decreased their expression levels of liver-specific mRNAs but not common mRNAs, but that expression levels remained the same as those of intact liver slices when the hepatocytes were re-aggregated.

There are many examples of systems in which one cell expresses a gene which eventually results in the release of a ligand that is picked up by an extracellular receptor on a neighbouring cell. This triggers a signal cascade within that neighbour that results in the activation or inhibition of expression of a different gene. In contrast the community effect involves a positive feedback loop involving the expression of the same gene in neighbouring cells which may result in an averaging effect and/or positive reinforcement over a group of cells [Bolouri and Davidson, 2010].

Obviously the feedback loop must be moderated to prevent an uncontrolled increase in levels, and to an infinite homogeneous mass of cells forming. The feedback loop may therefore also incorporate the expression of an inhibitor to limit expression levels.

2.9.5 Single cell measurements

Ensemble measurements of cells will often hide the presence of multiple subpopulations, which may only be detected as a function of multiple measurements. Often these are not obvious from viewing the data by eye, so in practice we should ideally take all measurements

from individual cells to ensure we can keep track of the correlations between variables [Altschuler and Wu, 2010].

The complexity of this depends on the type of cells being imaged. Many automated screens aim to plate out cells so they are widely separated, greatly easing the segmentation process. Others involve experiments where there are multiple interacting cells, and it is necessary to image the whole group in one go and to later work out the boundaries of each cell during the analysis stage. Closely related to this is the segmentation of tissue samples, for instance tumour biopsies, where multiple cell types are naturally combined together in a complex 3D matrix with connective tissue and fluids, and this is further complicated by the vast variability in the quality of stainings.

2.9.5.1 Single cell imaging

From the microscopy perspective the process of labelling and imaging multiple proteins in the same cell has been around for a while, but the traditional qualitative or manual analysis of protein levels does not lend itself to obtaining accurate statistical distributions of levels, or the detection of multiple subpopulations (or states) which may only become clear when multiple variables are simultaneously taken into account. These have only become practical with the development of automated analysis and visualisation tools.

2.10 Mitosis and cytokinesis

The cell is the basic hereditary unit of all organisms, containing everything needed to store and replicate the genetic information that codes for an organism. It is a microscopic factory, continually synthesizing and degrading proteins and interacting with its environment 24 hours a day. Mitosis is the process by which the unit of life replicates itself.

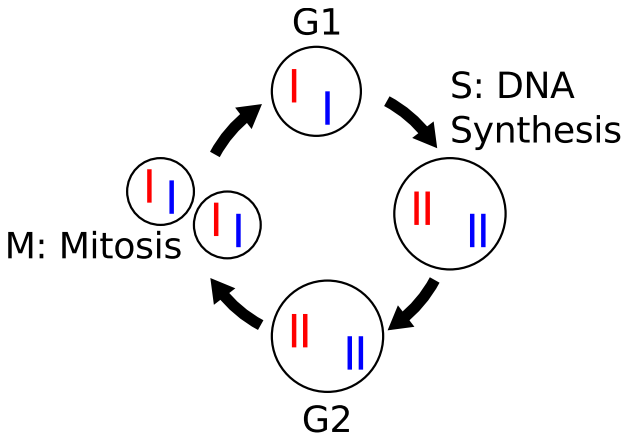


Figure 2.3: The eukaryotic cell cycle. Interphase is comprised of G1, S and G2 stages. During G1 and G2 the cell grows in size, whilst DNA is synthesized in S (two chromosomes, coloured red and blue, are shown). During M (mitosis) the cell partitions its duplicated chromosomes into two, before separating into two daughter cells. Cells which are no longer actively dividing move into a separate quiescent state G0.

2.10.1 The cell cycle

The cell cycle is split into four main stages, illustrated in Figure 2.3.

G1 and G0 The first of the gap stages, G1, accounts for the majority of the cell cycle in terms of time. During G1 a cell grows in size and accumulates substances required for the remaining stages before division including amino acids for protein synthesis, nucleotides for DNA replication and phospholipids for membrane formation. Environmental conditions are monitored, and if they are unfavourable the restriction checkpoint is triggered, preventing further cell division. Many cells will enter a variation of G1 called G0 in which all cell division activity ceases. This may be a response to unfavourable growth conditions, or it can be a natural state for a differentiated cell which will have no further requirement to divide.

S During S (synthesis) phase the DNA in the cell is duplicated. In most cells the centrosome is also duplicated in S phase, and will ultimately form the poles of the mitotic

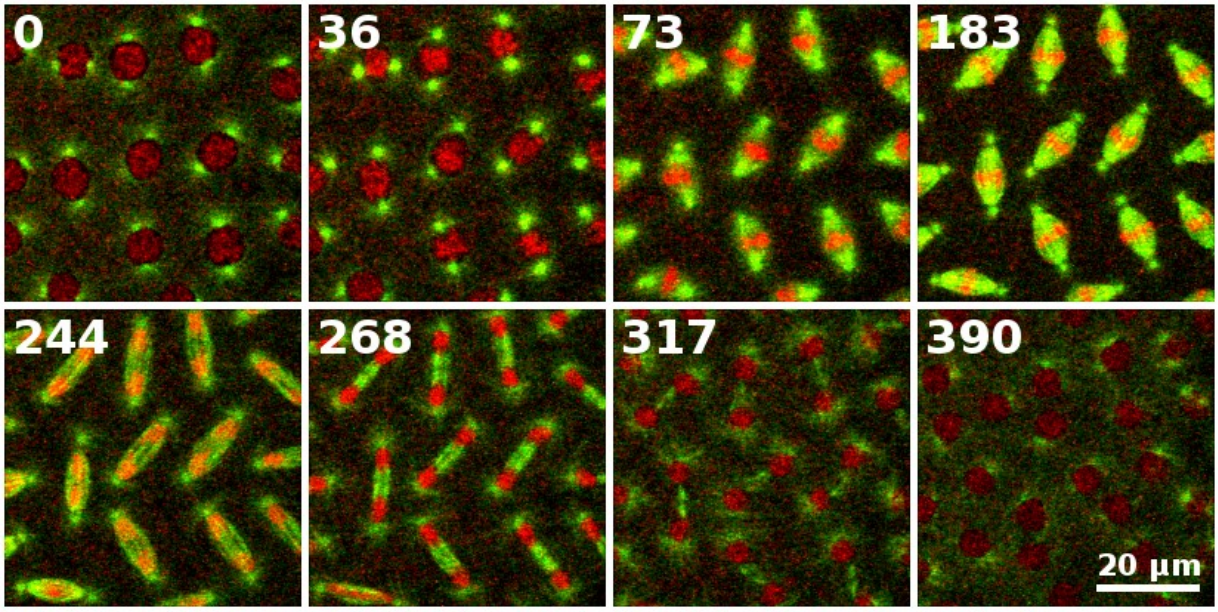


Figure 2.4: Fluorescent images of mitosis in a *Drosophila* embryo. Red: Histone-RFP (DNA). Green: Tubulin-GFP (microtubules). Times are given in seconds. 0: Prophase. Chromosome condensation is underway. Astral microtubules start to form from the poles. 36,73: Prometaphase. The nuclear envelope breaks down and microtubules attach to the chromosomes. 183: Metaphase. The chromosomes line up along the metaphase plate. 244,268: Anaphase. The two sets of chromosomes are pulled to opposite poles of the cell. 317,390: Cytokinesis. The chromosomes decondense and the nuclear envelope reforms.

spindle during mitosis.

G2 During the second gap stage, G2, the cell continues to prepare for mitosis, including checking the DNA for any replication errors.

2.10.2 Mitosis (M) and cytokinesis

These are the final but most dramatic of the processes within the cell cycle. The duplicated chromosomes split into two coherent sets, and the cell divides into two, see Figure 2.4. Mitosis is divided into five stages and is normally followed by cytokinesis, marking the end of one complete cell division cycle.

Prophase The transition from interphase to prophase is marked by the condensation of DNA from an indistinct mass throughout the nucleus into pairs of chromatids which are clearly visible under the microscope. The pair of centrosomes separates and moves to opposite ends of the cell, and the microtubule network undergoes a complete reorganisation from a semi-random network throughout the cell during interphase to the beginnings of the mitotic spindle in prophase.

Prometaphase The breakdown of the nuclear envelope marks the beginning of prometaphase. Microtubules which form the spindle can now grow into the nucleus, and attach themselves to the chromosomes at special structures called kinetochores. A pair of kinetochores is bound to each of the sister chromatids at a specialised region of condensed DNA called the centromere, which is also the point where the two chromatids are bound together by protein complexes termed cohesins. This is a critical stage in which each pair of chromosomes must be connected to opposing poles.

Metaphase The metaphase or spindle assembly checkpoint (SAC) prevents the onset of anaphase by delaying the activation of the anaphase promoting complex (APC) until all the chromosomes are correctly aligned and connected to opposing microtubules.

Anaphase The APC indirectly results in the cleavage of the cohesins which tie the paired chromosomes together, upon which they are rapidly and synchronously moved towards opposing poles through a combination of lengthening of the spindle and movement of the kinetochores along the microtubules. The mitotic spindle is reshaped into a structure called the central spindle, and becomes constricted at its midpoint.

Telophase In telophase the sets of daughter chromosomes decondense and the nuclear envelope reforms around them. A contractile ring of actin filaments and myosin motors

forms around the body of the cell, in line with the central spindle.

Cytokinesis The final stage of cell division is the separation of the cytoplasm by the pinching of the cell membrane at the contractile ring.

2.11 *Drosophila melanogaster* as a model organism

Model organisms are a vital tool in biomedical research, providing a simpler more tractable model of a system. They can provide a bridge between a reductionist approach which considers the interaction of individual proteins, and an integrative approach which looks at the effect of a gene at the level of a whole organism.

Drosophila melanogaster has been used as a model organism for over 100 years, especially in the field of genetics and developmental biology where a Nobel prize was won in 1995. A very large number of mutants have been identified and maintained, and annotated databases of the complete genome such as *FlyBase* are available [Crosby et al., 2007]. The rapid breeding cycle of *Drosophila* (around two weeks) allows experiments to be performed in a reasonable time-frame, and they are relatively cheap to maintain.

2.12 The Chromosomal passenger complex

The chromosomal passenger complex (CPC) is a complex of four core proteins which play an important role in the regulation of mitosis, including chromatin condensation, sensing tension as part of the spindle assembly checkpoint, correcting microtubule-kinetochore attachments, and governing the position of the cleavage furrow [Ruchaud et al., 2007]. In order to execute its multiple functions the CPC undergoes a highly dynamic localisation throughout mitosis, and disruption of any member leads to incorrect localisation and function of the complex, complicating attempts to probe its functional mechanism.

2.12.1 Overview of the complex

Aurora B is a serine/threonine kinase, and forms the enzymatic component of the CPC. The other three subunits, Incenp, Survivin/Deterin and Borealin/Dasra-B are typically viewed as having roles in regulating the activity and localisation of Aurora B, with the complex as a whole most likely being in the form of a heterotetramer [Jeyaprakash et al., 2007, Sessa et al., 2005].

As the cell proceeds through prometaphase the CPC concentrates at the centromeres, where it is essential for correct chromosome bi-orientation at metaphase as part of the tension sensing mechanism, and is also involved in breaking incorrect microtubule-kinetochore attachments.

Once the chromosomes have been successfully aligned in metaphase and the spindle checkpoint satisfied, the CPC transfers to the anti-parallel microtubules in the spindle midzone.

During cytokinesis the CPC has an essential role in determining and perhaps initiating the formation of the contractile ring at the cell cortex, though the exact mechanism remains to be determined.

2.12.2 Aurora B (Ipl1-aurora-like-kinase)

2.12.2.1 Early mitosis

During prophase and prometaphase the kinase activity of Aurora B is required for chromosome condensation, with the other CPC members being required for its localisation and activation. Aurora B phosphorylates Histone H3 at multiple sites [Giet and Glover, 2001, Goto et al., 2002], though levels of H3 phosphorylation do not appear to be directly correlated with actual levels of condensation [Adams, 2001].

At the same time phosphorylation of Histone H3 by another kinase, Haspin, is also

required for the targetting of the CPC to the centromeric region of the chromosome via Survivin [Kelly et al., 2010], where it plays a vital role in chromosome alignment. The overall picture is therefore unclear.

2.12.2.2 Spindle assembly checkpoint

The CPC is required for both maintaining the spindle checkpoint and for the separate process of correcting microtubule-kinetochore interactions [Vader et al., 2007], but as with most other aspects of the CPC the underlying mechanisms are still unclear.

There is evidence from other species to suggest that the binding kinetics of Aurora B and other CPC members vary depending on the state of the check point. Fluorescence recovery after photo-bleaching (FRAP) experiments have shown that the overturn of Aurora B and survivin at centromeres in *Xenopus* and HeLa cells greatly decreases post-metaphase, though there is disagreement over the rates of Aurora B turnover [Delacour-Larose et al., 2004, Ahonen et al., 2009]. This turnover is also dependent on Incenp, since injection of an antibody against the C-terminal of Incenp required for binding to Aurora B resulted in very little fluorescence recovery [Ahonen et al., 2009].

Liu et al. [2009] describes a possible mechanism for the sensing of tension in which Aurora B is constitutively active at the centromeres and phosphorylates kinetochore substrates such as CENP-A and the Ndc80/Hec1 complex leading to a destabilisation of the microtubule attachment [Kelly and Funabiki, 2009]. Correctly attached kinetochores result in strong tension and increased separation of the kinetochores, leading to a reduction in phosphorylation and thus increased stability. The separate process of correcting MT-kinetochore attachments is partially governed through the regulation of MT depolymerases, such as Mitotic centromere-associated kinesin (MCAK) [Wordeman et al., 2007].

2.12.2.3 Late mitosis

At anaphase Aurora B relocalises to the spindle midzone where it is required for central spindle formation [Giet and Glover, 2001], and appears to be essential for successful cytokinesis. It is able to phosphorylate MKLP1 and potentially other members of the centralspindlin complex leading to activation of the small GTPase Rho [Carmena, 2008], which in turn leads to localised actin polymerisation at the equatorial cell cortex and formation of the contractile ring.

A possible role for the CPC in anaphase-B spindle elongation has also been suggested. Micro-injection of an anti-Incenp antibody at anaphase leads to a reduction in the number of microtubule bundles in the spindle midzone and a decrease in pole-pole distance [Ahonen et al., 2009].

2.12.3 Incenp

The *Drosophila* orthologue of Incenp shows the expected CPC localisation pattern, with RNAi knockdown leading to reduced Histone H3 phosphorylation and failure of the chromosomes to congress at the metaphase plate [Adams, 2001]. A coiled-coil domain enables Incenp to bind to microtubules [Tseng et al., 2010], which appears to be vital for the role of the CPC in spindle formation.

The C-terminal of Incenp is highly conserved between *Drosophila* and other species, most notably in the IN-BOX which both binds and acts as a substrate for Aurora B. Aurora B is activated on binding to Incenp, and in turn phosphorylates Incenp, which leads to further activation of Aurora B in a positive feedback loop [Honda et al., 2003].

In humans the N-terminal of Incenp contains a centromere targetting domain, and is also a binding site for the other two passengers, Borealin and Survivin [Klein et al., 2006].

There is evidence for other roles for the CPC in spindle assembly [Colombié et al., 2008],

possibly independent of the centrosome and chromosome mediated pathways [Maresca et al., 2009].

2.12.4 Borealin

The *Drosophila* orthologue of Borealin was identified in a screen for genes required for cytokinesis by Eggert et al. [2004], who showed that it possessed the characteristic CPC localisation pattern throughout mitosis and was required for Histone H3 phosphorylation by Aurora B. RNAi knockdown leads to incorrect chromosome segregation, as with the other CPC emembers [Hanson et al., 2005]. It is the only passenger that has been shown to be capable of directly binding DNA [Gao et al., 2008], and in common with its human homologue co-precipitates with Incenp but not Aurora B.

Two nuclear localisation signals in human Borealin have been found, and analysis of truncation mutants suggests the C-terminal is required for localisation of Borealin to centromeres [Gassmann et al., 2004] but the N-terminal is required for binding to Incenp [Klein et al., 2006], so only full length Borealin can function correctly during prometaphase. RNAi knockdown of human Borealin leads to errors in kinetochore-microtubule attachments, implying a role for the CPC in localising Aurora B to the kinetochores. The functional role of Borealin during anaphase is less clear since errors originating in prometaphase may accumulate as mitosis progresses, and in addition the other passengers can localise to the central spindle with both N- and C- terminal truncations [Gassmann et al., 2004]. However, in *Drosophila* spermatocytes Australin, a paralogue of Borealin, is essential for central spindle formation [Gao et al., 2008].

2.12.5 Deterin/Survivin

Deterin (the *Drosophila* homologue of Survivin) was originally identified as a member of a family of proteins containing a baculovirus inhibitor of apoptosis (IAP) domain [Jones et al., 2000]. In vertebrates the localisation pattern of Survivin is the same as that of the other passengers and it has been shown to be a core member of the CPC, confirming its role as a mitotic regulator [Lens et al., 2006]. Its dual function as an apoptotic inhibitor has also been confirmed, but the mechanism of action is still unclear [Altieri, 2006].

Our understanding of the behaviour of Survivin as part of the CPC is also incomplete. The C-terminal (which contains a coiled-coil domain) is required for localisation of the CPC to the central spindle and midbody, whereas the N-terminal containing the BIR domain is required for centromere localisation [Lens et al., 2006]. This was confirmed by Yue et al. [2008] in chicken DT40 cells, but in contrast to other studies they concluded that it was possible to reach metaphase without any obvious defects in the absence of Survivin, conflicting with the findings of Kelly et al. [2010].

2.12.6 Putative CPC members

There are two additional proteins, TD-60 [Mollinari, 2003, Rosasco-Nitcher et al., 2008, Duncan, 2011] and EVI-5 [Faitar et al., 2005], which have been shown to associate with the core CPC in organisms other than *Drosophila*, though the nature of their associations are unclear and they may just be involved in passing interactions

2.12.7 Interactions with Polo

Polo (Plk1 in mammals) is another essential mitotic kinase which shares similarity in aspects of localisation to Aurora B. At prophase it localises to centrosomes where it contributes to spindle formation, but only after binding to Incenp which has been phosphory-

lated by Cdk1 [Goto et al., 2006], then relocates to the kinetochores in prometaphase and metaphase, before finally moving to the central spindle in anaphase and the midbody in telophase [Petronczki et al., 2008].

2.12.8 Central spindle

The central spindle forms during anaphase from the bundling of anti-parallel microtubules between the separating daughter chromosomes, and is vital in positioning the actomyosin based contractile ring [Glotzer, 2009].

The CPC is known to have a vital role, with Aurora B phosphorylating the centralspindlin complex, a heterotetramer consisting of MKLP1 which is phosphorylated by Aurora B, and MgcRacGAP which may also be phosphorylated by Aurora B, and eventually leads to localised actin polymerisation at the equatorial cell cortex. In humans Incenp is required for correct MKLP1 localisation in late mitosis [Zhu et al., 2005], but there is conflicting evidence of its relationships with the CPC in *Drosophila* [Adams, 2001, Giet and Glover, 2001, Cesario et al., 2006].

MKLP2 is also found on the central spindle and is required for the transferral of the CPC from centromeres to the central spindle in anaphase [Gruneberg et al., 2004, Cesario et al., 2006]. The question of exactly how this controls the position of the contractile ring is unknown, though a phosphorylation gradient caused by Aurora B may be important [Fuller et al., 2008].

Chapter 3
Chromosomal passenger complex experiments

The chromosomal passenger complex (CPC) is a complex of four proteins which plays a vital role in the regulation of mitosis (see Section 2.12). Fully inhibiting any of the members leads to severe mitotic defects making it difficult to probe the underlying mechanisms, but by using quantitative techniques we hope to gain insight by measuring the effects of partial inhibition.

This work is split into two chapters. Sections 3.2–3.7 in this chapter describe the experimental methods used to create the fluorescent fly lines for imaging, and the associated biochemistry to verify the lines. This is followed by the imaging protocol in Section 3.8 that was used to obtain the images for this work.

Following on from this Chapter 4 describes the results of the experimental work, and the analysis algorithms and results used to quantify the differences in expression levels and localisation following perturbation of the complex.

3.1 Collaborations

Part of the work related to this chapter was done in collaboration with others in the laboratory:

- *Borealin related-GFP* was cloned by a former researcher, Gao Shan [Gao et al., 2008].
- The anti-Incenp coiled-coil peptides (Section 3.8.3.2) were designed and created by Dr. Pete Jones.

Unless indicated the remaining experimental work was done by myself.

3.2 Overview

The primary aim of this work was to obtain fluorescently tagged endogenously expressed proteins of the four core members of the CPC: Borealin, Aurora B, Deterin (Survivin) and Incenp.

This involved the following principle stages:

1. Create a copy of (clone) the genes of interest (Section 3.4).
2. Attach a GFP tag, and prepare DNA samples for stable insertion into a fly line (Section 3.5).
3. Cross fly lines with a driver to induce *in-vivo* expression (Section 3.6).

Abbreviations used in this section:

Amp Ampicillin

dH_2O Distilled water

ddH_2O Double distilled water (Millipore Milli-Q)

Kan Kanamycin

PCR Polymerase chain reaction

RT Room temperature

3.3 Gene sequences

Nucleotide sequences for the four core CPC members in *Drosophila melanogaster* were obtained from Flybase (Table 3.1). Alignments of the *Drosophila* and human transcripts

Gene (synonym)	Symbol	Annotation ID	FlyBase ID	CDS length (bp)	Protein weight (kDa)
Aurora B (IplI-aurora-like kinase)	ial-PA	CG6620	FBpp0079769	990	38.3
Borealin (Borealin related)	Borr-PA	CG4454	FBpp0079416	948	34.7
Deterin (Survivin)	Det-PA	CG12265	FBpp0082849	462	17.5
Incenp (Inner centromere protein)	Incenp-PA	CG12165	FBpp0088035	2268	83.5

Table 3.1: CPC genes in *Drosophila*. CDS: Coding sequence, measured in base pairs (bp)

are provided in Appendix C. To save space genes and proteins will sometimes be referred to in abbreviated form: AurB, Bor, Det and Inc.

3.4 Entry vectors

3.4.1 Polymerase chain reaction

The four genes of interest were amplified using the blunt ended PFX Polymerase (Invitrogen), in accordance with the manufacturer's protocol (1 µl forward primer, 1 µl reverse primer, 1 µl template DNA, 5 µl amplification buffer, 1.5 µl dNTPs, 1 µl $MgSO_4$ (50 mM), 30 µl ddH_2O, 0.5 µl PFX). The thermal cycle was: initial denature 94 °C 5 min followed by 35 cycles of 94 °C 1 min denature, 57 °C 1 min anneal, 68 °C 1 min (*deterin*, *aurora B*) or 3 min (*incenp*) elongation, and a final elongation of 68 °C 10 min followed by holding at 4 °C. Primers (listed in Table 3.2) were ordered from MWG Biotech, designed with a 5' CACC overhang (required for the Gateway entry vector) and without a terminating stop codon, and were diluted to 50 µM before use.

Template DNA was either complementary DNA obtained by a colleague from 0 h to 3 h old wild-type embryos (*deterin*, *aurora B*), or a construct previously cloned by a colleague

Incenp-Fwd	CAC CAT GGA GGA CAT CTT GGG CGT GC
Incenp-Rev	GTA TTT CGG CAG CTC GGA ATA GC
Aurora B-Fwd	CAC CAT GAC GCT TTC CCG CGC GAA
Aurora B-Rev	ATT TCT GGC CGT GTT CTC CTT GC
Deterin-Fwd	CAC CAT GGA ATC GCC AGT GGT AAA CGA AG
Deterin-Rev	TTT TTG ATT ACG CGT AAA CTC ATC TAG AC

Table 3.2: Primers used to amplify the four genes to be cloned. Sequences are writen 5' to 3'. Fwd indicates a primer designed to anneal to the 5' end of the nucleotide region of interest, Rev indicates a primer designed to anneal to the reverse-complemented 3' end.

(FLAG-tagged *incenp*). PCR products were verified using DNA gel electrophoresis (1 % agar (Sigma), ethidium bromide at 0.5 µg ml^{-1} (Sigma)), and purified using the QIAquick Gel Extraction Kit (QIAGEN) following the manufacturer's instructions with the purified DNA being eluted into 40 µl ddH_2O. 5 µl of purified product was run on a gel to verify the size of the product and to provide an estimate of the concentrations of each product.

3.4.2 TOPO cloning

Entry vectors (Figure 3.1a) were created using the pENTR Directional TOPO Cloning Kit (Invitrogen) with the pENTR/D-TOPO vector in accordance with the manufacturer's instructions. Approximately 10 ng *deterin*, 10 ng *aurora B* and 30 ng *incenp* PCR products were used, and the reaction was incubated for 30 min RT.

The entry vector was transformed into DH5-α chemically competent *E. coli* (New England Biolabs). The entry vector was added into a 50 µl vial of cells and incubated on ice for 10 min. The cells were heat shocked in a 42 °C water bath for 30 s, then placed on ice for 2 min. 250 µl RT SOC medium (Invitrogen) was added to the tubes which were placed in a shaking incubator (200 rpm) 37 °C 1 h. 200 µl was spread onto selective LB/Kan (50 µg µl^{-1}) agar plates and incubated at 37 °C overnight, then stored at 4 °C.

Individual colonies were picked and incubated overnight (37 °C 200 rpm) in 3 ml selective LB/Kan (50 µg µl^{-1}) broth. Plasmid DNA was extracted and purified using the QIAprep

M13 (-20) Fwd	GTA AAA CGA CGG CCA G
M13 Rev	CAG GAA ACA GCT ATG AC
Incenp-SeqFwd1	AGC AAC TCC CAG TTG CTG GC
Incenp-SeqFwd2	AGC GCG AGA AAC TCA GTA CC
Incenp-SeqFwd3	ACT GGG TCC TGC GCC GAC GA
Incenp-SeqFwd4	ACA CTA CCA TTA AGC TCG GC
Aurora B-SeqFwd1	CCG CAC ATC CTG CGC CTG CT
GFPr	CTT CGG GCA TGG CGG ACT TG

Table 3.3: Primers used for the sequencing reactions. Primers in the top part of the table were used for sequencing the entry vectors, the primer in the bottom row was used for verifying the expression vector sequence.

Spin Miniprep Kit (QIAGEN) following the manufacturer's protocol. The size of the insert was verified with a diagnostic digest: 2 µl DNA, 0.5 µl EcoRV, 0.5 µl Not1, 2 µl Buffer 3, 0.2 µl BSA and ddH_2O to total 20 µl was incubated in a 37 °C water bath for 30 min (enzymes and buffer from New England Biolabs).

3.4.3 Sequencing

The entry vectors were verified by sequencing. The region containing the insert was PCR amplified using 0.5 µl BigDye Terminator v3.1 (Applied Biosystems), 3.2 µl (1 µM) primer, 2 µl sequencing buffer, 1.5 µl entry vector and 2.8 µl ddH_2O. The thermal cycle was 25 cycles of 96 °C 30 s denature, 50 °C 15 s anneal, 60 °C 4 min, followed by holding at 4 °C.

M13 forward and reverse primers (MWG Biotech) were used. Additional intermediate primers were also used for *incenp* and *aurora B* (Table 3.3).

Products were purified by the addition of 2 µl 125 mM EDTA and 1 µl 3 M NaOAc (pH 5.2). DNA was precipitated by adding 30 µl 100 % ethanol and incubating at RT 15 min. Samples were centrifuged at 13 000 rpm 15 min and the supernatant was removed. Pellets were washed with 200 µl 70 % ethanol followed by centrifugation at 13 000 rpm 5 min. Supernatant was removed and pellets were air dried, then sent for sequencing at an

Gene and sample number	Nucleotide position	Nucleotide change	Amino acid substitution
ial-7	213	G → A	
	222	C → T	
	258	A → G	
	267	G → A	
	378	C → A	
	483	A → G	
	489	G → A	
	606	C → T	
	648	C → T	
	675	G → A	
	726	C → T	
Det-3	14	T → G	5: V → G
	247	G → C	83: E → Q
	381	G → A	
Incenp-3	500	T → C	167: L → P
	543	T → C	
	598	C → T	
	601	G → A	201: E → K
	1172	A → G	391: N → S
	1407	G → —	470: Q → —
	1408	C → —	
	1409	A → —	
	2169	A → G	

Table 3.4: Mutations found in the sequence of the chosen entry clones. The position of any base changes is given together with the original and mutated nucleotide, and the corresponding amino acid position and change (if any). — indicates a deletion.

in-house facility.

Chromatograms were analysed using FinchTV (Geospiza) and compared with reference sequences. The samples showing minimal changes from the reference sequences were chosen, and are listed in Table 3.4.

Glycerol stocks (0.5 ml glycerol, 0.5 ml LB/Kan containing grown up transformed *E. Coli*) of the selected entry vectors were made and stored at −80 °C.

3.5 Expression vectors

The final expression vectors each containing one of the CPC genes fused to a GFP tag were created using the Gateway system [Hartley, 2000]. The destination vector (pPWG, C-terminal EGFP UASp promoter) was supplied as a −80 °C *E. Coli* DMSO stock by the Drosophila Genomics Resource Center at Indiana University, and was grown up in LB/Amp (50 µl ml^{-1}) at 37 °C overnight and purified using the QIAGEN Midiprep Kit following the manufacturer's protocol.

The UAS (universal activator sequence) promoter originates from yeast and is not present in *Drosophila*, allowing us to control the expression of the tagged genes [Brand and Perrimon, 1993].

3.5.1 LR reaction

The LR recombination reactions (Figure 3.1b) were performed as half-reactions in accordance with the manufacturer's instructions for the Clonase II enzyme (Invitrogen): 2 µl entry vector, 2 µl destination vector, 1 µl Clonase II, incubate overnight, then add 0.5 µl Proteinase K. Competent cells were transformed as before, plated onto LB/Amp (50 µl ml^{-1}) selective agar plates and incubated at 37 °C overnight. Colonies were picked and grown up overnight in LB/Amp (50 µl ml^{-1}) at 37 °C 200 rpm.

Expression vectors were extracted and purified using the QIAGEN Miniprep Kit as before. The vectors were verified using a restriction digest as before, using 2 µl DNA, 0.5 µl Xho1, 2 µl Buffer 2, 0.2 µl BSA and ddH_2O to total 20 µl.

Vectors were sequenced using the protocol previously given across the junction of the gene insertion and the C-terminal EGFP tag to check the insertion was correct, using the GFPr primer from Table 3.3.

The selected expression vectors (Deterin-pPWG-1, Incenp-pPWG-1, Aurora B-pPWG-

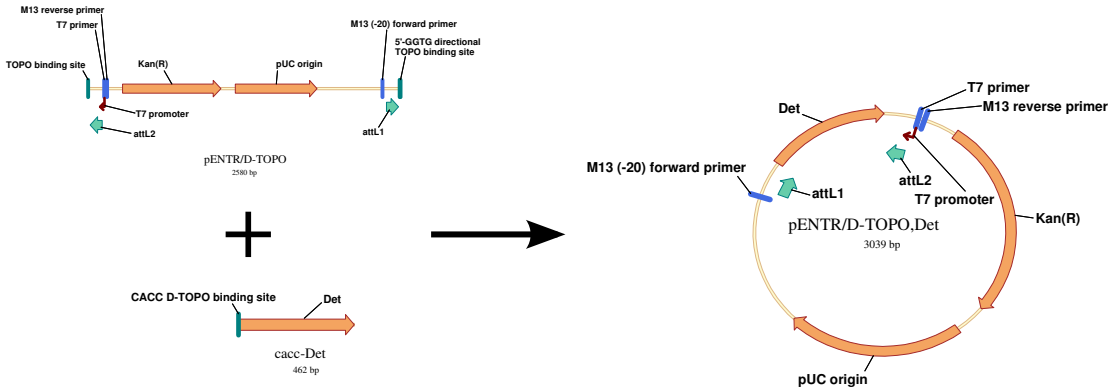
1) were grown up and purified using the QIAFilter Midiprep Kit (QIAGEN) following the manufacturer's protocol. Expression vectors for the four genes were prepared as 80 µl DNA, 200 µl 100 % ethanol and 4 µl NaOAc, and sent to Bestgene Inc (California, USA) for transformation into w1118 wild-type files.

3.6 Fly crosses

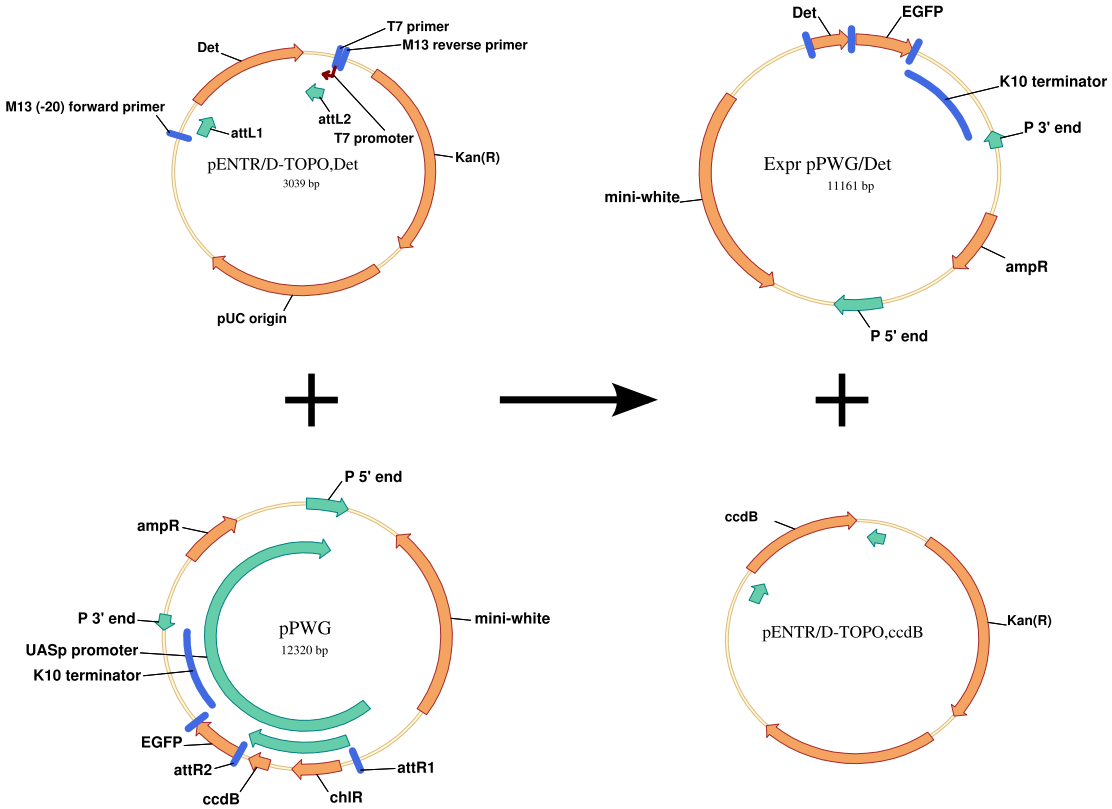
All fly work was performed at 25 °C unless otherwise stated. Fly stocks were maintained on standard yeast/dextrose media (1 % agar, 7.5 % dextrose, 5 % yeast, 6.7 % organic wholemeal flour, 0.25 % Nipagen, 0.3 % propionic acid).

Unbalanced transformed flies were supplied by Bestgene. The expression vector contained the *mini-white* reporter gene (or) allowing selection by eye colour, so double balanced stocks for insertions on the second chromosome could be created as shown in Figure 3.2a. An analogous set of crosses was done for insertions on the third chromosome.

The GFP lines contain a promoter which is only activated by an exogenous GAL4 driver; in this case a maternal-α-tubulin GAL4 line (MaT-GAL4) was used to induce expression in embryos [Rørth, 1998]. Stable lines constitutively expressing each of the GFPs were created using the crosses shown in Figure 3.2b. For insertions on the third chromosome a similar set of crosses using a MaT-GAL4 line expressing on the second chromosome was used.



(a) TOPO reaction



(b) LR reaction

Figure 3.1: Gateway entry and LR reactions for *deterin*. (a) The TOPO reaction creates a plasmid termed the entry vector, containing the cloned gene of interest (Deterin). (b) The LR reaction involves a recombination between the entry vector and the destination vector pPWG, leading to the formation of the final expression vector pPWG/Det, and a waste plasmid containing the lethal *ccdB* gene.

$$w;\frac{\text{or}}{+};\frac{+}{+}\quad ♂\quad \times\quad w;\frac{Tft}{CyO};\frac{MKRS}{TM6B}\quad ♀$$
$$\downarrow$$
$$w;\frac{\text{or}}{CyO};\frac{+}{TM6B}\quad ♂\quad \times\quad w;\frac{\text{or}}{CyO};\frac{+}{MKRS}\quad ♀$$
$$\downarrow$$
$$w;\frac{\text{or}}{CyO};\frac{MKRS}{TM6B}\quad ♀♂$$

(a) Double balancing

$$w;\frac{\text{or}}{CyO};\frac{MKRS}{TM6B}\quad ♂\quad \times\quad w;\frac{Tft}{CyO};MaT\text{-}GAL4\quad ♀$$
$$\downarrow$$
$$w;\frac{\text{or}}{CyO};\frac{MaT\text{-}GAL4}{TM6B}\quad ♀♂$$

(b) Promoter cross

Figure 3.2: Fly crosses. or indicates the *mini-white* gene that codes for red eyes, and is used as a marker for the inserted transgene.

3.7 Biochemistry

3.7.1 Preparation of tissues for biochemistry

3.7.1.1 Embryo collection

Adult flies were placed in a plastic cage containing air vents, with a fruit juice agar plate as a base. The plates were created by boiling 9 g agar (B. T. P. Drewitt, London), 10 g sucrose (Silver Spoon) and 300 ml dH_2O, adding 100 ml apple juice (Tesco Value), and pouring a thin layer of agar into 50 mm diameter petri dishes. A small blob of yeast paste (powdered yeast and dH_2O) was placed on the centre of the agar plate before attachment to the cage.

The cages were kept at 25° for the required period of time, the embryos were then brushed into a perforated dish, washed with dH_2O, and dechorionated by immersion in thin bleach (Tesco value) for 3 min to 5 min. Embryos were rinsed with dH_2O, dried and placed in a 1.5 ml tube. If not used immediately they were flash frozen in liquid nitrogen and stored at −80 °C.

3.7.1.2 Embryo extracts

30 µg to 80 µg of frozen embryos were homogenised in 50 µl PBST (PBS tablets (Sigma) dissolved in ddH_2O with 0.5 % Tween-20 (Fisher Scientific)) in a 1.5 ml tube, mixed with 50 µl 2× sample buffer, and incubated at 90 °C 10 min.

3.7.2 Western blotting

A 5 ml 10 % polyacrylamide resolving gel was made up from 1.25 ml 40 % acrylamide stock (BioRad), 1.25 ml 1.5 M TrisHCl pH 8.8, 50 µl 10 % sodium dodecyl sulphate (SDS, Sigma), 50 µl 10 % ammonium persulphate, 2.4 ml ddH_2O, followed by 5 µl TEMED (Sigma). A 2.5 ml stacking gel consisted of 0.3 ml 40 % acrylamide stock, 625 µl 1.5 M TrisHCl pH 6.8, 25 µl 10 % SDS, 25 µl 10 % ammonium persulphate, 1.525 ml ddH_2O, followed by 5 µl TEMED (Sigma).

Gels were loaded with 5 µl of ladder (PageRuler prestained, Fermentas) and 10 µl of each protein sample, and run at 150 V in 1× TGS running buffer (10× stock made up from 30 g Tris (Sigma), 144 g glycine (Sigma), 10 g SDS, ddH_2O to 1 l). Protein was transferred to a nitrocellulose membrane (Amersham Hybond ECL) at 250 mA in 1× transfer buffer (100 ml 10× TGS, 200 ml methanol, 700 ml ddH_2O).

The membrane was washed in PBST, incubated in blocking buffer (5 % non-fat milk powder (Tesco) in PBST) for 1 h on a rocker, and incubated overnight at 4 °C with the primary antibody in blocking buffer (1:333 mouse monoclonal anti-GFP, Santa Cruz Biotech-

nology). The membrane was washed 3 times in PBST for 15 min and the secondary antibody added in blocking buffer for 2 h (1:5000 rabbit anti-mouse HRP).

The membrane was washed 3 times in PBST 15 min, immersed in SuperSignal West Pico Chemiluminescent Substrate (Thermo Scientific) reagents and gently agitated for 2 min. The blot was then exposed on X-ray film (Thermo Scientific CL-XPosure) for 10 s or 1 min and developed.

3.8 Imaging

3.8.1 Embryo fixation

0 h to 4 h old embryos were collected and dechorionated as in Section 3.7.1.1, and immediately fixed by agitation in a 1.5 ml tube containing 500 µl heptane and 500 µl methanol. Fixed embryos sunk to the bottom and were pipetted into a new 1.5 ml tube, washed once with 1 ml methanol, followed by three 15 min washes with 1 ml PBST on a rocker to rehydrate the embryos.

The embryos were blocked in 1 ml 3 % bovine serum albumin (BSA) on a rocker for 20 min, the block was removed and replaced by the primary antibody (1:500 mouse anti-α-tubulin (DM1A, Sigma) in 1 ml block), and the embryos incubated overnight at 4 °C on a rocker.

This was followed by three 10 min 1 ml PBST washes, and incubation at room temperature for 2 h on a rocker with the secondary antibody (1:1000 Rhodamine Red-X goat anti-mouse IgG (Invitrogen Molecular Probes) in 1 ml block). The embryos were washed in 1 ml PBST for 10 min, the PBST removed, and the embryos suspended in 40 ul mounting media (Vectashield with DAPI H-1200, Vector Laboratories) which was then pipetted onto a microscope slide and a coverslip placed on top. Slides were stored in a light-tight box at

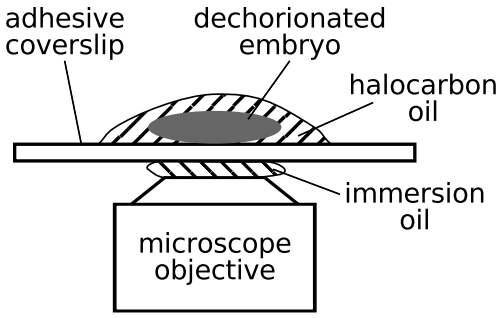


Figure 3.3: Microscope slide for live imaging of *Drosophila* embryos.

4 °C prior to imaging.

3.8.2 Live embryo preparation

Heptane glue was prepared in advance by dissolving the adhesive on double-sided cellulose tape (Scotch) in heptane.

A 20 × 50 mm microscope coverslip was prepared by drawing a line of heptane glue down the middle of the slide, and allowing it to air dry. 0 h to 1 h embryos were collected from a cage as in Section 3.7.1.1 but were not dechorionated. Instead they were aged for 1 h at 25 °C, and 10–15 embryos were dechorinated by rolling them on double-sided cellulose tape with tweezers. These embryos were stuck onto the prepared coverslip, and covered in halocarbon oil (Sigma) to prevent them from dehydrating (see Figure 3.3).

3.8.3 Microscopy

3.8.3.1 Live imaging

Images were acquired on a Zeiss LSM 510 meta laser-scanning confocal microscope. An initial set of timelapse images showing the localisation of the CPC components were acquired with the following settings:

X-Y spatial resolution: 512 × 512 pixels of size 0.28 µm

Z stack: 5 planes with 1 µm spacing

Temporal resolution: 5 s

Objective: 63×

Pixel time: 1.6 µs

Zoom: 1×

3.8.3.2 Microinjection into embryos

Three proteins were chosen for injection into embryos, all were designed and created by other members of the laboratory.

Anti-Borealin: Purified polyclonal antibody against full length Borealin [Gao et al., 2008], which was expected to functionally inhibit the interactions of Borealin with the CPC and its substrates.

Homologous coiled-coil domain: Recombinant coiled-coil domain taken from human Incenp [Tseng et al., 2010], which may either competitively bind to substrates of the CPC, or effectively dimerise with the same coiled-coil domain of Incenp, thus inhibiting the localisation and function of the CPC.

Predicted coiled-coil domain: The subsequence of *Drosophila* Incenp which is predicted by the *COILS* server [Lupas et al., 1991] to form a coiled-coil domain.

Proteins for injection were buffer exchanged into injection buffer (25 mM HEPES pH 7.6,10 mM KCl), and concentrated to 5 mg ml^{-1} to 10 mg ml^{-1}.

Embryos were prepared on a slide as previously described in Section 3.8.2, except they were partially dehydrated in open air after mounting for 5 min to 10 min before being covered with halocarbon oil.

The protein to be injected was briefly centrifuged (6000 rpm 30 s) to remove any solids, and a needle (Eppendorf Femtotips) was loaded with 2 µl to 5 µl of protein using a Microloader pipette tip (Eppendorf).

Injections were carried out using an Eppendorf Injectman NI2, with the microscope set to the 20× objective in widefield mode. A suitably staged embryo was focussed at its midplane, and the tip of the microinjection needle lowered into the same plane. The microscope stage was moved laterally until the embryo was pierced, pressure applied to the syringe until there was visual confirmation that some of the protein had entered the embryo, and the microscope stage moved back.

The embryo was then imaged as described in Section 3.8.3.1 with the following changes:

Z stack: 10 planes with 1 µm spacing

Temporal resolution: 10 s

These changes were intended to handle variation in the depth of the nuclei over one or more cycles of division— the cost and difficulty of injecting an embryo meant that it was preferable to sacrifice the temporal resolution in exchange for certainty in capturing the dynamics of the CPC throughout mitosis.

Chapter 4
Quantitative analysis of disruptions to the chromosomal passenger complex

In the previous chapter we described the experimental methods used to clone the CPC genes and tag them with GFP, and to subsequently express these proteins in embryos.

This chapter begins by showing that these methods successfully led to the creation of embryos which expressed each of the four GFP tagged CPC genes, along with a description of the localisation of these proteins throughout mitosis. This includes confirming that Deterin is the *Drosophila* orthologue of Survivin, which to our knowledge has not been demonstrated before.

The motivation for wishing to apply quantitative analysis methods to this data is given in Section 4.2, followed by the novel algorithms used to analyse this data. An early version of this work was presented at the IEEE International Symposium on Biomedical Imaging 2011. Finally we present some preliminary results in Sections 4.4–4.5.

4.1 Initial results

4.1.1 GFP expression

Expression of the GFP tagged proteins was successfully verified by western blotting embryo extracts, and probing with anti-GFP (Figure 4.1).

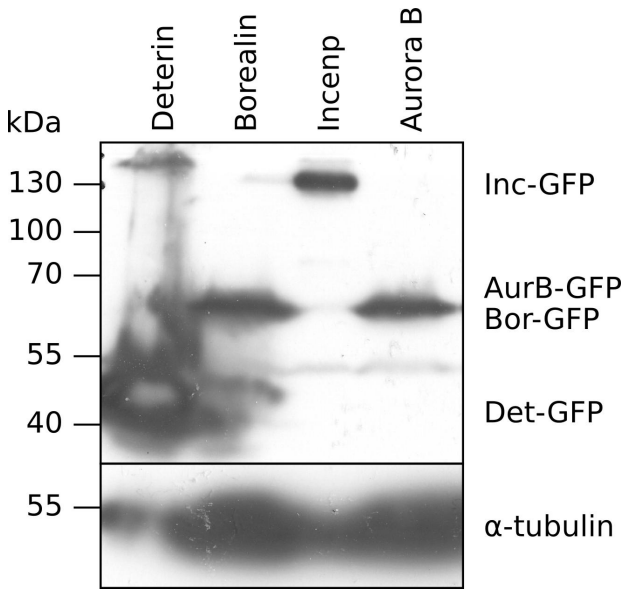


Figure 4.1: Western blot showing expression of GFP tagged CPC proteins in embryo extracts. The GFP tag has an approximate weight of 27 kDa, weights of the CPC members are given in Table 3.1. α-tubulin (50 kDa) was used as a loading control.

4.1.2 Live imaging

The *Drosophila* syncytium contains multiple nuclei which rise to the cortex and are amenable to confocal imaging. An example of the typical field of view can be seen in Figure 4.2, which shows the localisation of Deterin-GFP.

A closeup of another member, Aurora B-GFP, is shown in Figure 4.3. During prophase Aurora B appears to localise to the condensing chromatin, before increasing in concentration on the centromeres of the chromosomes in prometaphase, and eventually forming a line of dots at metaphase indicating the positions of the aligned chromosomes. However a small amount seems to remain on the chromosome arms. On entry to anaphase the CPC relocates to the spindle midzone as a series of streaks, corresponding to its localisation on the central spindle, before condensing down into the midbody and eventually fading away before entry into the next cycle of mitosis. During interphase it appears to have a low though non-zero presence within the nuclear envelope.

Aurora B-GFP, Deterin-GFP and Borealin-GFP show a similar localisation to each other in embryos (Figure 4.4) as is expected from previous observations of other species, and is also consistent with the idea that they primarily exist as a single complex. The embryos do not exhibit any obvious defects, which suggests that the presence of the GFP tag does not adversely affect the structure and therefore function of these proteins, although it is theoretically possible for the protein to localise correctly without being fully functional.

These images also confirm that Deterin is the *Drosophila* orthologue of Survivin. Existing work has either assumed it is a Survivin orthologue based on sequence homology, or studied its role as a potential regulator of apoptosis [Jones et al., 2000], but it has not to our knowledge been explicitly shown to behave as a CPC member in *Drosophila*.

In contrast, Incenp-GFP embryos showed severe defects, with no embryos hatching into larvae. There are several possible explanations for this, including:

- Over-expression of Incenp may have deleterious effects.
- The semi-random insertion site of the transgene may have disrupted the function of another gene.
- The sequence mutations (Table 3.4) may have disrupted the function of the protein.

4.1.3 Fixed images

To confirm that the CPC localises to the spindle and kinetochores, and to obtain a better idea of the Incenp-GFP localisation, fixed embryos were imaged with α-Tubulin and DAPI as additional markers.

A still of Aurora B-GFP at metaphase is shown in Figure 4.5, where it is clear that Aurora B appears as concentrated spots on the DNA. It is also clear that the fainter surrounding fluorescence does coincide with the spindle.

Figure 4.6 shows Incenp-GFP in a fixed embryo. Incenp-GFP appears to form a fine

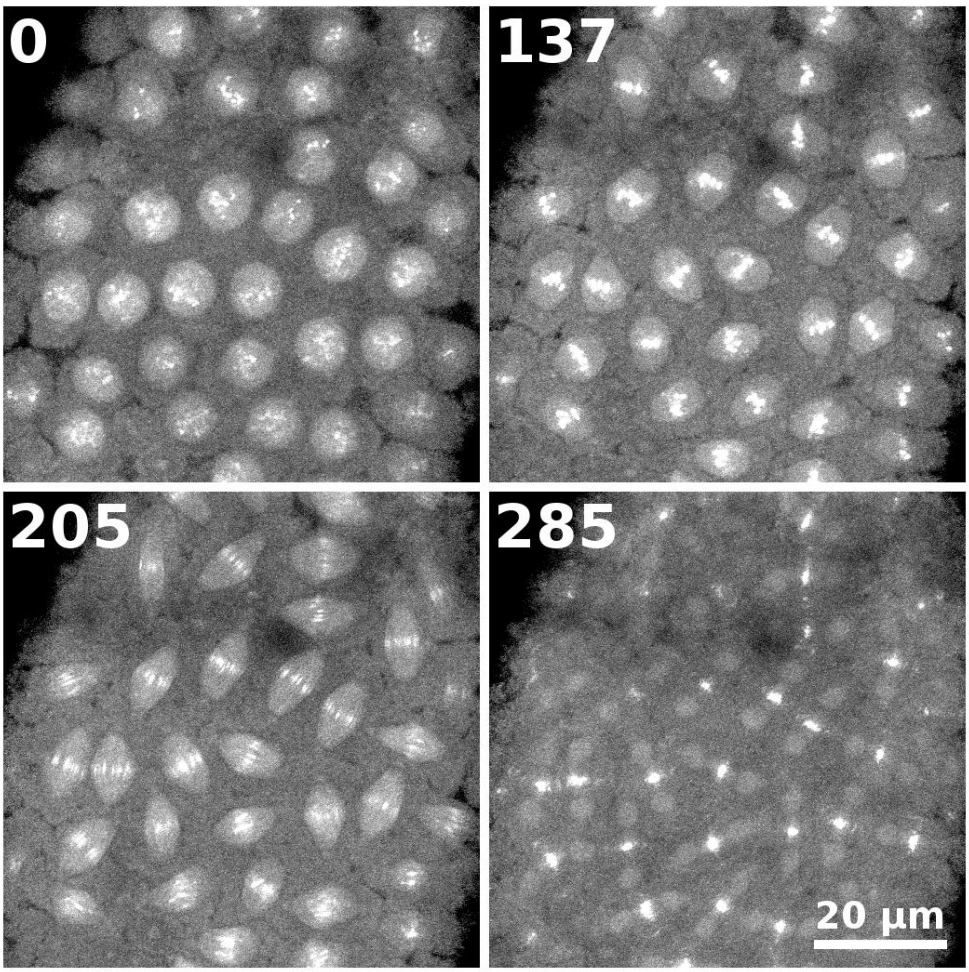


Figure 4.2: Localisation of Deterin-GFP during mitosis in a *Drosophila* syncytial embryo. This set of images illustrates the typical layout of the nuclei in an embryo. Times are given in seconds.

network of filaments reminiscent of microtubules during interphase, and comparison with the α-Tubulin localisation provides evidence of colocalisation, as previously observed by Mackay [1993].

The N-terminus of Incenp is sufficient for the centromeric and subsequent central spindle localisations of Incenp in HeLa cells whereas the C-terminus containing a coiled-coil domain is required for the binding of microtubules in interphase [Mackay, 1993, Mackay et al., 1998]. Since the GFP tag is at the C-terminal it therefore should not prevent Incenp from

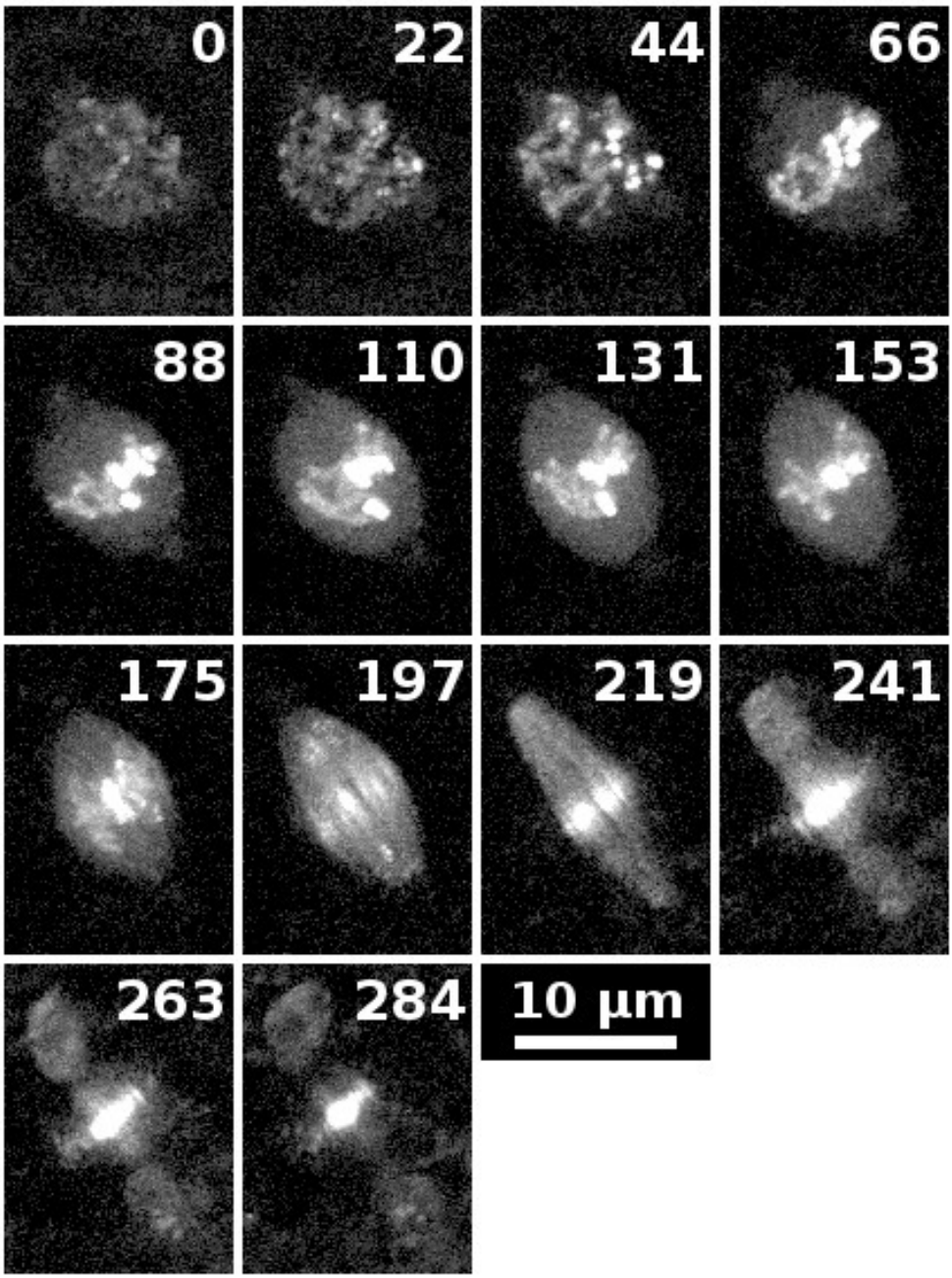
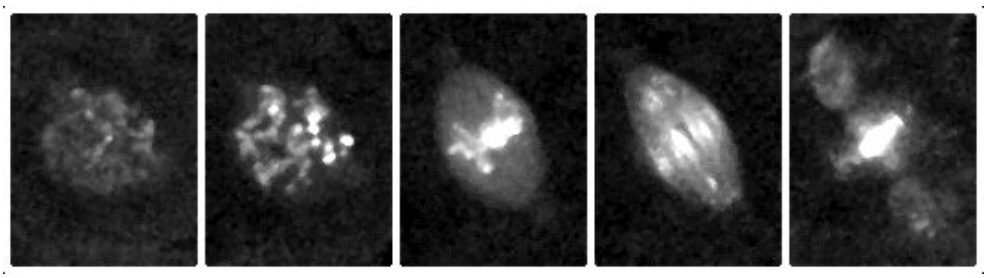
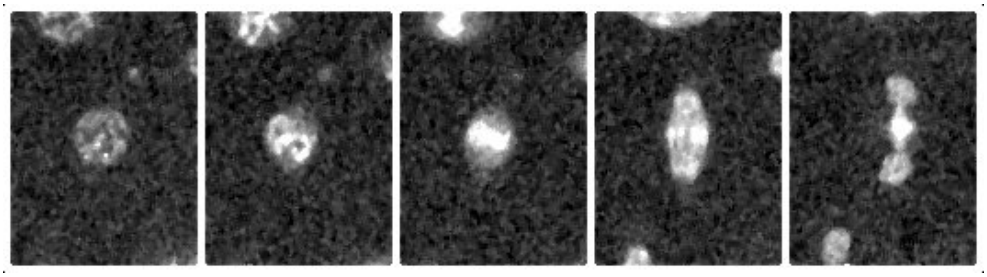


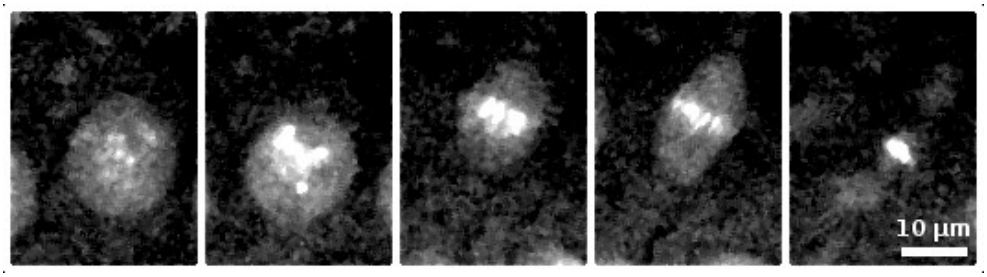
Figure 4.3: Closeup of the localisation of Aurora B-GFP during mitosis in a *Drosophila* syncytial embryo. Times are given in seconds. 0 s prophase, 153 s metaphase, 197 s anaphase, 284 s telophase.



(a) Aurora B



(b) Borealin



(c) Deterin

Figure 4.4: Localisation of Aurora B-GFP, Borealin-GFP and Deterin-GFP during mitosis. Left–right: prophase, prometaphase, metaphase, anaphase, telophase. Images have been denoised using a median filter, and intensity levels adjusted.

localising to the centromeres.

There is evidence that the relocalisation of the CPC to the central spindle requires the phosphorylation of Incenp by Cdk1 and Aurora B [Goto et al., 2006, Nakajima et al., 2011], so it is possible that phosphorylation of interphase Incenp is also required before the CPC can localise to centromeres.

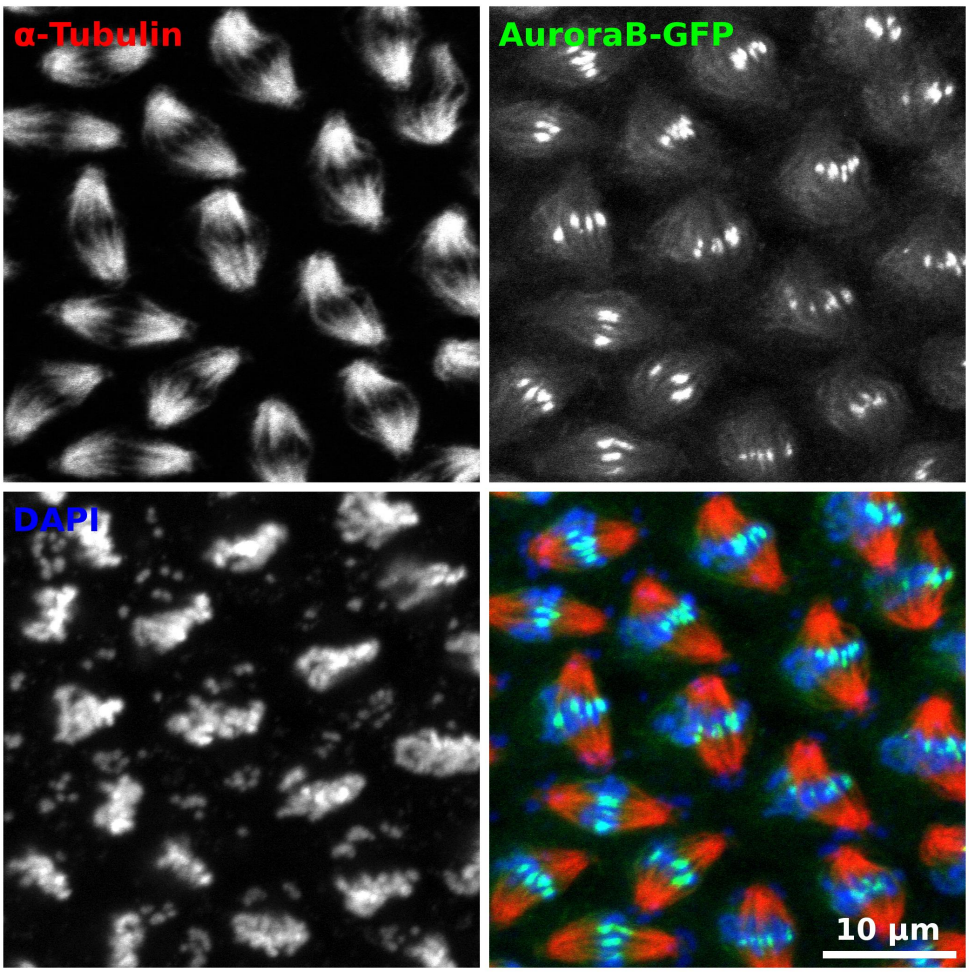


Figure 4.5: Aurora B-GFP metaphase localisation in fixed embryos labelled with α-Tubulin and DAPI. Images are a maximum projection of four z-planes 1 µm apart, with intensity levels adjusted.

4.1.4 Micro-injections

Since Aurora B is the catalytically active subunit of the complex and also resulted in the best timelapse images we concentrated on using Aurora B-GFP embryos for antibody injections. In this section we will present examples of embryos injected with an antibody and suggest some apparent differences which will be quantitatively analysed later.

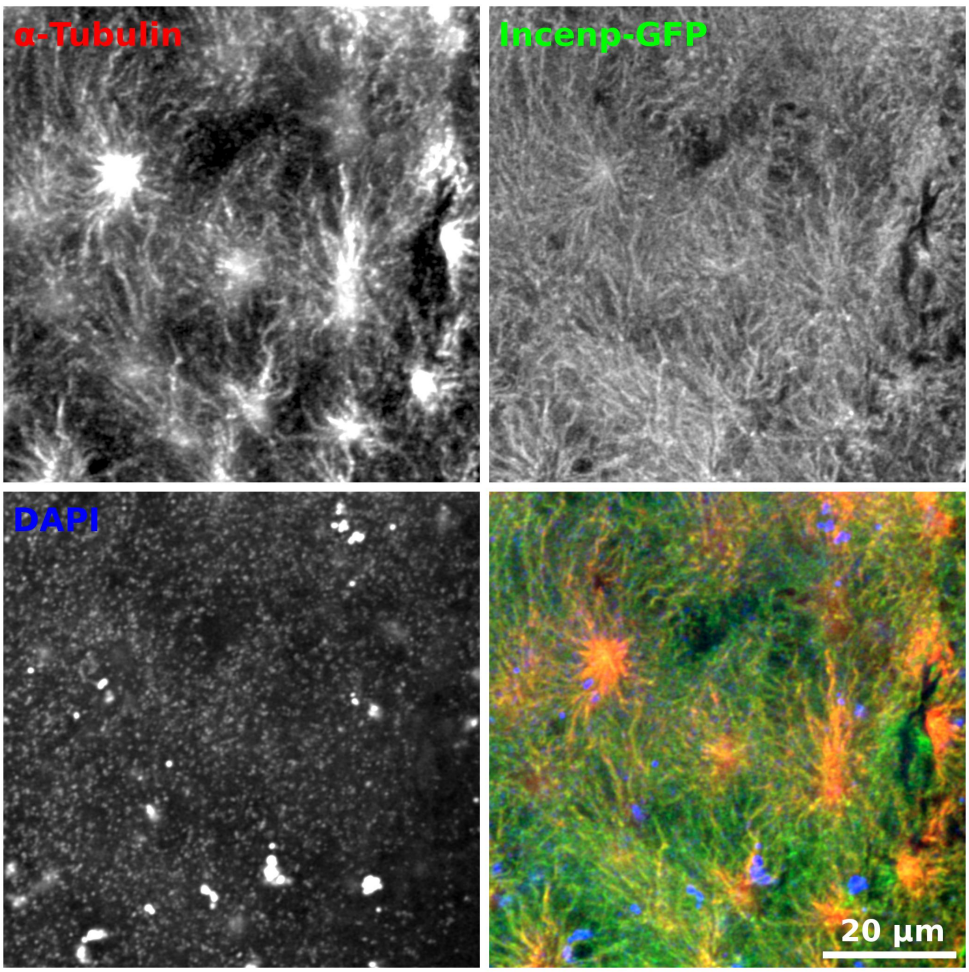
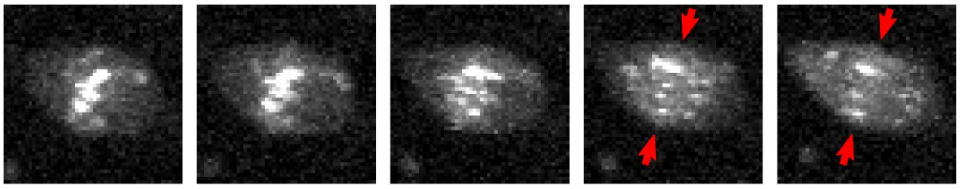


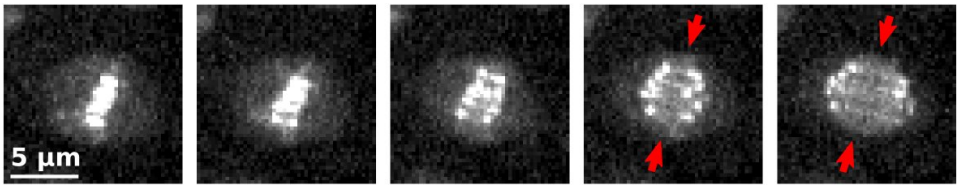
Figure 4.6: Incenp-GFP localisation in fixed embryos labelled with α-Tubulin and DAPI. Images are a maximum projection of two z-planes 1 µm apart, Gaussian filtered ($\sigma = 2$ pixels), and intensity levels adjusted.

4.1.4.1 Anti-Borealin injected into Aurora B-GFP

We expected that injection of the anti-Borealin antibody would lead to partial functional inhibition of Borealin, and hence disruption of the CPC. Visual examination of the images suggested a possible reduction in expression levels at the spindle midzone for a few seconds after the onset of anaphase (Figure 4.7).



(a) Control



(b) Anti-Borealin

Figure 4.7: Anti-Borealin micro-injected into an embryo expressing Aurora B-GFP. Consecutive frames (12 s apart) are shown starting at the metaphase-anaphase transition. An apparent difference in localisation between control and antibody injected embryo is visible around the central spindle (between the arrows).

4.1.4.2 Coiled-coil domains injected into Aurora B-GFP

In contrast to the observations with anti-Borealin, injection of the predicted coiled-coil domain (termed Pred-CC from now on) led to an apparent delay or failure of mitosis near the site of injection where the concentration of antibody would be greatest (Figure 4.8). Closer examination of the nuclei at the top right of the figure also suggests the levels of Aurora B on the chromosome arms are higher than normal.

Injection of the homologous coiled-coil peptide showed very similar results (not shown).

4.1.4.3 Predicted coiled-coil domain injected into dual labelled embryos

Although the injection of Pred-CC causes an apparent phenotype we should also try to establish whether it is purely an effect on the CPC, or whether there may be more general disruptions of mitotic structures. Given the key roles of Histone in structuring the DNA and Tubulin in forming the mitotic spindle observation of both these proteins simultaneously

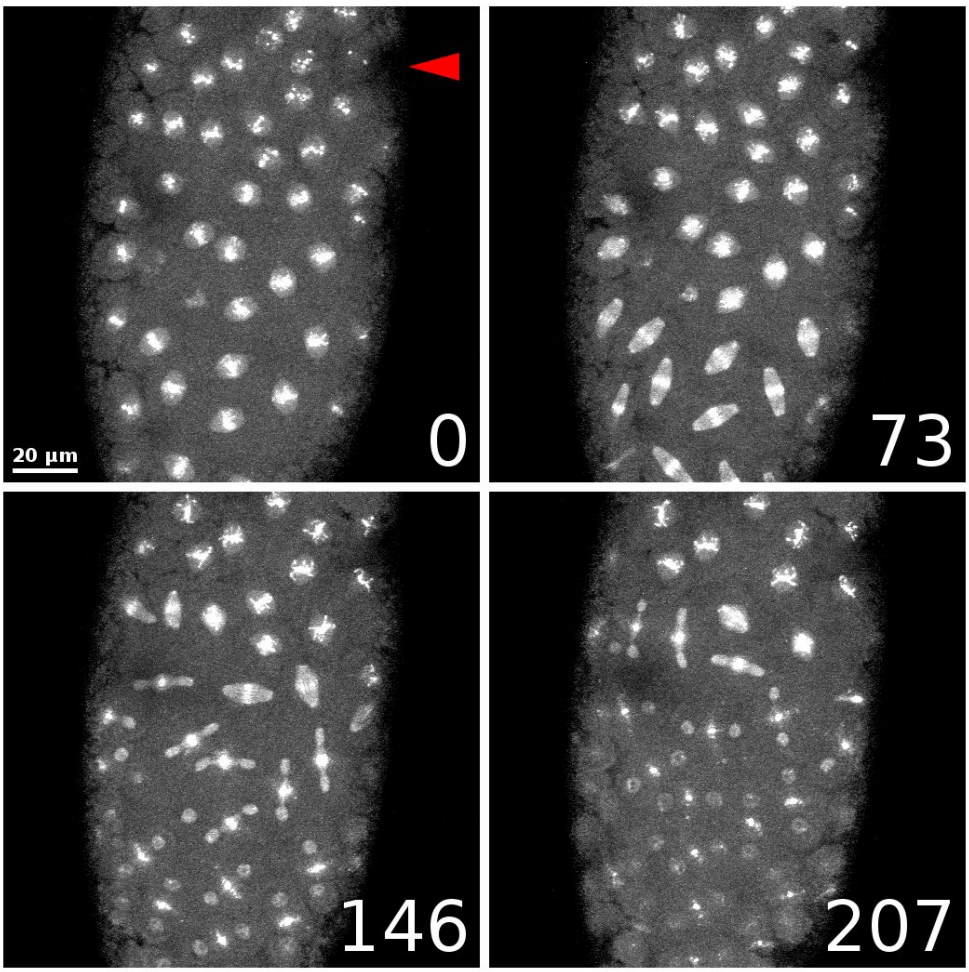


Figure 4.8: Predicted coiled-oil domain injected into an embryo expressing Aurora B-GFP. The embryo was micro-injected at the top right of the image (red arrowhead). Times are given in seconds.

using a dual-labelled fly line following injection was done, and is shown in Figure 4.9.

The injection appears to induce severe disruption of the embryo which is not seen in control injected embryos, ruling out the physical disturbance caused by the needle.

4.2 Motivation for further analysis

Some of the observations in Section 4.1 appear to be significant, but it is difficult to ascertain their importance, and we would like to quantify the differences. For instance, it

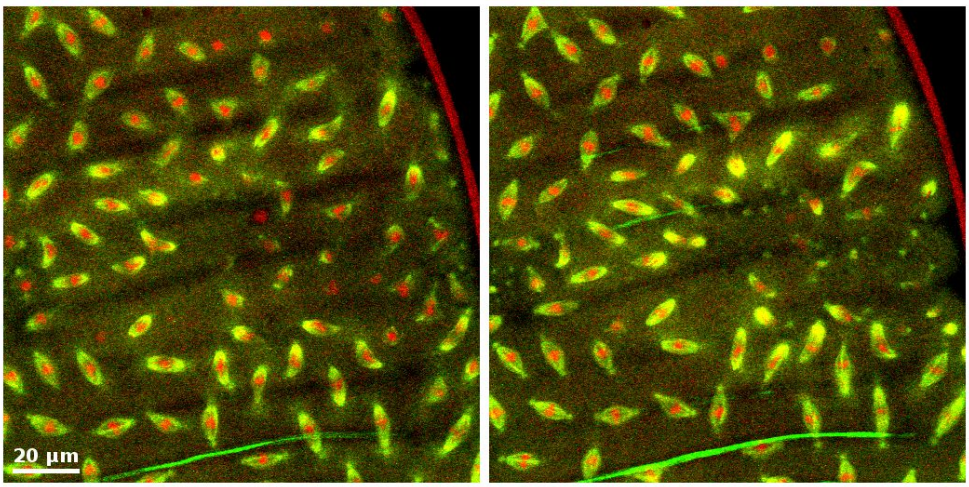


Figure 4.9: Predicted coiled-coil domain injected into dual labelled Histone-RFP (red) Tubulin-GFP (green) embryos. The embryo was micro-injected at the middle right of the image. The images were taken 72 s apart, both post-injection.

is difficult to say whether the apparent change in localisation of Aurora B-GFP at anaphase following injection of anti-Borealin is a real phenotype or just natural variation.

The injection of the coiled-coil domains also led to a clear phenotype, but given the difficulty in performing injections and the limited amount of peptide available it would be advantageous to identify in more detail the changes that have occurred.

However, we should also bear in mind that expert biologists still have powers of inference and pattern recognition which are unmatched by computers, and this suggests that semi-automated tools would be very useful for helping with the analysis of these images, either in terms of improving visualisation of the data to assist with comparisons of different samples, or later for extracting biologically useful measurements.

This would allow us to narrow down the possible underlying mechanisms of the disruption, and hopefully suggest a small set of focussed experiments to continue the research. Quantitative analysis should help to reduce the human bias that may cloud the analysis, but equally importantly the ability to extract more information from a limited dataset means researchers can make more efficient use of their time, and also of laboratory resources

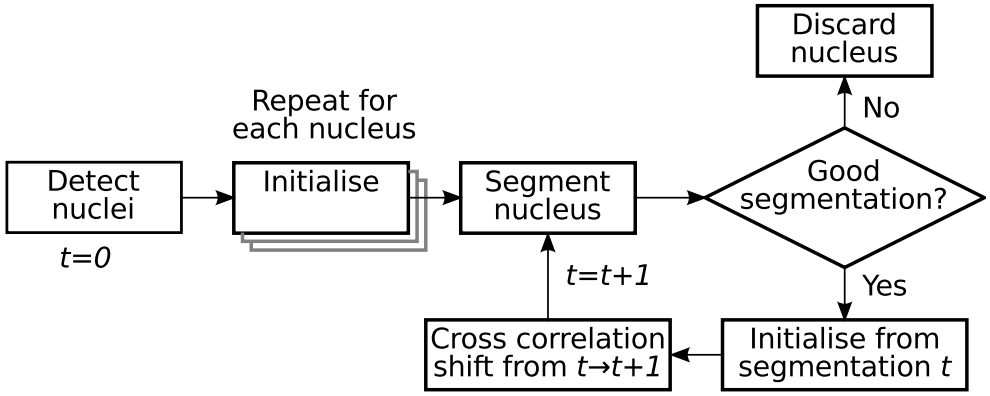


Figure 4.10: Flowchart of the CPC segmentation procedure.

and consumables.

4.3 The analysis tools

The data we are dealing with can vary significantly in terms of overall quality, even when the same experimental protocol was followed. Additional factors to consider are that we will be looking at normal and perturbed systems, the volumes of data are relatively small in comparison with a high-throughput screen, and it would be useful if the system could handle other similar problems.

The approach presented here therefore requires some user interaction, most notably in setting parameters or in discarding poor segmentations. A flowchart describing the segmentation procedure is shown in Figure 4.10, and a summary of the parameters used is given at the end of the section in Table 4.1.

4.3.1 Terminology

In the description of the algorithm we define some key parts of the image as follows:

Foreground: The nuclear and spindle regions which we are interested in.

Background: The remainder of the embryo other than the foreground.

External area: The area outside outside an embryo, i.e. empty areas of the microscope slide.

4.3.2 Detection

The purpose of the detection step is to locate the positions of visible nuclei in the embryo, and to provide an initialisation for the segmentation algorithm. Our detection step is designed to work with images from prophase to metaphase, in which blobs with intensity greater than the surroundings are present, and uses a series of morphological operations.

We first blur the image $I(\mathbf{x})$ by filtering with a circular binary structuring element of radius r, then dilate the resulting image with the same structuring element. This results in an image with flat plateaus at the approximate location of each nucleus, which can be found by searching for all local maxima. The size of the structuring element is set so that only one local maxima is found per nucleus without overlapping a neighbouring nucleus. In practice the detection stage is relatively insensitive to the exact size, so we set it to around half the expected size of a nucleus. See Figure 4.11a and b for an example.

Occasionally a false detection is made, typically when the cytoplasm shows more auto-fluoresence than normal. An additional problem is that since the cortex is curved whereas the imaging plane is flat some nuclei may be out of focus and of low intensity. We therefore iteratively discard any detections whose average intensity is less than $\mu - 1.5\sigma$ where μ and σ are the mean and standard deviation of the mean intensities of all remaining local maxima. This leaves us with N detected objects.

4.3.3 Segmentation

4.3.3.1 Intensity adjustment

Our aim is to segment a nucleus based on the faint spindle fluorescence, but as mitosis proceeds the CPC gathers on the kinetochores as a series of bright dots, so we first apply a sigmoidal filter to each image $I(\mathbf{x})$ to reduce the apparent brightness of the kinetochores:

$$I_f(\mathbf{x}) = \frac{1}{1 + e^{-(I(\mathbf{x})-\beta)/\alpha}}$$

where α controls the range of intensities which we wish to emphasize, and β is the median of the intensity range. This assumes that the embryo fills the majority of the frame. If the external area is significant then we would like to omit these pixels from the calculation of the median intensity, which we can achieve by ignoring intensities less than a threshold b. See Table 4.1 for typical parameter values.

4.3.3.2 Level set segmentation

Typical level set segmentation algorithms involve the time consuming solution of a partial differential equation which can achieve a sub-pixel solution. However, we can take advantage of a more efficient narrowband method, such as the fast level set algorithm of Shi and Karl [2008], previously used by Yau [2007] and Beaumont [2009], to speed up the analysis.

This fast level set algorithm is essentially the ultimate narrowband method, taking into consideration only the pixels immediately either side of the contour. At each iteration the contour can be shifted by one pixel, with regularisation incorporated by filtering the contour with a Gaussian kernel. See Appendix A.1 for more details.

The lack of edges between the background and spindle fluorescence means a region-based potential field must be used to govern the evolution of the level set [Chan and Vese,

2001]. This is modified slightly by using the local hybrid level set from Lankton and Tannenbaum [2008] which considers a neighbourhood of radius r_c around each point on the boundary for the region calculation.

The initialisation for the first frame is provided by the detection stage, with a frame at time t being initialised using the segmention from frame $t-1$.

See Figure 4.11c and d.

4.3.3.3 Failed segmentations

A nucleus may move out of the field of view over time, so any nucleus which hits the edge of the image is automatically discarded.

4.3.3.4 Large scale movement

The nuclei in a *Drosophila* embryo are liable to experience waves of movement in which all nuclei move by a relatively large distance, particularly during telophase, for biological reasons which remain unclear. This can pose a problem for the level set which is initialised from the segmentation at the previous frame.

However, since the movements of the nuclei are mostly geometrically translated and are highly correlated we can obtain a rough registration of two consecutive frames using cross-correlation (implemented using the fast Fourier transform for efficiency), which is then used to geometrically translate the initialisation appropriately.

4.3.3.5 Optional shape model

The signal-to-noise ratio of the images remains low, and the spindle fluorescence frequently fades out instead of having a well-defined edge, so during development the incorporation of a shape model was investigated to prevent the level set from expanding too far. The exact shape of the spindle varies over time and to keep the algorithm applicable to other

proteins with a similar though non-identical localisation we chose a simple model for the shape of the spindle, namely an ellipse, which we fitted using the method of moments. This method finds an ellipse whose first and second moments (centre of mass and centre of inertia) are equal to that of the binary segmented region. In addition we used an area prior which encourages the area of the segmented region to remain the same or increase slightly, in accordance with the observed behaviour of the spindle.

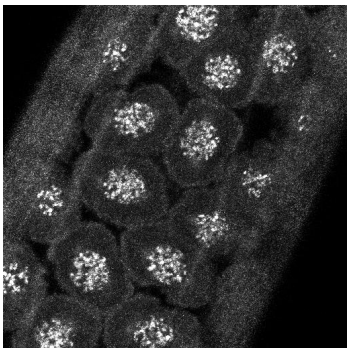
In practice the benefits offered by the shape model were outweighed by the time penalty, so this was not used in the later analysis.

4.3.4 Alignment of nuclei

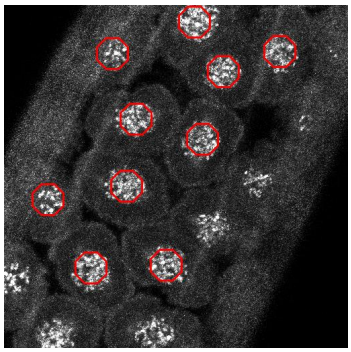
The segmentations for each nucleus could be used to obtain quantitative measurements, for example spindle length, area, and rate of movement. However, we are interested in visualising the uncharacterised and often unclear behaviour of a protein occuring within the spindle region, so to aid the visualisation of multiple nuclei we rotate all spindles into a horizontal plane.

This is possible because in late mitosis the spindle displays a strongly elongated shape, which we approximate by fitting an ellipse to the binary segmentation of each nucleus using the method of moments (Figure 4.11e). The long axis of the ellipse should coincide with the axis of the spindle, so the image can be rotated appropriately. During early mitosis the spindle region is circular, so the orientation of the ellipse at this stage is not useful.

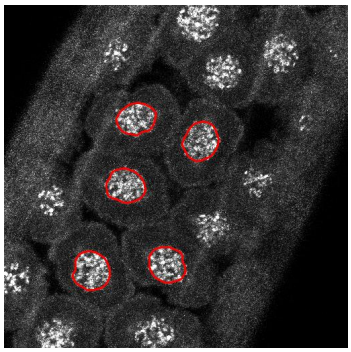
At this stage we have a useful visualisation tool in which all time points of one nucleus can be simultaneously viewed (Figure 4.11g), or alternatively multiple nuclei at the same time point can be inspected, with the possibility of animating all nuclei even if from different embryos.



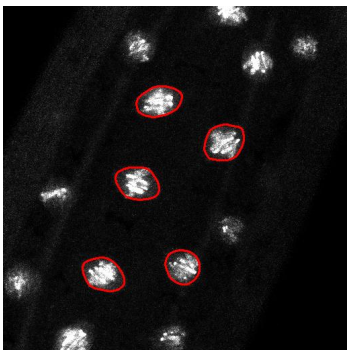
(a) Original



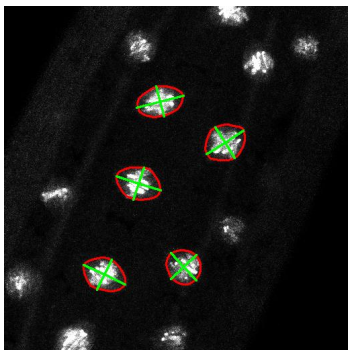
(b) Detection



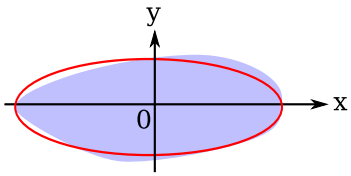
(c) Segmentation 1



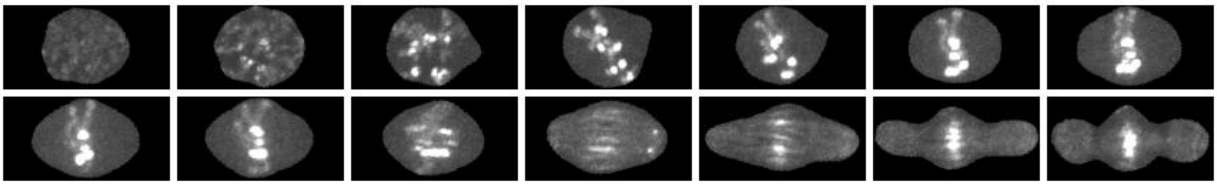
(d) Segmentation 10



(e) Ellipses



(f) Spindle axes



(g) Single nucleus over time

Figure 4.11: Nuclei detection, segmentation and alignment. (b) Red circles show the detected nuclei. (c)–(e) Red lines show the segmentation, green lines show the major and minor axes of the fitted ellipses. (g) A montage of one of the nuclei over time (20 s between frames). Note that the segmentations of the last two frames cannot be relied upon since the nucleus has divided into two, but the presence of the midbody means the level set has not correctly picked out the two daughter nuclei.

4.3.5 Kymographs

A kymograph is a projection of a temporally varying image into a single reduced dimensional image, in which multiple time points can be viewed at once in exchange for discarding information in one or more dimensions [Pereira and Maiato, 2010]. The term *kymograph* originates from the medical apparatus designed by Carl Ludwig for recording physiological processes such as blood presure over time.

A typical use would be monitoring the motion of particles such as cells along a blood vessel, where the path of the vessel is clear and static but locating the cell is extremely difficult. However, by plotting the intensity along the vessel at each time step it becomes easy to pick out the cell as it moves [Sato et al., 1997].

Recently kymographs have been created with non-static regions of interest, typically using a line or rectangular region positioned by tracking point markers. However, the segmentation step means we can now create kymographs from irregularly shaped regions.

We define the main spindle axis as the x-axis, and the y-axis as perpendicular, with the origin situated at the centre of the segmentation (Figure 4.11f).

The kymograph is created by applying a function F, for instance summation, to intensities along the y-axis which results in a 1D intensity plot as a function of position x along the spindle:

$$k_n(x,t) = \sum_y J_{n,t}(x,y)$$

where $J_{n,t}$ is one of the segmented and rotated nuclei from frame t. This is repeated for subsequent frames to form the complete kymograph $k_n(x,t)$ for nucleus n, an example of which is shown in Figure 4.12.

The removal of one dimension helps to emphasize trends in the protein dynamics which are difficult to detect in the original images, such as the presence of multiple distinct localisations as opposed to a single widely spread distribution, and the compact representation

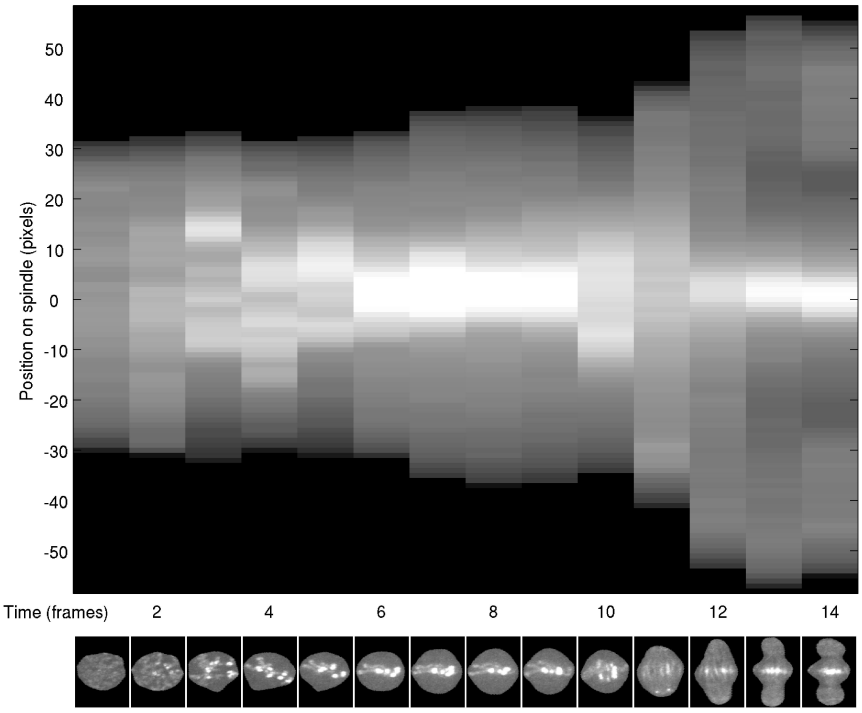


Figure 4.12: A kymograph created from one nucleus.

Parameter name	Symbol	Typical value
Radius of structuring element	r	15 pixels
Width of sigmoidal intensity filter	α	0.02 pixels
Background threshold for sigmoidal filter	b	0.1
Hybrid field search radius	r_c	32 pixels

Table 4.1: Summary of parameters used for detection and segmentation. All parameter values assume a normalised image with pixel values in the range $[0, 1]$.

of temporally varying data would facilitate the comparison of different samples.

4.3.6 Run times

Run-times for the full analysis are dominated by the segmentation which takes around 1.5 s per-nucleus per-frame on an Intel Core2 CPU T5600 with 2GB RAM purchased in 2006, with an additional 10 s to 20 s for the ellipse fitting and the creation of the kymograph. In

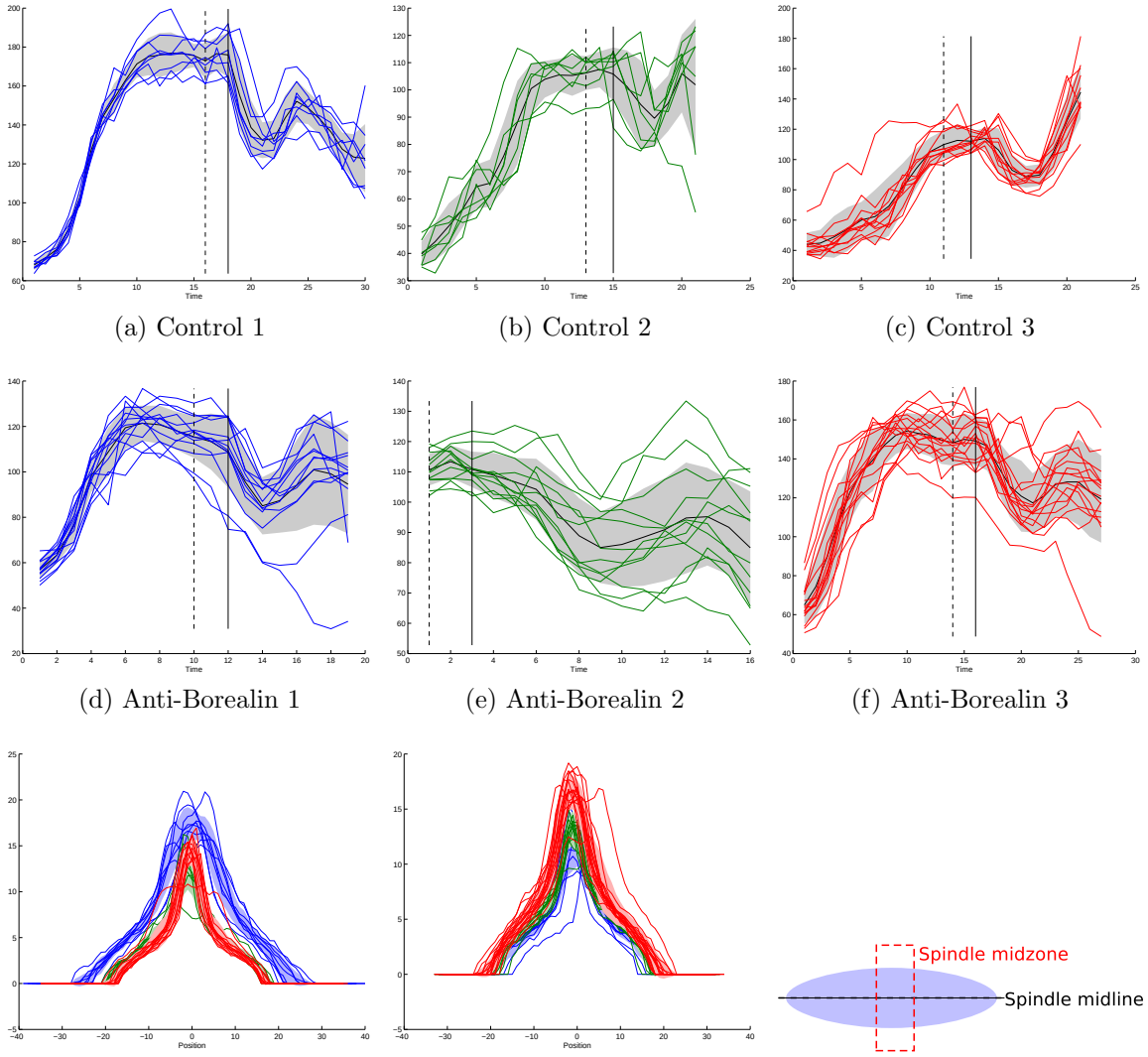
practice the run-times can be significantly reduced by analysing multiple nuclei simultaneously. For example the cross-correlation step accounts for 60% of the segmentation time, but only needs to be run once for all nuclei in the same image.

4.4 Anti-Borealin microinjections

In Section 4.1.4 we saw some evidence that the levels of Aurora B were decreased in the central midzone immediately after the start of anaphase. Since this decrease appears to occur between the separating chromosomes, which move along the spindle axis, we should be able to obtain the required measurement directly from the kymograph.

Figure 4.13 shows this information in two ways. Individual plots for individual embryos show a characteristic increase in intensity followed by a sharp decrease around 30 seconds (2–3 frames) after the metaphase-anaphase transition due to the chromosomes moving away from the centre with the CPC appearing to remain bound to them. This is followed by a second increase as the central spindle and subsequently the midbody forms between the separated chromosomes. A separate plot shows the intensity along the midline of the spindle shortly after metaphase.

We anticipated that the apparent dip in intensity in anti-Borealin injected embryos would show up as either a larger dip in intensity than control cells, or in a difference in spindle intensity profile shortly after metaphase. However this does not appear to be the case, at least with the embryos that were studied. The conclusion therefore is that either our observations were part of the natural variation that occurs, or that the effect is more subtle, and can not be quantified with the tools developed.



(a) Control 1 (b) Control 2 (c) Control 3

(d) Anti-Borealin 1 (e) Anti-Borealin 2 (f) Anti-Borealin 3

(g) Control spindle profile (h) Anti-Borealin spindle profile (i) Spindle measurements

Figure 4.13: Aurora B-GFP spindle intensity in control and anti-Borealin injected embryos. Times are given in frames (12 s intervals). (a)–(c) Mean intensity of the spindle midzone, defined as five pixels either side of the middle of the spindle (red dashed box in (i)), over time for three independent control embryos. Each line corresponds to a different nucleus within the same embryo, the grey shaded region shows the standard deviation over all nuclei. (d)–(f) Mean intensity for three embryos injected with anti-Borealin. The solid vertical line shows the approximate time of metaphase, the dashed line marks two frames (24 s) after metaphase. (g),(h) The intensity along the spindle (black dashed line in (i)) in each of the embryos at metaphase +2 frames. Colours correspond to the different embryos in (a)–(c) and (d)–(f). Note that the total time ranges shown varies between embryos, and that recording of (e) did not begin until metaphase.

4.5 Coiled-coil microinjections

Section 4.1.4 suggested there was a delay in mitosis near the site of injection of the coiled-coil peptides. The characteristic change in intensity profiles, particularly the drop post-metaphase (as seen in Figure 4.13), provides a potential way of temporally aligning the nuclei in an embryo.

If we make the assumption that all nuclei behave identically other than a variable initial time delay, we can use cross-correlation to calculate the time shift Δt required to match the intensity plots $y_i(t)$ and $y_j(t)$:

$$s(d) = \sum_{\tau=-L}^{L} y_i[n] y_j[d+n]$$

$$\Delta d = \underset{d}{\operatorname{argmax}}\, s(d)$$

where t, n and d are discrete indices since the timelapse images are acquired at fixed intervals, i and j are two different intensity plots.

However, the need to compare multiple nuclei from each embryo complicates this since there is no single reference nucleus. Instead we calculate all pairwise cross-correlations and time shifts $\Delta D(i,j)$ between the N nuclei and use the mean

$$m(i) = \frac{\sum_{j=1, j\neq i}^{N} \Delta D(i,j)}{N}$$

as the temporal offset for nucleus i.

A further modification is to calculate the weighted mean using the normalised cross-correlation coefficients.

Figure 4.14 shows an example of an embryo injected with the predicted coiled-coil domain at the top right of the image. As expected there is a clear trend, with nuclei near

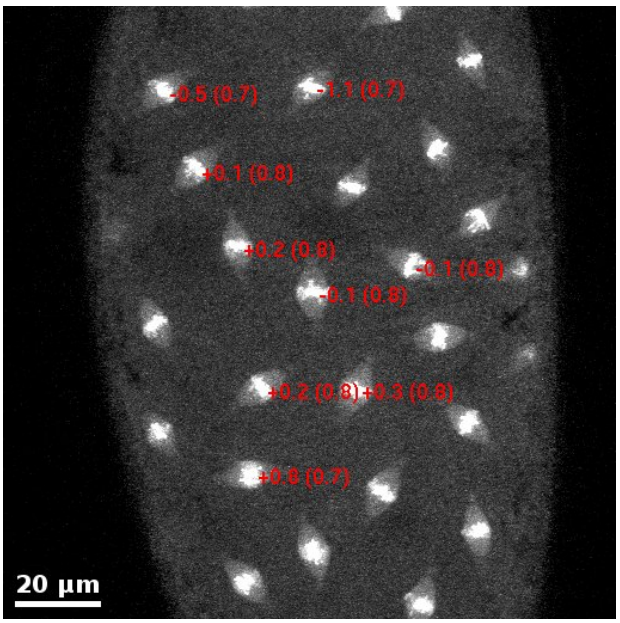


Figure 4.14: Relative time differences (in frames) in metaphase-anaphase transition in a PredCC injected embryo. Positive values indicate nuclei which entered anaphase ahead of others, negative values indicate those which lag behind. Cross-correlation coefficients are shown in brackets.

the site of injection being delayed. The cross-correlation coefficients are also shown, and in future could be used as a confidence value— a small value implies it was not possible to properly align the intensity profiles.

4.5.1 Possible effects of the coiled-coil peptides

It is interesting that the spatial localisation of Aurora B-GFP appeared to be normal following the injection of PredCC, yet the progression of mitosis was delayed. Assuming this is due to the interaction of the coiled-coil domain with Incenp or its substrates, rather than some other unexpected interactions, it implies the function of the CPC is only partially affected.

Ahonen et al. [2009] found that injecting a polyclonal anti-Incenp antibody into *Xenopus* cells led to reduced turnover of Aurora B at kinetochores which seemed to affect

microtubule-kinetochore attachments, and that the density of microtubules in the central spindle was reduced, both of which would be expected to cause a delay in mitosis. It is therefore possible that we are observing the same effect but with a more targeted peptide, which could prove to be significant. Clearly we cannot say for sure with the current data, but this is a strong case for continuing with these experiments.

4.6 Conclusions

Although we have been unable to find any significant differences on the current data set the analysis methodology is quite general and should be well posed for quantifying future experiments.

For instance, if we were to fluorescently label the injected coiled-coil peptides we should see a gradient as it diffused through the embryo, and we would expect this to be correlated with the observed delay.

The potential utility of this for analysing other markers can also be seen in a recent paper by Wang et al. [2011] in which phosphorylation gradients caused by the spatially varying activity of Aurora B activity in relation to the kinetochores were monitored.

4.7 Improvements

There are a number of ways in which this prototype software could be improved.

As already mentioned a shape model was rejected as part of the solution due to the increase in time required. This could be explored further, perhaps with a penalty for excessive deviation from the segmentation of the previous frame. Alternatively, a global area prior might be feasible, only requiring a change in the threshold used to determine the $+1$ or -1 speeds.

In terms of user-friendliness, this software would need to be developed further for use as a tool for biological research. It could be significantly improved with the addition of a GUI, and consideration of how to reduce the number of manually set parameters.

Finally this software was designed for use with 2D images, possibly obtained from a maximum projection of a 3D volume. Although the extension to a full 3D analysis might be beneficial it would be extremely challenging due to the low spatial resolution in the z-axis, and should therefore be avoided unless a compelling case is found.

Chapter 5
Clumped cell analysis

This chapter explains how novel analysis algorithms have been applied to study and quantify the behaviour of clumped cells. A fundamental problem in this type of work is the identification of boundaries for each cell. This is addressed in Sections 5.5–5.8 where we describe the development of an algorithm for the segmentation of tightly packed clumps of cells in which the key idea is to enhance specific boundary structures before using a multi-phase segmentation algorithm. We then describe the extraction of cell features and the novel application of random forests on cell-derived image features to characterise the behaviour of cells in Section 5.9. Finally we investigate a relatively new area— the use of spatial analysis methods to quantify interactions between cells.

The work described in this chapter is just the first stage in a larger project involving the development and application of quantitative analysis techniques to clinical data, with the ultimate goal of analysing tumour biopsies and tailoring treatments to an individual patient's needs, with the possibility of continual modification throughout the treatment protocol.

An early version of the work in Section 5.7 was presented at the International Conference of the IEEE Engineering in Medicine and Biology Society 2010.

5.1 Introduction

At the heart of many high throughput systems is the idea of culturing and manipulating hundreds or thousands of cells, but extracting and measuring them as if they were inde-

pendent samples. Cells are grown at a low enough density to enable the easy separation of each cell from its nearest neighbours, greatly simplifying the segmentation and measurement processes. This is ideal when analysing intracellular changes where it is sufficient to reference any changes to the original state of each cell, or when studying changes that are independent of other cells.

However, there is increasing interest in examining the behaviour of closely clumped cells. Many cell lines naturally grow in clumps with inter-cellular signalling pathways enabling individual cells to communicate with each other, and the overall behaviour of a clump of cells can differ markedly from that of an isolated cell.

The segmentation of these clumps is a challenging problem. The ideal situation would be to use a membrane specific label to clearly indicate the boundaries of cells. However, the difficulty arises in selecting a marker that consistently works in both control and treated (or perturbed) cells, is resistent to preceding stages of the fixing and labelling protocols which may affect some subcellular structures, and does not in itself disrupt the labelling of other structures. A further consideration is that the number of labels is limited by the availability of fluorophores, antibodies, and the need to separate multiple fluorescent channels.

In practical terms this has meant we were unable to find a suitable membrane marker for the experiments described in this chapter, and as a consequence the images we have to work with frequently contain cells with missing or unclear boundaries, so we will need to use an algorithm that can pick out weak boundary information where it is present, and will otherwise default to a suitable alternative.

We will be looking at temporal changes, so an obvious imaging protocol would be to use live timelapse imaging to give continous temporal information on each cell, since any heterogeneity observed at one time point cannot be directly correlated to that observed at another time point with fixed images. Unfortunately live imaging is significantly more

complicated than imaging fixed cells, both in terms of the preparation of samples, and in the time and equipment required to acquire the movies.

Live imaging requires *vital* dyes, antibodies or other markers which are non-toxic to a cell, but very few of these are available, many will have functional side-effects, and it is difficult to introduce them into the cell in a non-disruptive manner. Tagging of a protein with a fluorescent protein is often acceptable, but the generation of stable clones expressing the tagged protein is difficult, and the expression of the protein is usually in addition to that of the endogenous protein, i.e. the protein is overexpressed.

For these reasons we will only be looking at fixed cells.

5.1.1 Division of work

All laboratory work and microscopy was carried out by Claudia Buehnemann in the Dunn School of Pathology, University of Oxford. This is described in Section 5.3. My contribution to the work described in this chapter is all of the analysis of the images including the development of algorithms for segmenting clumps of cells, the extraction of features from each cell or image, and the application of machine learning and visualisation techniques.

5.2 Biological background to this chapter

The biological topic of this chapter is the IGF2 signalling pathway. There is a strong clinical motivation for wanting to characterise this pathway to a depth which would not be practical using traditional manual analysis techniques. However, it is not so complicated that results will be unverifiable, making it a good system for the development of our analysis tools.

5.2.1 The IGF signalling pathways

The insulin-like growth factor 1 (IGF1) cell signalling pathway has an important role in regulating cell growth, differentiation and proliferation (Figure 5.1). Two closely related ligands are involved in activating the pathway: IGF1 which is is expressed in adults, and IGF2 which is predominantly expressed during fetal development with lower expression levels post-natal and in adults [Denley et al., 2005]. The stability and availability of the ligands is modulated by several binding proteins (IGFBPs).

There are multiple receptors capable of receiving the IGFs. Intracellular signalling is triggered when an IGF binds to IGF1R. Another receptor, IGF2R, has a preference for IGF2 over IGF1 but does not seem to be involved in intracellular signalling, and instead appears to modulate cell signalling by sequestering free ligand.

The coupling of multiple proteins and feedback loops in the signalling pathway means a binding event at the cell membrane can result in spatially and temporally widespread changes via the phosphotidylinositol-3 kinase (PI3K) and mitogen-activated protein kinase (MAPK) pathways. These responses may persist for significantly longer than the original stimulant.

5.2.1.1 Cell-cell communication and the PI3K pathway

A breakdown in cell-cell communication, for instance through the breakdown of intercellular junctions, seems to be a factor in tumourigenesis. Contact inhibition normally prevents cells from proliferating uncontrollably, so a breakdown in communication could lead to an increase in growth and proliferation rates. The breakdown of intercellular junctions is also a prerequisite to metastasis, where a cell separates from a tissue and migrates elsewhere.

However, the role of the PI3K pathway in cell-cell communication remains unclear [Momiyama et al., 2003, Nelson and Chen, 2002, Ito et al., 2010].

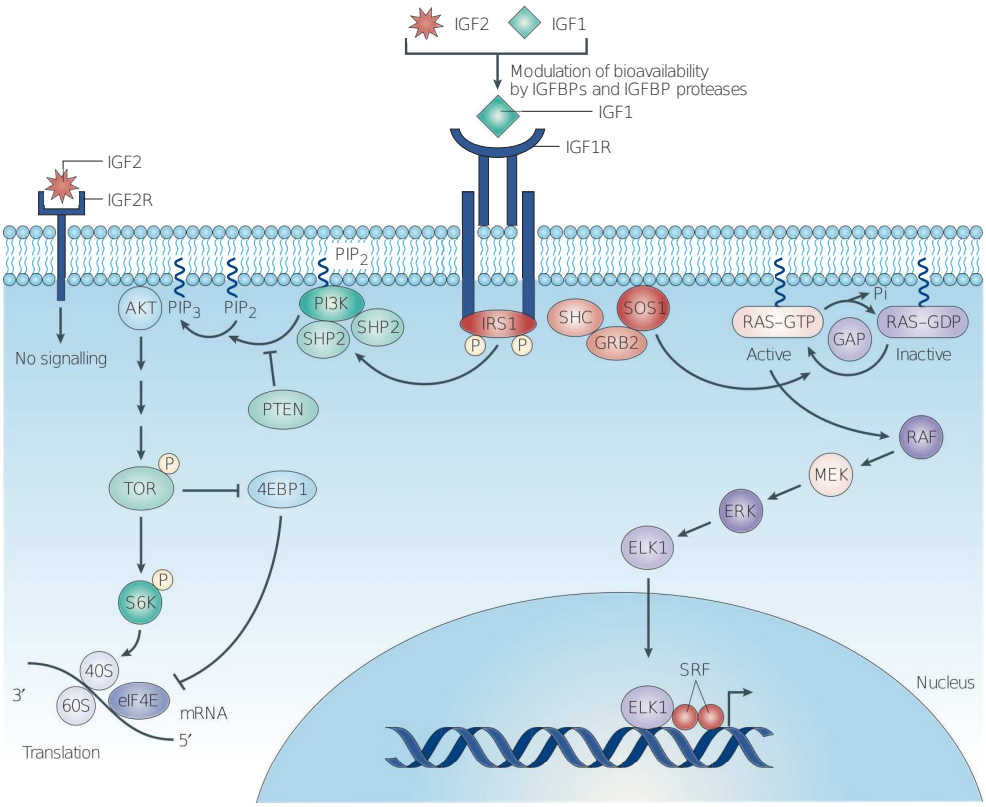


Figure 5.1: An overview of the insulin-like growth factor 1 receptor downstream pathway. The PI3K/Akt pathway is on the left, and the MAPK pathway on the right. Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Cancer [Pollak et al., 2004].

5.2.2 Signalling Markers

This project involves analysing the response of multiple downstream markers to IGF2. Five markers were used.

Phospho-Akt: An important kinase within the PI3K pathway which has a large number of substrates throughout the cell, regulating processes such as proliferation, survival and metabolism.

Phospho-S6: Directly targets the S6 ribosomal subunit, promoting cell growth by increasing translation of some mRNAs.

Egr1: Expressed in response to insulin in a manner that requires the MAPK pathway

but not the PI3K pathway [Keeton et al., 2003].

Foxo3a: Part of the FoxO family of forkhead transcription factors are found in the nucleus, where they appear to upregulate genes which can suppress growth, increase stress resistance, stop cell cycling or induce apoptosis.

Activation of the PI3K pathway results in the phosphorylation of FoxO and subsequent translocation into the cytoplasm, where it is targetted for proteosomal degradation [Tothova and Gilliland, 2007]. Egr1 and FoxO3A are therefore expected to behave in a reciprocal manner in response to IGF2.

5.3 Description of the data

All experiments were carried out on MCF-7 cells, a commonly used and well characterised tissue culture cell line that was originally isolated from the breast tissue of a patient with malignant adenocarcinoma. The popularity of this line is due to its expression of an estrogen receptor [Brooks et al., 1973] which makes it a good model for breast cancer, but it also expresses the IGF2 receptor, making it an ideal model for studying the IGF2 pathway [Peyrat and Bonneterre, 1992].

All of the images used in this chapter were obtained from the same experimental protocol:

1. Spread cells on a coverslip.
2. Incubate in cell media.
3. Add $0\,\mathrm{ng\,ml^{-1}}$ to $200\,\mathrm{ng\,ml^{-1}}$ IGF2 ligand to the cell media to stimulate the cells.
4. Incubate for 15 min to 120 min.
5. Fix cells and label with DAPI, phalloidin (F-actin antibody), and one other marker (an antibody). See Appendix D.1 for details.

Two sets of experiments were carried out.

5.3.1 Experiment 1: time dose

The purpose of this experiment was to explore the changes in cells that occurred over time after activation with IGF2, and to see how dependent this was upon the concentration of IGF2. Cells were treated with all combinations of the following doses and times:

IGF2 concentration ($\mathrm{ng\,ml^{-1}}$)	Incubation time (min)
0	15
25	30
50	60
75	120
100	
200	

$0\,\mathrm{ng\,ml^{-1}}$ indicates a control population. It is still essential to incubate these cells for the appropriate time since they may continue to develop and proliferate even without an external stimulant. $200\,\mathrm{ng\,ml^{-1}}$ IGF2 was chosen to saturate the expected response of cells, i.e. the PI3k/MAPk pathways should be fully activated at this concentration. The time points were chosen to give a representative snapshot of the changes in levels of all markers— some show a rapid change in expression levels following stimulation with IGF2, others take longer before a change can be seen.

5.3.1.1 Image acquisition

Four images for each time, dose and marker combination were acquired on a Zeiss 510 laser scanning confocal microscope using a 63× oil immersion objective. An initial set of images were supplied at a resolution of 512×512, later images were supplied at 2048×2048 but covered the same total physical area (the pixel size was reduced by an equivalent factor), so for consistency these were resampled to 512×512 before further processing. All images were visualised as three-channel RGB images:

Blue: DAPI (4',6-diamidino-2-phenylindole), a fluorescent stain which intercalates DNA. It can be used to label DNA in live cells but may have toxic effects. It shows maximum absorption in the UV range, and maximum emission in the blue spectrum of visible light.

Green: Phalloidin, binds F-actin, which forms part of the cytoskeleton. It prevents depolymerisation of F-actin and is therefore toxic to a cell. It is conjugated to a fluorophore for imaging.

Red: Signalling pathway marker, an antibody conjugated to a fluorophore which binds one of Egr1, Foxo3A, phospho-Akt, or phospho-S6.

An example image is shown in Figure 5.2.

5.3.1.2 Qualitative overview of images

A visual impression of the changes that occur in the levels of the four markers at different time points and IGF2 concentrations can be obtained from Figures 5.3, 5.4, 5.5 and 5.6.

For example, Egr1 (Figure 5.3) seems to be generally expressed at a low level in the cytoplasm in the control images, and of the three time points shown is maximally espressed in the nucleus after one hour following addition of IGF2. There is also an apparent correlation with IGF2 concentration. In contrast Foxo3a (Figure 5.4) appears to behave in a reciprocal manner, being initially expressed in the nucleus and translocating to the cytoplasm following IGF2 addition, with no obvious dependence on the IGF2 concentration (out of the concentrations shown).

Phospho-AKT (Figure 5.5) levels show a very rapid increase which persists for at least one hour, whereas phospho-S6 (Figure 5.6) seems to be expressed in a fraction of control cells, with the fraction increasing following IGF2 addition. It is also tempting to think that the control cells expressing pS6 might be clustered together, though visual inspection does not allow us to say whether this is likely to have occured by chance or not.

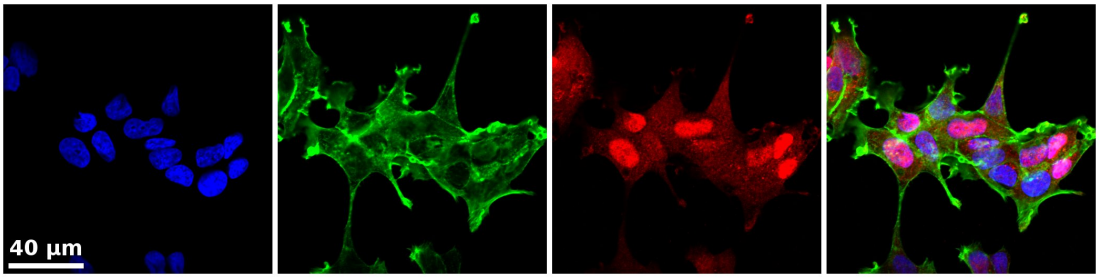


Figure 5.2: An example image of a MCF-7 cell stained with three markers. Left–right: DNA, cytoplasm, Egr1, merged image.

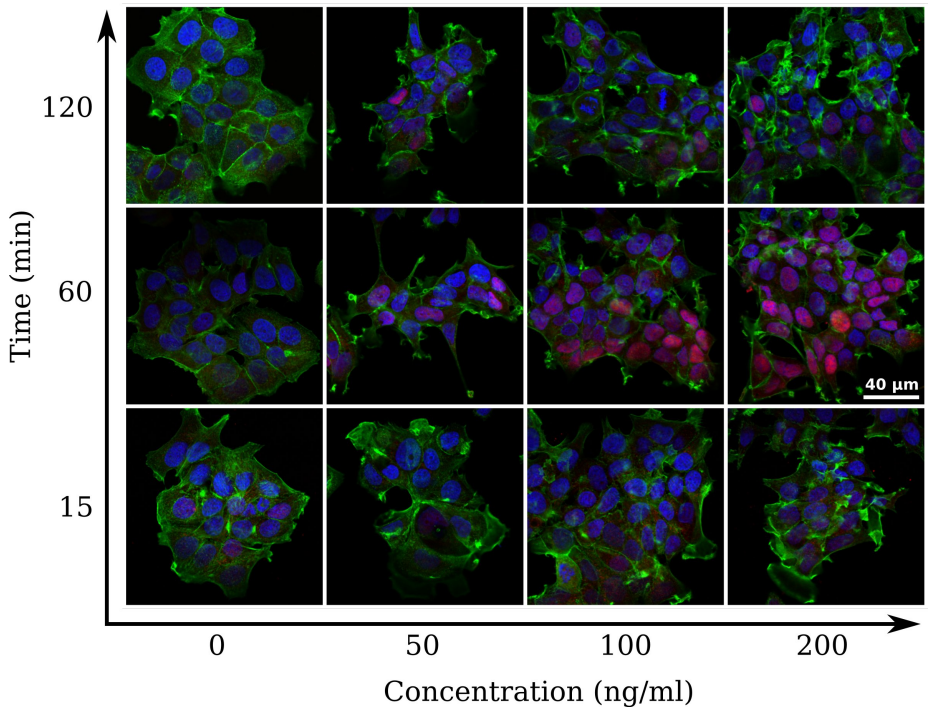


Figure 5.3: MCF-7 cells under different conditions with Egr1 labelled in red.

5.3.2 Experiment 2: Cell clumps

The second experiment was aimed at investigating the relationships between clump size and the responses of individual cells. Cells were treated with one of three different concentrations of IGF2 and labelled with one of two markers expected to behave in a reciprocal manner:

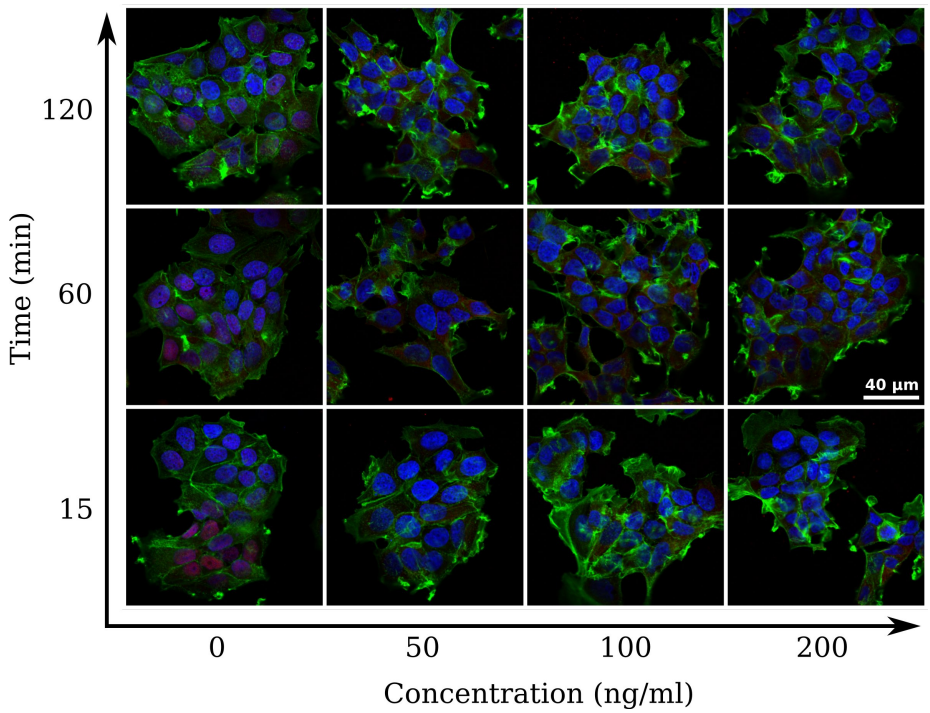


Figure 5.4: MCF-7 cells under different conditions with Foxo3a labelled in red.

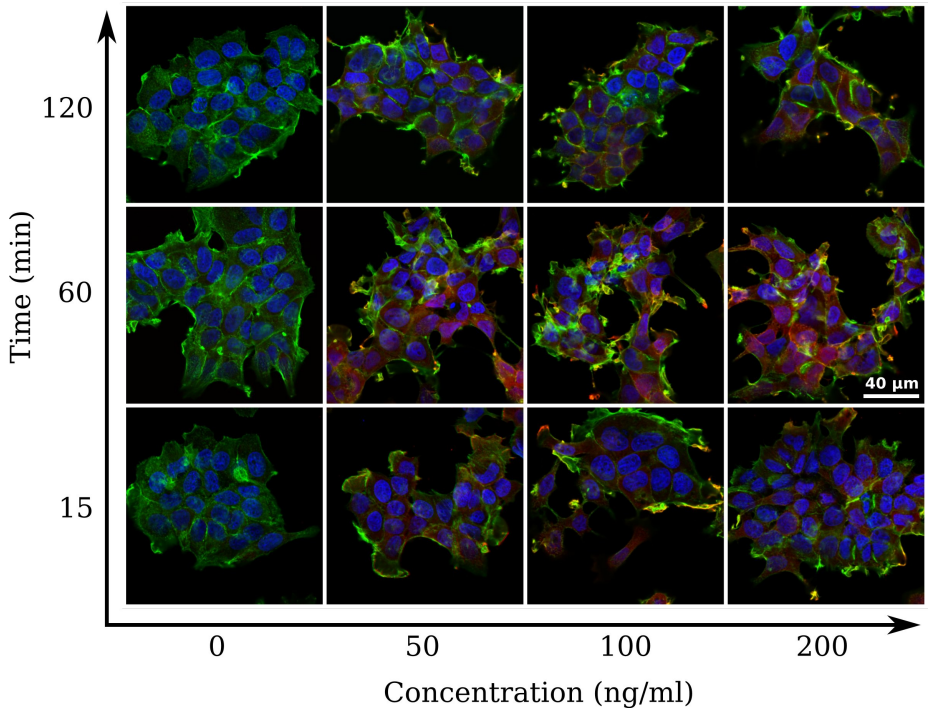


Figure 5.5: MCF-7 cells under different conditions with pAkt labelled in red.

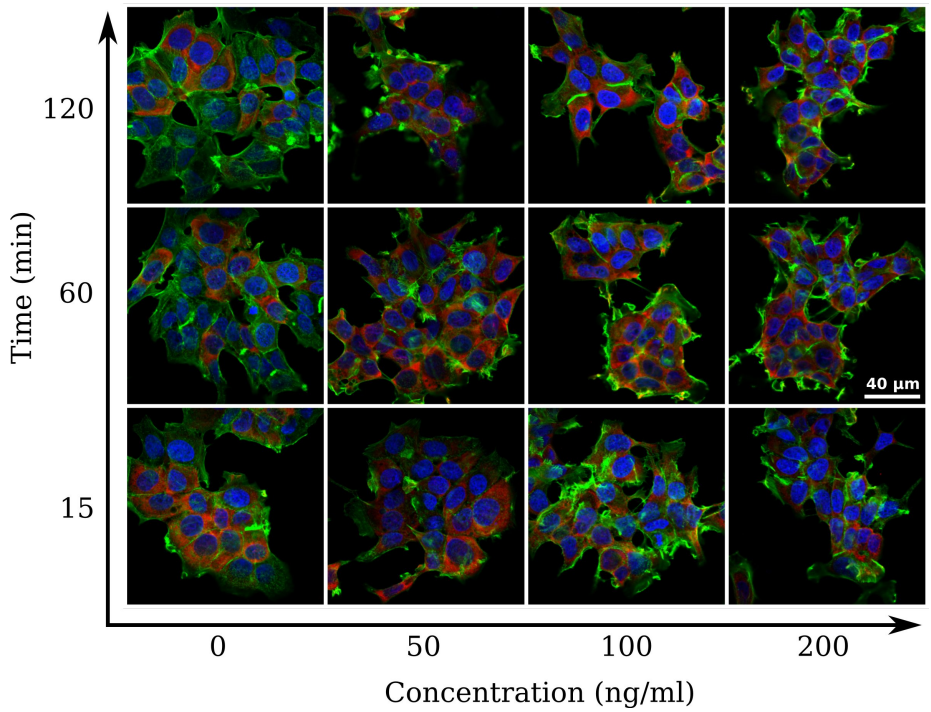


Figure 5.6: MCF-7 cells under different conditions with pS6 labelled in red.

IGF2 concentration ($\mathrm{ng\,ml^{-1}}$)	Marker
0	Egr1
50	Foxo3a
200	

Cells were fixed after one hour, and 20 images of each combination of IGF2 concentration and marker were acquired. Fields of view were chosen such that a clump of cells in a single layer filled the majority of the frame. A newer microscope, an Olympus FluoView FV1000 laser scanning confocal, became available so was used for this experiment, though using the same settings as for the images in Section 5.3.1.1.

5.4 Cell segmentation methods

5.4.1 Nuclei segmentation

In this section we will describe some existing methods used to segment nuclei. Nearly all methods assume that the nuclei will be approximately convex; given that it is normally a tightly packed bundle of DNA enclosed by the nuclear envelope this is a reasonable assumption. Some cell types also exhibit an overall concave shape, in which case they can be treated in the same way as far as segmentation is concerned.

Images of separated well-defined nuclei can be segmented with relative ease, for instance using adaptive thresholding (Otsu [1979] being a classic algorithm), morphological operations, or the straightforward application of levelsets or graph methods followed by the identification of disconnected segmented regions [Li et al., 2008b]. In these cases the core issue is that of distinguishing the nuclei from noisy and potentially varying background intensity levels.

This segmentation problem becomes significantly harder when dealing with closely spaced or touching nuclei where there may be no clear boundary. It then becomes necessary to encode some sort of shape information (either explicitly or implicitly) into the segmentation algorithm.

5.4.1.1 Watershed methods

The watershed algorithm lends itself naturally to the segmentation of multiple nuclei, though requires modification to avoid over-segmentation. If the nuclei are reasonably round and only touch slightly then the watershed transform should automatically find a boundary linking two convex points. However it is prone to oversegmentation so modifications may be required, for example the marker controlled watershed method [Malpica, 1997], adaptively

varying thresholds [Wählby et al., 2002, Yan et al., 2008], and/or a merging algorithm based on the shape or intensity features of the segmented fragments [Chen et al., 2006, Zhou et al., 2009].

5.4.1.2 Intensity convexity

In some types of images the intensity of the nuclei shows a convex (dome-like) profile in 2D. Although this becomes less well-defined in closely packed cells it may still be possible to take advantage of this information, as done by Li et al. [2008a] where a gradient vector field was calculated and smoothed using an elastic deformation model, followed by local adaptive thresholding of the diffused gradient field. Combinations of morphological operations and reconstruction have also been used [Schmitt, 2009b].

5.4.1.3 Splitting on concave points

One method of implementing clump splitting is to use what is intuitively the most obvious feature— the concavities in the combined boundaries of two touching convex objects, which depending on the application may be nuclei or whole cells.

G. Fernàndez and Zrÿd [1995] was one of the first to implement this idea for segmenting plant cells, using a scoring function derived from the length of a splitting line and the variation in gray level intensities along this line. Wang [1998] modified this to use a polygonal approximation of the boundaries of objects to smooth the transitions between individual pixels.

These methods can be further refined at the expense of additional complexity. For instance, Bai [2009] uses the concave points to divide a clump segmentation into segments and attempts to fit multiple ellipses to each segment, making the assumption that all nuclei are elliptical. This is followed by a selection and merging optimisation step.

Schmitt [2009a] explicitly use the curvature of the boundary to determine the best

splitting point. The boundary is considered in the 2D complex plane, and a Fourier decomposition used to provide an alternative representation of the boundary, and to also allow smoothing.

5.4.2 Clumped cell segmentation

As already mentioned many nuclei segmentation methods are applicable to cell segmentation. In particular, the many modifications to the watershed algorithm may enable clumped cells to be segmented successfully. Closely related to the watershed algorithm are modified Voronoi methods [Jones et al., 2005]. Here we mention a few others.

5.4.2.1 Level sets

Many of the multiregion level set implementations reviewed in Section 2.3 have been applied and/or adapted to the challenge of segmenting clumps of cells. For instance, coupled level sets [Nath et al., 2006b], level sets with a shape and size constraint [Padfield et al., 2009], and incorporation of topological constraints [Yu et al., 2008].

In general these methods require a seed, for instance a nucleus segmentation, to act as an initialisation and to control the number of regions. Yan et al. [2008] uses a combination of an edge and region based level set to segment clumped cells with very low contrast boundaries.

5.4.2.2 Graph methods

The ease with which multiple regions are incorporated into graph methods has naturally led to their use for multi-cell segmentation problems, some of which are briefly reviewed in Ta et al. [2009]. Ljosa and Singh [2006] and Kim and Genovesio [2009] implemented a simulated (stochastic) random walker for segmenting horizontal cells, a type of neuron in the retina which is characterised by long stretched arms, termed neurites, which may

overlap. The geodesic commute distance, which is the expected time for a random walker to travel between two locations considering all possible paths, was used by Du et al. [2010] for interactive segmentation of clumped cells.

Graph representations of images can also be used at a higher level for merging oversegmented regions in a superpixel approach [Ren and Malik, 2003]. Ta et al. [2007] used the extrema of an image as seeds for a watershed type related presegmentation, followed by the construction of a graph with a node at each oversegmented region. The number of nodes is obviously significantly smaller than the one node per pixel representation, so whereas pixel graphs are limited by computational requirements to a small local neighbourhood connectivity, higher order relations, or even a fully connected graph, become feasible.

5.5 Segmentation of clumped cells

The nuclei in each image are typically well-defined. Although there are several complicating issues which increase the difficulty of segmenting the nuclei it is still much easier than attempting to segment the cytoplasm of clumped cells directly, so we will base the final segmentation on the assumption that each cell has one nucleus which can be used as a marker.

The main focus of this chapter is the algorithm for segmenting the cytoplasm of clumped cells, in particular the key idea is to enhance poor quality features in the image before segmentating the clumps of cells. This requires a nucleus to act as a seed for each cell and the nucleus segmentation algorithm used here was empirically judged to be adequate for this, but future work could involve its optimisation.

A flowchart providing an overview of the complete segmentation algorithm is shown in Figure 5.7.

All parameters specified in the following sections assume that the pixels of each image

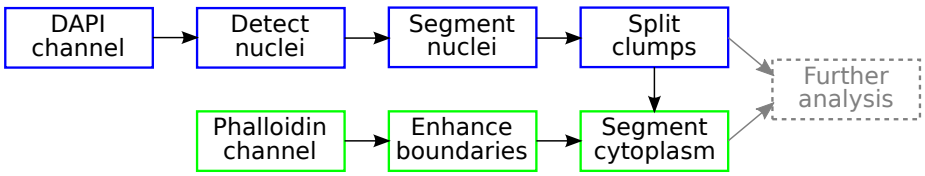


Figure 5.7: Overview of the segmentation process.

channel are in the range $[0, 1]$, size parameters use pixels as the base unit.

5.6 Nuclei segmentation (DAPI)

The nuclei in each image typically show good contrast in the DAPI channel against the background due to the concentration of DNA in the nucleus which emits a much greater fluorescence signal compared with the low levels of RNA which may be present in the cytoplasm, and autofluorescence.

However this is complicated by the inherent 3D structure of cells. Although they are plated out on a cover slip they may still partially overlap, and as a consequence it may not be possible to easily distinguish one nucleus from another. We therefore require a clump detection and splitting procedure.

5.6.1 Detection

In the detection stage we need to find the locations of nuclei, and provide an initialisation for the subsequent segmentation.

The majority of nuclei are found as follows:

1. Median filter the DAPI channel to remove speckle noise using a 5×5 pixel window.
2. Use Otsu's method to segment the image into two classes by finding an intensity threshold that maximises the between class variance [Otsu, 1979].

This occasionally fails to find some less intense nuclei, so we combine the Otsu segmentation with the regional maxima obtained using morphological operations as follows:

1. Median filter the DAPI channel as before.
2. Apply morphological opening and closing by reconstruction using a circular structuring element with radius 10 pixels.
3. Find regional maxima (which will be plateaus corresponding to nuclei).

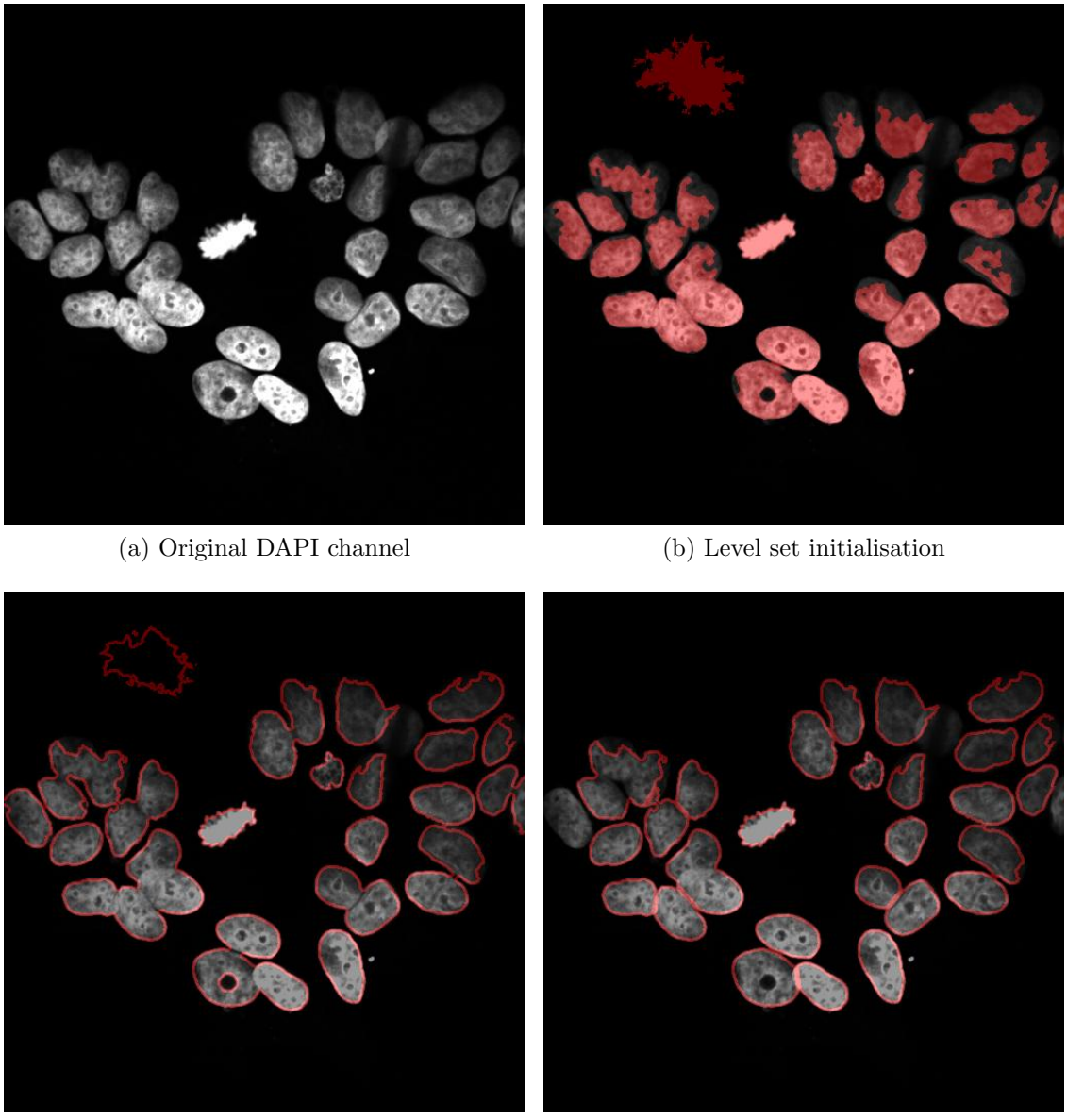
5.6.2 Segmentation

The detection step results in a very rough segmentation of the nuclei, which we refine using the fast level set method level set algorithm [Shi and Karl, 2008] with the hybrid potential field [Lankton and Tannenbaum, 2008] implemented in Section 4.3.3.2 and Appendix A.1. See Figure 5.8.

The parameters in this algorithm govern the balance between the image driven and regularisation contour evolutions, a judicious choice of which reduces the number of nuclei in close proximity which end up segmented as a single region.

- Number of image driven iterations = 10.
- Number of smoothing iterations = 3.
- Total number of iterations (an upper limit in the event that the level set does not converge) = 10.
- Window size of Gaussian kernel = 7 pixels.
- Width σ of Gaussian kernel = 3 pixels.

The number of smoothing iterations and the parameters of the Gaussian kernel should both control the regularisation in a similar way. Therefore when selecting the parameters we concentrated on optimising the numbers of image and smoothing iterations.



(a) Original DAPI channel

(b) Level set initialisation

(c) Level set segmentation

(d) Final segmentation after splitting

Figure 5.8: Main steps of the nuclei segmentation algorithm. Note the false detection in (b) which is removed in a later step.

5.6.3 Clump splitting

The level set cannot separate touching nuclei, so we use a post-segmentation step to find and split groups of nuclei loosely following the algorithm of Kumar [2006], which is based on the idea of searching for concavities [Wang, 1998].

The first step is to find candidate clumps which may be comprised of multiple nuclei, which is done by calculating the solidity for each segmented region:

$$s = \frac{\text{Perimeter}^2}{\text{Area}}$$

The minimum possible value of s is 4π, corresponding to a circle. Large values of s correspond to a convoluted and/or concave boundary, so we empirically choose a threshold of 5π, above which regions are treated as candidates for splitting. We also enforce a minimum area threshold for splitting.

Once we have identified candidates for splitting we find the convex hull $\mathcal{H}$ of the region, and calculate the minimum distance of each point d on the boundary $\mathcal{C}$ from the convex hull, together with the closest point on the hull.

The splitting algorithm involves searching for a straight cut line (see Figure 5.9), defined by two points $\mathbf{c_1}$ and $\mathbf{c_2}$ from $\mathcal{C}$ which maximises a score function calculated by combining the following measurements:

- Length of the cut line $l = \|\mathbf{c_1} - \mathbf{c_2}\|$.
- Distance of the ends of the cut line from the convex hull $e = d(\mathbf{c_1}) + d(\mathbf{c_2})$.
- Sharpness of the concave points at the ends of the cut line $p = \frac{e}{\bar{e}_1 + \bar{e}_2}$ where $\bar{e}_1$ is the mean value of d for the 5 boundary pixels in $\mathcal{C}$ either side of c_1, and similarly for $\bar{e}_2$.
- Angles between the cut line and the convex hull a_1, a_2.

Larger values of e and p are desirable since they correspond to starting and ending the

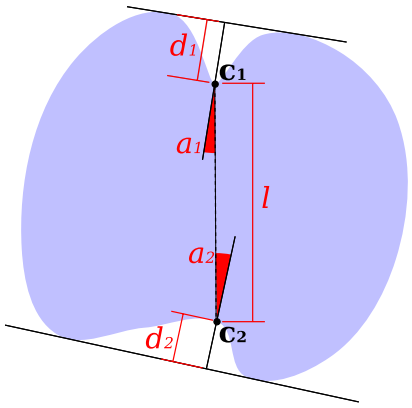


Figure 5.9: Measurements used to find the optimal split between two nuclei.

cut at deep concavities. The total length l of the cut should not be too short as this can lead to small protrusions being cut off. Intuitively we expect the cut line to be orthogonal to the convex hull.

Weightings for these parameters can be conveniently defined using the normal distribution $\mathcal{N}$, to give the following empirical formula for the score k:

$$k = e \times p \times (\mathcal{N}(l + 20, 14) + \mathcal{N}(l + e + 55, 30)) \times (\mathcal{N}(a_1, 0, \pi/4) + \mathcal{N}(a_2, 0, \pi/4))$$

These parameters were empirically determined from a visual screen of around ten images.

This splitting algorithm assumes that clumps of nuclei always have a non-convex boundary. Although there are some situations when this might not be the case these are also situations where experts would have difficulty in manually splitting them.

In the final step we discard any nuclei which are likely to be false detections. This includes low intensity patches or intense spots of fluorescense caused by foreign material. Any nuclei which appear to intersect with the edge of the image are marked since any measurements taken from them will be invalid, but it is necessary to take account of their presence when segmenting the cytoplasm.

5.7 Cytoplasm segmentation (phalloidin)

The phalloidin marker shows great variation in its ability to label external cell membranes. This is particularly noticable when comparing IGF2 treated against control cells— whereas in control cells it is frequently possible to make out boundaries between cells in a clump (albeit with some difficulty) it is almost impossible in some of the treated cells.

To segment these clumps we will use the segmented nuclei as a marker for each cell, and make the assumption that in the absence of any other information the boundary should lie at an equal distance between two nuclei. We will also use our prior knowledge that cell membranes should appear as curvilinear structures in an image to specifically enhance these structures, improving the quality of our segmentation.

5.7.1 External clump boundary

Cells with a free boundary (that is not in contact with another cell) tend to spread out on the coverslips, and consequently there is no distinct edge— instead the cytoplasm gradually fades from view as seen in Figure 5.10. This means the choice of an external boundary can be somewhat arbitrary.

The dark background external to the clump appears to be well-defined in all images, so we use a modified form of thresholding:

1. Gaussian filter the phalloidin image (window size = 3, $\sigma = 1$ pixels).
2. Threshold the intensity image at the 10th percentile + 0.01, where the second term ensures that the threshold is always greater than zero.
3. Remove small holes in the segmentation using morphological closing followed by opening.

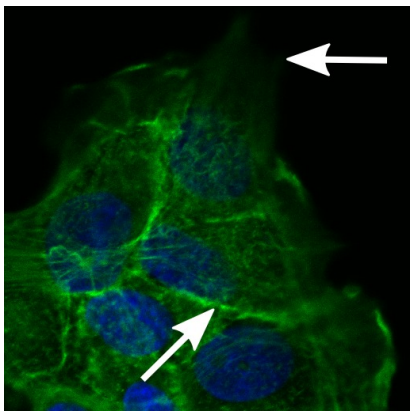


Figure 5.10: External cell boundaries tend to fade into the background compared with internal ones.

5.7.2 Boundary enhancement

The structures in the phalloidin image vary in width, from very fine filaments of actin which can occur within as well as at the boundary of cells, to thick bundles of fibres. A complicating factor is that the cytoplasm of neighbouring cells may overlap, in which case even if the boundary is well-marked by an intensity ridge there is still some ambiguity in the correct position of the boundary.

The boundaries can be considered as curvilinear structures, which are linear features or ridges that arise in a wide range of image analysis problems [Frangi et al., 1998, Steger, 1998, Wai et al., 2004, Zwiggelaar et al., 2004]. They can be considered as locally one-dimensional but may exhibit a curvature at a larger scale. In addition they have a finite width which can vary along the structure, and may exhibit branch points or junctions where multiple lines merge or cross.

The ridges we wish to detect are indistinct in the sense that they do not have two well-defined edges, so we should consider approaches based on finding the centreline of a curvilinear structure, such as the steerable filters developed by Freeman [1991] which provide an intuitive way of developing orientation sensitive feature detectors. A simple application of these filters involves convolving a non-isotropic template h at multiple rotations θ against an image $I(\mathbf{x})$, with the orientation $\hat{\theta}$ of a feature being determined by the

maximum response $\hat{r}$.

$$\hat{\theta}(\mathbf{x}) = \underset{\theta}{\operatorname{argmax}}(I(\mathbf{x}) * h(\mathbf{R}_\theta \mathbf{x}))$$
$$\hat{r}(\mathbf{x}) = I(\mathbf{x}) * h(\mathbf{R}_{\hat{\theta}} \mathbf{x})$$

$\mathbf{R}_\theta$ is the standard 2D rotation matrix, $\mathbf{x} = (x_1, x_2)^T$ is the 2D pixel location and $*$ denotes convolution. The magnitude of the response can be used as a confidence value.

Jacob and Unser [2004] describes a method for designing an optimal steerable filter based on the tradeoff between localisation accuracy and signal-to-noise ratio, and shows how the implementation can be made more efficient by careful choice of the operators used to build the filter. We use one of their suggested ridge detectors which is comprised of second order derivatives of Gaussians whose response is given by:

$$\begin{aligned} f(\mathbf{x}, \theta, i) = &\left(-\cos^2\theta + \frac{1}{3}\sin^2\theta \right) \frac{\partial^2 G_i}{\partial x_1^2} * I(\mathbf{x}) \\ &+ \left(-\frac{8}{3}\cos\theta\sin\theta \right) \frac{\partial^2 G_i}{\partial x_1 \partial x_2} * I(\mathbf{x}) \\ &+ \left(-\sin^2\theta + \frac{1}{3}\cos^2\theta \right) \frac{\partial^2 G_i}{\partial x_2^2} * I(\mathbf{x}) \end{aligned}$$

where i represents the scale of the filter and G_i is an isotropic 2D Gaussian with standard deviation σ_i. The equivalent filter for $\theta = 0$ is shown in Figure 5.11a.

The filter response at any orientation θ can be calculated from just the three second-order Gaussian partial derivatives, and by setting $\frac{\partial f}{\partial \theta} = 0$ the orientation giving the maximum response can be found. Each of the Gaussian derivatives are separable (meaning they can be calculated from two 1D convolutions), so this forms an efficient method of obtaining a measure of ridge strength in the image. The multiple ridge scales were chosen to detect the variable sized bundles of filaments in the images, in this case $\sigma = 1, 2, 4, 8$ pixels.

The filter responses are combined by taking a weighted sum

$$F(\mathbf{x}) = \sum_i \sigma_i^2 f_i(\mathbf{x})$$

where the weights account for the lower expected absolute response of the filter at larger scales, and also to place more emphasis on larger bundles of actin-labelled filaments.

Finally a potential function (also known as an edge stopping function) g governing the evolution of a level set is formed with a constant weighting term α.

$$g(\mathbf{x}) = \frac{1}{1 + \alpha F(\mathbf{x})}$$

An example of the intermediate filter responses used to create $g(\mathbf{x})$ is shown in Figure 5.11.

5.7.3 Multi-region level set segmentation

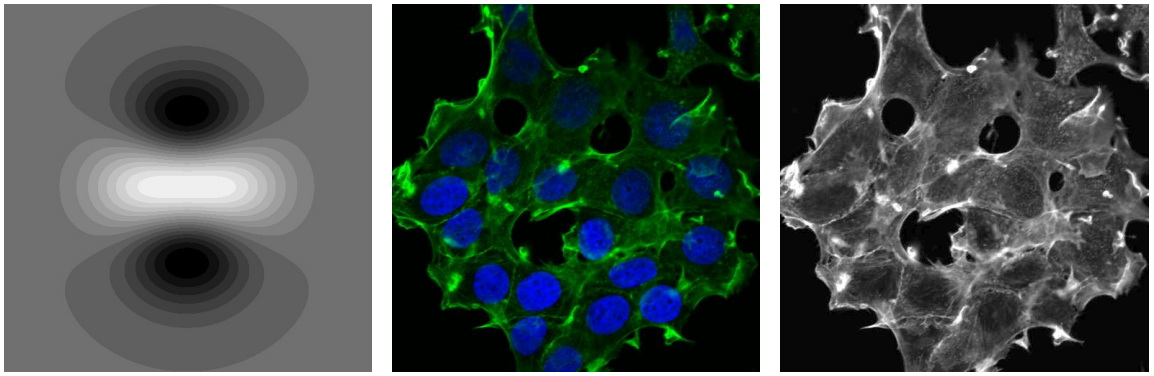
Since we are segmenting a clump comprised of multiple regions in which some neighbouring regions may not have a visible boundary our segmentation algorithm must include an interaction or competition term to prevent regions from encroaching on each other.

Our first implementation is based on the fast level set described in Section 5.6.2, but with one level set phase per cell for a total of M phases, and a different potential field. In each speed iteration the geodesic speed field S_k proposed by Caselles et al. [1997] is used to control the direction in which the level set should locally evolve in response to the potential field g:

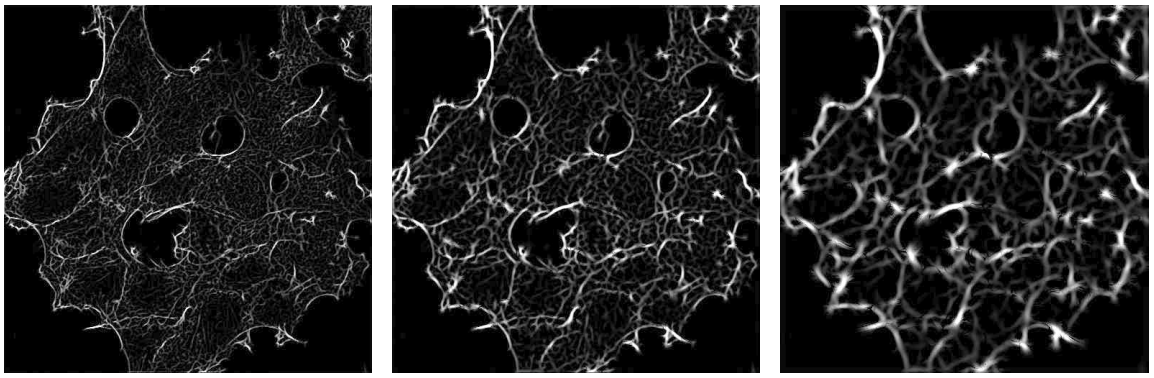
$$S(\mathbf{x})_k = g(\mathbf{x})V_0 + \nabla g(\mathbf{x}).\nabla \phi_k(\mathbf{x})$$

where V_0 is a constant expansion force and ϕ_k is the level set function for region k.

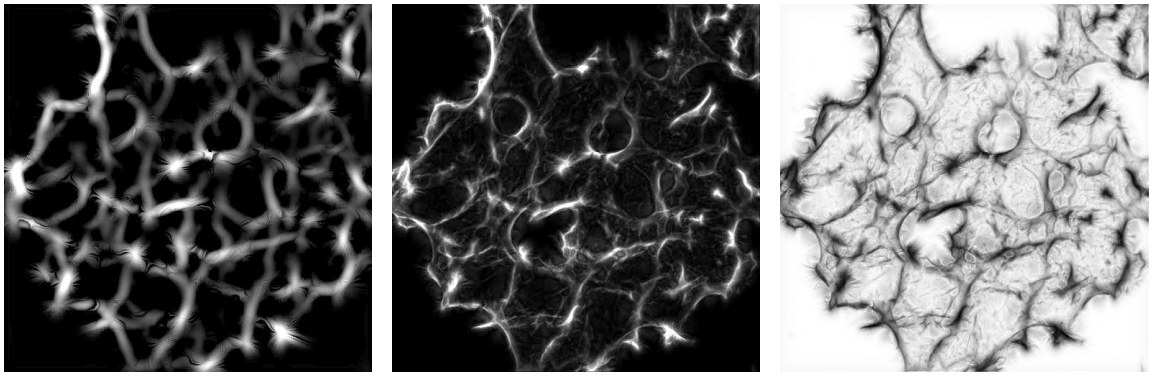
We define a region interaction indicator Q which is equal to the number of lists



(a) Ridge filter $\theta = 1$ (b) Original image (DAPI and phalloidin) (c) Original image (phalloidin only)



(d) $\sigma = 1$ (e) $\sigma = 2$ (f) $\sigma = 4$



(g) $\sigma = 8$ (h) Sum of responses (i) Potential field

Figure 5.11: The ridge filter responses and final potential field. (a) The impulse response of the ridge filter for $\theta = 0$. (b),(c) An example image used for segmentation. (d)–(g) The filter responses at each scale. (h) The weighted sum of filter responses $F(\mathbf{x})$. (i) The final potential field $g(\mathbf{x})$.

$(L_{in,k}, L_{out,k})$ for different regions k containing a pixel:

$$Q(\mathbf{x}) = \left| \{ \mathbf{x} \mid \mathbf{x} \in L_{in,k} \vee \mathbf{x} \in L_{out,k} \} \right|$$

where $|\cdot|$ denotes the cardinality of the set.

If $Q(\mathbf{x}) > 1$ then the level set speed for the competing regions is assigned as follows (recall the fast level set only requires the sign of the speed function):

$$S(\mathbf{x})'_k = \begin{cases} +1 & \text{if } S(\mathbf{x})_k = \max\{S(\mathbf{x})_1, \ldots, S(\mathbf{x})_M\} \\ -1 & \text{otherwise} \end{cases}$$

This interaction term applies to all level set iterations, whether image driven or smoothing.

5.7.4 Local Minima Avoidance

Level sets will find a local minima, and are therefore dependent on their initialisation. If a boundary exists but is some distance from the initialisation the level set may fail to find it. A further issue is the speed of the algorithm— although the evolution of individual contours is extremely fast the interaction calculations are a bottleneck.

Both of these issues can be mitigated by performing the segmentation at multiple scales. Since the level set is purely dependent on the calculated ridge potential field we downsample this field by a factor of 8 ($m = 3$) as follows:

1. Downsample the level set speed field g by a factor of 2^m to obtain $g_{(2^m)}$.
2. Downsample the initialisation A by a factor of 2^m to obtain $A_{(2^m)}$.
3. Run the level set algorithm initialised with $A_{(2^m)}$ to obtain a segmentation $B_{(2^m)}$.
4. Upsample $B_{(2^m)}$ by a factor of 2 to obtain $A_{(2^{m-1})}$.

5. Repeat steps 3 and 4 until a segmentation $B_{(1)}$ at the original scale is obtained.

Unfortunately a comparison of this fast level set method with manual segmentations (see Section 5.7.7) suggests it does not work as well as intended.

5.7.5 Random walker segmentation

As an alternative to the level sets algorithm we have also used the ridge potential field with the random walker algorithm [Grady, 2006]. The main benefits of the random walker over the level set algorithm are the decreased computational complexity, and the ability to obtain a confidence map of each segmentation.

5.7.5.1 Random walker algorithm

The random walker algorithm is a graph based image segmentation method that intrinsically handles multiple regions. The algorithm is initialised by seed points defined for each region to be segmented, and the outputs of the algorithm are the probabilities $p(\mathbf{x}, k)$ of each point $\mathbf{x}$ being associated with a seed point with label k. Since p is a probabilitiy

$$\sum_{k=1}^{K} p(\mathbf{x}, k) = 1$$

where K is the number of regions. See Figure 5.12 for an example.

There are several analogies to describe the meaning of the calculated probabilities. If a hypothetical random walker is released from node $\mathbf{x}$ which is not a seed, then $p(\mathbf{x}, k)$ is the probability the walker reaches a seed with label k before any of the others. The edge weights represent the difficulty that the walker experiences when trying to walk between two neighbouring nodes.

Alternatively, to use an electrical analogy, the edge weights are conductances and the seed points constant voltage sources. If seed points with label k are set to unit voltage and

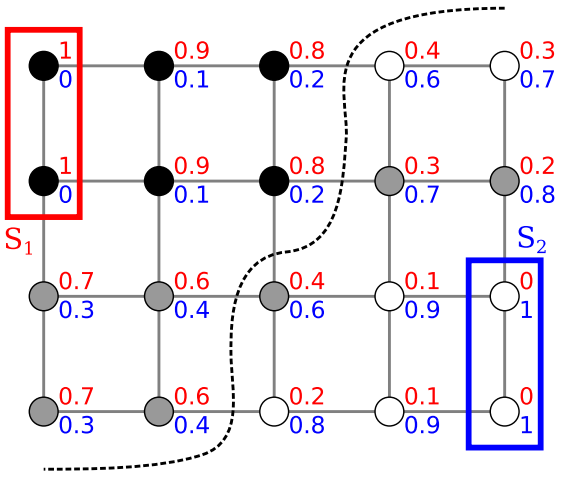


Figure 5.12: Schematic of a random walker binary segmentation. S_1 and S_2 are the initial seed labels. The numbers at each node are the random walker probabilities for each seed. The dashed line represents the segmentation obtained by partitioning nodes according to the seed with the maximum probability.

all other seeds to 0 V then $p(\mathbf{x}, k)$ is the voltage at x.

A hard segmentation $S(\mathbf{x})$ can be obtained by partitioning the nodes according to the label corresponding to the seed with the maximum probability:

$$S(\mathbf{x}) = \underset{k}{\operatorname{argmax}}(p(\mathbf{x}, k))$$

If all weights are the same then the segmentation will be equidistant between the seeds, as would be intuitively expected. The mathematical details of the algorithm are described briefly in Appendix A.2.

5.7.5.2 Node weights

We first use the potential field g as the input image to the default implementation of the random walker, so the node weights are calculated from this field[1]:

$$w_{ij} = \exp\left(-\beta \left|g_i - g_j\right|\right)$$

Secondly we modify the field to increase the relative influence of ridges which showed a strong response to the ridge detector. In the potential image strong ridges have a lower value (closer to 0) than flat areas, and smaller weights correspond to higher dissimilarity between two nodes. Hence to signify increased dissimilarity at strong ridges we can divide by the total magnitude of the potential field between the two nodes:

$$w'_{ij} = \exp\left(-\beta \frac{|g_i - g_j|}{g_i + g_j}\right)$$

5.7.6 Graph colouring

Both of these algorithms have a memory requirement proportional to the number of cells. Specifically, the level set method requires each cell to have a matrix representation of the level set function ϕ with dimensions equal to that of the image, and the random walker results in a separate probability matrix for each cell, again of the same dimensions as the image.

To reduce the memory requirements we can use the idea of graph colouring to allow disjoint regions belonging to different cells to be represented by the same phase. Graph colouring is the assignment of a colour c_i to each node v_i such that $c_i \neq c_j, \forall e_{ij} \in E$.

There is a classic proof that four colours is sufficient to colour any planar graph, and

[1] This is used in the example code supplied by Grady, his paper uses a slightly different formula: $w_{ij} = \exp\left(-\beta(g_i - g_j)^2\right)$

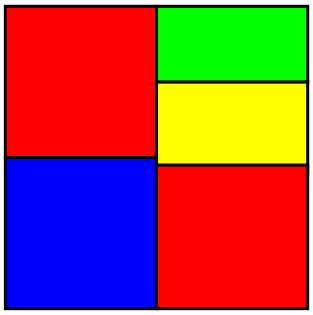


Figure 5.13: Four colours may be insufficient to colour an evolving graph. The close proximity of the two separate red regions means they may incorrectly merge into one as the segmentation algorithm evolves.

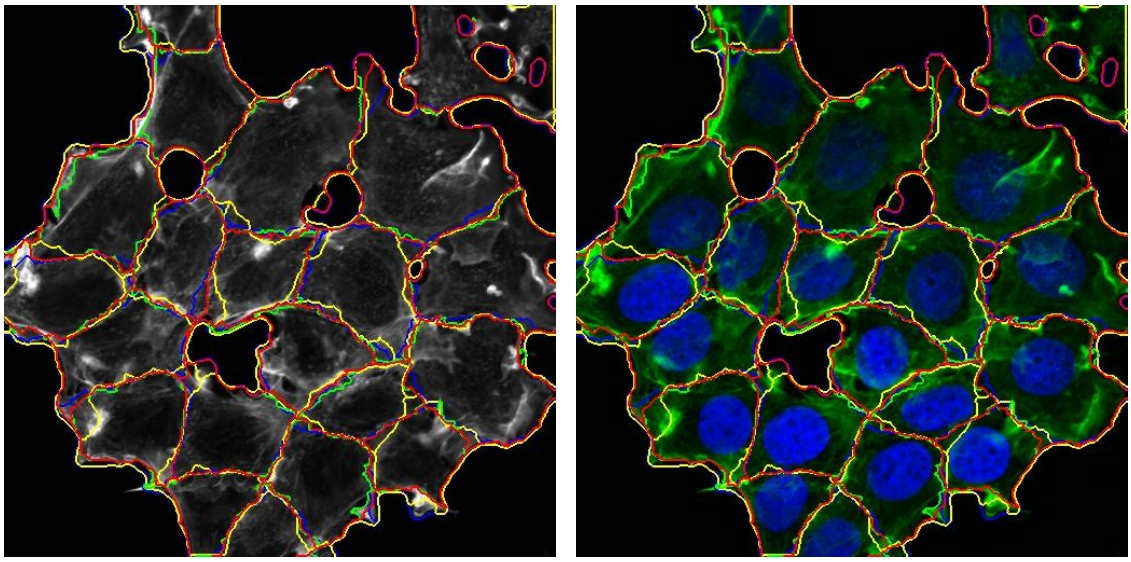
this is used by Nath et al. [2006a] in their level set method. In practice this assumes *a-priori* knowledge of which cells are adjacent, but obviously this is not known until we have obtained the segmentation.

A rough estimate can be found using a fast Voronoi method to create a neighbourhood graph of adjacent cells, and this often results in a suitable four-colour graph. However it fails when two cells are almost but not quite neighbours (Figure 5.13), since the full segmentation may result in them merging.

We therefore choose to find a graph colouring that ensures all first and second order neighbours in the Voronoi segmentation have a different colour, which for all practical purposes should be sufficient to avoid merging. There is no known upper bound for the number of colours required, but in practice it tends to reach a maximum at around 20 in an image of densely packed cells. This colouring could be further optimised by only considering second order neighbours within a certain distance.

This graph colouring will not affect the level set segmentation since the contour evolution is only dependent on a small area around each pixel. In theory it will affect the random walker, but in practice the probabilities at the distances involved are negligibly small.

An example segmentation for all methods is shown in Figure 5.14.



(a) Phalloidin (b) Phalloidin and DAPI

Figure 5.14: Example cytoplasm segmentation overlaid on the original image. Blue: Voronoi. Yellow: level set. Green: random walker (standard weights). Red: random walker (modified weights).

5.7.7 Manual comparison of all methods

For the purposes of judging the accuracy of these segmentation methods a set of images were additionally stained with β-catenin antibody, which *should* localise to cell membranes [Aberle et al., 1996] (see Figure 5.15). Ideally this marker would have been used for all images, but in practice it was not possible to design an experimental protocol in which the fidelity of this marker could be maintained in addition to that of the other required markers. Furthermore it is still not fully reliable as a membrane marker.

Images were manually segmented using a Wacom Cintiq 20WSX PenTablet, which allows an expert to accurately draw directly onto a screen with an electronic pen. Experts were instructed to segment all cells, making a judgment as to where the boundary should be in the case of cells lacking poor boundary information.

The images contained 9–41 cells, with the majority having 20–30 cells. Four images

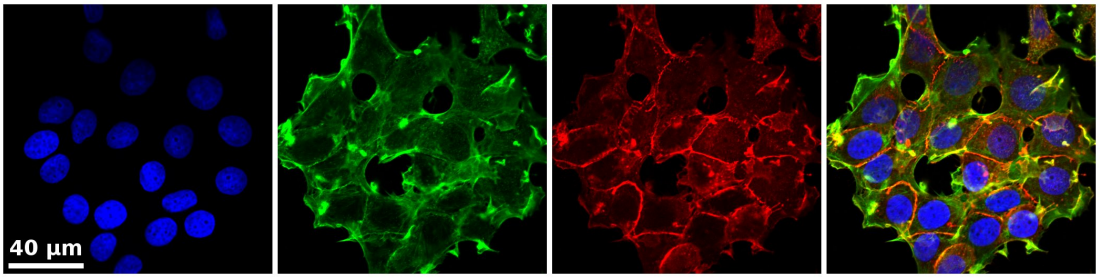


Figure 5.15: An example image showing the β-catenin staining of the membrane. Left–right: DNA, cytoplasm, β-catenin, merged image.

(labelled 1–4) were taken of cells treated with $50\,\mathrm{ng\,ml^{-1}}$ IGF2 for one hour, and four (labelled 10–13) were taken of control cells.

The modified Voronoi method developed by Jones et al. [2005] and implemented by Carpenter et al. [2006] was used on the original phalloidin channel as a benchmark for comparison with the algorithms developed here.

5.7.7.1 Similarity coefficients

There are several similarity indices that we could use to measure the accuracy of segmentation. As with most image segmentation comparisons we do not want to be unduly biased by pixels which are correctly classified as background since these will always vastly outnumber the number of foreground pixels, so we use two metrics, the Dice similarity coefficient (DSC) [Dice, 1945] and the Hausdorff distance (HD) [Huttenlocher et al., 1993].

$$\begin{aligned}\mathrm{DSC}(X,Y) &= \frac{|X \cap Y|}{|X| + |Y|} \\ \mathrm{HD}(X,Y) &= \max\left(h(X,Y), h(Y,X)\right) \\ h(X,Y) &= \max_{y \in \partial Y} \min_{x \in \partial X} \|x - y\|\end{aligned}$$

where X and Y are the two segmented regions to be compared, ∂X and ∂Y are their boundaries, $|\cdot|$ is the number of pixels in a region, and $\|x - y\|$ is the Euclidean distance between points x and y.

The DSC ranges from 0 when there is no overlap between two segmentations to 1 when the two segmentations perfectly coincide. The Hausdorff distance can be loosely thought of as the maximum deviation of one segmentation from another, so larger values indicate a less accurate segmentation. The Jaccard or Tannimoto index [Jaccard, 1908] is another commonly used metric, but this can be directly derived from the DSC so will not be considered.

Figure 5.16 shows the variation in DSC over a range of algorithm parameters for all methods. The Voronoi method appears to be reasonably stable for all values which is an advantage in the sense that little parameter tuning is required, and a disadvantage in that it cannot be directly adapted to a particular problem.

In contrast the methods based on the ridge potential image show more variation but have the potential to perform better, as seen from the DSC figures in Table 5.1, and the Hausdorff distances in Table 5.2. The random walker with the modified weight field performs particularly well.

Each image contains several cells, so to visualise the variation within each image Figure 5.17 shows the mean and standard deviations over all cells for each image using the parameter values from Table 5.1d.

5.7.7.2 Random walker connectivity

A final parameter we may wish to consider is the connectivity used in forming the random walker graph. Grady and Jolly [2008] found that increasing the connectivity led to a small improvement in the accuracy of the segmentation, though this was on a 3D image.

The standard 2D connectivity is 4-connected, where each pixel is connected by an edge

Image	Voronoi	Level set	Random walker 0	Random walker 1
1	0.85	0.84	0.86	0.86
2	0.82	0.84	0.82	0.85
3	0.81	0.82	0.86	0.86
4	0.80	0.77	0.82	0.82
10	0.84	0.85	0.84	0.86
11	0.82	0.85	0.84	0.85
12	0.88	0.88	0.90	0.92
13	0.84	0.84	0.85	0.86

(a) Expert 1

Image	Voronoi	Level set	Random walker 0	Random walker 1
1	0.84	0.82	0.85	0.85
10	0.81	0.83	0.83	0.84

(b) Expert 2

Image	Inter-expert
1	0.93
10	0.91

(c) Inter-expert

Voronoi	$\lambda = 0.025$
Level set	$\alpha = 6$
Random walker 0	$\beta = 35$
Random walker 1	$\beta = 35$

(d) Algorithm parameters

Table 5.1: Mean dice similarity coefficients (larger is better) for each image calculated between manual and automatic segmentations. (a) 8 images segmented by expert 1. (b) 2 images segmented by expert 2. The random walker results are shown for the standard (0) and the modified (1) weighting functions. (c) Inter-expert comparison. (d) The parameters used for each algorithm.

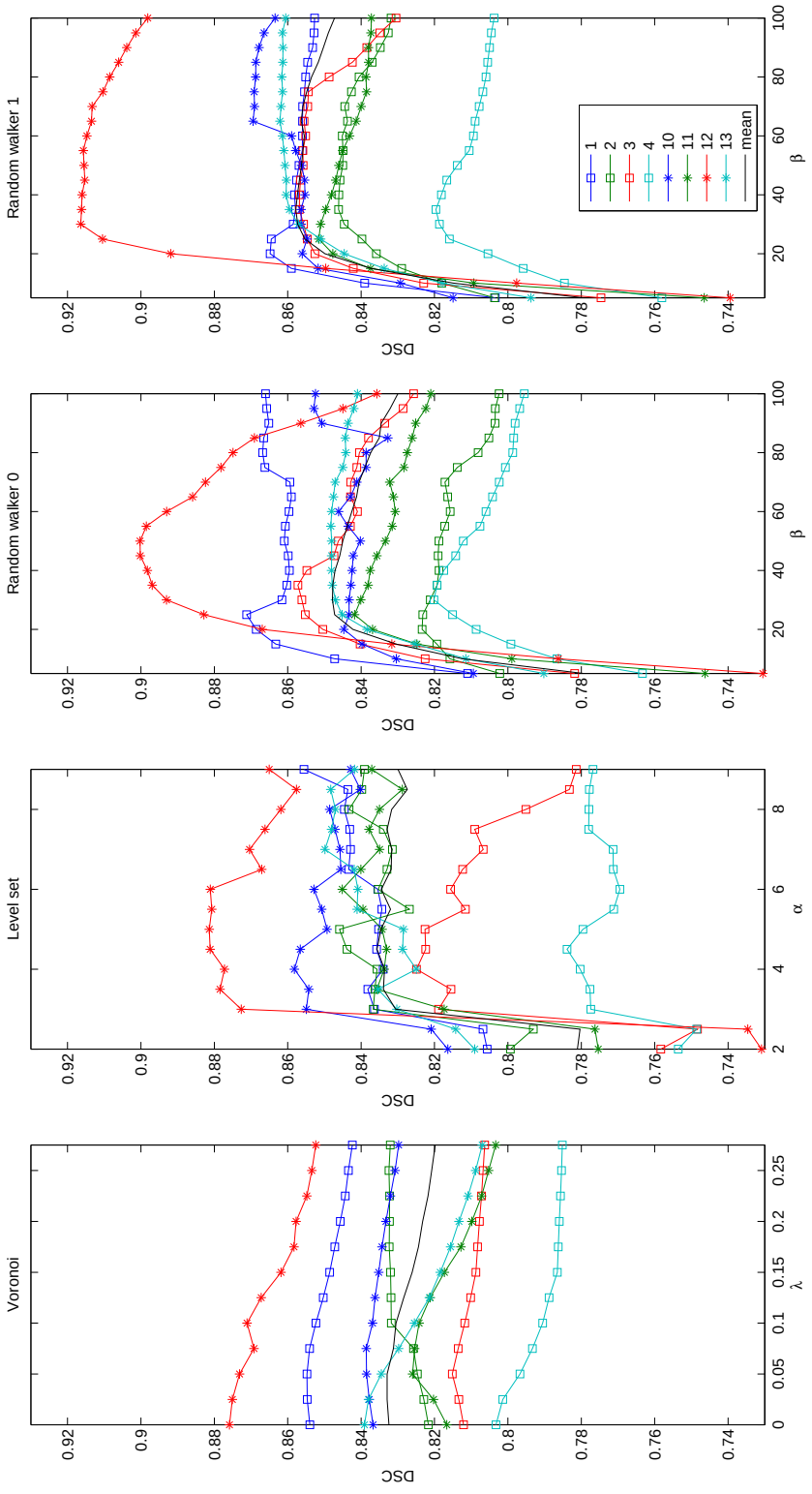


Figure 5.16: Comparison of manual and automated segmentations for different algorithm parameters using the DSC. Comparisons for eight images manually segmented by a single expert are shown. Images 1–4 were treated with IGF2, 10–13 are control images. The black line indicates the mean over all eight images.

Image	Voronoi	Level set	Random walker 0	Random walker 1
1	36	35	37	37
2	33	22	25	23
3	31	34	25	27
4	35	36	29	30
10	34	30	30	30
11	48	37	38	37
12	47	29	26	22
13	32	32	30	29

(a) Expert 1

Image	Voronoi	Level set	Random walker 0	Random walker 1
1	35	38	36	35
10	34	32	30	31

(b) Expert 2

Image	Inter-expert
1	26
10	17

(c) Inter-expert

Table 5.2: Mean Hausdorff distances (in pixels, smaller is better) for each image calculated between manual and automatic segmentations, and between different experts. See Table 5.1 for more details.

to its orthogonal neighbours. For comparison Table 5.3 compares the DSC for a 4-connected graph with an 8-connected one, which includes the orthogonal and diagonal neighbours. The connectivity appears to have negligible effect on the accuracy of the segmentation on these images.

5.7.7.3 Random walker probabilities

The random walker algorithm calculates the probability of each pixel belonging to each of the seed regions, and this potentially tells us how much confidence we can have in each segmentation. In particular we would hope that areas in which experts disagreed would

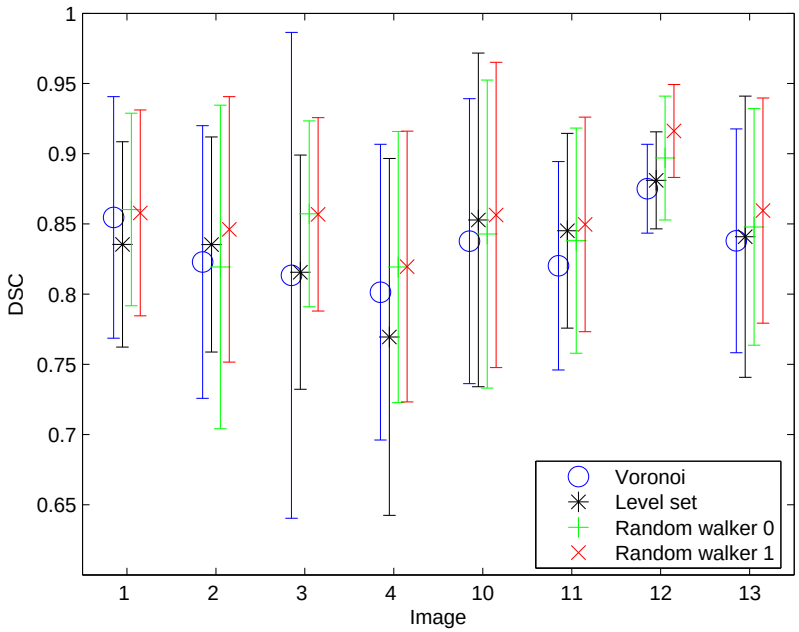


Figure 5.17: Comparison of manual and automated segmentations for selected parameter values using the DSC. Error bars indicate the standard deviations of the DSC over all cells in an image.

Image	Random walker 0		Random walker 1	
	4-connected	8-connected	4-connected	8-connected
1	0.86	0.87	0.86	0.86
2	0.82	0.82	0.85	0.84
3	0.86	0.86	0.86	0.86
4	0.82	0.81	0.82	0.82
10	0.84	0.86	0.86	0.87
11	0.84	0.85	0.85	0.85
12	0.90	0.89	0.92	0.91
13	0.85	0.85	0.86	0.86

Table 5.3: Dice similarity coefficients for the random walker segmentation using a 4-connected and 8-connected graph.

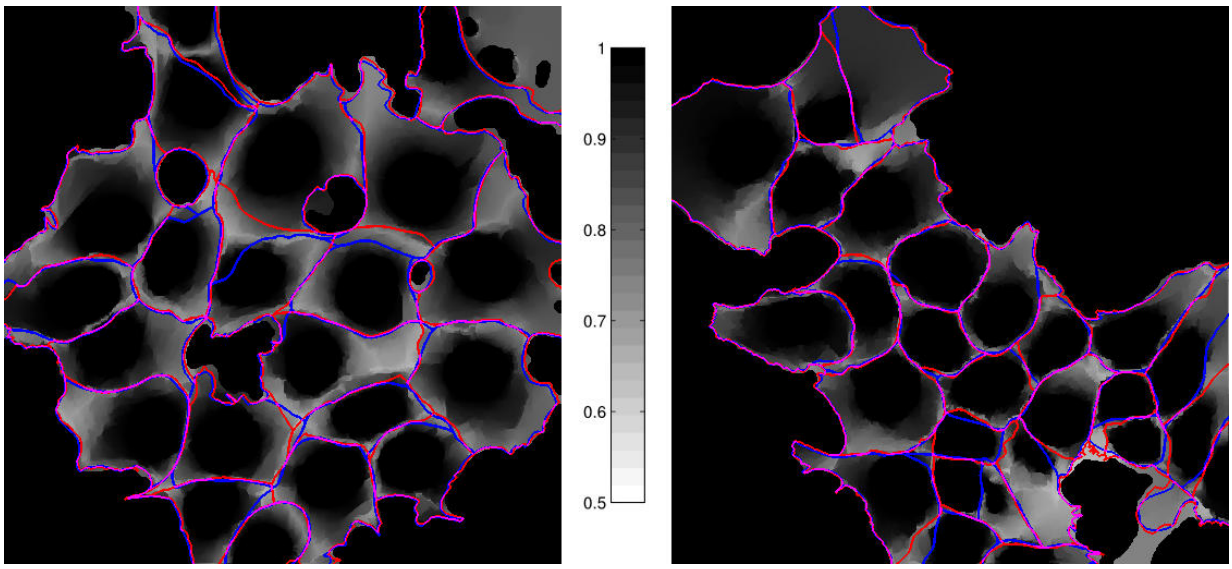


Figure 5.18: Probabilities from the random walker segmentation (shown in greyscale, 0.5–1), overlaid with two expert segmentations (red and blue).

be reflected in the confidence values. Unfortunately this does not appear to be the case (Figure 5.18). However we only have two images to compare, and this is certainly an area worth investigating in future.

5.7.8 Run times

Table 5.4 shows the approximate run-times for each stage of the cytoplasm algorithm, obtained using an Intel Core2 CPU T5600 with 2GB RAM purchased in 2006. The random walker appears to provide a good balance between computational complexity and segmentation performance.

5.8 Segmentation conclusions

Overally the fast level set does not perform as well as had been hoped, but the random walker performed much better. The random walker is also relatively fast and the use of graph colouring helps to keep the memory requirements to a manageable level.

	Time (seconds)
External boundary (Matlab)	0.4
Ridge detection (Matlab)	3.1
Voronoi (Matlab mex)	0.5
Level set (Matlab mex)	200.0
Random walker (Matlab)	5.0

Table 5.4: Run-times for the cytoplasm segmentation.

However there are a few minor problems with the random walker. Grady [2006] claims that the algorithm should never result in an unlabelled area. Intuitively this makes sense as every pixel should always have a "nearest" seed, but in practice occasionally a few pixels are not assigned a label. This could be a manifestation of the problem discovered by Cheng and Zhang [2011], who showed that in multi-region problems it is possible for areas to be unlabelled.

A further consideration is the tradeoff between regularisation and segmentation accuracy. It is notable that the parameters used for the Voronoi and random walker methods lead to boundaries which by eye are relatively jagged in comparison with the level set method. Changing these parameters can give a smoother segmentation, at the expense of reduced accuracy as quantified by the similarity coefficients.

It should be borne in mind that we chose to provide the experts with additional information in the form of the β-catenin label which was not given to the segmentation algorithms since we wanted to compare the performance of the algorithms on our real experimental data with that which could be obtained in an ideal situation.

5.9 Analysis of cells

Now that we have a segmentation for each cell we can start to extract measurements of protein levels from each cell, and begin to analyse the heterogeneity present in a population.

Expert biologists already have knowledge of some of the main differences to be expected under different experimental conditions, and may easily pick up on these features in the images from a visual examination. However, this expert knowledge can sometimes be a hindrance— it may cause a bias in the evaluation of images, so ideally an unbiased system for extracting information from images is desirable.

The first task is to obtain quantitative measurements which can be interactively explored, and we would then like to extend this to automatically picking out features of interest.

5.9.1 Feature extraction

A feature vector $\mathbf{x_i} = \{x_{i,1}, \ldots, x_{i,K}\}$ was created for each cell i using the measurements detailed here.

5.9.1.1 Cell features

All dimensions are given in terms of pixels. All intensity features were created from images with pixels in the range $[0, 1]$.

- Shape features each for the nucleus and cytoplasm ($2 \times 5 = 10$ features):
  - Area.
  - Convex area (obtained by finding the convex hull).
  - Perimeter.
  - Eccentricity of fitted ellipse.

  - Solidity (ratio of area to convex area).

- Shape features obtained by comparing the nucleus and cytoplasm segmentations (5 features):
  - Ratio of nucleus to cytoplasm areas.
  - Maximum, minimum, mean and standard deviation of the distances between all points on the nucleus boundary to the closest point on the cytoplasm boundary—this will be related to position of the nucleus relative to the cytoplasm.
- Intensity features in the nucleus for DAPI and the marker, and in the cytoplasm for phalloidin and the marker ($5 \times 2 \times 2 = 20$ features):
  - Total (integrated) intensity.
  - Mean intensity.
  - Standard deviation.
  - Maximum.
  - Minimum.
- Additional marker intensity ratios (2 features):
  - Ratio of the mean nuclear marker intensity to the mean cytoplasmic marker intensity.
  - Ratio of the integrated nuclear marker intensity to the integrated cytoplasmic marker intensity.
- Additional features obtained by defining an inner and outer region of the cytoplasm, and an outer region of the nucleus, as all pixels within 10 pixels of the boundaries ($3 + (4 \times 3) + 3 + 3 = 21$ features):
  - Area of the outer nucleus region.
  - Area of the inner cytoplasm region.

- Area of the outer cytoplasm region.
- Maximum, minimum, mean and standard deviation of marker intensity in each of the three regions.
- Ratio of the mean intensities between all pairs of the three regions
- Ratio of the sum intensities between all pairs of the three regions

This gave a total of 58 features for each cell.

5.9.2 Image features

The features already mentioned were calculated for each individual cell. However, we are interested in the aggregated distribution of all these features, so we need a method for combining the cell features into a set of per-image features.

There is no single way to do this. We would like a way of characterising the variation of each feature across a population of cells, so we choose to use basic summary statistics for each feature k:

$$\text{Mean } \mu_k = \frac{\sum_{i=1}^{N} x_{i,k}}{N}$$

$$\text{Standard deviation } \sigma_k = \sqrt{\frac{\sum_{i=1}^{N}(x_{i,k} - \mu_k)^2}{N-1}}$$

$$\text{Skewness } \gamma_{1,k} = \frac{\sum_{i=1}^{N}(x_{i,k} - \mu_k)^3}{N\sigma_k^3}$$

$$\text{Kurtosis } \gamma_{2,k} = \frac{\sum_{i=1}^{N}(x_{i,k} - \mu_k)^4}{N\sigma_k^4}$$

Maximum

Minimum

where N is the number of cells.

The first two are standard population features which are suitable when a feature is

assumed to be normally distributed. The addition of the second two means we no longer have to make this assumption.

An alternative would be to fit a parametric distribution to each feature, but this would require a potentially time-consuming fitting procedure in addition to a preliminary analysis to determine the best fitting distribution for each feature.

A possible disadvantage of this method is that we may loose important information about correlations between features at the cell level. For example, suppose that feature A is negatively correlated with B in one sample, but positively correlated in another. If features A and B cover the same range in both samples then the simple summary statistics above would not pick up a difference between the two. However, our decision to include marker ratios should adequately cover any second order interaction effects.

5.9.3 Random forests classification of cells

Machine learning schemes have been used in several studies to automatically classify cells based on phenotype. Two general methods are support vector machines (SVMs) and random forests. Both are able to handle data sets with thousands of features, many of which may be uninformative, and do not require a time-consuming feature selection approach common to many other classification algorithms [Díaz-Uriarte and Alvarez de Andrés, 2006, Statnikov et al., 2008].

However random forests also has the advantage of not requiring a difficult parameter optimisation step, and can also handle multi-class (categorical) data directly, hence our reason for choosing it. See Section 2.7 for background information, including introductory references.

The classification results in this section are presented in the form of confusion matrices, which show the number of correct and incorrect predictions for each true class, and are also colour coded to show the proportion of predictions for each class. See Appendix B.1

for an explanation.

5.9.3.1 Image classification: Egr1

The first test of the random forest classifier used the Egr1 stained clump data set consisting of 20 images of each concentration (0, 50 and 200 $\mathrm{ng\,ml^{-1}}$, see Section 5.3.2). Stratified leave-one-out cross validation was used: 19 images of each concentration were used as a training set with the remaining one image from each concentration used to test the classifier; this was repeated a further 19 times so that all images were tested once.

Table 5.5a shows the classification results obtained by using all computed image features.

The choice of Egr1 as a marker was made by an expert as it was expected to give a visible change in response over the concentration range, so it is therefore of interest to see how well the classifier performs in the absence of the marker. This is done by removing all marker related image features (any image feature which is directly or indirectly dependent on a marker related measurement), and the classification results are shown in Table 5.5b.

Clearly the removal of the marker leads to a deterioration in performance, but the classifier still appears to pick up useful information from the DNA and phalloidin channels.

The next question is how dependent are the results on the distribution of each feature, as opposed to individual cell features. This was considered by attempting to classify each cell independently. The cross-validation method used for whole-image classification was again used, except that at each fold all individual cell features were used.

The reason for ensuring all cells from an image were in either the training or testing set but not split across both was that although all images should be acquired under identical conditions in practice there may be some common bias within an image. If the classifier subsequently picks out a feature that is only common to a particular image this will be reflected in reduced classification performance instead of an unwarranted successful

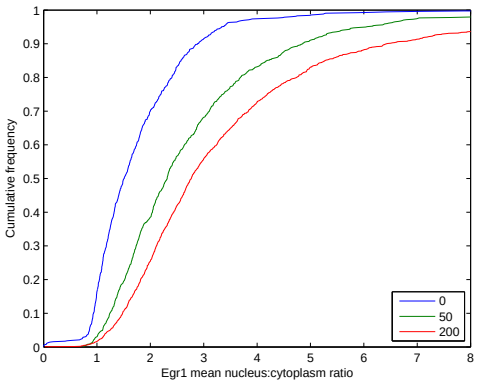


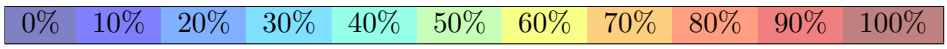
Figure 5.19: CDF plot of the ratio of the mean nuclear EGR1 levels to the mean cytoplasmic EGR1 level in cells treated with 0, 50 and 200 ng ml^{-1} IGF2.

classification.

The results can be seen in Table 5.6a and b. When all features are used the classifier performs quite well on the control and 200 ng ml^{-1} cells, but quite poorly on the 50 ng ml^{-1} cells. Intuitively this could be because the control and 200 ng ml^{-1} doses are designed to provide cells which are resting and fully activated respectively, and so the distributions should be quite narrow in range. The 50 ng ml^{-1} cells should show a range of responses which encompasses the two states, so a classifier of individual cells performs poorly whereas one which takes into account the whole distribution can distinguish between the three cases.

If this is repeated after excluding all marker features the classifier performs very poorly, and seems to default to assigning cells to the 200 ng ml^{-1} class. This is likely to indicate a lack of informative information, as evidenced by studying the features deemed most important.

If the assumption that the overall state of the 50 ng ml^{-1} cells should lie between the control and 200 ng ml^{-1} cells is correct then we would expect the empirical CDF of informative features for the 50 ng ml^{-1} cells to lie in between those of the other two classes, as seen for example in Figure 5.19.



Predicted ng ml^{-1}	True ng ml^{-1} 0	50	200
0	20	2	0
50	0	18	0
200	0	0	20

(a) With Egr1

Predicted ng ml^{-1}	True ng ml^{-1} 0	50	200
0	13	3	1
50	4	15	3
200	3	2	16

(b) Without Egr1

Predicted ng ml^{-1}	True ng ml^{-1} 0	50	200
0	20	0	0
50	0	16	5
200	0	4	15

(c) With Foxo3a

Predicted ng ml^{-1}	True ng ml^{-1} 0	50	200
0	18	3	2
50	1	15	5
200	1	2	13

(d) Without Foxo3a

Table 5.5: Automatic classification of whole images of Egr1 or Foxo3a stained cells, using features from all three channels, and excluding marker features.

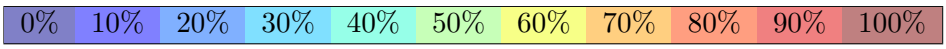
5.9.3.2 Image classification: Foxo3a

The classifier was also trained and tested on the Foxo3a stained images, with the results shown in Tables 5.5c,d for images and Table 5.6c,d for cells. As before discrimination between control and treated cells is good, but performance is reduced between cells treated with 50 or 200 ng ml^{-1} IGF2.

5.10 Feature importance

Visual examination of the images shows several features, particularly the nuclear expression levels, which help to distinguish between the different concentrations. This can be verified using the feature importance scores from the random forest classification, which could potentially suggest other features to be investigated.

Tables 5.7 and 5.8 show the 20 most important features for the individual cell classifier. These importance values are the means of the 20-fold cross validation procedure described above. As expected the marker intensity features are prominent, either in the nucleus or



		True ng ml^{-1} 0	50	200
Predicted ng ml^{-1}	0	620 (81.4%)	168 (21.7%)	107 (7.0%)
	50	51 (6.7%)	268 (34.6%)	129 (8.5%)
	200	91 (11.9%)	338 (43.7%)	1283 (84.5%)

(a) With Egr1

		True ng ml^{-1} 0	50	200
Predicted ng ml^{-1}	0	337 (44.2%)	132 (17.1%)	138 (9.1%)
	50	68 (8.9%)	260 (33.6%)	154 (10.1%)
	200	357 (46.9%)	382 (49.4%)	1228 (80.8%)

(b) Without Egr1

		True ng ml^{-1} 0	50	200
Predicted ng ml^{-1}	0	811 (98.2%)	11 (1.6%)	11 (1.3%)
	50	2 (0.2%)	350 (51.0%)	216 (26.2%)
	200	13 (1.6%)	325 (47.4%)	597 (72.5%)

(c) With Foxo3a

		True ng ml^{-1} 0	50	200
Predicted ng ml^{-1}	0	510 (61.7%)	166 (24.2%)	193 (23.4%)
	50	129 (15.6%)	287 (41.8%)	168 (20.4%)
	200	187 (22.6%)	233 (34.0%)	463 (56.2%)

(d) Without Foxo3a

Table 5.6: Automatic classification of individual Egr1 or Foxo3a stained cells, using features from all three channels, and excluding marker features. Figures in brackets are column-wise percentages indicating the proportion of cells in each predicted class.

in variants of the nucleus to cytoplasm ratio. It is important to note that several features are correlated, for instance higher mean expression levels of a marker would be expected to be accompanied by a higher maximum value.

	Feature	Importance $\times 10^{-3}$
1	Marker ratio: nuclear outer mean:cytoplasm inner mean	50.655
2	Marker: nuclear mean	46.484
3	Marker: nuclear stdev	42.088
4	Marker: nuclear max	28.665
5	Marker: nuclear outer mean	24.279
6	Marker: nuclear outer max	22.109
7	DNA: nuclear sum	21.041
8	DNA: nuclear mean	20.321
9	DNA: nuclear stdev	17.621
10	DNA: nuclear max	17.331
11	Marker: nuclear outer stdev	16.752
12	DNA: nuclear min	13.628
13	Marker ratio: cytoplasm inner mean:cytoplasm outer mean	11.980
14	Marker: nuclear outer sum	11.722
15	Marker: nuclear sum	11.232
16	Cytoplasm: cytoplasm min	10.669
17	Marker ratio: nuclear outer sum:cytoplasm inner sum	10.016
18	Marker ratio: nuclear sum:cytoplasm sum	9.908
19	Marker ratio: nuclear mean:cytoplasm mean	8.965
20	Marker: cytoplasm outer max	7.551

Table 5.7: 20 most important features for classifying individual Egr1 stained cells (Table 5.6a). .

5.10.1 Time-dose classification

The classification and feature importance methodology can also be applied to the time-concentration data set (Section 5.3.1) to find temporally varying features. Unfortunately we are restricted by the low numbers of images available for each time/dose combination, typically around four (though in some cases fewer images are available).

One option is to combine multiple groups of images (based on the advice of a collaborator), and assume the changes that occur between the groups will not unduly affect the classification. We tested two classifiers as follows.

- Group the $50\,\mathrm{ng\,ml^{-1}}$, $75\,\mathrm{ng\,ml^{-1}}$ and $100\,\mathrm{ng\,ml^{-1}}$ images which should show a mod-

	Feature	Importance $\times 10^{-3}$
1	Marker ratio: nuclear outer mean: cytoplasm inner mean	137.939
2	Marker: nuclear mean	55.933
3	Marker ratio: nuclear mean:cytoplasm mean	42.810
4	Marker: nuclear min	29.571
5	Marker: nuclear outer mean	22.286
6	Marker ratio: nuclear outer mean:cytoplasm outer mean	19.381
7	DNA: nuclear min	18.060
8	Marker: cytoplasm outer min	17.425
9	Marker: nuclear sum	16.758
10	Marker: nuclear outer min	16.568
11	Marker: cytoplasm min	15.332
12	Marker: nuclear outer sum	11.418
13	Marker ratio: nuclear outer sum:cytoplasm inner sum	10.503
14	Marker: cytoplasm inner stdev	7.464
15	Marker: cytoplasm stdev	7.043
16	Marker: cytoplasm inner min	6.361
17	Marker: cytoplasm min	5.843
18	Marker ratio: cytoplasm inner mean:cytoplasm outer mean	5.107
19	DNA: nuclear stdev	5.003
20	Marker ratio: nuclear outer sum:cytoplasm outer sum	4.913

Table 5.8: 20 most important features for classifying individual Foxo3a stained cells (Table 5.6c). .

erate level of activation, classify against all time points. Results are shown in Table 5.9.

- Group the 15 min and 30 min images, making the assumption that changes in markers have started to occur by this point. Results are shown in Table 5.10.

Beginning with the time point classification, it appears that the classifiers using any of the four markers are predictive, implying that temporal changes have occured in the cells. Although the numbers of images are very small we can speculate about how we might use these results in a future study with more data. For instance, whereas Egr1 (Table 5.9a) shows good classification for all time points Foxo3a (Table 5.9b) shows poorer classification, which may imply the temporal behaviour of this marker is more variable,

and thus more difficult to classify. The pAkt classifier (Table 5.9c) seems to have some difficulty in distinguishing the 15 min and 120 min time points, which may imply the cells are returning to their initial state after this time. The pS6 classifier performs poorly at 30 min which may be due to the particularly low number of images for this point, though it also hints at a possible lack of activity at such an early time point, and at later time points it is difficult to distinguish between 60 and 120 min, suggesting its activity may have reached a peak.

Moving on to the dosage classifiers on the combined 15 min and 30 min images (Table 5.10 it is clear that most markers do not provide sufficient information to distinguish between the different concentrations at these times, though the control cells can generally be picked out implying there is some weak activity.

As expected the feature importances for these classifiers are dominated by marker related features. However looking beyond these some interesting features can be picked out. For example, the area of the nuclei appears to change over time, as illustrated in Figure 5.20. At non-zero concentrations the plots of the 60 and 120 min cells is generally to the left of the 15 and 30 min cells, indicating a decrease in expected nuclear size over time. IGF2 causes cells to proliferate, and these plots might suggest a high density of cells causes them to squash up, or it could be that many cells have undergone mitosis and are therefore smaller.

5.11 Classification conclusions

The main output of this section is a classification framework for the analysis of image features derived from individual cell features. The classifier has successfully distinguished between cells activated to different degrees, and has also shown some discrimination ability on a second dataset where the number of images is very limited.

0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

		True min			
		15	30	60	120
Predicted min	15	14 (87.5%)	1 (9.1%)	0 (0.0%)	0 (0.0%)
	30	2 (12.5%)	9 (81.8%)	0 (0.0%)	0 (0.0%)
	60	0 (0.0%)	1 (9.1%)	13 (100.0%)	0 (0.0%)
	120	0 (0.0%)	0 (0.0%)	0 (0.0%)	11 (100.0%)

(a) Egr1

		True min			
		15	30	60	120
Predicted min	15	5 (45.5%)	2 (16.7%)	0 (0.0%)	1 (9.1%)
	30	3 (27.3%)	8 (66.7%)	2 (16.7%)	1 (9.1%)
	60	1 (9.1%)	2 (16.7%)	8 (66.7%)	2 (18.2%)
	120	2 (18.2%)	0 (0.0%)	2 (16.7%)	7 (63.6%)

(b) Foxo3a

		True min			
		15	30	60	120
Predicted min	15	7 (77.8%)	0 (0.0%)	0 (0.0%)	1 (10.0%)
	30	0 (0.0%)	9 (81.8%)	1 (7.7%)	0 (0.0%)
	60	0 (0.0%)	2 (18.2%)	12 (92.3%)	0 (0.0%)
	120	2 (22.2%)	0 (0.0%)	0 (0.0%)	9 (90.0%)

(c) pAkt

		True min			
		15	30	60	120
Predicted min	15	12 (100.0%)	1 (14.3%)	0 (0.0%)	0 (0.0%)
	30	0 (0.0%)	1 (14.3%)	0 (0.0%)	1 (9.1%)
	60	0 (0.0%)	2 (28.6%)	8 (66.7%)	4 (36.4%)
	120	0 (0.0%)	3 (42.9%)	4 (33.3%)	6 (54.5%)

(d) pS6

Table 5.9: Automatic time point classification of whole images stained with one of four markers.

		True ng ml^{-1}					
		0	25	50	75	100	200
Predicted ng ml^{-1}	0	2 (50.0%)	1 (12.5%)	1 (9.1%)	0 (0.0%)	1 (12.5%)	0 (0.0%)
	25	0 (0.0%)	3 (37.5%)	7 (63.6%)	1 (12.5%)	0 (0.0%)	4 (50.0%)
	50	0 (0.0%)	0 (0.0%)	1 (9.1%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
	75	2 (50.0%)	1 (12.5%)	0 (0.0%)	3 (37.5%)	2 (25.0%)	1 (12.5%)
	100	0 (0.0%)	1 (12.5%)	0 (0.0%)	2 (25.0%)	4 (50.0%)	2 (25.0%)
	200	0 (0.0%)	2 (25.0%)	2 (18.2%)	2 (25.0%)	1 (12.5%)	1 (12.5%)

(a) Egr1

		True ng ml^{-1}					
		0	25	50	75	100	200
Predicted ng ml^{-1}	0	8 (100.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (12.5%)
	25	0 (0.0%)	5 (62.5%)	4 (50.0%)	3 (42.9%)	0 (0.0%)	1 (12.5%)
	50	0 (0.0%)	1 (12.5%)	3 (37.5%)	1 (14.3%)	0 (0.0%)	1 (12.5%)
	75	0 (0.0%)	0 (0.0%)	1 (12.5%)	1 (14.3%)	0 (0.0%)	1 (12.5%)
	100	0 (0.0%)	2 (25.0%)	0 (0.0%)	1 (14.3%)	6 (75.0%)	2 (25.0%)
	200	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (14.3%)	2 (25.0%)	2 (25.0%)

(b) Foxo3a

		True ng ml^{-1}					
		0	25	50	75	100	200
Predicted ng ml^{-1}	0	8 (100.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
	25	0 (0.0%)	7 (100.0%)	1 (16.7%)	1 (16.7%)	1 (12.5%)	1 (14.3%)
	50	0 (0.0%)	0 (0.0%)	2 (33.3%)	2 (33.3%)	1 (12.5%)	0 (0.0%)
	75	0 (0.0%)	0 (0.0%)	1 (16.7%)	0 (0.0%)	1 (12.5%)	1 (14.3%)
	100	0 (0.0%)	0 (0.0%)	1 (16.7%)	0 (0.0%)	1 (12.5%)	2 (28.6%)
	200	0 (0.0%)	0 (0.0%)	1 (16.7%)	3 (50.0%)	4 (50.0%)	3 (42.9%)

(c) pAkt

		True ng ml^{-1}					
		0	25	50	75	100	200
Predicted ng ml^{-1}	0	7 (58.3%)	1 (14.3%)	2 (33.3%)	1 (14.3%)	0 (0.0%)	0 (0.0%)
	25	1 (8.3%)	3 (42.9%)	1 (16.7%)	0 (0.0%)	1 (16.7%)	2 (33.3%)
	50	0 (0.0%)	1 (14.3%)	0 (0.0%)	3 (42.9%)	1 (16.7%)	2 (33.3%)
	75	0 (0.0%)	0 (0.0%)	1 (16.7%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
	100	4 (33.3%)	1 (14.3%)	2 (33.3%)	1 (14.3%)	4 (66.7%)	0 (0.0%)
	200	0 (0.0%)	1 (14.3%)	0 (0.0%)	2 (28.6%)	0 (0.0%)	2 (33.3%)

(d) pS6

Table 5.10: Automatic dose classification of whole images fixed and stained with one of four markers after 15 min or 30 min (both timepoints combined into one).

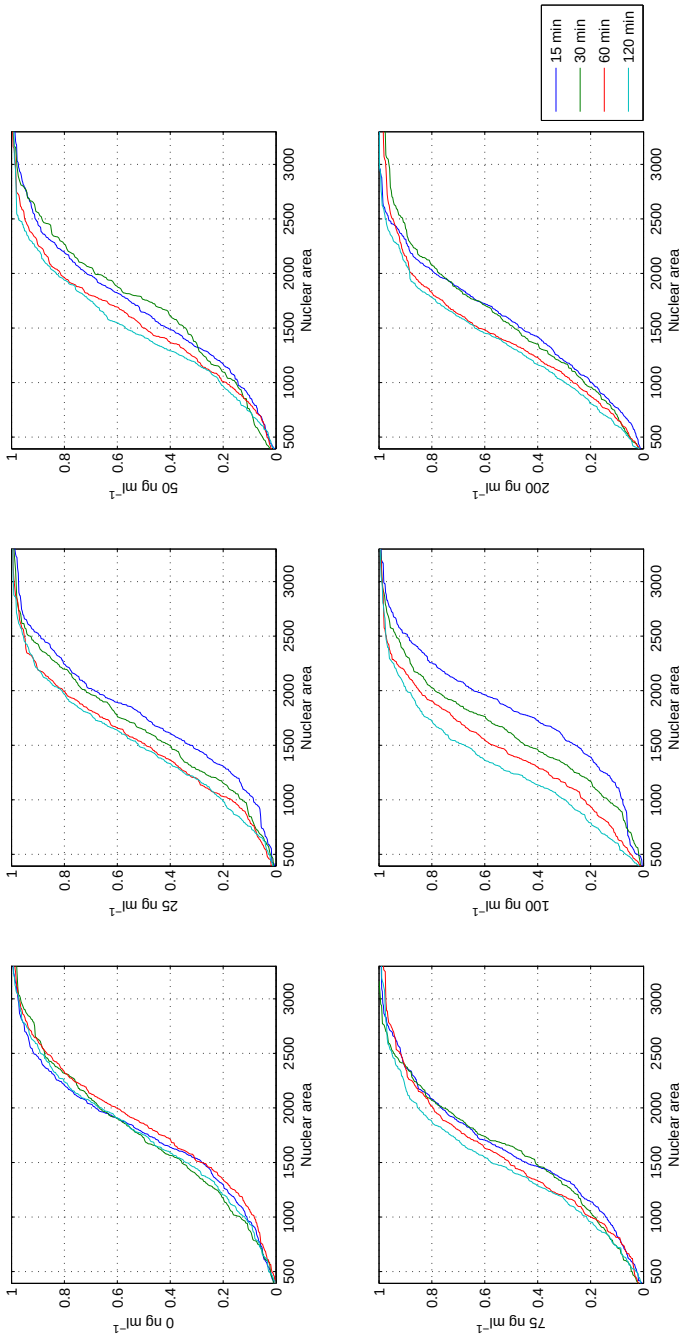


Figure 5.20: CDF plots of the nuclear size for different concentrations and time points, ignoring any markers. The concentration of IGF2 ($0\,\mathrm{ng\,ml^{-1}}$ to $200\,\mathrm{ng\,ml^{-1}}$) is indicated on the left of each plot (control is at the top left). Line colours indicate the time in minutes.

The random forest feature importances showed that the top ranking features were the ones that intuitively make sense, and are easily interpretable since the cell features used for the classification was deliberately restricted.

This classification framework can now be applied to other more difficult problems.

5.12 Spatial analysis

So far we have looked at individual features within cells, and the distribution of those features across populations of cells. We would now like to consider whether there are dependencies between cells, and whether the spatial configuration of clumps of cells affects the behaviour of cells.

We are still restricted to fixed images. Although live imaging would be a possibility for these experiments it is not extensible to other types of sample which we might wish to analyse in future, for instance tissue biopsy samples.

Note that due to limited data this should be considered a pilot study which demonstrates a method that could be applied to a larger data set in future.

5.12.1 Analysis of spatial interactions

To date there has been limited research into quantitatively analysing spatially-correlated interactions between cells, especially when considering interactions between more than two cells.

Weyn et al. [1999] was one of the first to use neighbourhood graphs to analyse histopathology images, which are characterised by very dense cells with variable staining quality. These graphs were created by placing a node at each detected nucleus, finding the Voronoi partition of the nodes and adding an edge between all touching Voronoi regions.

This was subsequently generalised to use a neighbourhood function which could include cells that were not direct neighbours to form what was termed a *cell-graph*, from which network parameters were extracted and used to classify images of biopsy tissue as healthy or cancerous [Gunduz et al., 2004], or to compare the heterogeneity in spatial density of tissues [Fernandez-Gonzalez et al., 2005].

In essence these approaches only use the locations of cells, and aim to characterise the

overall structure of a tissue. Additional information can be included by classifying cells into one of a limited number of types (for example mature cells, precursor cells, extracellular matrix), and either building separate cell graphs for each tissue type [Bilgin et al., 2010], or assigning a colour to the edges of the graph depending on the types of cell on that edge [Altunbay et al., 2010].

Su et al. [2009] looks at a binary state (for example profilerating or resting) associated with each cell and searches for changes in the variation of the output with distance using a histogram based method which they term local cell metrics (LCM)— in other words it describes how the probability of two cells having the same classification varies with distance.

5.12.2 Geographic measures of spatial autocorrelation

Looking at other research domains, the problem of quantifying spatially varying data is one of the issues handled by geographers and epidemiologists. Several spatial statistics measures can be found in the geographical information systems literature [Unwin, 1996], and are typically used to analyse changes between geographic regions, for example to see whether districts of high poverty in a city tend to be closely sited or are distributed randomly throughout the city, or to look at the prevalence and spread of diseases. Geographical measures such as the semivariogram, Moran's I, and Geary's C, have been used as a way of characterising the texture of tissues in X-ray images [Silva et al., 2004, da Silva et al., 2008, Braz Junior et al., 2009], though using spatially binned pixels rather than cells.

5.12.2.1 Semivariogram

The LCM idea appears to be conceptually similar to the semivariogram [Clark, 1979]. The semivariogram γ is defined as the expected squared difference of a stochastic process $Z(\mathbf{x})$

between two locations $\mathbf{x_1}$ and $\mathbf{x_2}$

$$2\gamma(\mathbf{x_1}, \mathbf{x_2}) = E\left(|Z(\mathbf{x_1}) - Z(\mathbf{x_2})|^2\right)$$

The factor of 2 is for mathematical convenience. If the process is stationary and isotropic then the semi-variogram becomes a function of the distance h between two points

$$\gamma(\mathbf{x_1}, \mathbf{x_2}) = \gamma'(h)$$

h can be any distance measure, for instance the Euclidean norm $\|\mathbf{x_1} - \mathbf{x_2}\|$.

5.12.2.2 Moran's I

Moran's I is a measure of global spatial autocorrelation which gives a value of around +1 for perfect correlation and around −1 for perfect dispersion [Moran, 1950]

$$I = \frac{N}{\sum_{i,j} w_{ij}} \frac{\sum_{i,j} w_{ij} \overbrace{(Z_i - \bar{Z})(Z_j - \bar{Z})}^{\text{Product of differences from mean}}}{\sum_i (Z_i - \bar{Z})^2} \tag{5.1}$$

where $\bar{Z}$ is the mean value of Z, N is the number of edges in W, and w_{ij} are the elements of an interaction matrix W. w_{ij} may just indicate the presence (1) or absence (0) of an edge between two nodes, or alternatively it may contain a weight indicating the strength of interaction between two nodes.

W does not have to be symmetrical— for example one common weight is based on the ratio of the shared boundary between regions S_i and S_j to the total boundary

$$w(i, j) = \frac{|\partial S_i \cap \partial S_j|}{|\partial S_i|}$$

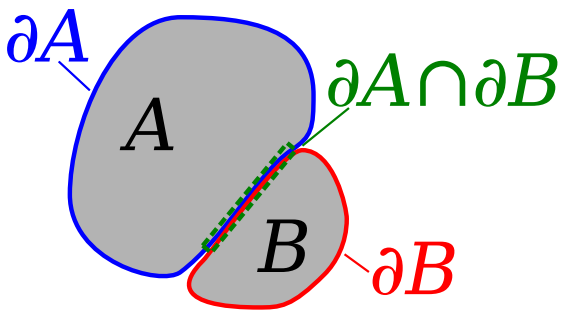


Figure 5.21: An example of an interaction weight based on the shared boundary. ∂A and ∂B are the boundaries of regions A and B, $\partial A \cap \partial B$ is the subset of the boundaries which are shared between A and B. A weight (which is non-symmetric) for the influence of B on A can be defined as the length of the shared boundary divided by the total boundary of A.

It is also common to normalise each row of W

$$\sum_j W_{ij} = 1$$

so that row i contains the proportion of contributions of each of the other regions j to region i. See Figure 5.21 for a graphical description.

The expected value of I assuming no correlation is $\frac{-1}{N-1}$. The theoretical variance is more complicated, and is given in Appendix A.3.

A possible disadvantage of Moran's I is that the normalisation is done over a single graph. This is the natural thing to do with geospatial data where only a single set of measurements are taken, but may not be so suitable when Moran's I has to be calculated for multiple graphs.

5.12.2.3 Other measures

A closely related measure is Geary's C, which should be more sensitive to local correlations since it is calculated using differences between pairs of nodes instead of between a node and the mean [Geary, 1954].

Morans' I and Geary's C are both global measures— they assume the sample area is homogeneous, and a single statistic is output for each area. In some cases there may be

local correlations which are masked by the overall global calculation, so local indicators of spatial association (LISA) can also be used [Anselin, 1995, Getis and Ord, 1992, Ord and Getis, 1995].

In spatial regression problems spatial dependencies can be modelled using a Markov random field (MRF), with a weighting parameter which is estimated as part of the estimation process [Kazembe and Mpeketula, 2010]. However, it requires a model to be fitted so cannot be used to check for spatial dependencies on its own.

5.12.2.4 Examples

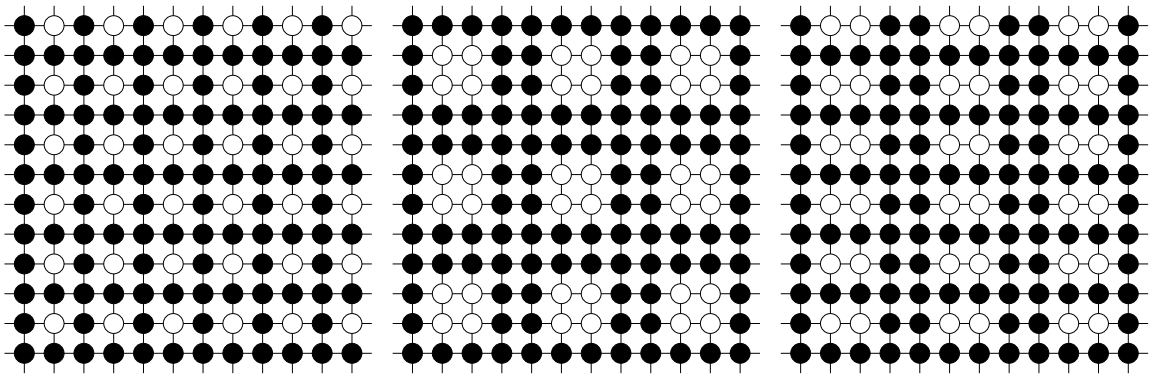
The meaning of these statistics can be understood by an analysis of artificial data. Figure 5.22 shows examples of networks with highly dispersed cells (negative Moran's I), clustered (positive Moran's I) and equivalent to random. Note that Moran's I is calculated from relations between cells, and the lack of a significant value does not mean there is no spatial structure, as can be seen from the regular pattern in 5.22c.

Figure 5.23 shows an example of a semivariogram. The periodic nature of the data in 5.23a can be clearly seen.

Local measures are useful when trying to identify small areas of abnormal activity amongst a large population. However, in our study cells are ostensibly derived from the same parent population, and the aim is to characterise the overall behaviour of cells in a clump. This suggests employing the global measures such as Moran's I and the semivariogram.

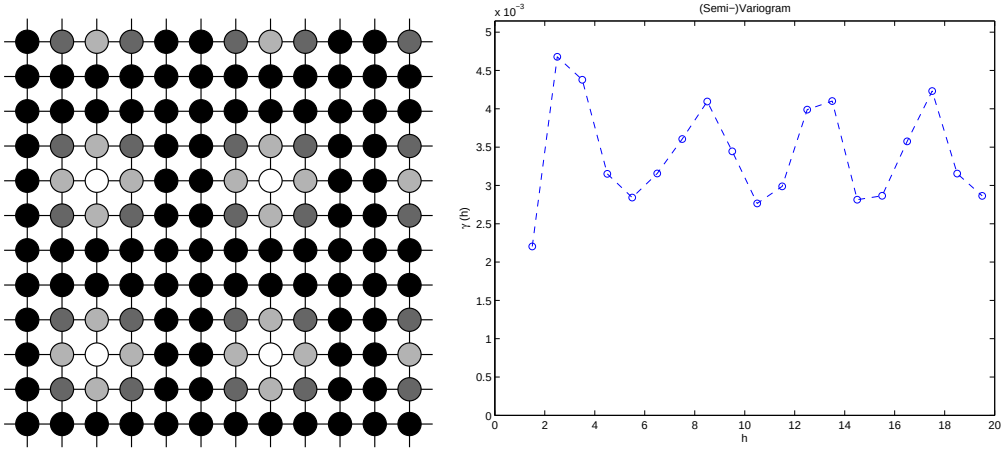
5.12.3 Clump sizes

In our work clumps were found by converting the segmentation of cells into a graph, with a node at the centre of mass of each cell, and an edge wherever two nodes shared a common boundary. A clump was defined as a connected subcomponent.



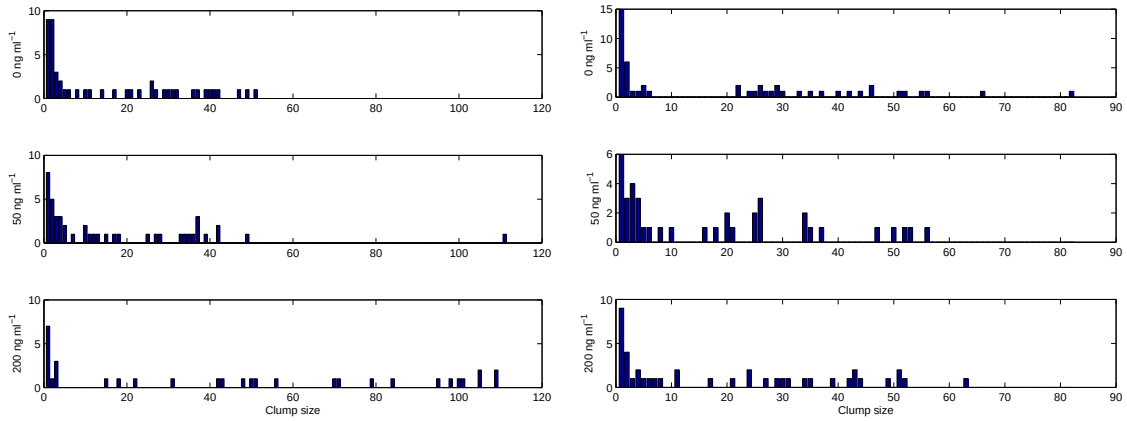
(a) Separated cells, $I = -0.33$ (b) Clustered cells, $I = 0.33$ (c) Partially clustered cells, $I = 0.00$

Figure 5.22: Moran's I calculated on an artificial network of cells. Cells are in one of two states (black or white). Green lines indicate links between cells, the total number of cells is the same. (a) Sparsely distributed cells. I is negative, indicating a more dispersed distribution than would be expected at random. (b) Clusters of four cells. I is positive, indicating a more clustered distribution than would be expected at random. (c) Clusters of two cells. I is close to the value expected for a random distribution.



(a) Gaussian smoothed cell responses (b) Semivariogram $E\left(|Z(\mathbf{x_1}) - Z(\mathbf{x_2})|^2\right)$

Figure 5.23: Semivariogram for an artificial network of cells. (a) A network formed of isolated cells which have been Gaussian smoothed, which means neighbouring cells take on a value more similar to each other. (b) Semivariogram showing the periodic variation in the expected difference as a function of distance between cells.



(a) Egr1 labelled images (b) Foxo3a labelled images

Figure 5.24: Distribution of clump sizes for the Egr1 and Foxo3a stained images. There are a large number of single cells in all images, which are usually on the edge of an image, possibly belonging to a neighbouring clump which is outside the field of view.

The distribution of clump sizes is shown in Figure 5.24. It is clear that a large number of clumps are actually single cells or pairs of cells, which in practice are frequently due to parts of a neighbouring clump encroaching on the field of view. The remaining clump sizes are spread out up to around a clump size of 50, after which the distribution of sizes is too spread out to assume the samples are from the underlying continuous distribution. The 200 ng ml^{-1} IGF2 Egr1 stained images appear to be biased towards larger clump sizes compared with Foxo3a images— this is due to experimental variation rather than being a property of the marker. The only difference between the two sets of images is in the choice of marker staining, which is performed after the cells have been fixed.

5.12.4 Cell distances

A possible definition of the interaction matrix W for describing the expected strength of interaction has already been made, but this assumes interactions can only occur between direct neighbours, for instance through gap junction signalling.

In order to consider other factors, for instance diffusion of signalling molecules, it is

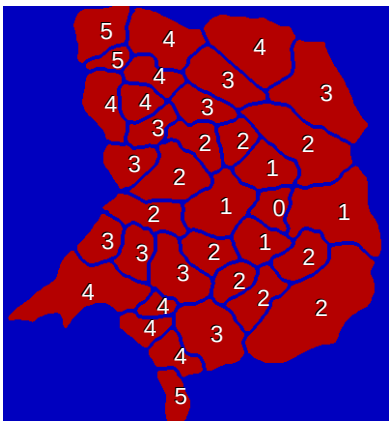


Figure 5.25: Example of a cell-cell distance measure, showing the distance of each cell from the cell labelled 0.

necessary to create some other measure which takes account of the physical distance between cells. An immediate complication is that we are considering cells at quite a high spatial resolution, so using the centre of mass as the points from which cell-cell distances are calculated may not be optimal.

An alternative is to consider the number of cells in between any two cells— this is equivalent to finding the shortest path between two nodes with the edge distances of all neighbouring cells set to one (Figure 5.25). If information is only successively propagated between neighbouring cells rather than through diffusion then this should be a good metric. This does not take into account cell size or the geometry of the clump but for short range interactions it should be acceptable as a first step for an exploring the data.

5.12.5 Moran's I: Egr1 and Foxo3a

Moran's I (Equation 5.1) can be calculated over any clump consisting of at least two cells, and for any cell feature.

The features of most interest are the nuclear and cytoplasmic levels of the two markers, Egr1 and Foxo3a, since these show the most obvious changes in appearance following the addition of IGF2.

Figure 5.26 shows the normalised values for Moran's I (z-scores) calculated on the mean

nuclear or cytoplasmic levels of Egr1 or Foxo3a in each cell, as a function of the size of the clump containing those cells, and with clumps split according to the concentration of IGF2.

The z-scores were calculated using the theoretical mean and variance (Appendix A.3), which correlated well with the statistics calculated using a permutation test (not shown).

Most clumps appear to have positive correlation, suggesting neighbouring cells are more likely to show similar levels of Egr1. As a rough guide the ± 2 z-scores are the limits beyond which $p < 0.05$ in a conventional analysis, though we should be aware of the multiple comparisons issue before drawing any conclusions.

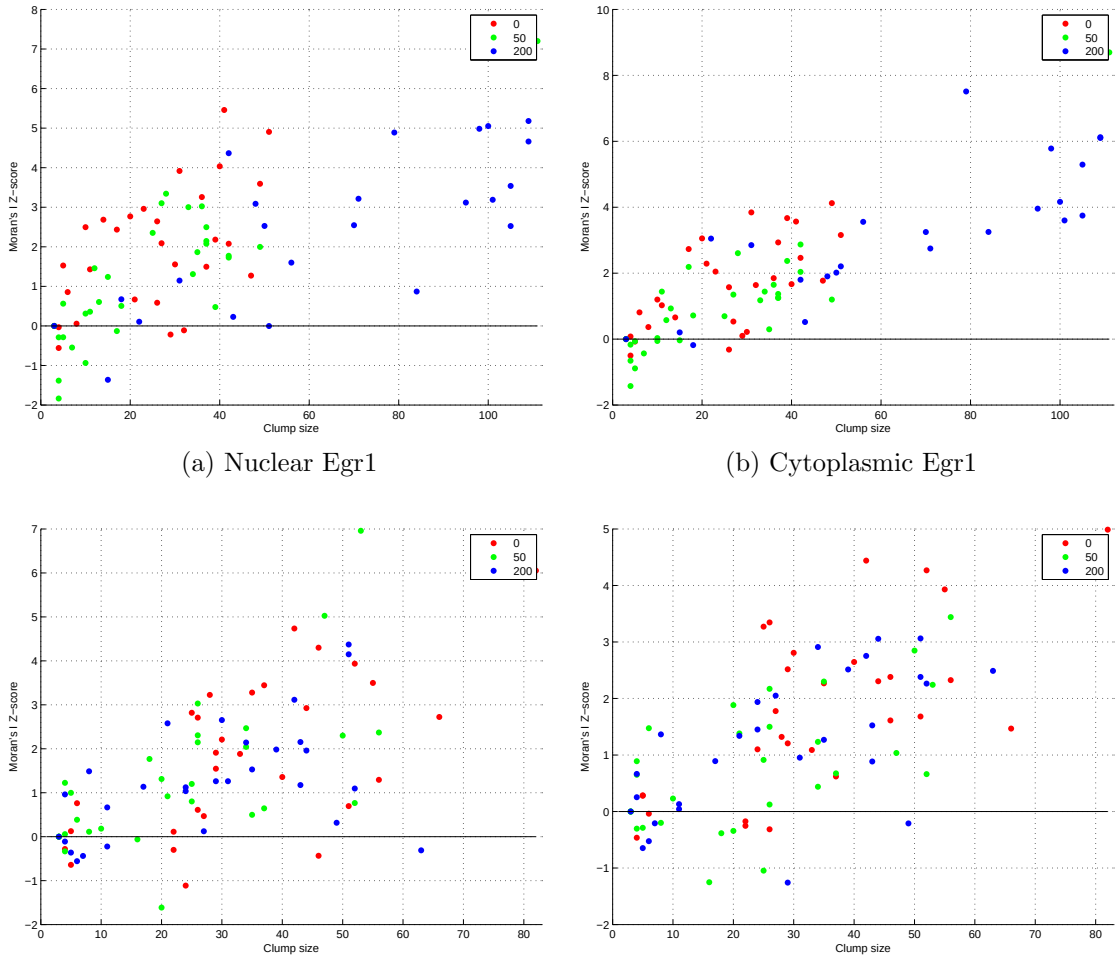
For comparison a feature which does not appear to be influenced by neighbours, the nucleus area, is shown in Figure 5.27.

5.12.6 Semi-variogram: Egr1 and Foxo3a

Moran's I suggested a tendency for neighbouring cells to show similar levels of Egr1 and Foxo3a. The semi-variogram, which shows the expected squared difference of a feature between pairs of cells as a function of their separation, should provide another way to see this. We would expect cells which are in close proximity to have a smaller difference compared with cells with a greater distance between them.

We first need to define a distance metric between cells, and we use a model where each cell is assumed to interact only with its touching neighbours irrespective of the length of shared boundary. This is equivalent to setting all distances in the adjacency matrix to one, and the distance metric is the number of cell boundaries which must be crossed to form a path between a given pair of cells.

The semivariograms for the Egr and Foxo3a mean nuclear and cytoplasmic levels are shown in Figure 5.28. A detailed version of 5.28a is shown in 5.28e, in which the individual squared difference values $(Z(\mathbf{x_1}) - Z(\mathbf{x_2}))$ are plotted to give an idea of the spread of values,



(a) Nuclear Egr1

(b) Cytoplasmic Egr1

(c) Nuclear Foxo3a

(d) Cytoplasmic Foxo3a

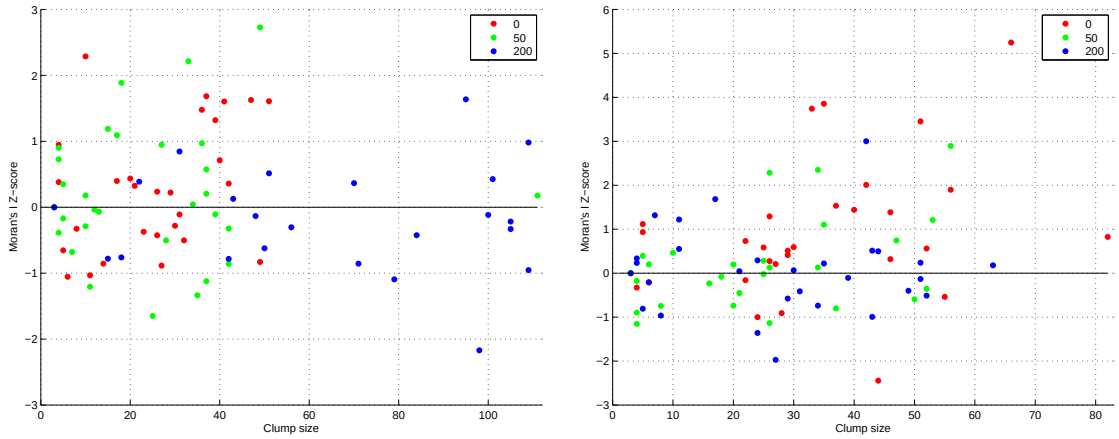
Figure 5.26: Moran's I calculated for mean Egr1 or Foxo3a levels in the nucleus or cytoplasm. The plots show normalised values (z-scores) calculated over each clump, in samples treated with 0, 50 or 200 ng ml^{-1} IGF2.

along with the standard deviation which is clearly very large. This is to be expected due to the large number of possible pairs of cells.

The standard error of the mean

$$se = \frac{s}{\sqrt{n}}$$

could be used where s is the standard deviation of $\gamma(h)$ and n the number of samples. The value of n to use requires careful consideration— it is not the number of pairwise



(a) Egr1 nuclear area (b) Foxo3a nuclear area

Figure 5.27: Normalised values (z-scores) of Moran's I calculated for the area of the nuclei in the Egr1 and Foxo3a images at different concentrations.

comparisons since they are not independent. A lower bound for n is the number of clumps containing cells a distance h apart, so this can be used as an estimate for $s_{\bar{x}}$.

A further complication is that larger clumps will have a larger number of comparisons which can bias the calculation of s, but we can use the idea of jackknifing [Efron et al., 1982] to compensate for this.

Jackknifing is the calculation of a statistic $\hat{\theta}_{(i)}$ from multiple jackknife samples $\{\tilde{\mathbf{X}}_\mathbf{i}\}_{i=1\ldots n}$ which are formed by leaving out one sample at a time

$$\begin{aligned}\tilde{\mathbf{X}}_\mathbf{i} &= \{X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n\} \\ \hat{\theta}_i &= f(\tilde{\mathbf{X}}_\mathbf{i})\end{aligned}$$

By setting the statistic $\hat{\theta}_{(i)}$ to be the mean of each jackknife sample we can obtain the

standard error of the mean [Efron et al., 1982]

$$\hat{se}_{jack} = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)}\right)^2}$$

$$\hat{\theta}_{(\cdot)} = \sum_{i=1}^{n}\frac{\hat{\theta}_{(i)}}{n}$$

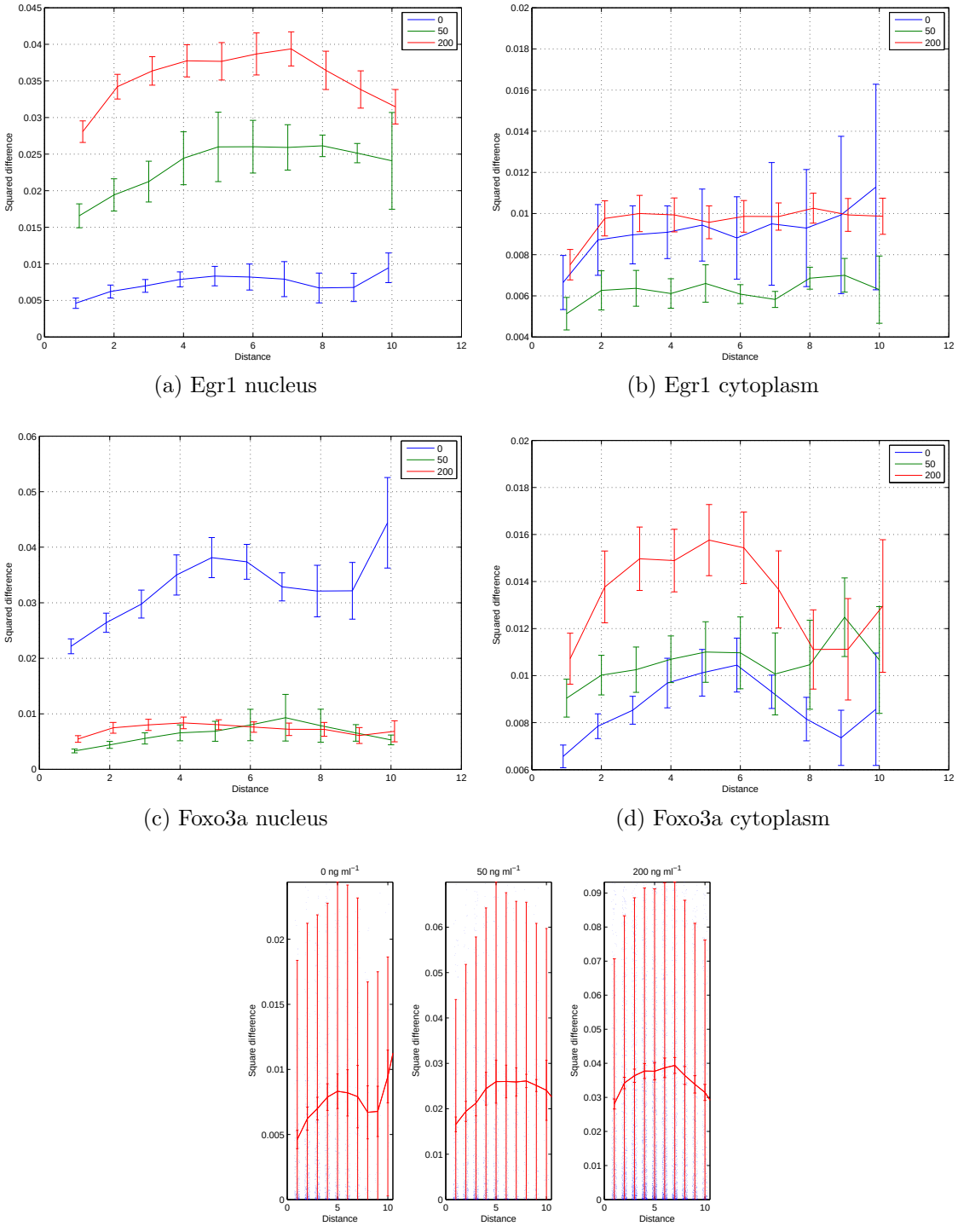
5.12.7 Spatial analysis conclusions

There appears to be evidence that the expression levels of Egr1 and Foxo3a are not randomly distributed across a clump of cells. The apparent positive value of Moran's I suggests a positive correlation between neighbouring cells, i.e. neighbouring cells are likely to show similar levels to their neighbours compared to cells further away.

Our definition of clumps does not fully take into account that some cells may intersect with the edge of the image, so the detected clump size may be inaccurate. In practice most clumps were fully included within the image, and the large number of single and pairs of cells in Figure 5.24 are due to isolated cells, or other clumps predominantly outside the field of view.

A further limitation is that only considering the clump size does not fully take into account our intuitive meaning of "clumpiness", which should embody some notion of the connectedness of each cell. For instance a dense clump of 20 cells would be expected to show more interactions and hence behave differently than 20 cells spread out in a line.

The agreement between the theoretical mean and standard deviation and that obtained from resampling suggests that there is no obvious bias in the chosen analysis method, though it does not exclude other possible sources of error, for instance inhomogeneous illumination, or poor segmentation leading to merging of neighbouring regions. This is likely to be a bigger issue with the cytoplasm than the nucleus which tends to be well-



(a) Egr1 nucleus

(b) Egr1 cytoplasm

(c) Foxo3a nucleus

(d) Foxo3a cytoplasm

(e) Egr1 nucleus detail

Figure 5.28: Semi-variograms of the Egr1 and Foxo3a mean nuclear intensity in all clumps, treated with 0, 50 or 200 ng ml^{-1} IGF2. Error bars are standard errors of the mean calculated using jackknifing. (e) Detail of the semivariogram for (a). Blue points show 10% randomly chosen pairwise differences, with dithering along the distance axis to improve visualisation of the points. Error bars show the standard deviation.

defined. There will also be edge effects arising from the spreading of cells on the edge of a clump. Neighbouring cells on the edge are likely to both spread out compared with neighbouring cells in the centre of the clump, and this may lead to an unwarranted positive value for Moran's I.

Chapter 6
Survival analysis of Ewing's sarcoma tissue microarrays

This chapter further develops the classifier from Chapter 5, and applies it as part of a novel analysis of Ewing's sarcoma in relation to clinical outcome data.

Sections 6.1–6.4 give an introduction to Ewing's sarcoma and the data to be analysed, and provide an overview of the hypotheses to be considered. This is followed by a brief introduction to survival analysis in Section 6.6.

The cell features that were extracted from the images, and their conversion into patient features which were used for much of the analysis, are described in Section 6.5.

A typical clinical-based approach to analysing biomarker data is shown in Section 6.9 before moving on to the main contributions of this work in Section 6.10, which are the use of random forests and related algorithms to analyse and visualise cell-based microscopy data.

Part of this work was presented at Microscopic Image Analysis with Applications in Biology, 2011, Heidelberg.

6.1 Ewing's sarcoma

Ewing's sarcoma is a rare form of cancer which mainly affects children and young adults. The survival rate is around 65% though it varies greatly with age— children aged 15 years or younger have an expected five year survival of around 75% compared with around 50% for age 15–19 years [Smith et al., 2010, Figure 7b].

The tumour is found in bones or soft tissues, and is caused by a chromosomal translocation (an abnormal event in which two chromosomes incorrectly recombine) which results in the formation of a fusion protein between the protein EWS and an ETS related transcription factor. This fusion protein is believed to act as an aberrant activator or repressor of proteins involved in regulating cell growth and differentiation [Toomey et al., 2010].

The mechanisms through which the protein gives rise to tumourigenesis are unclear, and the low occurrance of the tumour has meant that it has been difficult to advance research and understanding of the disease. This disease therefore presents a challenging problem in which the application of quantitative methods should enable us to extract more information from our limited data.

6.2 Hypothesis

A number of confocal images of tissue microarray (TMA) sections from multiple Ewing's sarcoma patients were available, together with some outcome data. The aim was to find links between the observed images and the eventual outcome data. Specifically, the goal was to test the following hypotheses:

- Is it possible to automatically predict the outcome of a case (in particular the final diagnosis or patient survival time) from the images?
- Are there subgroups amongst cases with the same outcome which show significantly different biological characteristics?
- Are there particular image features which are useful in distinguishing between outcomes, and what is their biological significance?

6.3 Collaborations

All of the laboratory work including the microscopy was carried out by Claudia Buehnemann in the Dunn School of Pathology, University of Oxford. The image processing steps including the segmentation and extraction of basic cell intensity features were designed and carried out by Haiyue Yu, also in the Dunn School of Pathology.

The remainder of the analysis as described in this chapter was carried out by myself.

6.4 Overview of the datasets

The TMAs were formed from tumour biopsies taken from Ewing's sarcoma patients at three different study centres: Valencia (Spain), Düsseldorf (Germany), and Leiden (The Netherlands). The tissue samples from each patient were fixed and stained with a set of fluorescent markers, and imaged on a confocal microscope by a collaborator, Claudia Buehnemann. Up to six images from different regions of the tissue sample were taken from each TMA, but due to problems in the sample preparation some images were unsuitable for further processing, and were excluded. Details of the experimental protocol are given in Appendix E.

Unfortunately a complete set of outcome data was not available for all patients, and after removal of those samples 119 patients remained. Table 6.1 shows a breakdown by dataset of the patients that were considered in this analysis. The metastasis status of two patients was unknown, but since the analysis of the survival and metastasis outcomes was done separately they were included.

Dataset	Survival outcome	Metastasis outcome
Düsseldorf	42	42
Leiden	15	13
Valencia	62	62
Total	119	117

Table 6.1: Number of patients with known survival outcome, and the numbers of those patients for which their metastasis status is also known.

6.4.1 Variation within the data

The experimental and biological variations within the data sets can be considered at multiple levels, as shown in Table 6.2. The biological variations shown are what we are interested in, i.e. do these biological differences correspond to a definable patient outcome? The experimental variations are more of a problem, particularly since we will be comparing different images and patients with each other. These may lead to non-informative and distracting biases or correlations in the data, and it may be necessary to normalise the images.

6.5 Image features

The staining and imaging protocols for all datasets were identical. Each image contained four channels, each showing a different marker, an example of which is shown in Figure 6.1.

DAPI: Labels DNA in the nucleus

CD99: Membrane bound glycoprotein expressed in 95% of Ewing sarcoma

pS6: A protein in the PI3K signalling pathway

Readout marker: One of Ki67, Foxo3a, or Egr1.

	Principal biological variations	Principal experimental variations
Cells in one image	Multiple cell types present. Stochastic variation in gene expression levels amongst identical cells.	Inhomogeneous microscope illumination (should be negligible).
Images from one patient	Images taken from different regions of the biopsy which has an anisotropic distribution of cells and local regulatory factors.	Staining inhomogenities. Different microscope conditions (should be negligible).
Patients in one dataset	Different tumour sub-types. Different tumour environments.	Tissue collection and preparation procedures may vary slightly between patient. Microscope conditions vary between patients.
Datasets	None in principle, in practice different studies may involve populations with different genetic and environmental backgrounds.	Tissue handling protocols may vary greatly between centres. Distribution of treated patients (in terms of outcome, tumour class, patient type such as sex or age) may be different.

Table 6.2: Hierarchy of principal sources of variation in the analysed images.

The readout markers were expected to be informative. Ki67 levels are typically heightened in proliferating cells, for instance during metastasis [Brown and Gatter, 2002], Foxo3a and Egr1 are both part of the PI3K signalling pathway which shows increased activity in some cancers [Pollak et al., 2004]. Experimental limitations restricted us to staining for just one of the three readout markers at a time, so we effectively have 3 sub-datasets.

For convenience we will refer to the readout marker in each image as the *readout* in the remainder of this section. Table 6.3 summarises the number of images and cells in each of these sub-datasets.

Appendix E.2 outlines the segmentation algorithms used to find the boundary of the two main compartments of a cell— the nucleus and the cytoplasm. The extracted data included the following basic image features for each cell:

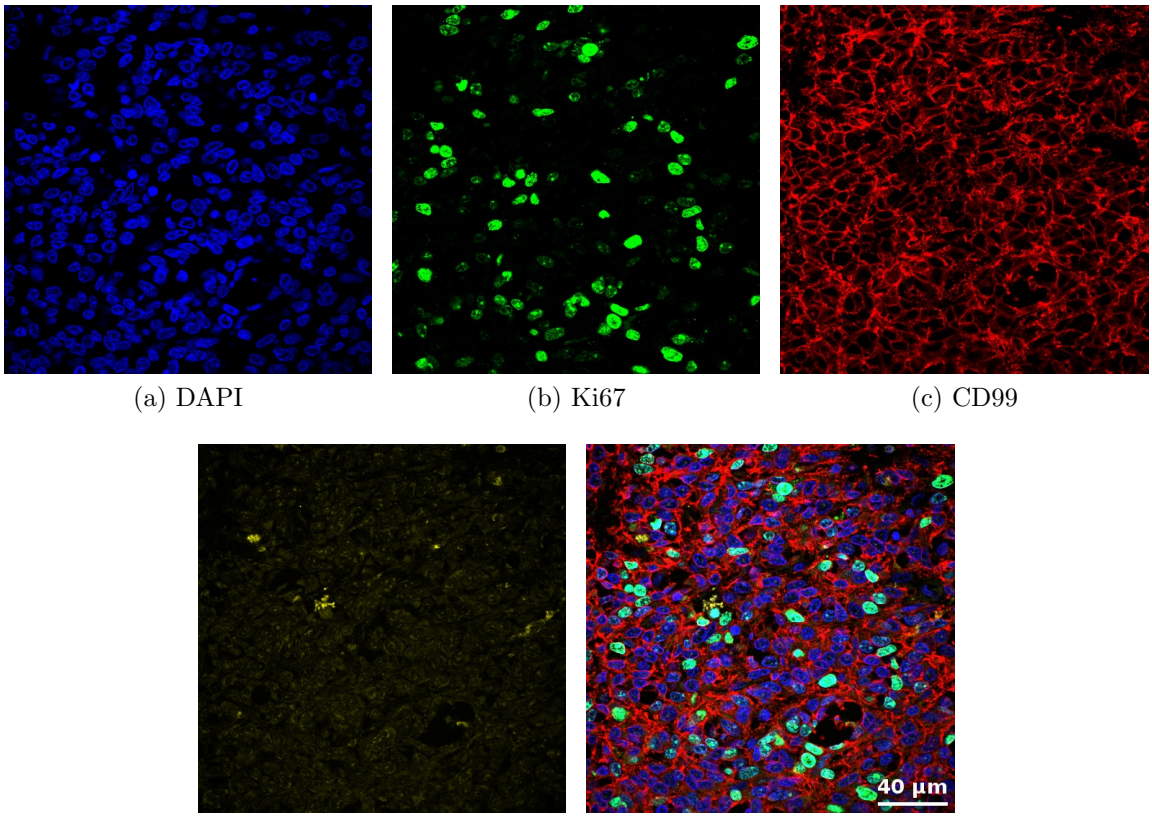
	Egr1		Foxo3a		Ki67	
Dataset	Images	Cells	Images	Cells	Images	Cells
Düsseldorf	139	32339	159	40364	210	44301
Leiden	68	20407	79	24267	94	26842
Valencia	186	43926	200	45501	196	41357
Total	393	96672	438	110132	500	112500

Table 6.3: Numbers of images and cells in each dataset. The images in this table correspond to the patients with known outcomes in Table 6.1. Multiple images for each patient were available.

- Areas (2 features):
  - Segmented nucleus area (pixels).
  - Segmented cytoplasm area (pixels).
- For each marker (DAPI, CD99, pS6, readout) in each compartment (nucleus, cytoplasm) the following intensity features (10 features per marker):
  - Total (integrated) intensity.
  - Mean intensity.
  - Standard deviation.
  - Skewness.
  - Kurtosis.

The latter four intensity features were intended to characterise the distribution of intensities of each marker within the nucleus and cytoplasm of each cell. All intensity features were calculated from an 8 bit image mapped to the range 0–1. Note that although DAPI has a predominantly nuclear localisation it may also bind RNA molecules in the cytoplasm.

This gave a total of 42 features for each cell, which were all supplied by a collaborator, Haiyue Yu. The remainder of this section describes the original method that was developed to analyse this data, and forms my contribution to the work.



(a) DAPI (b) Ki67 (c) CD99

(d) pS6 (e) Merge

Figure 6.1: An example of a TMA confocal image. The four channels are shown separately and in a merged image.

6.5.1 Derived image features

40 additional ratio features were derived from the primitive features:

- Single marker intensity ratios (8 features)
  - Ratio of the total intensity of a marker in the nucleus to the cytoplasm (DAPI, CD99, pS6, readout).
  - Ratio of the mean intensity of a marker in the nucleus to the cytoplasm.
- Dual marker intensity ratios in a single compartment (12 features)
  - Total nuclear intensity ratios between all pairs of markers (CD99/DAPI, DAPI/

readout, readout/pS6, pS6/CD99, CD99/readout, DAPI/pS6).

- Total cytoplasmic intensity ratios between the all pairs of markers.

- Dual marker intensity ratios in different compartments (24 features)
  - Total nuclear/cytoplasmic intensity ratios between all pairs of markers (CD99/DAPI, DAPI/CD99, DAPI/readout, readout/DAPI, readout/pS6, pS6/readout, pS6/CD99, CD99/pS6, CD99/readout, readout/CD99, DAPI/pS6, pS6/DAPI).
  - Mean nuclear/cytoplasmic intensity ratios between all pairs of markers.

All intensity ratios were $\log_2$-transformed so that relative changes in the numerator and denominator had equal weight. For example, doubling the numerator or denominator will change a ratio by a factor of 2 or 1/2, whereas the $\log_2$ ratio will be changed by $+1$ or -1.

Along with the basic features this gave a total of 82 features per cell.

6.5.2 Summary features

As described in the previous chapter (Section 5.9.2) we calculate basic summary features over all cells for each patient:

- Mean.
- Standard deviation.
- Skew.
- Kurtosis.

For legibility some figure labels in this chapter use abbreviated feature names. These are explained in Appendix E.4.

6.6 Survival analysis

Survival analysis in a clinical context typically involves the comparison of survival times of multiple groups of patients, and the quantification of the effect of covariates on survival.

We will mostly be interested in covariates that are potential biomarkers, measurable parameters which indicate the physiological or disease state of an organism. These can include protein levels, genetic variations, physical measurements, dynamic measurements of an injected label or tracer, etc.

Several introductory tutorials to survival analysis are available, see for example Clark et al. [2003] and subsequent articles in the series, so here we will briefly describe some of the main analysis techniques.

6.6.1 Survival times and censoring

Survival time means the time until an *event* occurs, for instance death, though it can refer to any defined event. If multiple mutually exclusive events are possible, for instance death due to different causes, a competing risks analysis is required, but here we will restrict ourselves to a single event.

Ideally the survival time of every patient would be known, but in clinical studies this is almost never the case. Patients may leave a study for a multitude of reasons, for instance moving away, dying of unrelated causes, or a study may reach its planned conclusion but since patients are enrolled at different times they will have different survival times. All these patients are termed *censored*, or more strictly speaking *right-censored* since they have a known start time but unknown time of the event.

6.6.2 Survival analysis methods

6.6.2.1 Kaplan-Meier estimate

The usual method of presenting survival data is therefore the Kaplan-Meier estimate [Kaplan and Meier, 1958], a non-parametric discrete estimate of the survivor function $S(t_j)$ which gives the probability of surviving to a time t_j after diagnosis taking into account censored patients whose survival time does not correspond to a death. This function is calculated by ordering all survival or event times $t_1, \ldots, t_k$ in increasing order, and defining the recursive relation

$$S(t_j) = S(t_{j-1}) \left(1 - \frac{d_j}{n_{j-}}\right),$$

where n_{j-} is the number of patients alive immediately prior to time t_j, d_j is the number of events occurring at time t_j, and $S(t_{j-1})$ is the proportion surviving at time t_{j-1}. Since this function is only evaluated when an event occurs the Kaplan-Meier plot has a stepped appearance. The censoring process is assumed to be random and should show no correlation with survival time, otherwise the Kaplan-Meier estimate is invalid.

A rough estimate of the variance can be obtained using Greenwood's formula [Greenwood, 1926],

$$Var\{S(t)\} = S(t) \sum_{t_j<t} \frac{d_j}{n_j(n_j - d_j)}$$

A related function is the hazard function $h(t)$ which is the instantaneous probability of an event at time t,

$$h(t) = \frac{\frac{dS(t)}{dt}}{-S(t)}$$

$$S(t) = \exp\left[-\int_o^t h(\tau)d\tau\right]$$

6.6.2.2 Cox proportional hazards model

The Cox proportional hazards model is a regression model that relates one or more predictor variables to a relative change in the hazard function $h(t)$ [Cox, 1972],

$$h(t) = h_o(t) \exp\{b_1 x_1 + \cdots + b_M x_M\}$$

It is considered to be a semi-parametric model since the link function between the predictor variables $x_1, \ldots, x_M$ and the change in hazard function is specified, but the underlying baseline hazard function $h_o(t)$ is unspecified and does not need to be known as far as calculating the effect of a variable is concerned. The baseline hazard represents the hazard function when all covariates are zero, and the coefficients $b_1, \ldots, b_M$ determine the effect of each covariate.

The key assumption which gives the model its name is that each covariate can be combined multiplicatively on the hazard, i.e. each covariate behaves proportionally irrespective of the value of any other covariate. Note that this does not preclude interactions between variables, however they must be incorporated through additional interaction terms, for example of the form $x_i x_j$.

The coefficients are also known as hazard ratios, since $b_i > 1$ implies an increased risk of an event with increasing x_i whereas $b_i < 1$ implies a decreased risk.

6.6.2.3 Log-rank test

The presence of censoring means that the usual hypothesis tests for comparing probabilistic distributions cannot be used. Instead survival distributions are often compared using the non-parametric log-rank test, also known as the Mantel-Cox test.

The log-rank test calculates the number of expected events in each group assuming both groups have the same survival distribution (the null hypothesis), using a similar idea

to the Kaplan-Meier estimate [Bland and Altman, 2004].

The number of expected deaths in group k is

$$E_k = \sum_{i=1}^{i=T} \frac{n_{k,i} d_i}{n_i},$$

where n_i is the total number alive immediately prior to time i, d_i is the total number of deaths occurring at time i, and $n_{k,i}$ is the number at risk (the number alive immediately prior to time i) in group k.

The test statistic is

$$\chi^2 = \sum_{k=1}^{G} \frac{(O_k - E_k)^2}{E_k},$$

which, provided $\sum d_i$ and $\sum n_i$ are large, should follow a χ^2 distribution with $G-1$ degress of freedom where G is the number of groups.

One disadvantage of the log-rank test is that it may fail to detect a difference if the hazard functions cross. For example, if group 1 initially has higher mortality than group 2, but over longer timescales group 2 has a higher mortality, the total number of deaths may be the same in both groups but the survival distributions are not the same.

6.6.2.4 Nonlinear methods

Difficulties arise when the assumption of proportional hazards or non-crossing hazard functions does not hold. A commonly used albeit crude way of handling this is to categorise a continuous covariate, for example by equally spaced percentiles, or by using a clinically meaningful threshold. Alternatively a threshold can be selected which gives the so-called "optimal" p-value, for example from the log-rank test. This can lead to highly biased results due to the multiple testing problem— multiple thresholds are tried which leads to artificially low p-values by chance, though correction factors can be applied [Altman et al., 1994].

Regardless of the thresholding method used it should be obvious that a lot of information is thrown away, and there may be significant variation within each group, so more advanced methods involve modelling the hazard using non-linear functions such as fractional polynomials [Royston, 1999] or locally fitted splines [Harre et al., 1988], or non-parametric methods such as random survival forests [Ishwaran et al., 2008].

6.6.2.5 Cancer survival terminology

We will briefly mention the main terms used when describing survival in cancer patients.

Overall survival (patient): In the context of an individual patient this is the time of the last known status. This may be the time of death, or it may be the time at which the patient left the study in which case the patient is known to have survived to at least that time.

Overall survival (population): In the context of a population overall survival is the proportion of patients with the diseases who survive to a given threshold, for example 3, 5 or 10 years. The threshold is often set to a level such that patients reaching that point are likely to be cured.

Event free survival (patient): This is the time between a treatment and a subsequent event of interest such as the occurrence of metastasis. If the only event that happened was death then this will be equal to the overall survival for the patient.

6.7 Image normalisation

The goal of image normalisation is to remove biases and variation due to experimental error whilst keeping the true biological variation.

One source of experimental error are the lighting inhomogenities present in some microscopes. This is not an issue with confocal microscopes, as verified in Appendix E.3.

Normalisation between images is more complicated. The tissue samples forming the three datasets were obtained from different study centres, and there is a possibility that the tissue handling protocols and conditions were slightly different.

The fixing, staining and imaging protocols for all samples were completed in the same Oxford laboratory by the same researcher at approximately the same time, but there will still be some variation in the staining procedures, and when coupled with the inherent sample variability by the time the sample reaches the microscope it is usually necessary to adjust the image acquisition settings, especially the gain which controls the sensitivity of the detector.

The images were normalised using the AQUA thresholding method [Camp et al., 2002] as part of the cell feature extraction process developed by Haiyue Yu.

6.7.1 Comparison of feature values in each dataset

One way to check for differences between datasets is to compare the distributions of feature values.

Figure 6.2 show the differences between the three datasets in terms of the areas, total intensities, and mean intensities, for one of the readout markers (Egr1) in the nucleus and cytoplasm of individual cells. The plots suggest most feature values are not grossly dissimilar between datasets, but we cannot confidently say that the variation in distributions is purely down to natural variability.

An alternative way of visualising the combined distribution of cell features is to use principal components analysis (PCA), a dimensionality reduction tool which searches for linear combinations of features so that modes of maximum variation in the data can be found. This allows us to project the data into two or three dimensions.

PCA is sensitive to the relative scales of each feature, and since different types of feature are analysed (intensities, areas) it is necessary to apply a normalisation operation. A z-score $z_i^f = (x_i^f - \mu^f)/\sigma^f$ for cell i is calculated for each feature f of each cell, where μ^f and σ^f are the mean and standard deviation of that feature across all cells in the whole data set.

The first two principal components are shown as density plots in Figure 6.3. These are 2D histograms in which the number of samples is proportional to intensity, and each dataset has been coded in a different colour. There appears to be some difference amongst the cell populations in each of the datasets.

PCA can also provide an indication of the variation between images from the same patient by calculating image level features, and comparing the variation between images from the same patient with that between images from different patients. As seen in Figure 6.4 images from the same patient tend to cluster together in the PCA reduced space. Although tumours can be very heterogeneous we would intuitively expect images from different regions of the same tumour to show higher similarity compared with those from a different tumour, so this plot is consistent with all images from one patient having been acquired under similar conditions.

We will make the assumption that all images from a single patient are drawn from the same distribution in order to combine the three readout datasets into one by calculating the patient summary features from section 6.5 over all images containing that marker. For instance, the DAPI patient features are calculated using the DAPI channels from all Egr1, Foxo3a and Ki67 images, but the readouts will obviously be calculated from a subset of

the images from a patient.

If PCA is performed on these patient-level features we start to see evidence of a systematic difference between the datasets. In Figure 6.5 the Düsseldorf and Valencia patients do not fully overlap, and the Leiden patients appear in the middle. In the following analysis we aim to establish whether there is a difference when analysing the combined dataset as one compared with analysing the separate datasets independently.

6.8 Distribution of outcomes

One of the challenges in analysing data from a rare cancer is that study sizes are often relatively small. This means results may not reach the levels of conventional statistical significance, but should instead be seen as indicative of areas for further study.

Figure 6.6 summarises the main patient outcomes in each of the datasets. Note that the survival times indicate the time of the last known patient state (alive/dead), not the time of death. Metastasis is included as it is one of the major predictors in cancer survival. This information is also shown in Table 6.4 along with the numbers of patients who did or did not die at their event free survival time.

Perhaps of greater value are the Kaplan-Meier plots shown in Figure 6.7. The size of each step (corresponding to a death) is proportional to the number of patients alive immediately prior to that time, termed the *number at risk*. This means these plots should be interpreted with care since a seemingly large drop may be due to just one patient— to assist with this the number at risk is indicated on the x-axis below the survival time.

Figure 6.7b which shows the aggregated survival for patients from all datasets is consistent with the expected survival for Ewing's sarcomas, with a typical long term survival rate of around 65%. As can be seen in Figure 6.7a the survival statistics are notably different for the Leiden dataset. Whether this is purely down to chance or due to some other factor

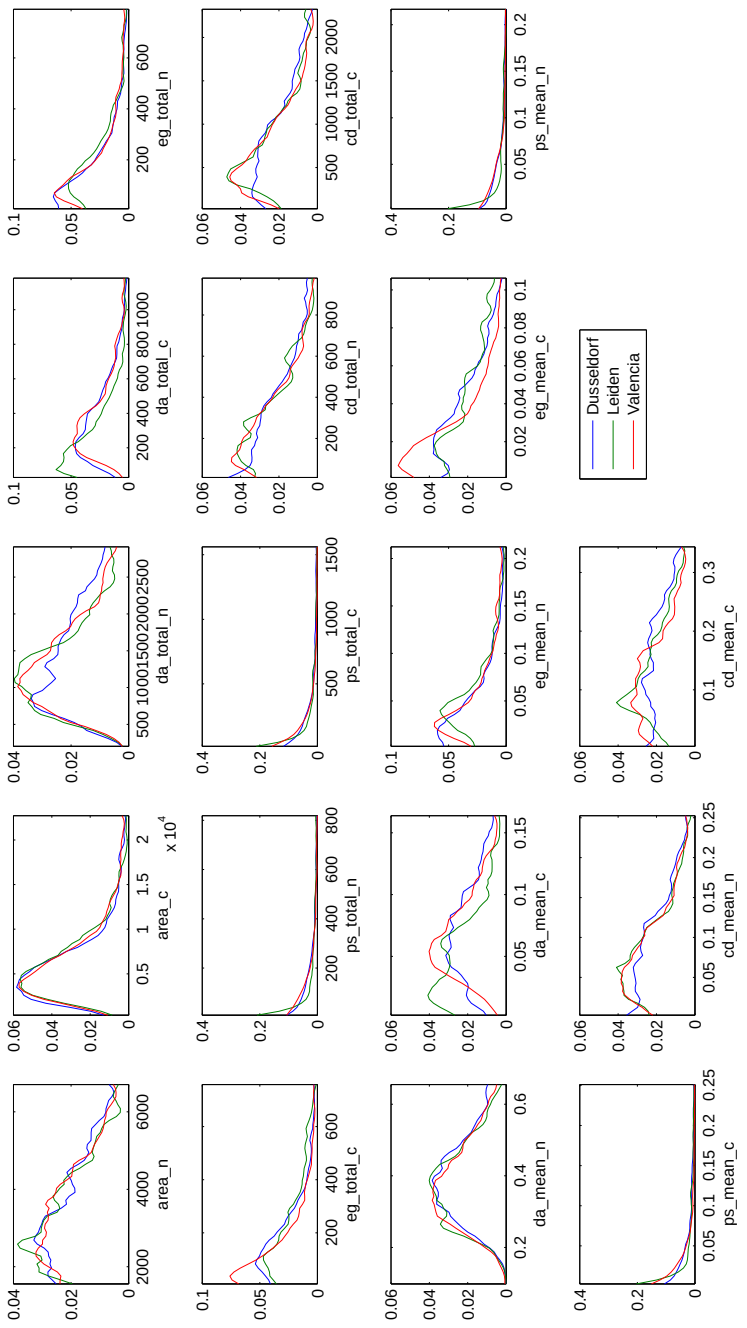
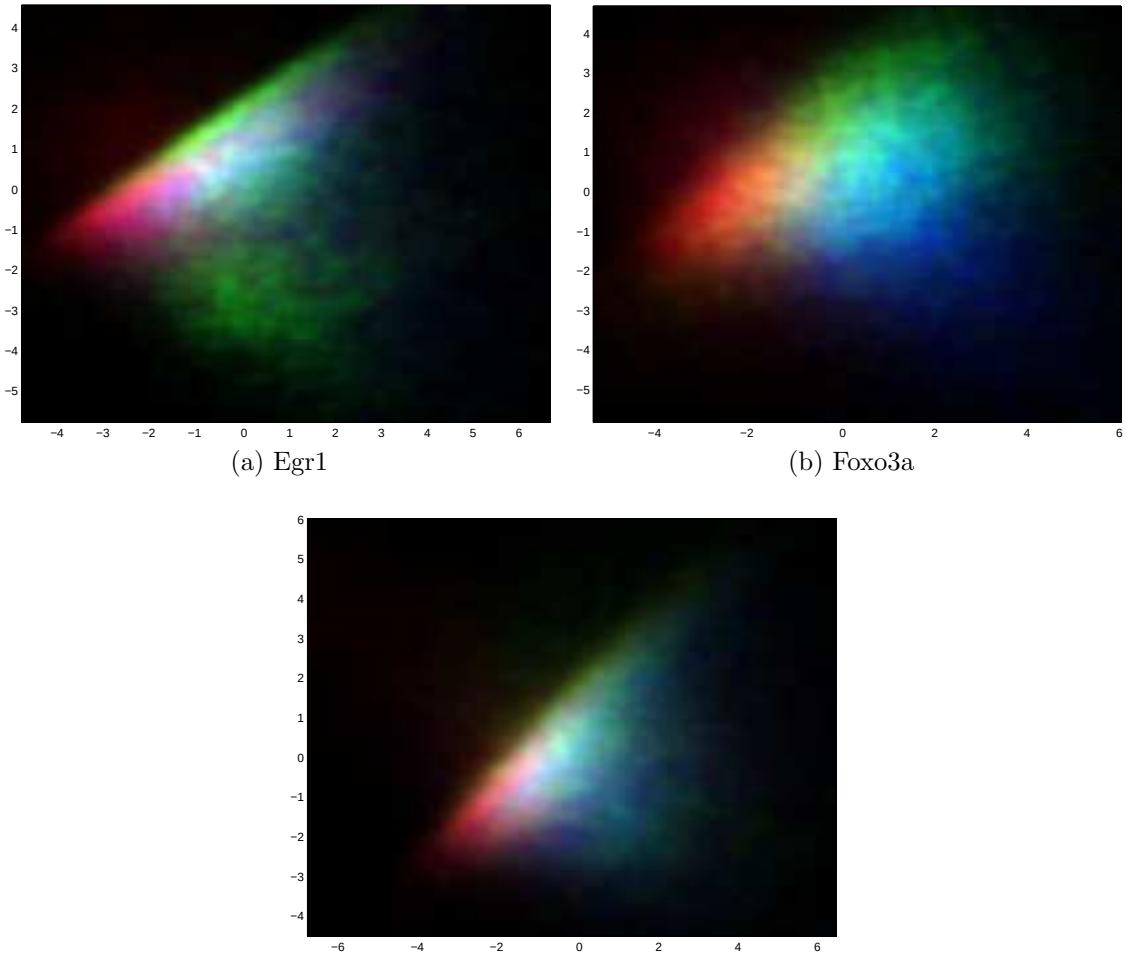


Figure 6.2: Histograms of all area, mean and total intensity cell features in the Egr1 labelled images, grouped by dataset.

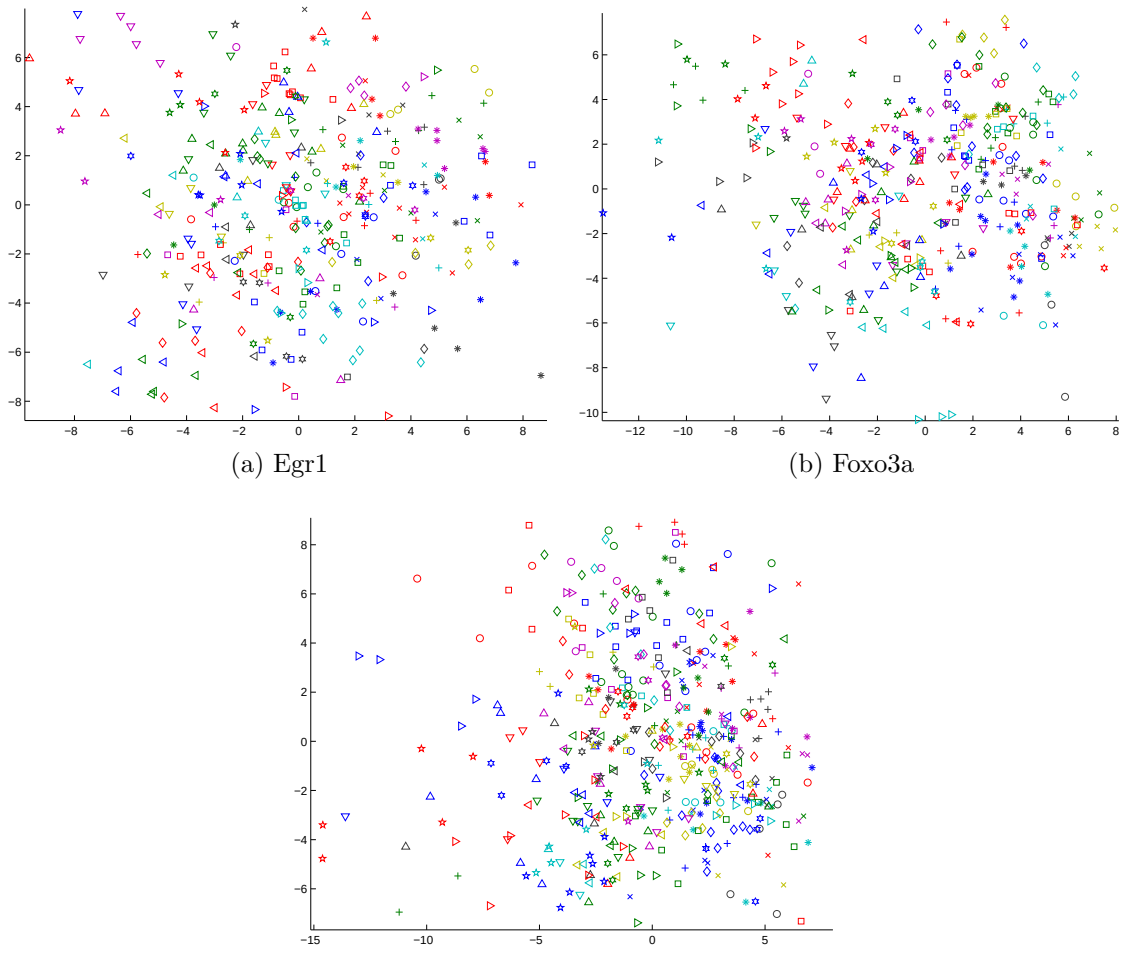


(a) Egr1

(b) Foxo3a

(c) Ki67

Figure 6.3: Density plots (2D histograms coded using intensity) of the first two principal components for all cell features for each of the three readout markers. Cells are colour coded depending on which dataset they originate from: Blue: Düsseldorf, Green: Leiden, Red: Valencia. Areas of overlap between the datasets are indicated by combining the appropriate colours, so for example perfect overlap between the three datasets is indicated by white.



(a) Egr1 (b) Foxo3a

(c) Ki67

Figure 6.4: Scatter plots of the first two principal components for all images from each patient. Each patient is shown using a unique colour-symbol combination to help visualise the variation amongst images from a single patient.

is unknown— for now we will assume that there is no hidden external factor responsible for this difference and that we can assume all patients in all datasets are comparable.

6.9 Ki67 biomarker

Published studies suggest that Ki67, which is a proliferation marker, is strongly correlated with survival in several tumour types, including some evidence that this is the case in

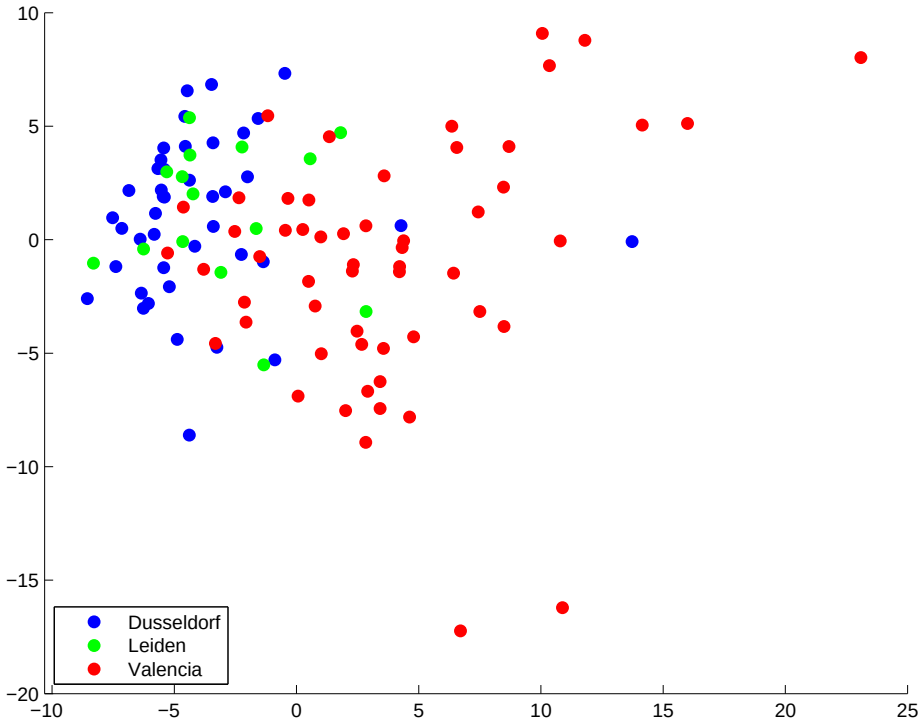
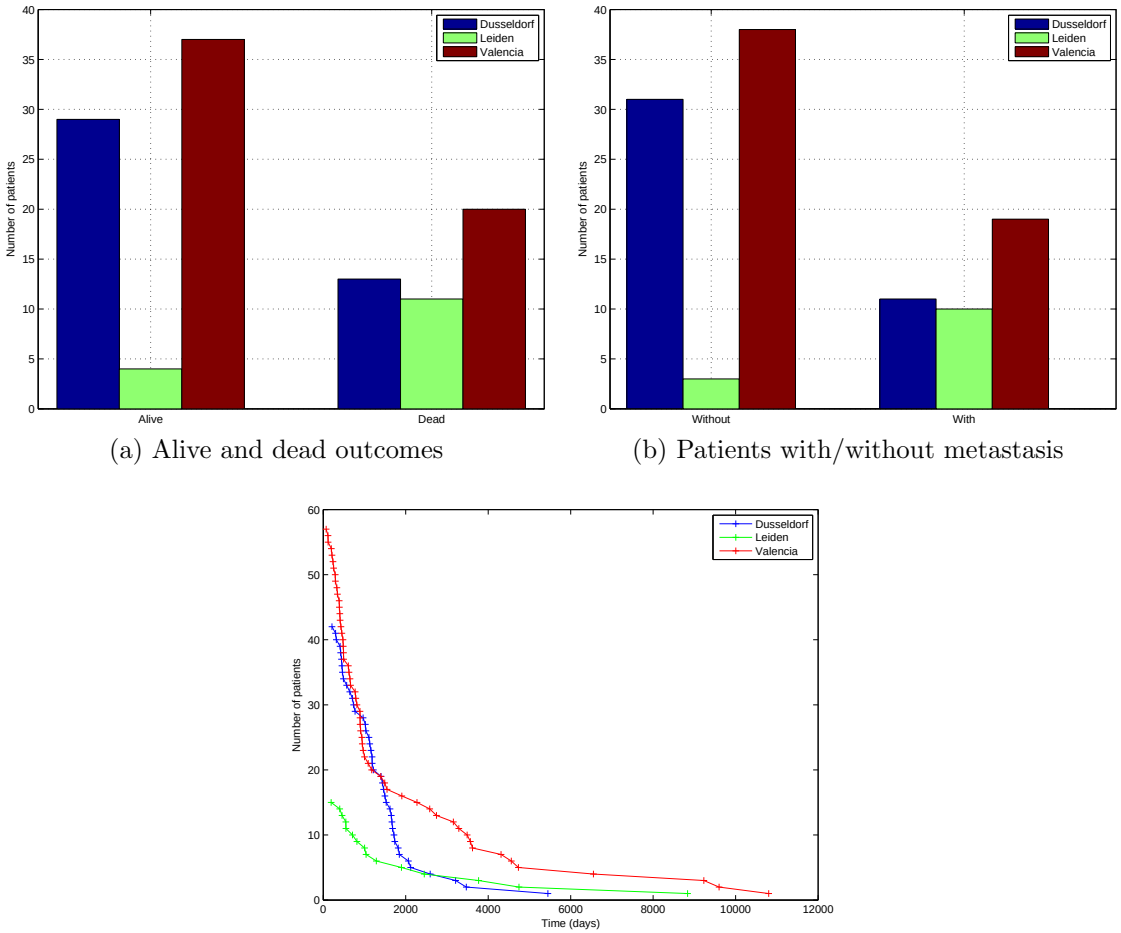


Figure 6.5: Scatter plot of the first two principal components of the patient-level features.

	Overall survival		Event free survival		
	Alive	Dead	Alive (no event)	Alive (event)	Dead (no event)
Düsseldorf	29	13	27	13	2
Leiden	4	11	2	12	0
Valencia	37	20	29	24	4
Totals	70	44	58	49	6

Table 6.4: Number of patients alive/dead at overall and event free survival. (No event) means the EFS and overall survival were the same i.e. no event occurred in the intervening period. Note that EFS was unavailable for one Leiden patient.



(a) Alive and dead outcomes

(b) Patients with/without metastasis

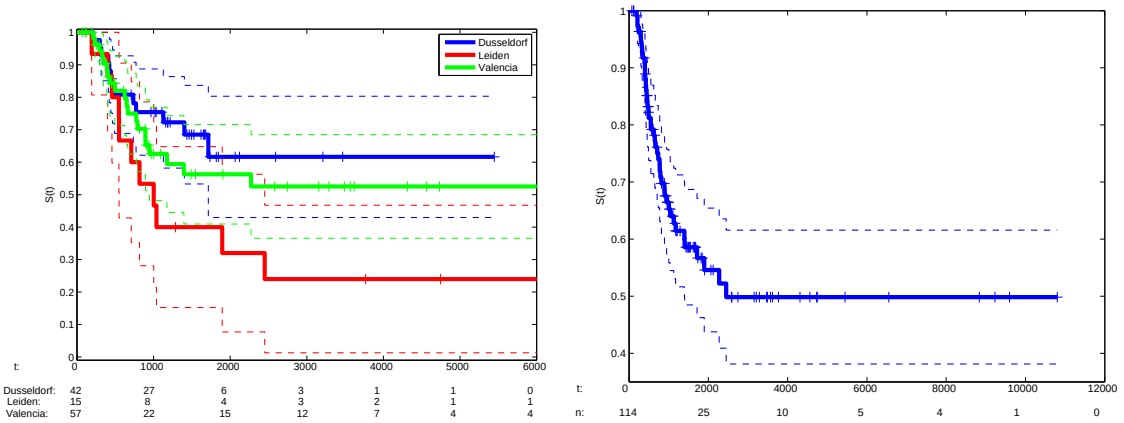
(c) Overall survival time

Figure 6.6: Distributions of the main patient survival outcomes in the three datasets.

Ewing's sarcoma [Amir, 2002].

In this section we examine this hypothesis using conventional approaches by defining the Ki67 index as the proportion of cells in a patient expressing Ki67, and investigating whether this is correlated with patient prognosis.

In doing so we highlight some of the problems that occur with the application of conventional approaches, and this provides the motivation for our later work in Section 6.10.



(a) Survival function for each of the three datasets.

(b) Survival function for the combined datasets, truncated at 3000 days (note the y-axis has been truncated).

Figure 6.7: Kaplan-Meier survival functions for the three datasets, and for the combined datasets. Tick marks indicate where patients were censored. The x-axis labels show the survival time in days, followed by the number at risk for each of the datasets. Dashed lines indicate the 95% confidence interval.

6.9.1 Identification of Ki67 positive cells

There are several factors to be considered before cells can be labelled as expressing (+) or not expressing (-) a biomarker. As previously mentioned in Section 6.7 there are several normalisation issues to consider. An attempt at partially correcting these was made at the feature extraction stage, but the combination of the wide range in biological heterogeneity, a relatively small number of subjects, potentially different sample preparation protocols, and microscope setup adjustments, means there may still be experimental errors present in the images.

A further complication arises from the presence of multiple cell types in the tissue samples, including lymphocytes and non-tumourous fibroblasts. The former are notable for their high (almost saturating) non-specific nuclear fluorescence in all antibody labelled channels and low cytoplasmic fluorescence in some of the acquired images. The latter are more of a challenge. Around 95% of cancer cells in Ewing's sarcomas express CD99 in

their cytoplasm [Bernstein et al., 2006], so we may use the presence of CD99 as a way of identifying tumour cells in this initial analysis.

Ideally the pre-processing steps should ensure the levels of the CD99 and Ki67 markers are comparable between patients, but in practice the background and non-specific localisations of the markers varies greatly. At this stage it is preferable to avoid any overly complex threshold definitions, but we can make use of our prior expectations that CD99 should be predominantly cytoplasmic and Ki67 predominantly nuclear to define global thresholds q_{cd}, q_{ki} based on the differences between the mean nuclear $\mathbf{x}_{cd,n}, \mathbf{x}_{ki,n}$ and cytoplasm $\mathbf{x}_{cd,c}, \mathbf{x}_{ki,c}$ levels, and so create binary indicator variables for whether cells are CD99+ ($\mathbf{y}_{cd}$) and Ki67+ ($\mathbf{y}_{ki}$) as follows.

$$\mathbf{y}_{cd}(i) = \begin{cases} 1 & \text{if } \mathbf{x}_{cd,c}(i) - \mathbf{x}_{cd,n}(i) \geq q_{cd} \\ 0 & \text{otherwise} \end{cases} \tag{6.1}$$

$$\mathbf{y}_{ki}(i) = \begin{cases} 1 & \text{if } \mathbf{x}_{ki,n}(i) - \mathbf{x}_{ki,c}(i) \geq q_{ki} \\ 0 & \text{otherwise} \end{cases} \tag{6.2}$$

The Ki67 index is then defined as the proportion of CD99+ cells that are Ki67+:

$$K_{q_{cd},q_{ki}} = \frac{\sum_i \left(\mathbf{y}_{cd}(i) \cap \mathbf{y}_{ki}(i)\right)}{\sum_i \mathbf{y}_{cd}(i)} \tag{6.3}$$

where the sums are taken over all cells in a subject.

Examination of the joint distributions of CD99 and Ki67 in the nucleus and cytoplasm (Figure 6.8) in each cell does not suggest an obvious method for defining a threshold. This issue is not unique to our analysis— the problem of how to define a threshold is an unresolved question in the field of biomarker research.

In order to illustrate the difficulties with defining a suitable threshold we tested mul-

tiple combinations of CD99 and Ki67 thresholds by checking for the significance of a Cox proportional hazards model fitted to the Ki67 index, resulting in multiple definitions of the Ki67 index for each pair of thresholds as shown in Figure 6.9a.

In studies aimed at defining a measurement for clinical use it is common to go one step further and categorise patients into a finite number of groups, such as high and low Ki67 indices. The rationale behind this is that in a clinical setting it is easier to assign patients into one of a small number of discrete categories and define a recommended treatment regime for each group than it is to deal with a continuous spectrum. We therefore partition the patients into two groups around all possible cutpoints c in $K_{q_{cd},q_{ki}}$:

$$\text{Group 1: } K_{q_{cd},q_{ki}} < c \tag{6.4}$$

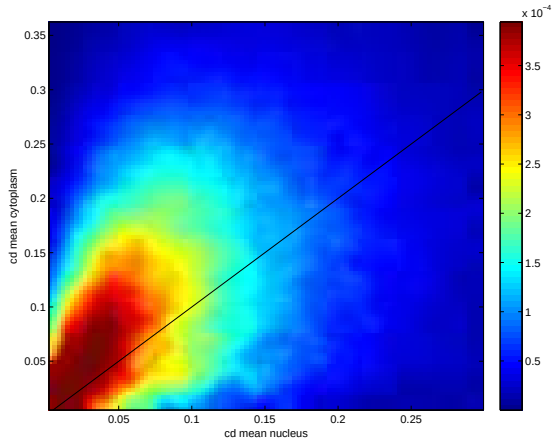
$$\text{Group 2: } K_{q_{cd},q_{ki}} >= c \tag{6.5}$$

and show a plot of the log-rank test values as a function of c in Figure 6.9b, along with a Kaplan-Meier plot for the most significant split in Figure 6.9c. The final significance value is $p = 8.6 \times 10^{-5}$. It is also possible to divide the patients into three groups, for instance using an X-tile type approach [Camp et al., 2004] (not shown).

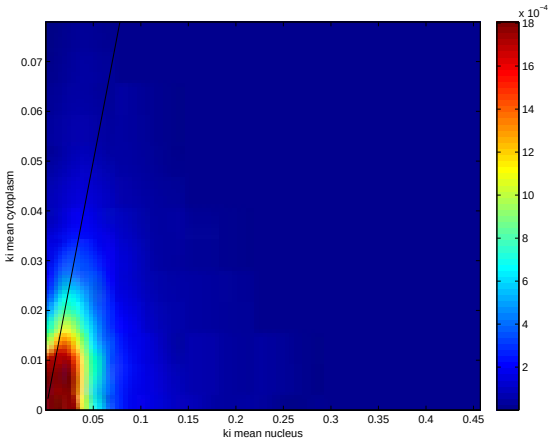
It is clear that this "optimal p-value" method is unreliable, as vehemently argued by Altman et al. [1994]. This is only partly due to the low number of patients. The log-rank test is not at all robust to small changes in the definition of the cutpoint, and it is not uncommon for the threshold with the highest log-rank test value to result in a very unbalanced partitioning of patients.

6.9.2 Manual grouping of images

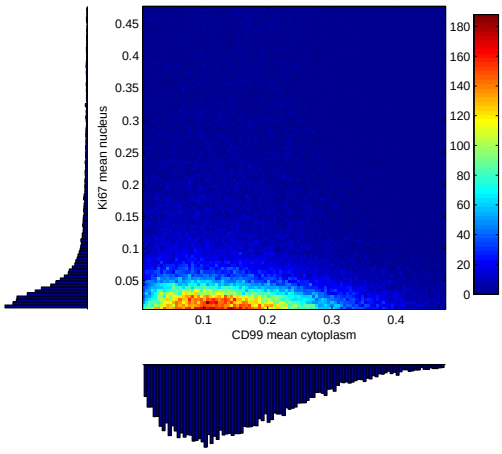
We can compare our stratification results with those obtained by a manual grading of the images by an expert. The biopsy images from a subset of 53 patients were manually graded



(a) Mean CD99 nucleus and cytoplasm

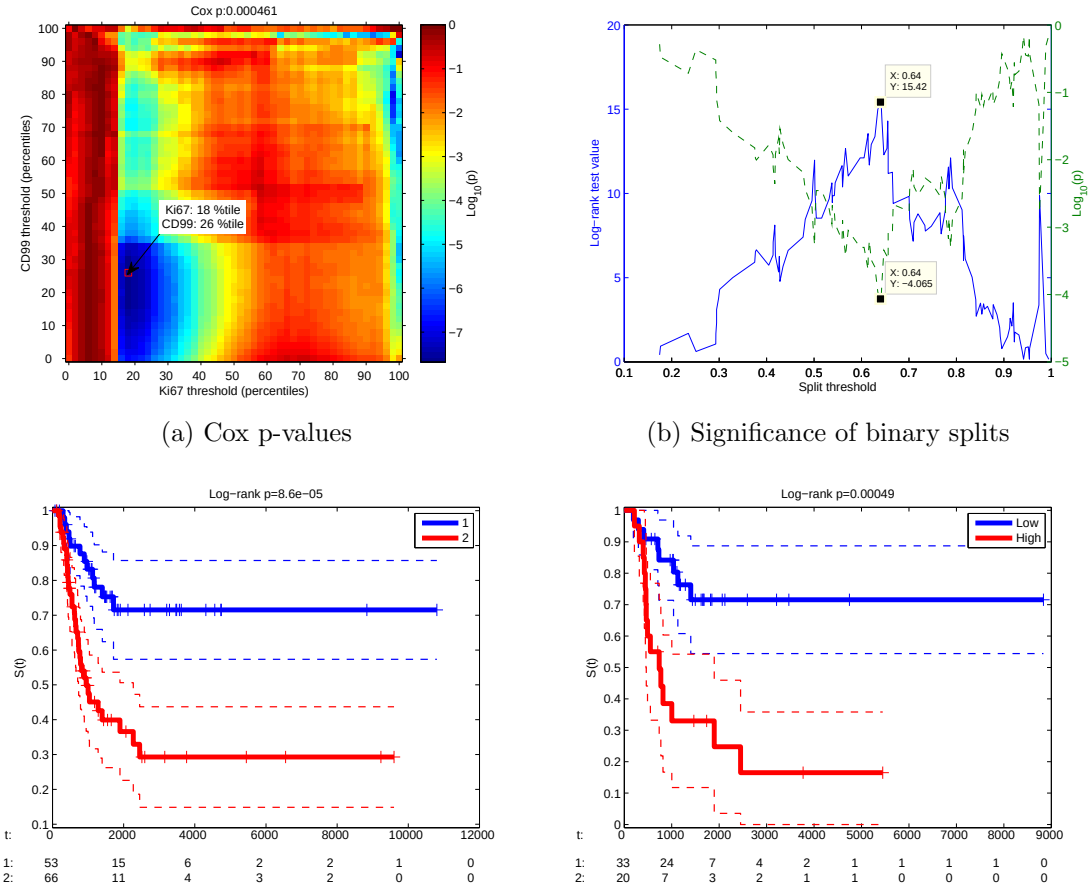


(b) Mean Ki67 nucleus and cytoplasm



(c) Mean CD99 cytoplasm and mean Ki67 nucleus

Figure 6.8: Density plots of the mean CD99 and Ki67 nucleus and cytoplasm levels in all cells. (a) and (b) show the joint mean nucleus and mean cytoplasm levels of each marker, the black line is the line of equality. The axes are set to show values lying between the 1st and 95th percentiles— the distribution of Ki67 has a much longer tail, resulting in the different scales to CD99. (c) shows the joint distribution of the mean CD99 cytoplasm level and the mean Ki67 nucleus level.



(a) Cox p-values

(b) Significance of binary splits

(c) Most significant split

(d) Manual grading

Figure 6.9: Determination of the "optimal" nucleus-cytoplasm-difference threshold for defining CD99+ and Ki67+ cells, compared with manual grading. (a): The $\log_{10}(p)$ significance values of a Cox proportional hazards model fitted to the Ki67 index for each pair of CD99 and Ki67 thresholds. 50 thresholds for each of the two markers were tested, giving a total of 2500 pairs. The "most significant" pair of thresholds ($p = 4.6 \times 10^{-4}$) is highlighted. (b): The log-rank test values for all possible binary partitions of patients based on the Ki67 index define by the pair of thresholds highlighted in (a). The split threshold which gives the most significant log-rank score is highlighted ($p = 8.6 \times 10^{-5}$). (c): The Kaplan-Meier plot for the split highlighted in (b). (d) Kaplan-Meier plot showing patients manually graded into low or high Ki67 levels.

as showing low or high Ki67 expression.

Based on this we can form the Kaplan-Meier plot shown in Figure 6.9d, which suggests a highly significant difference (log-rank $p = 5 \times 10^{-4}$), comparable with that obtained in Figure 6.9c using the exhaustive threshold search.

Note that only good quality images were graded so the difference between the two classes may be exaggerated, but it does show that there is useful information present in the images.

6.9.3 Conclusions

This section has illustrated a conventional way of analysing a putative biomarker, involving the setting of multiple thresholds to split patients into a small number of categories with different survival characteristics. Choosing thresholds based on the optimisation of statistical significance criteria is not robust, as shown here and by others, and suffers from problems of multiple hypothesis testing.

This motivates us to consider alternative methods of determining whether these biomarkers are informative or not in the following sections.

6.10 Random forest classification

The random forest classifiers described in this chapter are based on the ones developed in Section 5.9.3. However, we begin by using unsupervised clustering [Shi and Horvath, 2006] to see if there is any inherent structure in the data, before using supervised random forests to predict the occurrance of metastasis, and patient survival.

6.10.1 Unsupervised clustering

Our motivation for first trying unsupervised random forests is the work of Seligson et al. [2005], where an unsupervised random forest was used to find subgroups of patients based on the levels of multiple expression markers in a tissue microarray, and these subgroups were shown to have different recurrence rates.

Unsupervised clustering uses the proximity matrix P of the random forest classifier, created from the N samples used to train the forest. Matrix P is created by passing all data samples (in our case the feature vector for each patient) down the trees in the forest and recording in element $P_{i,j}$ the number of times sample i appears in the same terminal tree node as sample j. Therefore the greater the value of $P_{i,j}$ the greater the similarity between samples i and j.

In order to use this on unlabelled data a synthetic data set is created by randomly sampling from the sample values of each feature, but independently of any other feature. The synthesised data set therefore contains features which individually have the same distribution as the feature in the original set, but in which all correlations between features have been broken. The original and synthesised data sets are treated as two classes and used to "train" the classifier to distinguish between the original and random data.

We are interested in the part of the proximity matrix which shows similarities between each of the original samples. This similarity can be used with a visualisation approach

such as multidimensional scaling (MDS) to produce low dimensional plots in which similar points appear closer together [Borg and Groenen, 1997].

In order to obtain a cluster we can use an implementation of the k-medoids algorithm such as the partitioning around medoids algorithm (PAM) [Kaufman and Rousseeuw, 2005, chapter 2]. Medoids can be thought of as a multi-dimensional equivalent to the median, and the k-medoids algorithm works along principles similar to that of the k-means algorithm. When this is tried on the patient features the samples again appear to group themselves by dataset (Figure 6.10).

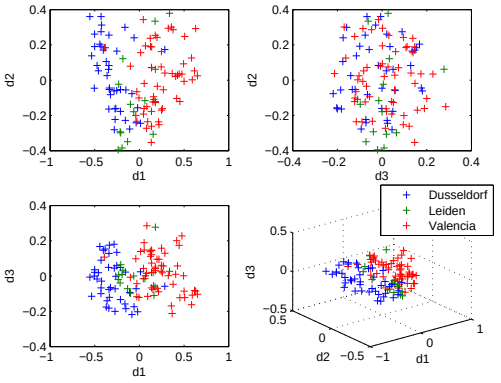
6.10.2 Supervised classification

Since the overall survival time is a continuous variable an obvious approach to analysing the data would be to fit a regression model, and given the potentially large number of features including many uninformative ones random forests regression would be a leading choice. However, care must be taken when formulating the problem of predicting survival times as a regression or classification problem due to the presence of censored data.

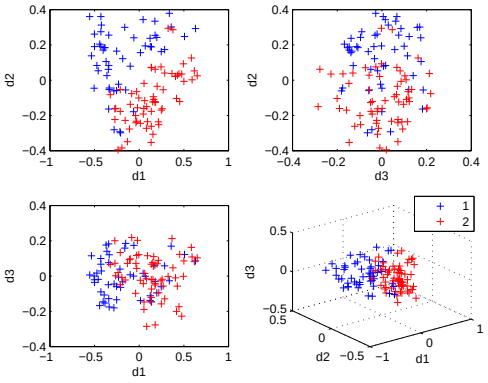
Discarding all censored patients is one possibility, and since the censoring process should be random this will not bias the results. The problem is that this will lead to a significant reduction (up to 65%) in the number of samples since we only know the lifetimes of patients who have died.

This leaves us with two main options for analysing survival times using a random forests approach:

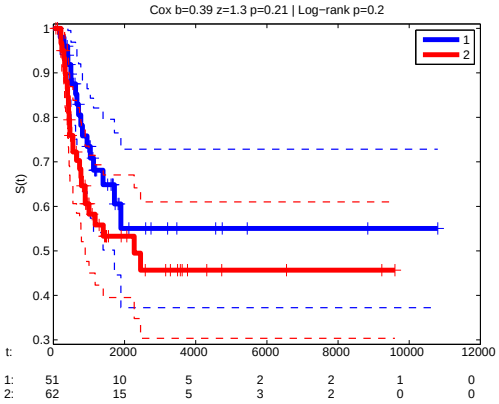
1. Classify patients against another outcome measure that is clinically useful, such as whether metastasis has occurred.
2. Use a variant of random forests, random survival forests [Ishwaran et al., 2008], to analyse the survival data directly.



(a) MDS coloured by dataset



(b) Two unsupervised clusters



(c) Kaplan-Meier plot of the two clusters

Figure 6.10: Unsupervised clustering of all patients. (a). Multidimensional scaling plot of patients created from the proximity matrix obtained using the unsupervised random forest procedure. The first three dimensions (d1, d2, d3) of the MDS are plotted against each other. Patients have been coloured by dataset. (b). Unsupervised clustering of all patients into two clusters determined using the PAM algorithm. (c). Kaplan-Meier plots of the patients in the two clusters.

Another potential option would be to classify patients based on whether they survive beyond a time threshold or not. Although conceptually simple any subjects that were censored before this threshold would be discarded, and furthermore random survival forests effectively uses this idea but considers all time thresholds at once.

6.10.3 Metastasis classification

Metastasis is a clinical indicator which means a tumour has spread from its original location to other parts of the body, for instance via blood or lymphatic vessels. We will treat its presence or absence as a binary indicator:

0: No metastasis

1: Metastasis present

In addition to classifying on the patient features obtained from all cells we will also look at whether considering only CD99+ cells is beneficial (based on the analysis in Section 6.9), and for comparison only CD99- cells.

6.10.3.1 Training and testing

Training and testing was performed using leave-one-out cross-validation, using the Matlab port version 0.02 of the R Random Forest package [Jaiantilal, 2009, Liaw and Wiener, 2002]. We are using a large number of patient features which can lead to poor performance of the random forest, so we need to consider some form of feature reduction. Several selection algorithms of varying complexity are available, but since we have limited amounts of data there is no great benefit in trying to find an optimal subset of features, so instead we use the following approach:

1. $F \leftarrow \{\text{all features}\}$

2. Train a random forest using F

3. Calculate feature importances using the permutation measure

4. $F' \leftarrow F$ excluding features with importance < 0

5. Train a random forest using the reduced feature set F'

6. Repeat steps 3–5 a further 6 times (see Section 6.10.5)

The random forest parameters were set as the following:

Number of trees: 200

Number of candidate features: $\sqrt{\text{number of features}}$

The classification results are summarised in Figure 6.11 and Table 6.5 using the sensitivity (fraction of metastases correctly identified) and specificity. The balanced accuracy was also calculated which ranges from 0 to 1 with 0.5 the value to be expected by chance, along with the overall error rate. See Appendix B.1 for definitions of these terms.

These figures should be interpreted by examining the balanced accuracy in conjunction with the difference between the sensitivity and specificity. A large difference indicates poor performance which could occur if the classifier defaults to assigning everything to the same class. To take an extreme example, if the classifier were to assign everything as positive this would result in a sensitivity of 1 but a specificity of 0.

We should therefore focus on the classifiers which have a sensitivity and specificity greater than 0.5, such as Düsseldorf and Valencia using all cells. Since these are individual datasets we are dealing with relatively few samples, so it is useful to look at the actual confusion matrices, which are shown in Table 6.6.

A test of significance using Pearsons's χ^2 test gives the following p-values:

Düsseldorf	0.069
Valencia	0.072
Leiden	0.400
Combined	0.062

The Düsseldorf and Valencia values are not too far from the conventional significance threshold of 0.05, and given the small numbers of patients we can conclude that it would be worth investigating this in a larger study.

It is notable from Table 6.5 that the performance of the classifier on the combined datasets was low, suggesting that the data normalisation procedure did not appear to

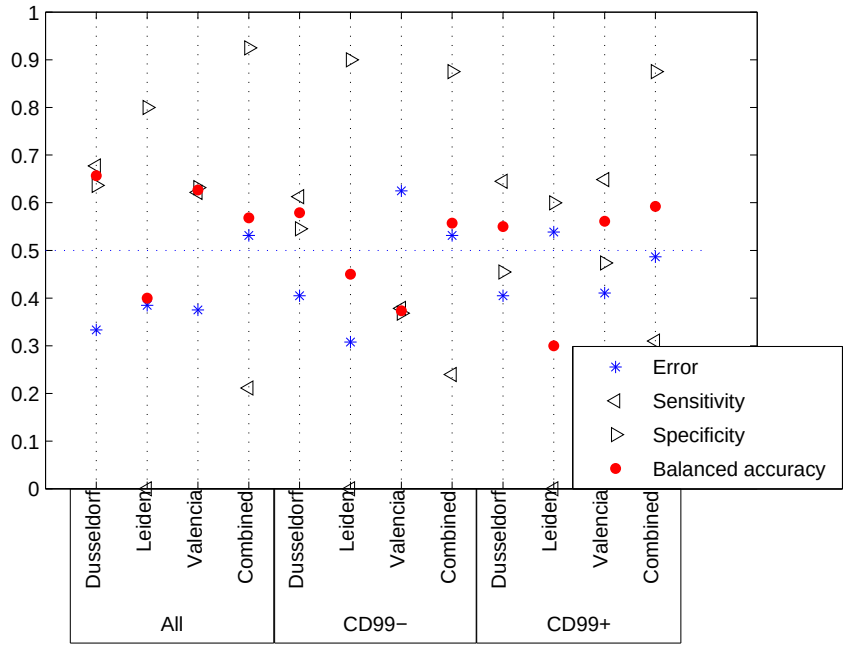


Figure 6.11: Results of the random forest metastasis classification. The overall error rate, sensitivity, specificity and balanced accuracy are shown for the three separate datasets, and the combined dataset. The results for all cells, CD99+ cells and CD99- cells are also shown. See Table 6.5 for the exact values.

correct for the variations in data collection and the experimental protocol. It also shows the importance of the initial exploration of the data using PCA to ascertain whether the datasets were directly comparable.

For comparison the classification rates calculated from the out-of-bag (OOB) samples during the training of a random forest without cross-validation are shown in Figure 6.12. The OOB estimates are often regarded to be as good as an independent test set for predicting the performance of a random forest [Breiman, 2001], but here the OOB estimates appear to underestimate the error rates.

We should therefore verify that the random forest OOB error estimates are comparable with the more computationally intensive cross validation measure before relying on them in any future studies.

		Overall error	Sensitivity	Specificity	Balanced accuracy
All	Düsseldorf	0.33	0.68	0.64	0.66
	Leiden	0.38	0.00	0.80	0.40
	Valencia	0.38	0.62	0.63	0.63
	Combined	0.53	0.21	0.93	0.57
CD99-	Düsseldorf	0.40	0.61	0.55	0.58
	Leiden	0.31	0.00	0.90	0.45
	Valencia	0.63	0.38	0.37	0.37
	Combined	0.53	0.24	0.88	0.56
CD99+	Düsseldorf	0.40	0.65	0.45	0.55
	Leiden	0.54	0.00	0.60	0.30
	Valencia	0.41	0.65	0.47	0.56
	Combined	0.49	0.31	0.88	0.59

Table 6.5: Results of the random forest metastasis classification. All rates were calculated using leave-one-out cross-validation for the three separate datasets, and for the combined dataset. Patient features calculated using all cells, and only either CD99+ or CD99- cells, were tested. The highlighted rows show classifiers which performed best (highest specificity and sensitivity).

		True	
		0	1
Predicted	0	21	4
	1	10	7

(a) Düsseldorf

		True	
		0	1
Predicted	0	23	7
	1	14	12

(b) Valencia

		True	
		0	1
Predicted	0	0	2
	1	3	8

(c) Leiden

		True	
		0	1
Predicted	0	15	3
	1	56	37

(d) Combined

Table 6.6: Confusion matrices showing the performance of the metastasis classifiers trained on all cells in each dataset.

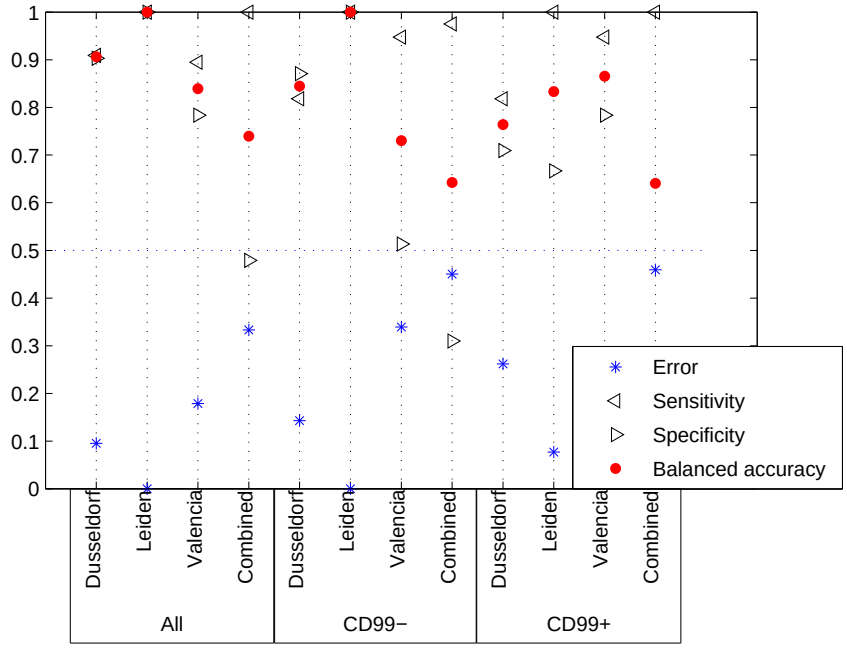
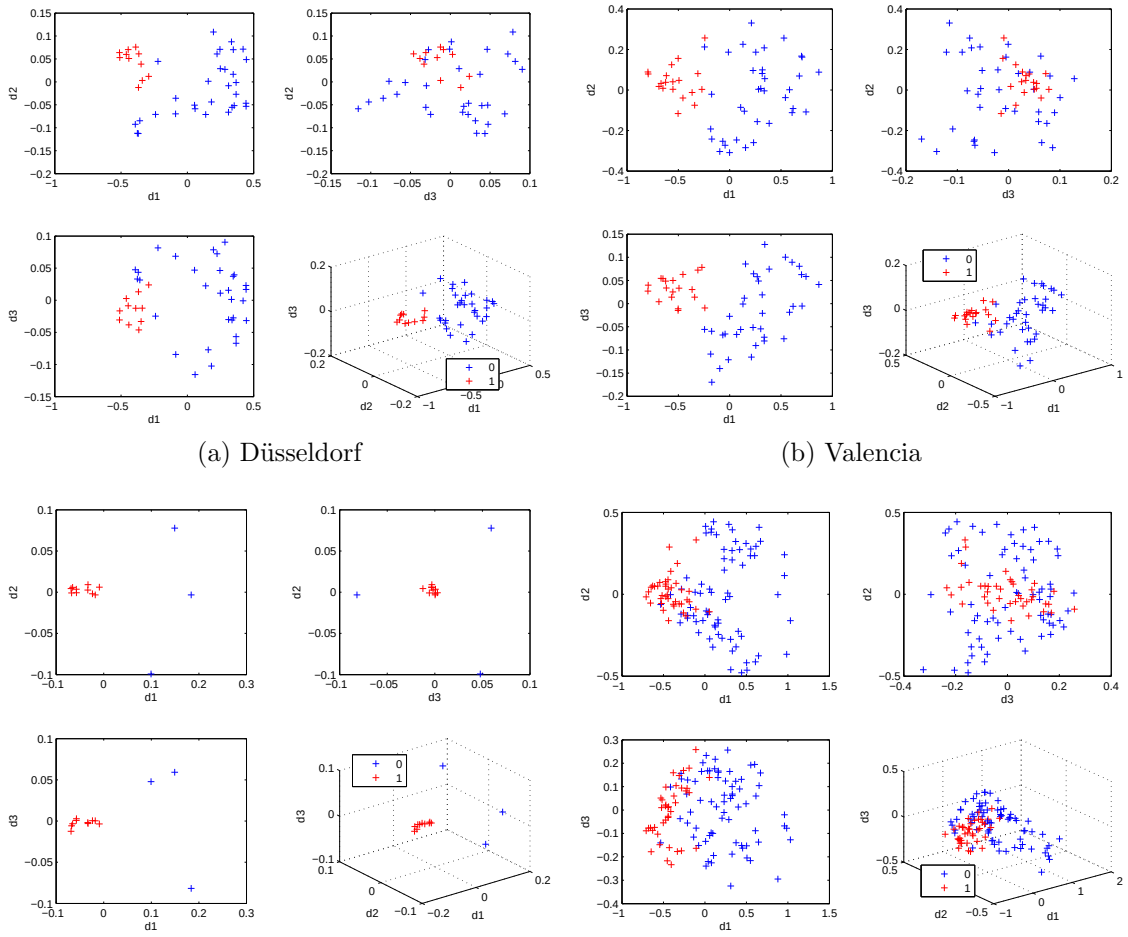


Figure 6.12: The random forest out-of-bag error estimates. The overall error rate, sensitivity, specificity and balanced accuracy were calculated using the OOB samples during the training of a random forest without cross-validation.

6.10.3.2 Similarity within the data

The random forest proximity matrix provides a measure of similarity between all pairs of samples, and can be visualised using a multi-dimensional scaling (MDS) plot. It can roughly be thought of as projecting the random forest classifier's internal view of the data into a lower dimension, in a manner analogous to PCA.

MDS plots for the classifiers using all cells, in separate and combined datasets, are shown in Figure 6.13. The metastasis and non-metastasis patients appear to form close groups within each of the separate datasets, whereas they overlap significantly in the combined dataset.



(a) Düsseldorf (b) Valencia

(c) Leiden (d) Combined

Figure 6.13: MDS plots of the random forest proximity matrix for the metastasis classifiers using all cells. The first three axes of the MDS are shown, labelled d1, d2 and d3. 0: No metastasis, 1: Metastasis.

6.10.4 Feature importance

The classification results indicate some of the images contain something informative with regards to metastasis, so we can examine the feature importances to see what measurements may be biologically relevant. Figure 6.14 shows an example of the top 25 features for the combined all cells classifier.

In order to reduce the number of displayed features across all classifiers to a manageable

	Düsseldorf	Leiden	Valencia	Combined
Kurtosis (mean nuclear DAPI:mean cytoplasmic Ki67)	0.001	0.113	0.504	0.004
Standard deviation (skew cytoplasmic CD99)	0.731	0.017	0.011	0.008
Skew (mean nuclear:mean cytoplasmic Ki67)	0.008	0.242	0.366	0.014

Table 6.7: Significance values from a two-sided Kolmogorov-Smirnov test between the distributions of each feature in patients with and without metastasis.

level the top 10 features from each of the 12 classifiers were selected, leading to a total of 79 features (note many of these will be unimportant, but they help to give context to the importance of other features). The importance of each of these 79 features in each classifier was used to form Figure 6.15.

Ideally we want features which are informative in more than one dataset, though the limited number of samples means we should not immediately discount those which are primarily important in only one.

Figure 6.16 shows the distributions (amongst all patients) of three selected features which may be important as judged from Figure 6.15. We can perform a Kolmogorov-Smirnov goodness-of-fit test between the metastasis and non-metastasis distributions; the p-values for a two-sided test are shown in Table 6.7.

Careful examination of the plots (Figure 6.16) for each feature in metastasis versus non-metastasis patients in each dataset suggests the shift in distribution (left or right) is consistent amongst all datasets. This implies that although the differences for individual datasets might not be statistically significant in aggregate we have evidence of an important feature.

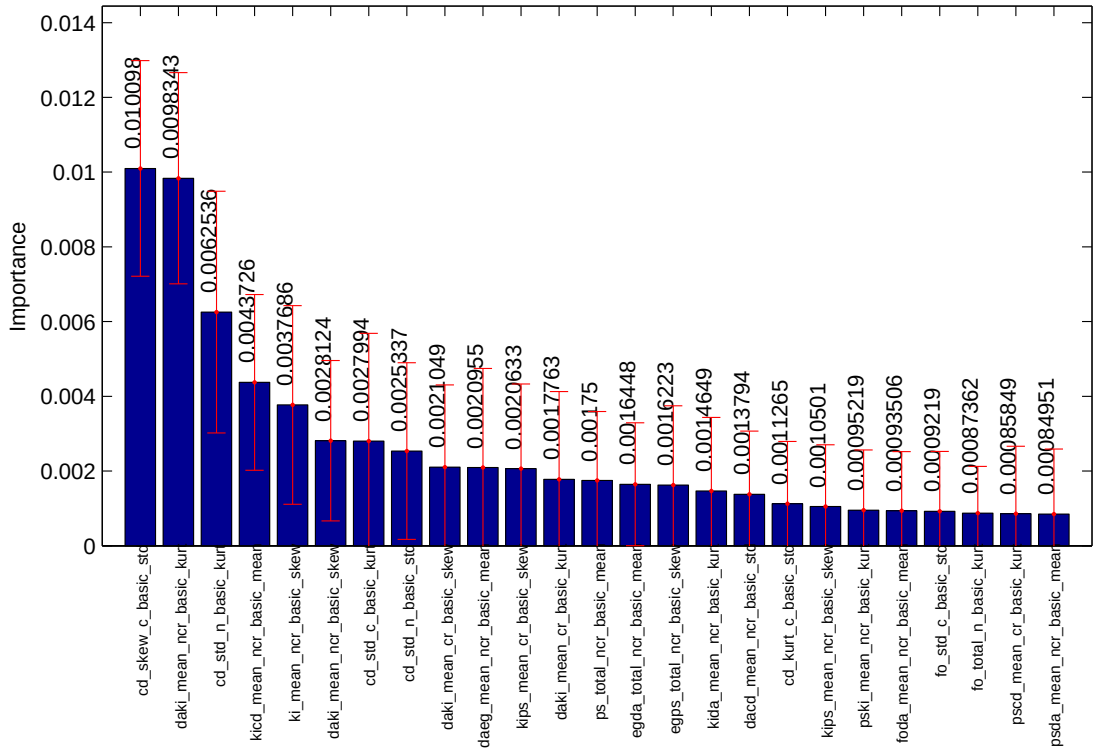


Figure 6.14: Feature importances for the combined dataset metastasis classifier.

6.10.4.1 Testing the feature importance

In order to support use of the random forest to suggest informative features we can deliberately augment a dataset with a feature that is known to be highly correlated with our target. For instance, we can augment the metastasis classifier with event free survival (EFS)— since metastasis is associated with a severe increase in severity of the disease it is likely to either be counted as an event, or be associated with other events. We deliberately chose to augment the all cell combined dataset classifier which had low classification accuracy to illustrate the fact that we can still make use of the properties of the random forest irrespective of its performance, which is shown below:

Overall error	Sensitivity	Specificity	Balanced accuracy
0.51	0.30	0.83	0.56

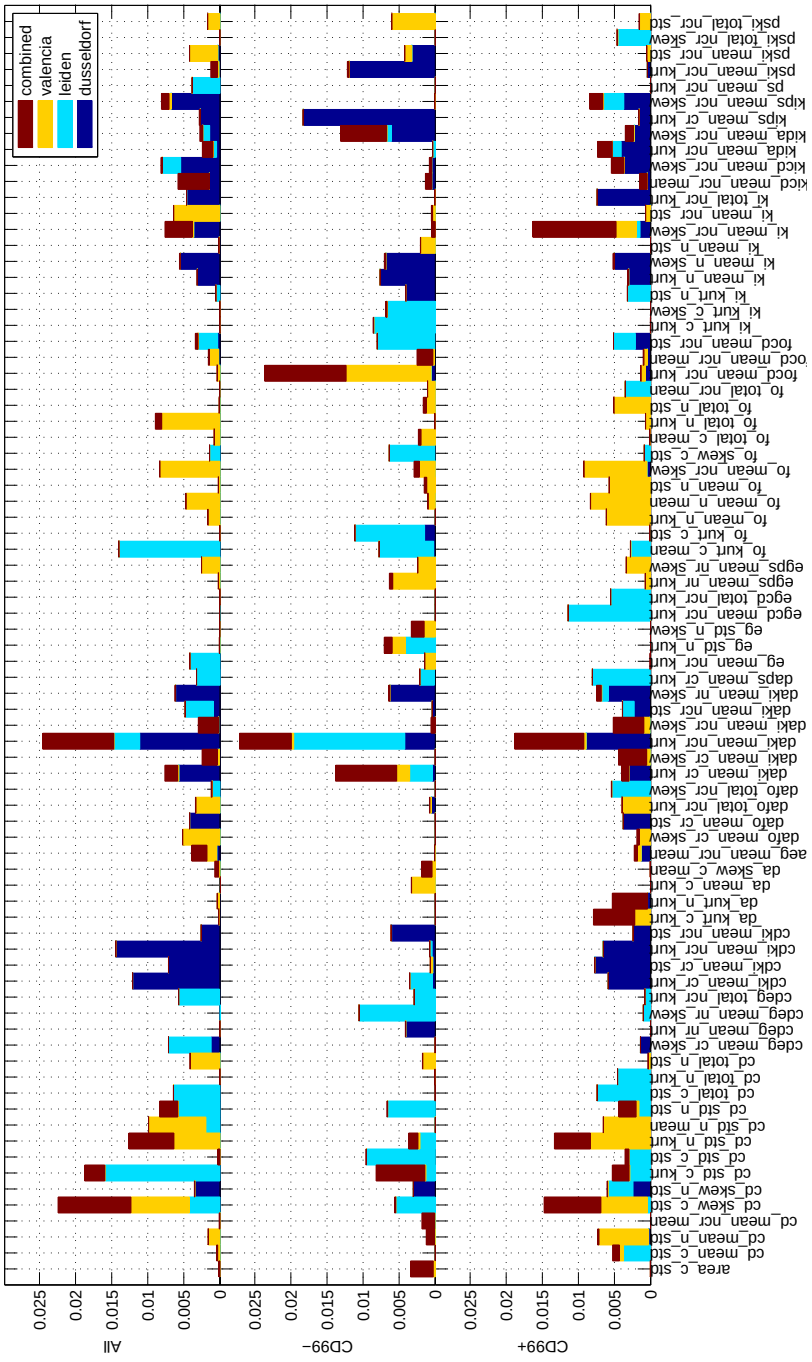
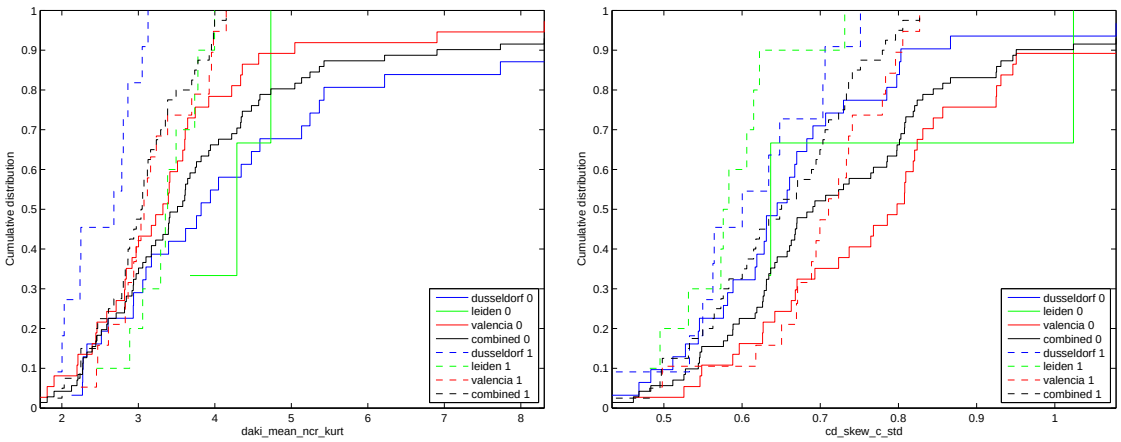
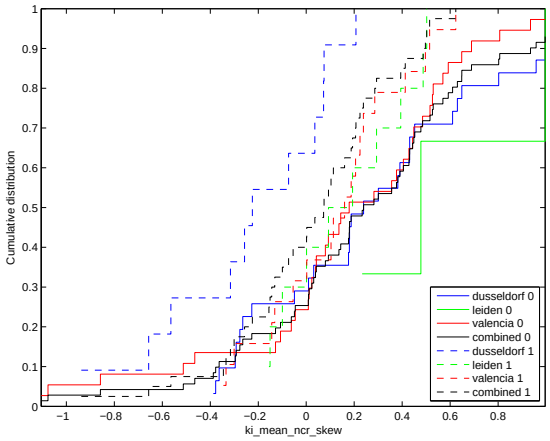


Figure 6.15: Feature importances for the metastasis classifiers. The three plots show classifiers built using all cell, CD99+ cells, and CD99- cells. For each of these plots a stacked bar chart shows the cumulative feature importance from the separate datasets, and the combined dataset. See text for details of the choice of subset of features.



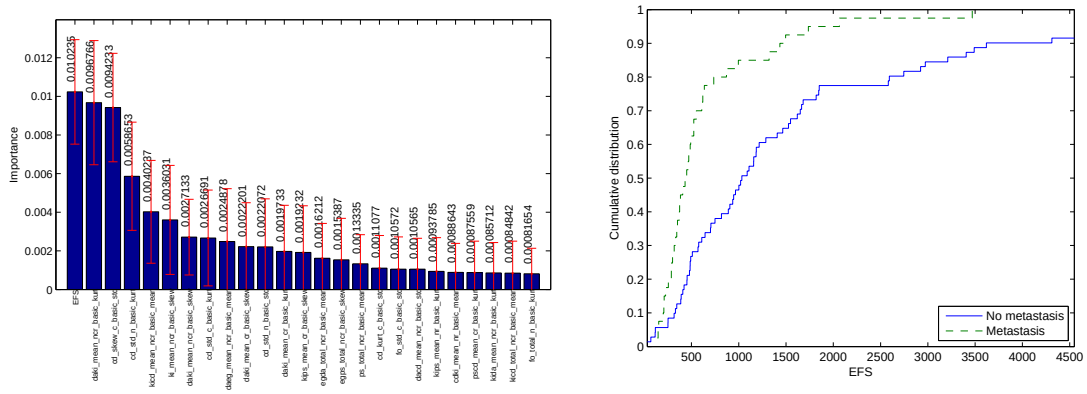
(a) Kurtosis of (mean nuclear DAPI):(mean cytoplasmic Ki67) ratio

(b) Standard deviation of the skew of the cytoplasmic CD99



(c) Skew of the mean Ki67 nucleus:cytoplasm ratio

Figure 6.16: Distributions of three potentially informative features split by metastasis and dataset using all cells. Plots are colour coded by dataset, solid lines show the distribution in patients without metastasis, and dashed lines show the distribution when metastasis is present.



(a) Feature importances.

(b) Cumulative plot of EFS.

Figure 6.17: EFS feature importance in the augmented metastasis classifier. (a): EFS feature importance. (b): Cumulative plot of EFS in patients with and without metastasis.

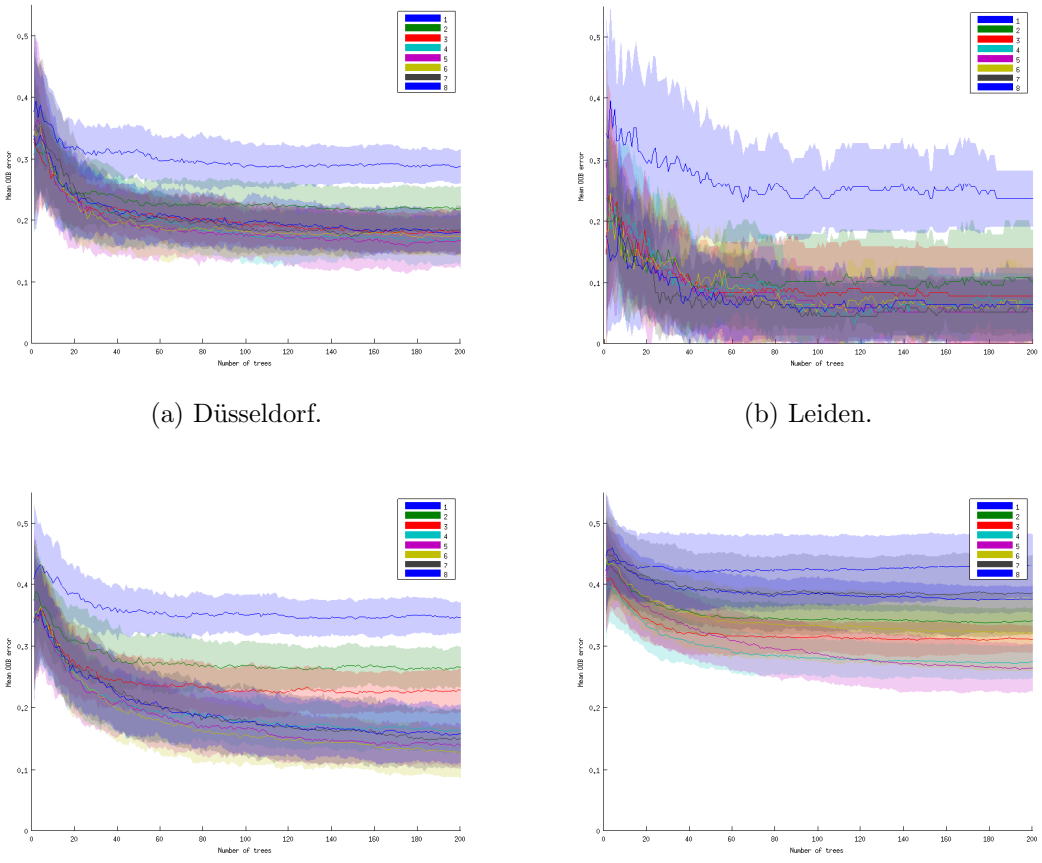
It turns out that the inclusion of EFS does not increase the overall performance (compare with the all cells combined classifier in Table 6.5). The feature importances for this classifier (Figure 6.17a) show that EFS is about as informative as the most importance feature in the non-augmented classifier, which is supported by the cumulative plot in Figure 6.17b.

An additional benefit of this exercise is that we can use the importance of the augmented feature as a reference point for other features.

6.10.5 Parameter selection

There are two main parameters in our metastasis classifier— the number of trees and the number of feature reduction iterations.

Figure 6.18 shows how the OOB error varies as the number of trees is increased, and as the number of feature reduction iterations is varied. It is clear that the error has stabilised after around 80 trees in all cases, and that the number of iterations appears to have little effect above four, though there is a possible slight increase in OOB error for greater than five.



(a) Düsseldorf.

(b) Leiden.

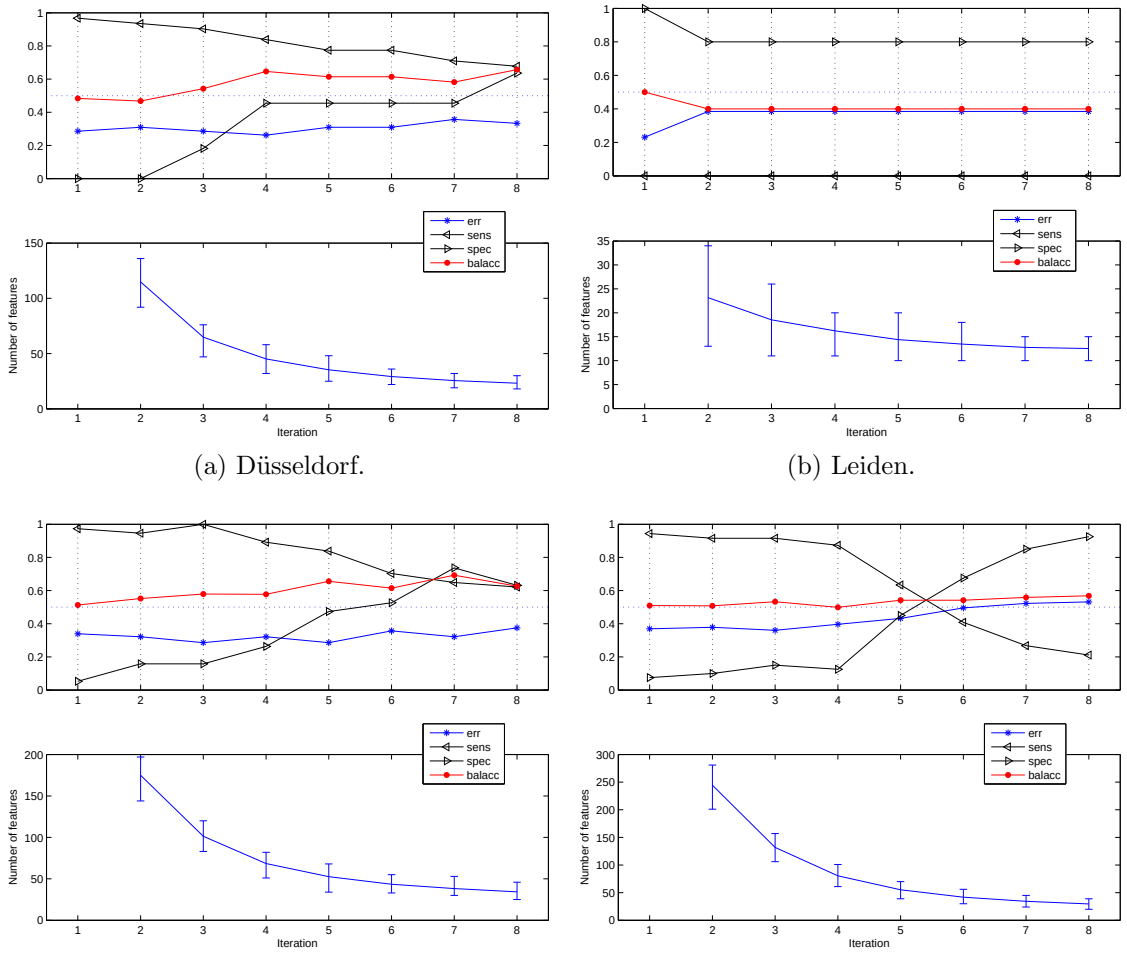
(c) Valencia.

(d) Combined.

Figure 6.18: Random forest OOB error rates as a function of the number of trees for the classifiers trained on all cells. Line colours indicate the number of feature reduction iterations, shaded band shows the standard deviation over all cross-validation folds.

However, these OOB error rates may not be entirely reliable, so we can also observe the effect of changing the number of feature reductions on the cross-validation error, as shown in Figure 6.19. It is clear that the Leiden and combined datasets perform poorly regardless of the number of features, not surprising given the low performance shown earlier.

The Düsseldorf and Valencia datasets do not appear to have a significant change in error rate, and the sensitivity and specificity are no longer so widely spread. This very different behaviour from the OOB error rates might imply a lack of data has led to overfitting, with the OOB error rates failing to adequately account for this.



(a) Düsseldorf.

(b) Leiden.

(c) Valencia.

(d) Combined.

Figure 6.19: Random forest error rates as a function of the number of trees and feature reduction iterations for the classifiers trained on all cells. The overall error rate, sensitivity, specificity, and balanced accuracy calculated using cross-validation are shown. Also shown are the mean numbers of features at each iteration, error bars show the maximum and minimum number, across all cross-validation folds. Note that the first iteration always used all features.

6.10.6 Random survival forests

A recent addition to the list of tools available for survival analysis field are random survival forests (RSF), in which the original random forests algorithm has been modified to handle survival data in the form of survival times and event status [Ishwaran et al., 2008]. Node splits are based on a criteria which maximises the difference in survival between the samples in the two daughter nodes, taking into account censoring.

The overall error for the RSF is quantified using Harrell's concordance index [Harrell Jr. et al., 1982], a non-parametric test which compares all pairs of samples to see which one is predicted to survive longer, and whether this comparison holds true for the actual survival times (see Appendix B.2).

Since we have a large number of features we can also take advantage of the *variable hunting* feature selection algorithm included in the R implementation by Ishwaran [2010], which iteratively adds features based on their importance as judged by their depth in the trees. It is designed to handle microarray datasets where the number of samples is significantly smaller than the number of predictors, and in theory should provide a model built on a greatly reduced feature set.

Initially we build a model using all features, and compare it with models built using the variable hunting selection algorithm in all cells, CD99+ cells, and CD99- cells using Harrell's concordance index, in the separate and combined datasets. The error rates are shown in Table 6.8, and summarised in Figure 6.20. An error rate of 0 means perfect prediction, and 0.5 is equivalent to random guessing.

The feature importances for each classifier built using variable hunting are shown in Figure 6.21. Note that these plots only show the features after the built-in feature selection process, in contrast to Section 6.10.3.

Based on Table 6.8 the Düsseldorf data set appears to contain the most information

		OOB (all features)	CV (var hunt)	OOB (var hunt)	CV (simple)	OOB (simple)
All	Düsseldorf	0.37	0.45	0.11	0.33	0.17
	Leiden	0.36	0.55	0.24	0.42	0.24
	Valencia	0.45	0.55	0.26	0.56	0.24
	Combined	0.49	0.46	0.28	0.44	0.25
CD99-	Düsseldorf	0.33	0.54	0.15	0.32	0.08
	Leiden	0.38	0.51	0.23	0.41	0.15
	Valencia	0.46	0.60	0.25	0.57	0.23
	Combined	0.70	0.46	0.27	0.42	0.27
CD99+	Düsseldorf	0.36	0.45	0.15	0.38	0.18
	Leiden	0.38	0.37	0.23	0.47	0.22
	Valencia	0.46	0.57	0.26	0.46	0.22
	Combined	0.48	0.44	0.31	0.40	0.25

Table 6.8: Error rates for the random survival forest survival prediction, calculated using Harrell's concordance index on either the out-of-bag (OOB) samples or the test samples from leave-one-out cross-validation (CV). Columns refer to random survival forests built as follows. All features: forest built using all features. Var hunt: forest built using variable hunting feature selection. Simple: forest built using eight iterations of the simple feature selection described in section 6.10.3.

with regards to predicting survival, so we will focus on those results. Figure 6.22 shows the selected feature importances for that classifier.

We can attempt to view the effects of these features using the conventional approach of fitting a Cox proportional hazards model and a Kaplan-Meier plot (Figure 6.23). Note that the significance values for the Cox model are smaller than the log-rank significance, indicating that simply dividing the patients into three equal groups does not capture the effect of the feature.

However, this assumes a linear or proportional relationship between a feature and survival, whereas one of the main reasons for using random forests and its variants are that they can model non-linear interactions between variables and the response.

Ishwaran et al. [2008] suggests a new measure of survival, *ensemble mortality*, which

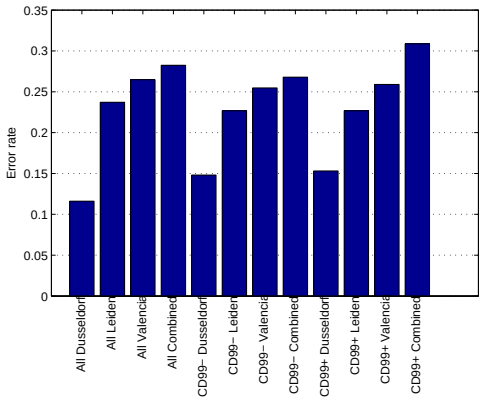


Figure 6.20: Error rates for each of the random survival forest classifiers. Classifiers were trained on all cells, CD99+ cells, and CD99- cells, with patients either divided into separate datasets or combined. Errors rates were calculated using the OOB samples.

is the total number of deaths that would be expected in the observed group of patients if they all had a particular value of a covariate. For an example see Figure 6.24, which shows the predicted marginal mortality for the selected features [Friedman, 2001].

Based on these plots we can get a feel for how the selected features influence mortality. For example, a positive skew of the mean Ki67 cytoplasmic levels and nucleus:cytoplasm ratios appears to be correlated with lower mortality.

Positive skew means the distribution of Ki67 over a patient's cells has a long tail on the right, implying that most cells have a low Ki67 value. A slight complication in the interpretation of these results is that Ki67 is typically expected to localise in the nucleus—further work will be needed to ascertain whether this is a genuine phenotype.

6.10.6.1 Verification of the random survival forest feature selection

As a final test of the random survival forest we deliberately insert informative features (obtained from the patient metadata) into the data sets. The error rates for the combined dataset with all cells is shown below.

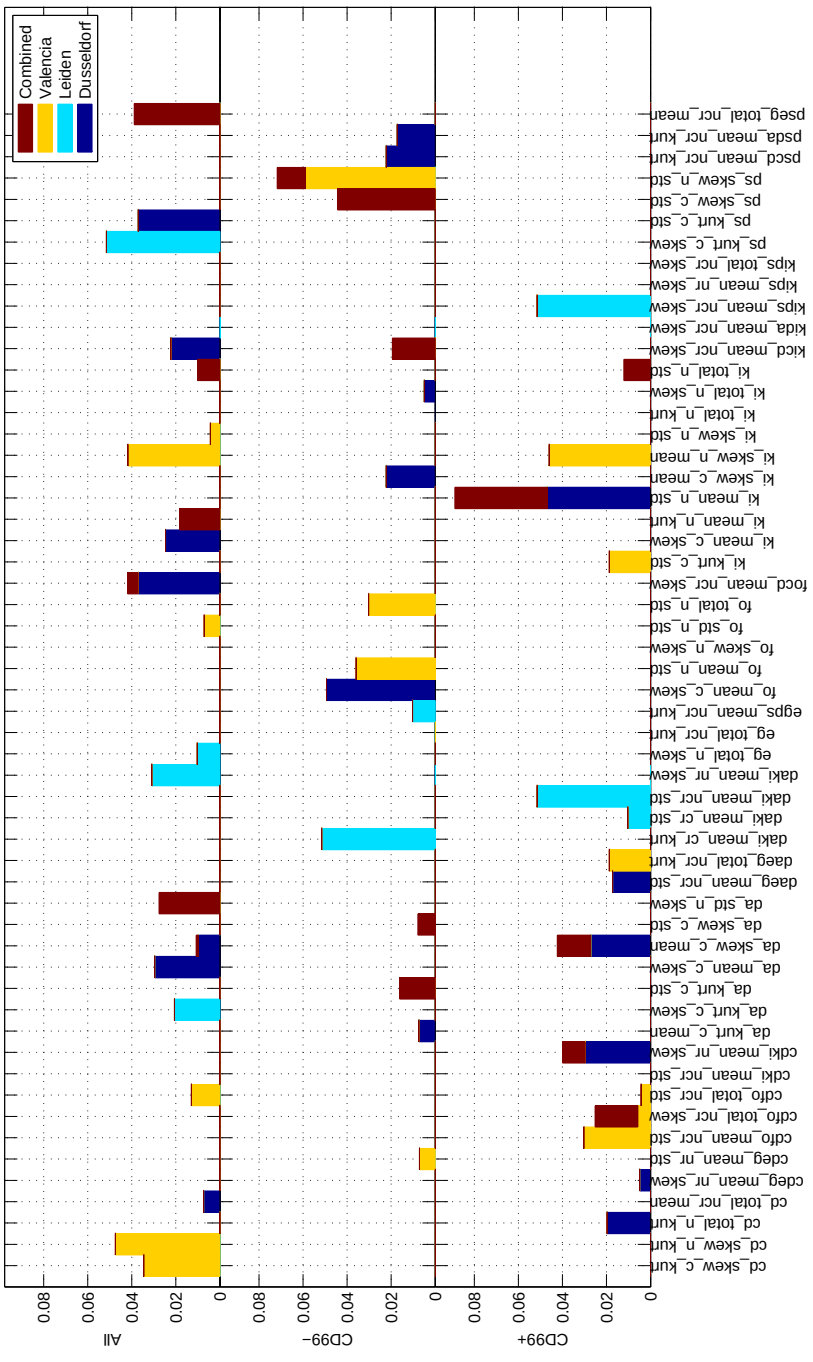


Figure 6.21: Feature importances for the random survival forest classifiers built using the variable hunting feature selection algorithm. Note that importances are only shown for features which were selected by the algorithm. For each of these plots a stacked bar chart shows the cumulative feature importance from the separate datasets, and the combined dataset.

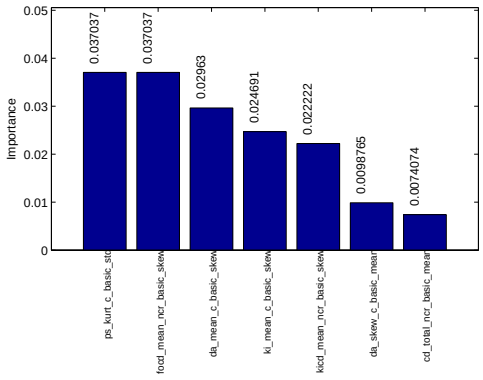
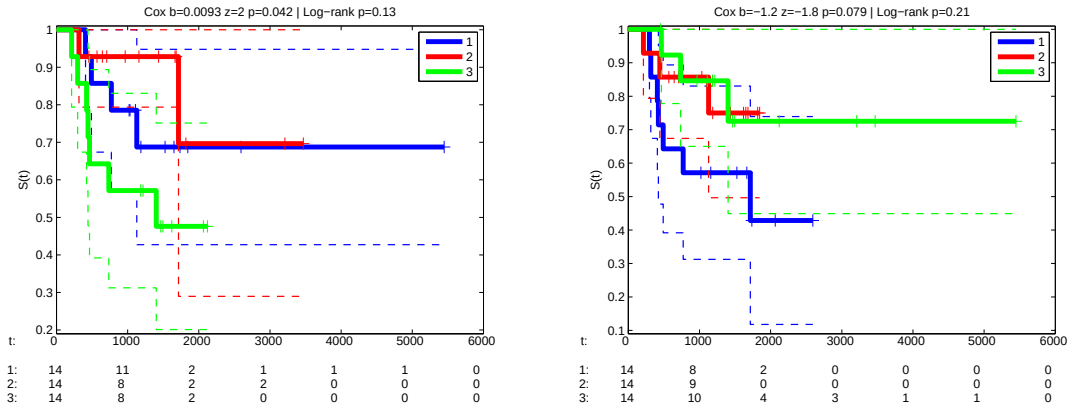


Figure 6.22: Feature importances for the all cell Düsseldorf classifier.



(a) Standard deviation of the kurtosis of cytoplasmic pS6.

(b) Skew of (mean nuclear Foxo3a):(mean cytoplasmic CD99) ratio.

Figure 6.23: Kaplan Meier plots showing the effects of two selected features on the Düsseldorf dataset. Patients have partitioned into three equally sized groups depending on whether they have a low (1), medium (2), or high (3) value of the indicated feature. (a). Cox p-value (unpartitioned data): 0.042, log-rank: 0.13. (b). Cox p-value (unpartitioned data): 0.080, log-rank: 0.21.

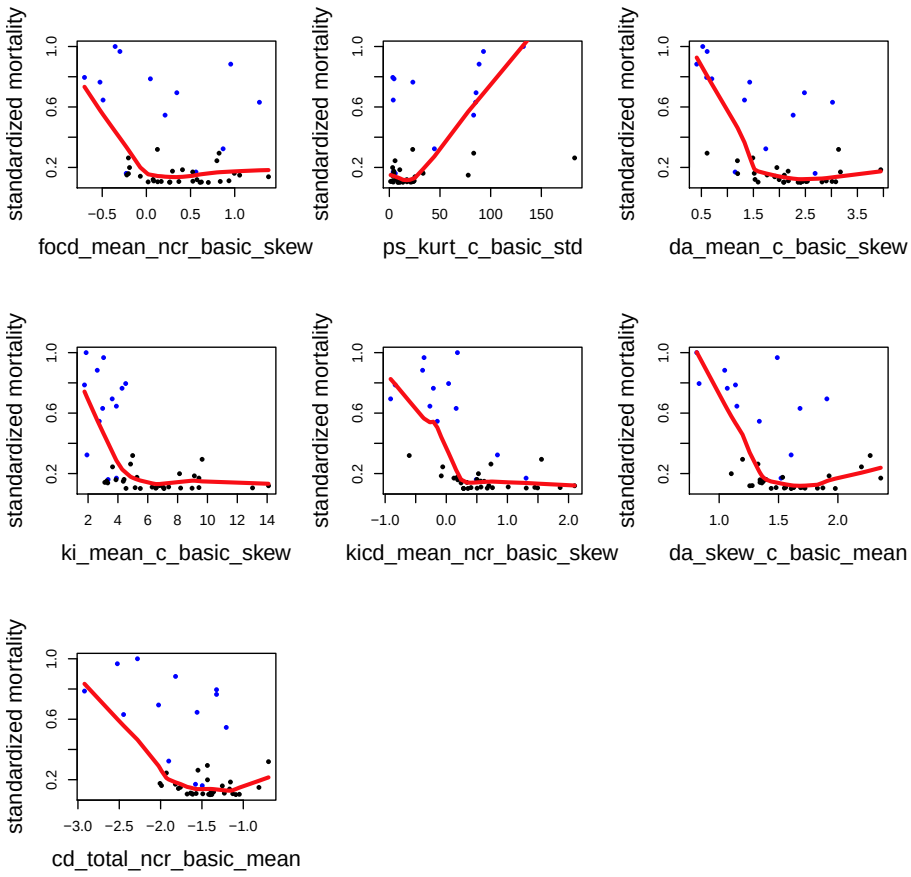


Figure 6.24: Predicted relative mortality as a function of selected features in the Düsseldorf dataset. Blue dots are observations where an event (death) occurred, black dots are censored observations.

	Error rate
Metastasis	0.25
EFS	0.13
Metastasis and EFS	0.12

The addition of metastasis brings down the error rate as expected, and the further addition of EFS brings it down even more as expected. The corresponding feature importances are shown in Figure 6.25, and the random survival forest does indeed identify EFS as being the most relevant feature.

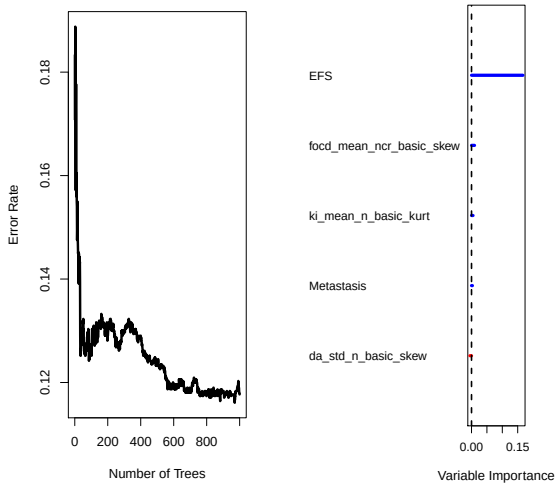


Figure 6.25: Classifier error and feature importance for the dataset augmented with metastasis and EFS.

Interestingly the addition of EFS means metastasis is no longer seen as more informative than the other unrelated features, even though it clearly should be. This is possibly a shortcoming in the feature selection or importance rating.

6.11 Conclusions

In this chapter we started with an example of a typical approach to biomarker interpretation, and illustrated some of the problems in the somewhat arbitary definition of thresholds. We then tried multiple machine learning approaches to look for patterns and correlations with metastasis and survival outcome, with some success.

The normalisation of images between datasets was an unresolved problem as shown by the differing results between the classifiers trained on independent datasets with a maximum balanced accuracy of 0.66 and 0.63 for the Düsseldorf and Valencia datasets respectively, compared with 0.57 for the combined one. More importantly the sensitivity and specificity (Figure 6.11) were very similar for Düsseldorf and Valencia indicating rela-

tively good discrimination between both metastasis and non-metastasis patients, whereas the combined classifier showed a large difference indicating generally poor performance.

Unfortunately the very small size of the Leiden dataset meant we could not obtain useful results from it. The relatively small sizes of the other two datasets mean that it is possible our results have occurred by chance as we have been looking at a large number of features, though we have enough evidence to say a further study would be worthwhile.

An unplanned benefit of analysing the datasets independently was that we were able to corroborate the feature importances from one dataset with another. The great variation between and within datasets meant that the classifiers came up with different feature importances in each one, but Figure 6.16 showed that when we selected a feature that was strongly important in one dataset and qualitatively compared the change in distribution of that feature between patients with and without metastasis, all datasets showed a similar shift.

Finally we used random survival forests to see if we could predict patient survival. The OOB error estimates suggest we were able to do so, particularly with the Düsseldorf dataset which had an error of 0.11. However we should be careful not to rely on the OOB estimates too much, as evidence from the metastasis classifier suggested that the cross-validation error rates were significantly higher than those obtained using the OOB data.

The intepretation of the random survival forest is still an area to be investigated. Example plots showing the marginal dependance of one important feature on patient mortality were given, but it is unclear how this can be taken further.

Chapter 7
Conclusions and future work

We set out to explore the applications of quantitative image analysis to cell biology, and did so with multiple projects ranging from vital low level subcellular research to clinical applications of image segmentation and machine learning.

The practicalities and difficulties of biological research were shown in Chapters 3 and 4, which together form an example of the whole process of starting from basic biological tools and ending up with quantitative data. Unfortunately the quantitative results were not as significant as hoped, but the tools developed will in future be used for the analysis of related datasets. From the biological side the cloning of the CPC members has left a valuable tool for future research into mitosis, and in particular Deterin was shown to be the *Drosophila* orthologue of human Survivin.

In chapter 5 we developed a novel segmentation algorithm for clumped cells which was validated and used to extract cell-based image features. One of the distinguishing features of cancer is the large heterogeneity between cells, and this chapter showed how the distribution of various markers amongst cells could be examined. The second contribution of this chapter was a machine learning framework to help with the analysis of responses of single cells, including an example of how the feature importances from a random forest classifier could be used to guide the examination of biologically relevant features. Finally a preliminary application of analysis methods for spatially-correlated data was carried out. The relatively small amount of data, and uncertainty in how to control for variation between images, meant that we could not draw any strong conclusions. However the motivation for the work, namely the desire to analyse how cells in a clump influence each other, still

remains, and the ideas described here should provide the basis for future experiments.

Finally chapter 6 once again compared the potentially biased human grading of images with taking a quantitative analysis approach. Importantly this chapter demonstrated the application of methods developed in the previous chapter to clinical data, with known patient survival outcomes. Unfortunately the amount of data available for analysis was very limited, but in common with many other studies of rare diseases this is not something that can be easily remedied. The lack of normalisation between datasets was a problem, and it remains unclear as to how this should be handled.

As a whole this thesis has demonstrated the application of quantitative imaging techniques at multiple levels of biomedical research, and shown how human biases can sometimes unduly influence the analysis of results. It has also shown how basic and applied research can be drawn together, with significant overlap between all projects, supporting the move towards translational research in the biomedical field.

7.1 Future work

Some of the problems and limitations that were encountered have already been highlighted within each chapter. However, there are some bigger challenges which have arisen from this thesis, as well as potential new avenues of research, and they are outlined here.

7.1.1 Spatio-temporal subcellular protein localisation

One of the problems of analysing the CPC is that either changes were so small that we were unsure whether they were real or not, or so large that it was not possible to understand what was happening. One aspect of the CPC that would be nice to monitor is the binding dynamics of members of the complex.

Fluorescence recovery after photobleaching experiments [Sprague and McNally, 2005]

were tried without much success and were omitted from this thesis, but more recently other tools such as photoconvertible proteins, which can be switched from green to red by a targetted beam of UV light [McKinney et al., 2009], would allow selective labelling of a subpopulation of a protein. A harder though much more versatile approach would be to engineer a photoactivatable version of one of the CPC members, allowing it to be turned on and off at will in particular locations, as demonstrated by Wu et al. [2009] with the protein Rac1, which regulates actin cytoskeletal dynamics.

In terms of developing the analysis tools, one issue faced by biologists is the need to compare multiple cells which may vary greatly in shape. Our kymograph tool rotated the nuclei into a common axis, but we may be able to make use of non-rigid registration as used in the medical imaging field [Rueckert et al., 1999], particularly when examining distorted cells.

7.1.2 Spatially-correlated interactions between cells

The work described in Section 5.12 looked at the dependencies between cell heterogeneity and spatial location in a clump. Research in this area is very limited, but now that we have demonstrated the basic ideas required for this analysis future work should look at more appropriate machine learning or statistical measures to identify spatially-correlated patterns.

A natural extension of this work would be to look for patterns in the Ewing's sarcoma TMAs, providing valuable information for computational biologists involved in modelling the growth of tumours.

7.1.3 Machine learning framework for partially normalised data

One of the major limitations in our analysis of the TMAs was the inability to normalise the samples from different study locations, so the development of improved normalisation algorithms for fluorescently labelled histopathological images is therefore one area of future work.

However it is likely that systematic differences between datasets will still exist, and since we have limited amounts of data the issue of combining multiple sets of data cannot be avoided. If we can make the assumption that the underlying population of patients is the same for all datasets then in principle the effects of each source of error in Table 6.2 could be modelled. Bayesian networks are an attractive option [Heckerman, 1996], but the choice of model, and more importantly how imaging data can be integrated into the framework, needs to be investigated.

Appendix A
Segmentation and analysis

A.1 Fast level set algorithm

The fast level set algorithm of Shi and Karl [2008] represents the contour in the form of two lists of pixels:

L_{in} contains pixels marking the inner edge of the contour.
L_{out} contains pixels marking the outer edge of the contour.

The contour (Figure A.1) is moved outwards by moving a pixel from L_{out} to L_{in}, adding any new pixels which now adjoin the boundary to the appropriate list, and removing any pixels in either list which are no longer on the boundary. An analogous procedure is used to move the contour inwards.

The evolution of the contour is divided into separate image driven iterations and smoothing (or regularisation) iterations. During the image driven iterations a speed field is calculated, and the contour is shifted inwards or outwards by one pixel depending on the sign of the speed (this is in contrast to a full level set method where the magnitude of the speed is also taken into account). During smoothing iterations a Gaussian kernel is used to smooth the internal representation of the higher order level set function ϕ, and the pixel lists are updated appropriately.

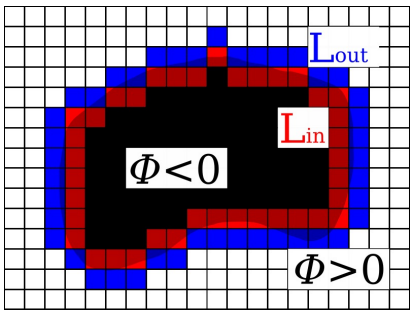


Figure A.1: The fast level set representation of an image.

A.2 Solution of the random walker algorithm

The solution of the random walker segmentation problem is briefly outlined here, see Grady [2006] for the complete solution.

Construct a graph from an image by creating nodes $v_i \in V$ at each pixel, and defining edges $e_{ij} \in E$ with weight w_{ij} between all pairs of neighbouring nodes, where i and j are used to index nodes. Create the $N \times N$ Laplacian matrix L_{ij} where N is the number of nodes (also the number of pixels)

$$L_{ij} = \begin{cases} d_{ij} & \text{if } i = j \\ -w_{ij} & \text{if } v_i \text{ and } v_j \text{ are joined by an edge } e_{ij} \\ 0 & \text{otherwise} \end{cases}$$

$$d_{ij} = \sum_i w_{ij}$$

Since we only need to solve for the unlabelled (unseeded) nodes the seeded nodes can be removed from the Laplacian to leave L_U, and a $K \times S$ matrix B encoding the Dirichlet boundary conditions is created where K is the number of labels (independent regions) and

S is the number of seed nodes (multiple nodes may have the same seed label).

$$B_{ki} = \begin{cases} 1 & \text{if } v_i \text{ is a seed with label } k \\ 0 & \text{otherwise} \end{cases}$$

Let x_U and x_M be the random walker potentials (probabilities) and the unseeded and seeded nodes respectively. The combinatorial Dirichlet problem has the solution

$$L_U x_U = -B^T x_M$$

The potentials x_U corresponding to seed (label) k are found by defining

$$m_i^s = \begin{cases} 1 & \text{if } v_i \text{ has label } k \\ 0 & \text{otherwise} \end{cases}$$

And solving

$$L_U x^k = -B^T m^k$$

for each label k to find the potentials x^k. This can be done efficiently using a sparse matrix solver.

A.3 Moran's index

The expected value $E[I]$ and variance $Var[I]$ of Moran's I calculated from N samples are given by [Goodchild, 1988]:

$$E[I] = \frac{-1}{N-1} \qquad Var[I] = E[I^2] - E[I]^2$$

where

$$S_0 = \sum_{i,j} w_{ij} \qquad D = \frac{\sum_i (Z_i - \bar{Z})^4}{\left(\sum_i (Z_i - \bar{Z})^2\right)^2}$$

$$S_1 = \frac{\sum_{i,j} (w_{ij} + w_{ji})^2}{2} \qquad C = (N-1)(N-2)(N-3)S_0^2$$

$$S_2 = \sum_k \left(\sum_i w_{ik} + \sum_j w_{kj} \right)^2 \qquad B = D\left((N^2 - N)S_1 - 2NS_2 + 6S_0^2\right)$$

$$E[I^2] = \frac{A - B}{C} \qquad A = N\left((N^2 - 3N + 3)S_1 - NS_2 + 3S_0^2\right)$$

Appendix B
Supervised classification

B.1 Confusion matrices

Confusion matrices show the number of samples predicted to be in each class against the true class labels.

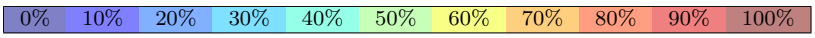
		Actual	
		0	1
Predicted	0	True negative (TN)	False negative (FN)
	1	False positive (TP)	True positive (TP)

For the special case of binary classification the classes may be referred to as true and false, for instance when classifying samples based on whether they are positive or negative for a biomarker, and the sensitivity and specificity can be used to summarise the overall performance of the classifier.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad \text{Error rate} = \frac{\text{FP} + \text{FN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}} \qquad \text{Balanced accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2}$$

To help with the reading of large confusion matrices colour coding can be used, as seen in Table B.1.



		True		
		1	2	3
Predicted	1	3 (100.0%)	0 (0.0%)	0 (0.0%)
	2	0 (0.0%)	3 (100.0%)	0 (0.0%)
	3	0 (0.0%)	0 (0.0%)	3 (100.0%)

(a) Perfect classification

		True		
		1	2	3
Predicted	1	2 (66.7%)	0 (0.0%)	0 (0.0%)
	2	1 (33.3%)	2 (66.7%)	1 (33.3%)
	3	0 (0.0%)	1 (33.3%)	2 (66.7%)

(b) Average classification

Table B.1: Example confusion matrices for good and reasonable classifications. Percentages are calculated column-wise, that is they show the percentage of true class 1 samples that were correctly classified as class 1, and incorrectly classified as class 2 or 3.

B.2 Harrell's concordance index

Harrell's concordance index is a non-parametric test that can be used to compare survival times [Harrell Jr. et al., 1982]. It is an extension of the Mann-Whitney U test which takes censoring into account.

Let $t_1, \ldots, t_N$ be the survival times of N subjects, and $c_0, \ldots, c_N$ be the $[0, 1]$ censoring indicators where 1 indicates a censored survival time. The index is calculated from all pairs of subjects $i, j = 1, \ldots, N$, $i \neq j$ in which it is possible to order the pair using the survival times, meaning any pairs in which both subjects were censored ($c_i = 1, c_j = 1$) are ignored.

A score S and counter M are initialised to 0, incremented for each pair as follows:

True survival	Predicted survival	Increment S	Increment M
$t_i > t_j, c_j = 0$	$\hat{t}_i > \hat{t}_j$	1	1
or	$\hat{t}_i = \hat{t}_j$	0.5	1
$t_i == t_j, c_i = 1, c_j = 0$	$\hat{t}_i < \hat{t}_j$	0	1
$t_i = t_j, c_i = 0, c_j = 0$	$\hat{t}_i = \hat{t}_j$	1	1
	$\hat{t}_i \neq \hat{t}_j$	0.5	1
otherwise		0	0

Harrell's concordance index is then found by normalising by the number of valid com-

parisons:

$$C = \frac{S}{M}$$

B.3 Feature importance characteristics of random forests

As outlined in Breiman [2001] random forests are very good at identifying informative features and avoiding over-fitting. The actual performance will depend on the characteristics of the dataset in question, but it is still illustrative to look at some of the feature discrimination properties of random forests.

B.3.1 Regression: synthetic data

A data set was synthesised based on the *Friedman #1* method used by Breiman [1996]. A 1000-feature synthetic data set X was constructed by drawing each variable $\mathbf{x_i}$ from a uniform $[0, 1]$ distribution, and 1000 samples were drawn. An output variable y was created from the first five features $x_1, \ldots, x_5$ as follows:

$$y = 10 \sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + \eta$$

where η is a normally distributed zero-mean unit-variance random noise term. The remaining 995 features are therefore uninformative. A plot of y against each of the first 5 features is shown in Figure B.1, along with one of the uninformative features, and it is clear that the dependence between any one of the features and the output is very weak.

10-fold cross validation was used to train and test a random forest regressor, initially with a training set of 900 and test set of 100 samples. We also wanted to see the performance of random forests on very small data sets, particularly where the number of samples is outnumbered by the number of features, so we repeated the cross-validation after swapping

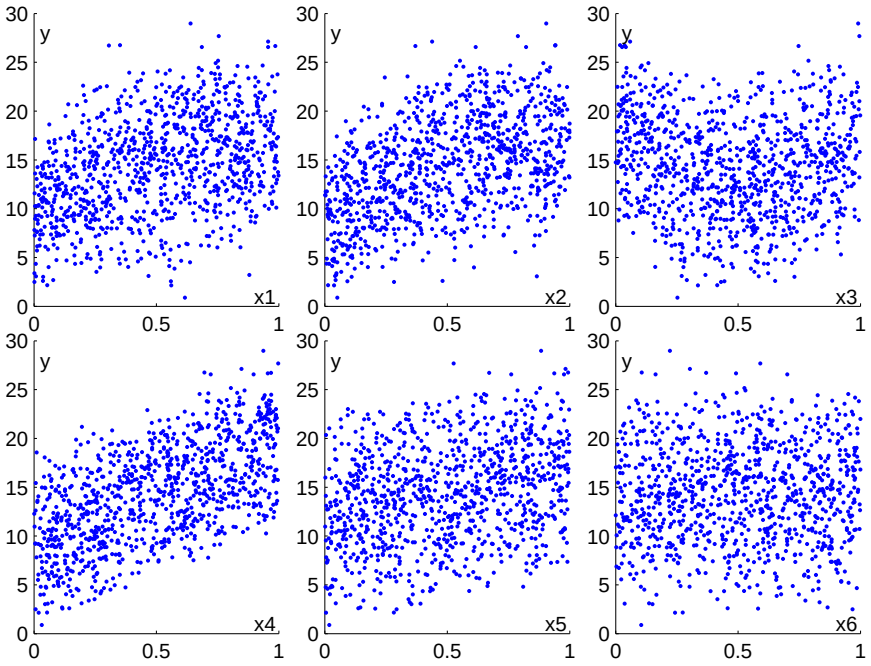
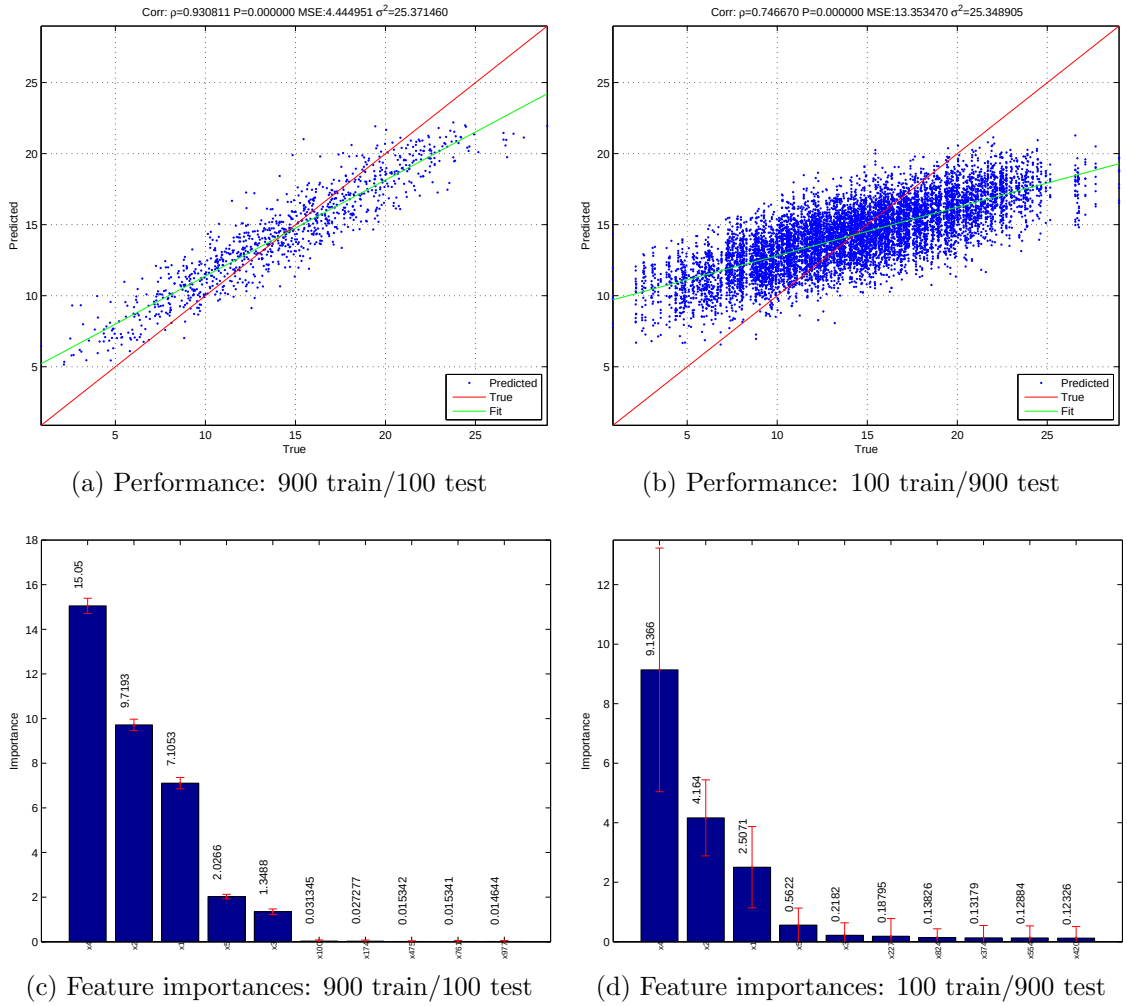


Figure B.1: Contribution of the first six synthetic variables to the output variable y. $x_1, \ldots, x_5$ contribute to y, x_6 is an example of the remaining 995 uninformative noise terms.

the training and test sets (100 training and 900 test samples at each fold). The results are shown in Figure B.2.

As expected the larger training set leads to a smaller regression error. Of greater interest are the feature importances in B.2c and d. The random forest trained using the larger set picks out all five relevant features to be above the background level, but it is also notable that the random forest trained on 100 samples does fairly well despite the number of features being 10 times greater. Strictly speaking since the mean importances are calculated from all cross-validation folds which means all 1000 samples are taken into account, but the standard deviations show that even if a single training set of 100 samples were used we would still expect the top three features to be identified.



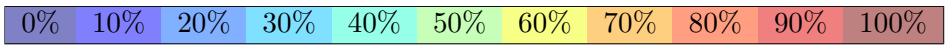
(a) Performance: 900 train/100 test

(b) Performance: 100 train/900 test

(c) Feature importances: 900 train/100 test

(d) Feature importances: 100 train/900 test

Figure B.2: Results of testing the random forest regression on different numbers of samples. (a) Regression performance from the cross validation using 900 training and 100 test samples. The true y is plotted against the predicted y, for a perfect regression all points should lie on the red line. (b) Regression performance using 100 training and 900 test samples.
(c) Aggregated feature importances from (a). Error bars show the standard deviation across all cross validation runs. $x_1, \ldots, x_5$ are the features we aim to pick out as informative. (d) Aggregated feature importances from (b).



		True		
		0	1	2
Predicted	0	26.2	2.6	0.0
	1	0.0	0.0	0.0
	2	73.8	97.4	100.0

(a) 900 train/100 test

		True		
		0	1	2
Predicted	0	68.0	43.6	21.6
	1	0.0	0.0	0.0
	2	32.0	56.4	78.4

(b) 100 train/900 test

Table B.2: Results of testing the random forest classification on different numbers of samples using 10-fold cross-validation. Each column shows the percentage of the true class frequencies that were predicted in each class.

B.3.2 Classification: Synthetic data

The performance of a random forest classifier on low numbers of samples in a high-dimensional data set was also considered. The synthetic output y was convered into a classification problem by creating a new target variable y_c ($\lfloor \cdot \rfloor$ means round down):

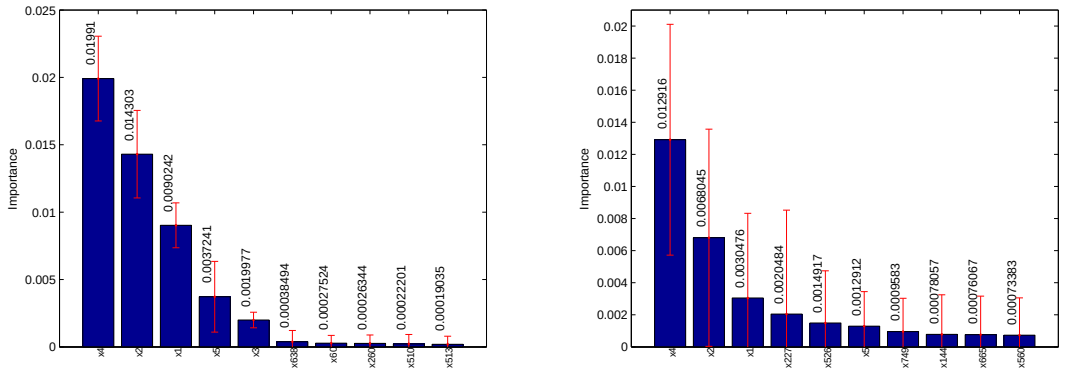
$$y_c = \lfloor y/10 \rfloor$$

This gives a three class data set:

Class	Frequency
0	214
1	647
2	139

The same cross validation procedure as for regression was used.

The classification results in Table B.2 show that the classifier is effectively useless, yet it still appears to be able to pick out the relevant features as seen in Figure B.3a where all five features are ranked first, and Figure B.3b where at least the first two would be expected to be identified without cross-validation.



(a) Feature importances: 900 train/100 test (b) Feature importances: 100 train/900 test

Figure B.3: Feature importance from testing the random forest classification on different numbers of samples. The feature importances were aggregated over all cross validation runs, the error bars show the standard deviations.

Appendix C

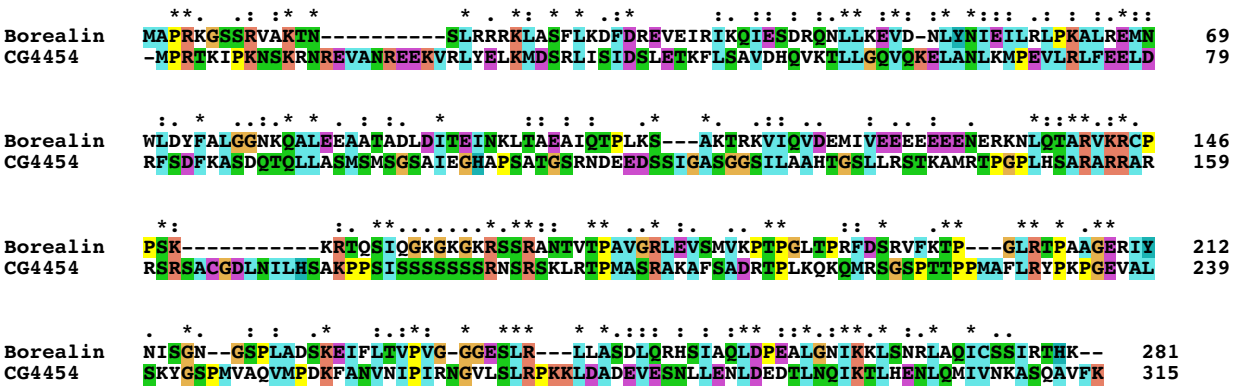
Protein Alignments

Alignments between the human reference sequences (first) and the corresponding *Drosophila melanogaster* orthologs (second) were calculated using ClustalX [Larkin et al., 2007].

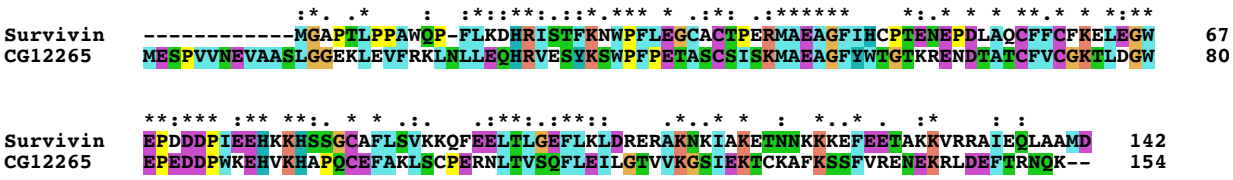
C.1 Aurora B, CG6620

```
                *: .. .:.    *:  *  *:.:*:   ::**:.  .* :**              .:: * *     .::  ***:*
Aurora-B    MAQKENSYPWPYGRQTAPSGLSTLPQRVLRKEPVTPSALVLMSRSNVQPTAAPGQKVMENSSGTPDILTRHFTIDDFEIG   80
CG6620      MTLSRAKHAN---RNHLPHLLAKVPE--EHQEPIKNMCLKMMS---------------HDAYGQP----YDWSPRDFEMG   56

                  **:****.******::**::**:**:**.::.*  *::*: *******::*:**:**** .:*:*. **** ** *..*
Aurora-B    RPLGKGKFGNVYLAREKKSHFIVALKVLFKSQIEKEGVEHQLRREIEIQAHLHHPNILRLYNYFYDRRRIYLILEYAPRG   160
CG6620      AHLGRGKFGRVYLARERHSHYLVAMKVMFKEELRKGCVQRQVLREIEIQSRLKHPHILRLLTWFHDESRIYLALEIASEG   136

                **:*.*: :    *** *:*.   ::*:** *** ::*****:****:**  ..:**:******.*:*. :*:*:*****
Aurora-B    ELYKELQKS--CTFDEQRTATIMEELADALMYCHGKKVIHRDIKPENLLLGLKGELKIADFGWSVHAPSLRRKTMCGTLD   238
CG6620      ELFKHLRGAPNHRFDEPRSAKYTYQVANALNYCHLNNVIHRDLKPENILLTSTDDLKLADFGWSAHTPNNKRRTLCGTLD   216

                ******::*. :::.** **:*:****::** ***** * :.** :* ::::.:*: :. *.::**. ***::.. *:.*
Aurora-B    YLPPEMIEGRMHNEKVDLWCIGVLCYELLVGNPPFESASHNETYRRIVKVDLKFPASVPMGAQDLISKLLRHNPSERLPL   318
CG6620      YLPPEMVDGNSYDDSVDQWCLGILCYEFVVGCPPFESNSTESTYSKIRRMEISYPSHLSKGCKELIGGLLRKESKGRITL   296

                .:* :* **:*. :.* *  .    . :..*
Aurora-B    AQVSAHPWVRAN-SRRVLPPS----ALQSVA--   345
CG6620      VDVMTHYWVKAGMAERELQLQKRERGKENTARN   329
```

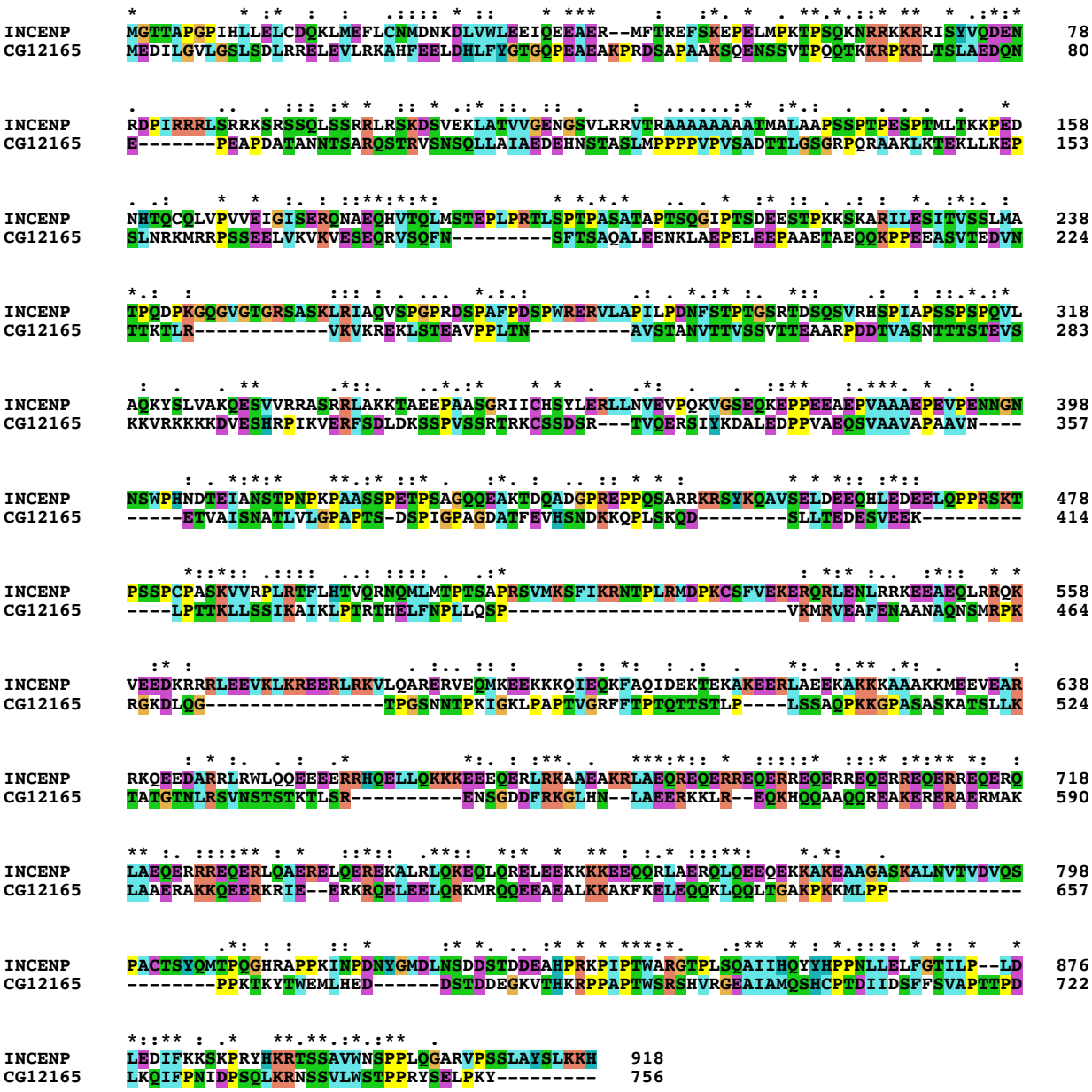
C.2 Borealin, CG4454



C.3 Survivin, CG12265



C.4 Incenp, CG12165



Appendix D
MCF7 experimental work

The experimental protocols described in this section were carried out by Claudia Buehnemann in the Dunn School of Pathology, University of Oxford.

D.1 MCF7 fixing and staining

Cells on coverslips were fixed in 4 % PFA for 10 min to 15 min, washed in 5× TBS, permeabilised/blocked for 1 h in TBS/0.5 % Triton/10 % serum, and incubated with the primary antibody in TBS/0.3 % Triton/3 % serum overnight 4 °C. The coverslips were then washed 5× TBS 3 min and incubated with secondary antibody (goat anti-rabbit Alexa 594 1:300 in TBS) 2 h RT. They were washed 3× TBS 3 min, incubated with Phalloidin Alexa 488 1:50 in TBS for 2 h RT, washed 3× TBS, incubated with DAPI 1:10000 10 min, and washed 3× PBS. Cells were mounted with Fluorgel at 4 °C overnight and sealed with nail varnish.

Appendix E

TMA data preparation

E.1 TMA experimental protocols

Fixing and staining was carried out by Claudia Buehnemann in the Dunn School of Pathology, University of Oxford.

TMA cores were dewaxed in Xylen 2× 5 min, rehydrated in ethanol 2× 100 %, 95 %, 80 %, 70 % and 50 % 3 min each. They were rinsed in cold dH_2O, incubated in 1 % Tween 20 for 30 min RT, and rinsed in dH_2O. Antigen retrieval was done using citrate pH 6.0 in a pressure cooker (125 °C 2 min then 85 °C 10 min, followed by cooling in retrieval solution for 20 min to 30 min. They were permeabilised/blocked in TBS/0.5 % Tween/10 % serum for 1 h and incubated in primary antibody, made up from TBS/0.3 % Tween/3 % serum with one of rabbit anti-Ki67 1:300, rabbit anti-Egr1 1:50, rabbit anti-pMAPK 1:50 or rabbit anti-FoxO3a 1:100, overnight at 4 °C.

5× 3 min TBS washes were followed by the secondary antibody (goat anti-rabbit Alexa 488 1:200) 2 h RT, 5× 3 min TBS washe, blocking again in TBS/0.5 % Tween/10 % serum for 30 min and incubation with the second primary antibody (mouse anti-CD99 1:100) overnight as with the first primary.

5× 3 min TBS washes were performed again, followed by incubation with the second secondary antibody (goat anti-mouse Alexa 555 1:300) 2 h RT, 3× 3 min TBS washes, blocking again for 30 min in TBS/0.5 % Tween/10 % rabbit serum, incubation with pS6 1:30 in TBS, 3× 3 min TBS washes, DAPI 1:10000 10 min, and 5× 3 min TBS washes.

E.2 TMA Image acquisition and analysis

Image acquisition was carried out by Haiyue Yu and Claudia Buehnemann in the Dunn School of Pathology. The image analysis which is briefly summarised in this section was carried out by Haiyue Yu.

Images of the stained TMA cores were acquired using an Olympus FluoView FV1000 confocal using a 60× oil objective (numerical aperture 1.35), and the pinhole set to acquire a 1 µm thick section. Each image had dimensions of 2048 × 2048 pixels, with a pixel size of 0.103 µm.

Images were denoised by median and Gaussian filtering. Nuclei were segmented from the DAPI channel, using Otsu's method to initialise a level set using a hybrid speed field, and refined using a watershed method to split clumps. For the cytoplasmic segmentation thresholding was used to find the outer boundaries of cells. The cytoplasm of individual cells was segmented by starting with a watershed tesslation and iteratively updating the segmentation based on the intensity gradient of the CD99 marker. Segmentation results were validated against manual segmentations from three experts.

E.3 Microscope bias field

Some types of microscopes, especially wide-field ones, can sufer from inhomogeneous sample illumination which effectively causes a bias field in the acquired image. This should not occur in a correctly configured confocal microscope, and to verify this average images $\bar{I}^c(\mathbf{x})$ and average segmentions $\bar{S}(\mathbf{x})$ for each channel c of each image $k = 1, \ldots, N$ in each of the

TMA datasets were calculated:

$$\bar{I}^c(\mathbf{x}) = \frac{1}{N}\sum_{k=1}^{N} I_k^c(\mathbf{x})$$
$$\bar{S}(\mathbf{x}) = \frac{1}{N}\sum_{k=1}^{N} S_k(\mathbf{x})$$

where $S_k(\mathbf{x})$ is the $\{0, 1\}$ segmentation indicator matrix.

An overall average image normalised by the segmentation can then be calculated and used to check for the presence of illumination inhomogeneities:

$$\bar{I}_s^c(\mathbf{x}) = \frac{\bar{I}^c}{\bar{S}}$$

The results are shown in Figure E.1. The full average images show a clear bias towards the centre of the images, but this is because the microscope would have been positioned so that the main mass of cells were in the centre of the field of view. After normalisation by the average segmentation there is no obvious remaining bias.

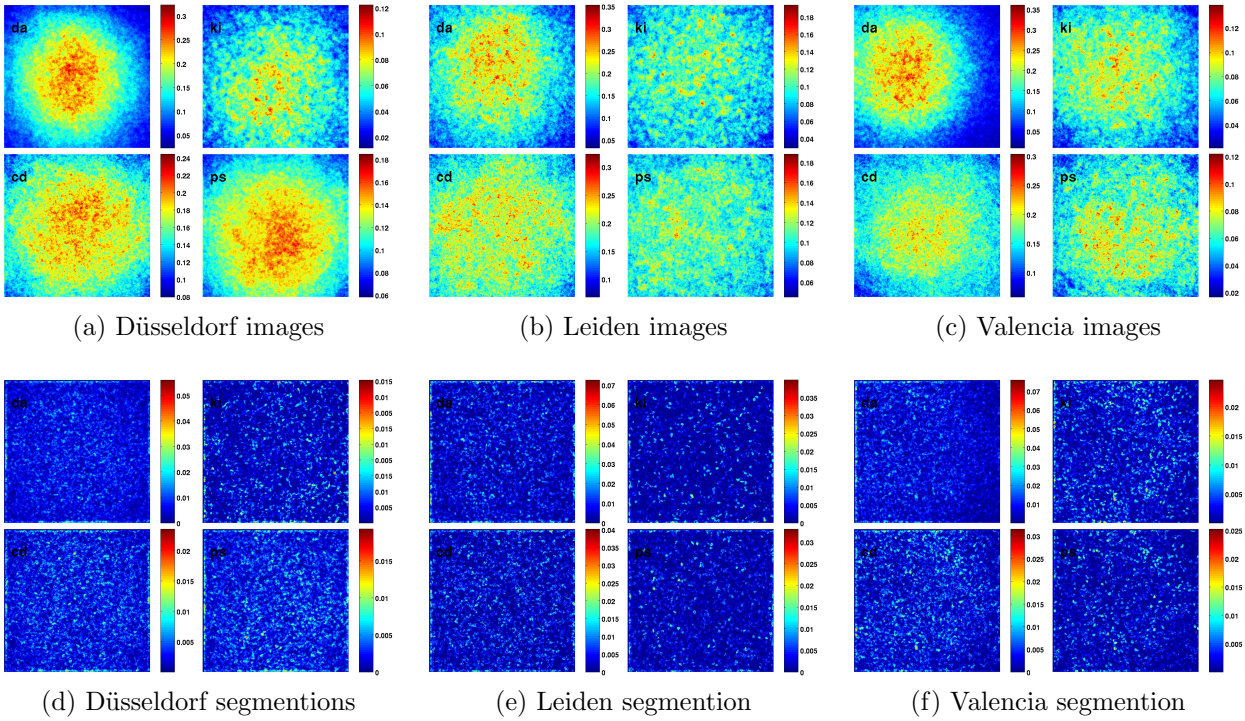
E.4 Feature naming convention

For simplicity features may be referred to in an abbreviated form, which is as follows.

Markers are given a two character abbreviation:

da: DAPI	**eg:** Egr1	**ki:** Ki67
cd: CD99	**fo:** Foxo3a	**ps:** pS6

In the case of ratios between two different markers two markers will be given, for example “daeg” is a ratio between DAPI and Egr1.



(a) Düsseldorf images (b) Leiden images (c) Valencia images

(d) Düsseldorf segmentions (e) Leiden segmention (f) Valencia segmention

Figure E.1: Average intensity images and segmentations for each channel in the Düsseldorf, Leiden and Valencia datasets.

Type of cell feature:

total: Total (integrated) intensity

mean: Mean intensity

std: Standard deviation of intensity

skew: Skewness of intensity

kurt: Kurtosis of intensity

Compartments are indicated by a single character:

c: Cytoplasm

n: Nucleus

Ratios:

cr: Ratio of two markers in the cytoplasm

nr: Ratio of two markers in the nucleus

ncr: Ratio of a marker in the nucleus to a marker in the cytoplasm

Patient features have a suffix indicating the summary statistic calculated over all cells in that patient:

mean: Mean feature value in all cells

skew: Skewness of feature value

std: Standard deviation of feature value

kurt: Kurtosis of feature value

References

H. Aberle, H. Schwartz, and R. Kemler. Cadherin-catenin complex: Protein interactions and their implications for cadherin function. *Journal of Cellular Biochemistry*, 61(4): 514, 1996. doi: 10.1002/(SICI)1097-4644(19960616)61:4¡514::AID-JCB4¿3.0.CO;2-R.

M. Abramoff. Image processing with imagej. *Biophotonics International*, 11(7):36–42, 2004.

R. R. Adams. Essential roles of drosophila inner centromere protein (incenp) and aurora b in histone h3 phosphorylation, metaphase chromosome alignment, kinetochore disjunction, and chromosome segregation. *The Journal of Cell Biology*, 153(4):865, 2001. doi: 10.1083/jcb.153.4.865.

L. J. Ahonen, A. M. Kukkonen, J. Pouwels, M. A. Bolton, C. D. Jingle, P. T. Stukenberg, and M. J. Kallio. Perturbation of incenp function impedes anaphase chromatid movements and chromosomal passenger protein flux at centromeres. *Chromosoma*, 118(1): 71–84, 2009. doi: 10.1007/s00412-008-0178-0.

D. C. Altieri. The case for survivin as a regulator of microtubule dynamics and cell-death decisions. *Curr. Opin. Cell Biol.*, 18(6):609–15, 2006. doi: 10.1016/j.ceb.2006.08.015.

D. G. Altman, B. Lausen, W. Sauerbrei, and M. Schumacher. Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. *JNCI Journal of the National Cancer Institute*, 86(11):829, 1994. doi: 10.1093/jnci/86.11.829.

S. J. Altschuler and L. F. Wu. Cellular heterogeneity: do differences make a difference? *Cell*, 141(4):559–63, 2010. doi: 10.1016/j.cell.2010.04.033.

D. Altunbay, C. Cigir, C. Sokmensuer, and C. Gunduz-Demir. Color graphs for automated cancer diagnosis and grading. *IEEE Trans Biomed Eng*, 57(3):665–74, 2010. doi: 10.1109/TBME.2009.2033804.

G. Amir. Expression of p53 gene product and cell proliferation marker ki-67 in ewing's sarcoma: Correlation with clinical outcome. *Human Pathology*, 33(2):170, 2002. doi: 10.1053/hupa.2002.31475.

Y. Amit and D. Geman. Shape quantization and recognition with randomized trees. *Neural Computation*, 9(7):1545, 1997. doi: 10.1162/neco.1997.9.7.1545.

L. Anselin. Local indicators of spatial association-lisa. *Geographical Analysis*, 27(2):93, 1995. doi: 10.1111/j.1538-4632.1995.tb00338.x.

X. Bai. Splitting touching cells based on concave points and ellipse fitting. *Pattern Recognition*, 42(11):2434, 2009. doi: 10.1016/j.patcog.2009.04.003.

L. Beaumont. *Investigating the High-Throughput Statistical Characterisation of Neoplastic Cells in Microscopy for Improving Cancer Treatment Planning*. PhD thesis, Institute of Biomedical Engineering, University of Oxford, Oxford, OX3 7DQ, August 2009.

M. Bernstein, H. Kovar, M. Paulussen, R. L. Randall, A. Schuck, L. A. Teot, and H. Juergens. Ewing's sarcoma family of tumors: current management. *Oncologist*, 11(5):503–19, 2006. doi: 10.1634/theoncologist.11-5-503.

C. C. Bilgin, P. Bullough, G. E. Plopper, and B. Yener. Ecm-aware cell-graph mining for bone tissue modeling and classification. *Data mining and knowledge discovery*, 20(3): 416–438, 2010. doi: 10.1007/s10618-009-0153-2.

C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. ISBN 0387310738.

J. M. Bland and D. G. Altman. The logrank test. *BMJ*, 328(7447):1073, 2004. doi: 10.1136/bmj.328.7447.1073.

H. Bolouri and E. H. Davidson. The gene regulatory network basis of the "community effect," and analysis of a sea urchin embryo example. *Dev. Biol.*, 340(2):170–8, 2010. doi: 10.1016/j.ydbio.2009.06.007.

I. Borg and P. Groenen. *Modern multidimensional scaling: theory and applications*. Springer series in statistics. Springer, 1997. ISBN 9780387948454.

Y. Boykov and M.-P. Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in n-d images. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 1, page 105, 2001. doi: 10.1109/ICCV.2001.937505.

A. H. Brand and N. Perrimon. Targeted gene expression as a means of altering cell fates and generating dominant phenotypes. *Development*, 118(2):401–15, 1993.

G. Braz Junior, A. Cardoso de Paiva, A. Corrêa Silva, and A. Cesar Muniz de Oliveira. Classification of breast tissues using moran's index and geary's coefficient as texture signatures and svm. *Comput. Biol. Med.*, 39(12):1063–72, 2009. doi: 10.1016/j.compbiomed.2009.08.009.

L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123, 1996. doi: 10.1023/A:1018054314350.

L. Breiman. Random forests. *Machine Learning*, 45(1):5, 2001. doi: 10.1023/A:1010933404324.

L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Chapman & Hall, New York, 1984. ISBN 0-412-04841-8.

A. Brock, H. Chang, and S. Huang. Non-genetic heterogeneity–a mutation-independent driving force for the somatic evolution of tumours. *Nat. Rev. Genet.*, 10(5):336–42, 2009. doi: 10.1038/nrg2556.

S. Brooks, E. Locke, and H. Soule. Estrogen receptor in a human cell line (mcf 7) from breast carcinoma. *Journal of Biological Chemistry*, 248(17):6251–6253, 1973.

D. C. Brown and K. C. Gatter. Ki67 protein: the immaculate deception? *Histopathology*, 40(1):2, 2002. doi: 10.1046/j.1365-2559.2002.01343.x.

G. J. Buttrick, L. M. A. Beaumont, J. Leitch, C. Yau, J. R. Hughes, and J. G. Wakefield. Akt regulates centrosome migration and spindle orientation in the early drosophila melanogaster embryo. *J. Cell Biol.*, 180(3):537–48, 2008. doi: 10.1083/jcb.200705085.

P. D. Caie, R. E. Walls, A. Ingleston-Orme, S. Daya, T. Houslay, R. Eagle, M. E. Roberts, and N. O. Carragher. High-content phenotypic profiling of drug response signatures across distinct cancer cells. *Mol. Cancer Ther.*, 9(6):1913–26, 2010. doi: 10.1158/1535-7163.MCT-09-1148.

R. L. Camp, G. G. Chung, and D. L. Rimm. Automated subcellular localization and quantification of protein expression in tissue microarrays. *Nat. Med.*, 8(11):1323–7, 2002. doi: 10.1038/nm791.

R. L. Camp, M. Dolled-Filhart, and D. L. Rimm. X-tile: a new bio-informatics tool for biomarker assessment and outcome-based cut-point optimization. *Clin. Cancer Res.*, 10 (21):7252–9, 2004. doi: 10.1158/1078-0432.CCR-04-0713.

M. Carmena. Cytokinesis: the final stop for the chromosomal passengers. *Biochem. Soc. Trans.*, 36(Pt 3):367–70, 2008. doi: 10.1042/BST0360367.

A. E. Carpenter, T. R. Jones, M. R. Lamprecht, C. Clarke, I. H. Kang, O. Friman, D. A. Guertin, J. H. Chang, R. A. Lindquist, J. Moffat, P. Golland, and D. M. Sabatini. Cellprofiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol.*, 7(10):R100, 2006. doi: 10.1186/gb-2006-7-10-r100.

V. Caselles, R. Kimmel, and G. Sapiro. Geodesic active contours. *International Journal of Computer Vision*, 22(1):61, 1997. doi: 10.1023/A:1007979827043.

J. M. Cesario, J. K. Jang, B. Redding, N. Shah, T. Rahman, and K. S. McKim. Kinesin 6 family member subito participates in mitotic spindle assembly and interacts with mitotic regulators. *J. Cell. Sci.*, 119(Pt 22):4770–80, 2006. doi: 10.1242/jcs.03235.

T. F. Chan and L. A. Vese. Active contours without edges. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, 10(2):266–77, 2001. doi: 10.1109/83.902291.

W. Chen, M. Reiss, and D. Foran. A prototype for unsupervised analysis of tissue microarrays for cancer research and diagnostics. *IEEE Transactions on Information Technology in Biomedicine*, 8(2):89, 2004. doi: 10.1109/TITB.2004.828891.

X. Chen, X. Zhou, and S. T. C. Wong. Automated segmentation, classification, and tracking of cancer cell nuclei in time-lapse microscopy. *IEEE transactions on bio-medical engineering*, 53(4):762–6, 2006. doi: 10.1109/TBME.2006.870201.

M.-M. Cheng and G.-X. Zhang. Connectedness of random walk segmentation. *IEEE Trans Pattern Anal Mach Intell*, 33(1):200–2, 2011. doi: 10.1109/TPAMI.2010.138.

F. Chibon, P. Lagarde, S. Salas, G. Pérot, V. Brouste, F. Tirode, C. Lucchesi, A. de Reynies, A. Kauffmann, B. Bui, P. Terrier, S. Bonvalot, A. Le Cesne, D. Vince-Ranchère, J.-Y. Blay, F. Collin, L. Guillou, A. Leroux, J.-M. Coindre, and A. Aurias. Validated prediction of clinical outcome in sarcomas and multiple types of cancer on the basis of a gene expression signature related to genome complexity. *Nat. Med.*, 16(7): 781–7, 2010. doi: 10.1038/nm.2174.

I. Clark. *Practical geostatistics / [by] Isobel Clark*. Elsevier Applied Science, London :, 1979. ISBN 085334843.

T. G. Clark, M. J. Bradburn, S. B. Love, and D. G. Altman. Survival analysis part i: basic concepts and first analyses. *Br. J. Cancer*, 89(2):232–8, 2003. doi: 10.1038/sj.bjc.6601118.

D. F. Clayton, A. L. Harrelson, and J. E. Darnell. Dependence of liver-specific transcription on tissue organization. *Mol. Cell. Biol.*, 5(10):2623–32, 1985.

A. A. Cohen, N. Geva-Zatorsky, E. Eden, M. Frenkel-Morgenstern, I. Issaeva, A. Sigal, R. Milo, C. Cohen-Saidon, Y. Liron, Z. Kam, L. Cohen, T. Danon, N. Perzov, and U. Alon. Dynamic proteomics of individual cancer cells in response to a drug. *Science*, 322(5907):1511–6, 2008. doi: 10.1126/science.1160165.

N. Colombié, C. F. Cullen, A. L. Brittle, J. K. Jang, W. C. Earnshaw, M. Carmena, K. McKim, and H. Ohkura. Dual roles of incenp crucial to the assembly of the acentrosomal metaphase spindle in female meiosis. *Development*, 135(19):3239–46, 2008. doi: 10.1242/dev.022624.

J.-A. Conchello and J. W. Lichtman. Optical sectioning microscopy. *Nat. Methods*, 2(12): 920–31, 2005. doi: 10.1038/nmeth815.

C. Conrad and D. W. Gerlich. Automated microscopy for high-content rnai screening. *J. Cell Biol.*, 188(4):453–61, 2010. doi: 10.1083/jcb.200910105.

C. Couprie, L. Grady, L. Najman, and H. Talbot. Power watersheds: A unifying graph-based optimization framework. *IEEE Trans Pattern Anal Mach Intell*, 2010. doi: 10.1109/TPAMI.2010.200.

D. R. Cox. Regression models and life tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972. doi: 10.2307/2985181.

M. A. Crosby, J. L. Goodman, V. B. Strelets, P. Zhang, and W. M. Gelbart. Flybase: genomes by the dozen. *Nucleic Acids Res.*, 35(Database issue):D486–91, 2007. doi: 10.1093/nar/gkl827.

E. C. da Silva, A. C. Silva, A. C. Paiva, and R. A. Nunes. Diagnosis of lung nodule using moran's index and geary's coefficient in computerized tomography images. *Pattern Analysis and Applications*, 11(1):89, 2008. doi: 10.1007/s10044-007-0081-y.

C. Decaestecker, X. M. Lopez, N. D'Haene, I. Roland, S. Guendouz, C. Duponchelle, A. Berton, O. Debeir, and I. Salmon. Requirements for the valid quantification of immunostains on tissue microarray materials using image analysis. *Proteomics*, 9(19): 4478–94, 2009. doi: 10.1002/pmic.200800936.

M. Delacour-Larose, A. Molla, D. A. Skoufias, R. L. Margolis, and S. Dimitrov. Distinct dynamics of aurora b and survivin during mitosis. *Cell Cycle*, 3(11):1418–26, 2004. doi: 10.4161/cc.3.11.1203.

A. Denley, L. J. Cosgrove, G. W. Booker, J. C. Wallace, and B. E. Forbes. Molecular interactions of the igf system. *Cytokine Growth Factor Rev.*, 16(4-5):421–39, 2005. doi: 10.1016/j.cytogfr.2005.04.004.

R. Díaz-Uriarte and S. Alvarez de Andrés. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7:3, 2006. doi: 10.1186/1471-2105-7-3.

L. R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26 (3):297, 1945. doi: 10.2307/1932409.

S. Doyle, S. Agner, A. Madabhushi, M. Feldman, and J. Tomaszewski. Automated grading of breast cancer histopathology using spectral clustering with textural and architectural image features. In *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, page 496, 2008. doi: 10.1109/ISBI.2008.4541041.

C.-J. Du, M. Marcello, D. G. Spiller, M. R. H. White, and T. Bretschneider. Interactive segmentation of clustered cells via geodesic commute distance and constrained density weighted nyström method. *Cytometry A*, 77(12):1137–47, 2010. doi: 10.1002/cyto.a.20993.

T. Duncan. *Investigating the function of Drosophila MAPs Msd1 and dTD-60 in mitotic spindle assembly.* PhD thesis, University of Oxford, 2011. URL `http://ora.ox.ac.uk/objects/uuid:3ed4021f-2ccc-4821-b7c5-40b06d5639b7`.

B. Efron, S. for Industrial, and A. Mathematics. *The jackknife, the bootstrap, and other resampling plans / Bradley Efron.* Society for Industrial and Applied Mathematics, Philadelphia, Pa. :, 1982. ISBN 0898711797.

U. S. Eggert, A. A. Kiger, C. Richter, Z. E. Perlman, N. Perrimon, T. J. Mitchison, and C. M. Field. Parallel chemical genetic and genome-wide rnai screens identify cytokinesis inhibitors and targets. *PLoS Biol.*, 2(12):e379, 2004. doi: 10.1371/journal.pbio.0020379.

S. L. Faitar, J. T. S. Dabbeekeh, T. A. Ranalli, and J. K. Cowell. Evi5 is a novel centrosomal protein that binds to alpha- and gamma-tubulin. *Genomics*, 86(5):594–605, 2005. doi: 10.1016/j.ygeno.2005.06.002.

R. Fernandez-Gonzalez, M. Barcellos-Hoff, and C. Ortiz-de Solorzano. A tool for the quantitative spatial analysis of complex cellular systems. *IEEE Transactions on Image Processing*, 14(9):1300, 2005. doi: 10.1109/TIP.2005.852466.

T. Fortmann, Y. Bar-Shalom, and M. Scheffe. Sonar tracking of multiple targets using joint probabilistic data association. *IEEE Journal of Oceanic Engineering*, 1983. doi: 10.1109/JOE.1983.1145560.

A. F. Frangi, W. J. Niessen, K. L. Vincken, and M. A. Viergever. Multiscale vessel enhancement filtering. *Medical Image Computing and Computer-Assisted Intervention - MICCAI 1998*, 1496:130, 1998. doi: 10.1007/BFb0056195.

W. Freeman. The design and use of steerable filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(9):891, 1991. doi: 10.1109/34.93808.

J. H. Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1): 1, 1991. doi: 10.1214/aos/1176347963.

J. H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189, 2001. doi: 10.1214/aos/1013203451.

B. G. Fuller, M. A. Lampson, E. A. Foley, S. Rosasco-Nitcher, K. V. Le, P. Tobelmann, D. L. Brautigan, P. T. Stukenberg, and T. M. Kapoor. Midzone activation of aurora b in anaphase produces an intracellular phosphorylation gradient. *Nature*, 453(7198): 1132–6, 2008. doi: 10.1038/nature06923.

M. K. G. Fernàndez and J. P. Zrÿd. A new plant cell image segmentation algorithm. *Lecture Notes in Computer Science*, 974/1995:229–234, 1995. doi: 10.1007/3-540-60298-4“¨263.

S. Gao, M. G. Giansanti, G. J. Buttrick, S. Ramasubramanyan, A. Auton, M. Gatti, and J. G. Wakefield. Australin: a chromosomal passenger protein required specifically for drosophila melanogaster male meiosis. *J. Cell Biol.*, 180(3):521–35, 2008. doi: 10.1083/jcb.200708072.

R. Gassmann, A. Carvalho, A. J. Henzing, S. Ruchaud, D. F. Hudson, R. Honda, E. A. Nigg, D. L. Gerloff, and W. C. Earnshaw. Borealin: a novel chromosomal passenger required for stability of the bipolar mitotic spindle. *J. Cell Biol.*, 166(2):179–91, 2004. doi: /10.1083/jcb.200404001.

R. C. Geary. The contiguity ratio and statistical mapping. *The Incorporated Statistician*, 5(3):115, 1954. doi: 10.2307/2986645.

A. Getis and J. K. Ord. The analysis of spatial association by use of distance statistics. *Geographical Analysis*, 24(3):189, 1992. doi: 10.1111/j.1538-4632.1992.tb00261.x.

R. Giet and D. M. Glover. Drosophila aurora b kinase is required for histone h3 phosphorylation and condensin recruitment during chromosome condensation and to organize the central spindle during cytokinesis. *J. Cell Biol.*, 152(4):669–82, 2001. doi: 10.1083/jcb.152.4.669.

E. Glory and R. F. Murphy. Automated subcellular location determination and high-throughput microscopy. *Dev. Cell*, 12(1):7–16, 2007. doi: 10.1016/j.devcel.2006.12.007.

M. Glotzer. The 3ms of central spindle assembly: microtubules, motors and maps. *Nat. Rev. Mol. Cell Biol.*, 10(1):9–20, 2009. doi: 10.1038/nrm2609.

M. F. Goodchild. *Spatial autocorrelation*, volume 47 of *Concepts and techniques in modern geography, 0306-6142*. Geo Books, Norwich, 1988. ISBN 0860942236.

G. Goshima and J. M. Scholey. Control of mitotic spindle length. *Annu. Rev. Cell Dev. Biol.*, 26:21–57, 2010. doi: 10.1146/annurev-cellbio-100109-104006.

H. Goto, Y. Yasui, E. A. Nigg, and M. Inagaki. Aurora-b phosphorylates histone h3 at serine28 with regard to the mitotic chromosome condensation. *Genes to Cells*, 7(1):11, 2002. doi: 10.1046/j.1356-9597.2001.00498.x.

H. Goto, T. Kiyono, Y. Tomono, A. Kawajiri, T. Urano, K. Furukawa, E. A. Nigg, and M. Inagaki. Complex formation of plk1 and incenp required for metaphase-anaphase transition. *Nat. Cell Biol.*, 8(2):180–7, 2006. doi: 10.1038/ncb1350.

L. Grady. Random walks for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 28(11):1768–83, 2006. doi: 10.1109/TPAMI.2006.233.

L. Grady and M.-P. Jolly. Weights and topology: A study of the effects of graph construction on 3d image segmentation. *Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention*, 5241:153, 2008. doi: 10.1007/978-3-540-85988-8˙19.

M. Greenwood. The natural duration of cancer. *Reports on Public Health and Medical Subjects*, 33:1–26, 1926.

U. Gruneberg, R. Neef, R. Honda, E. A. Nigg, and F. A. Barr. Relocation of aurora b from centromeres to the central spindle at the metaphase to anaphase transition requires mklp2. *J. Cell Biol.*, 166(2):167–72, 2004. doi: 10.1083/jcb.200403084.

C. Gunduz, B. Yener, and S. H. Gultekin. The cell graphs of cancer. *Bioinformatics*, 20 Suppl 1:i145–51, 2004. doi: 10.1093/bioinformatics/bth933.

M. N. Gurcan, L. Boucheron, A. Can, A. Madabhushi, N. Rajpoot, and B. Yener. Histopathological image analysis: A review. *IEEE reviews in biomedical engineering*, 2:147–171, 2009. doi: 10.1109/RBME.2009.2034865.

J. B. Gurdon. A community effect in animal development. *Nature*, 336(6201):772–4, 1988. doi: 10.1038/336772a0.

N. A. Hamilton and R. D. Teasdale. Visualizing and clustering high throughput sub-cellular localization imaging. *BMC Bioinformatics*, 9:81, 2008. doi: 10.1186/1471-2105-9-81.

X. Han, C. Xu, and J. Prince. Topology preserving level set method for geometric deformable models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25 (6):755, 2003. doi: 10.1109/TPAMI.2003.1201824.

K. K. Hanson, A. C. Kelley, and M. Bienz. Loss of drosophila borealin causes polyploidy, delayed apoptosis and abnormal tissue development. *Development*, 132(21):4777–87, 2005. doi: 10.1242/dev.02057.

F. E. Harre, K. L. Lee, and B. G. Pollock. Regression models in clinical studies: Determining relationships between predictors and response. *JNCI Journal of the National Cancer Institute*, 80(15):1198, 1988. doi: 10.1093/jnci/80.15.1198.

F. E. Harrell Jr., R. M. Califf, and D. B. Pryor. Evaluating the yield of medical tests. *Journal of the American Medical Association*, 247(18):2543–2546, 1982. doi: 10.1001/jama.1982.03320430047030.

J. L. Hartley. Dna cloning using in vitro site-specific recombination. *Genome Research*, 10 (11):1788, 2000. doi: 10.1101/gr.143000.

D. Heckerman. A tutorial on learning with bayesian networks. Technical Report MSR-TR-95-06, Microsoft Research, One Microsoft Way, Redmond, WA 98052, 1996.

R. Honda, R. Körner, and E. A. Nigg. Exploring the functional interactions between aurora b, incenp, and survivin in mitosis. *Mol. Biol. Cell*, 14(8):3325–41, 2003. doi: 10.1091/mbc.E02-11-0769.

R. Hooke. *Micrographia: or Some physiological descriptions of minute bodies made by magnifying glasses With observations and inquiries thereupon.* John Martyn, printer to the Royal Society, London, 1665.

D. Huttenlocher, G. Klanderman, and W. Rucklidge. Comparing images using the hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(9): 850, 1993. doi: 10.1109/34.232073.

I. H. G. S. C. IHGSC. Initial sequencing and analysis of the human genome. *Nature*, 409 (6822):860–921, 2001. doi: 10.1038/35057062.

M. Isard. Condensation-conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29(1):5, 1998. doi: 10.1023/A:1008078328650.

H. Ishwaran. High-dimensional variable selection for survival data. *Journal of the American Statistical Association*, 105(489):205, 2010. doi: 10.1198/jasa.2009.tm08622.

H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3):841, 2008. doi: 10.1214/08-AOAS169.

S. Ito, T. Hyodo, H. Hasegawa, H. Yuan, M. Hamaguchi, and T. Senga. Pi3k/akt signaling is involved in the disruption of gap junctional communication caused by v-src and tnf-α. *Biochem. Biophys. Res. Commun.*, 400(2):230–5, 2010. doi: 10.1016/j.bbrc.2010.08.045.

P. Jaccard. Nouvelles recherches sur la distribution florale. *Bulletin de la Sociète Vaudense des Sciences Naturelles*, 44:223–270, 1908.

M. Jacob and M. Unser. Design of steerable filters for feature detection using canny-like criteria. *IEEE transactions on pattern analysis and machine intelligence*, 26(8):1007–19, 2004. doi: 10.1109/TPAMI.2004.44.

A. Jaiantilal. Classification and regression by randomforest-matlab. Available at `http://code.google.com/p/randomforest-matlab`, 2009.

A. A. Jeyaprakash, U. R. Klein, D. Lindner, J. Ebert, E. A. Nigg, and E. Conti. Structure of a survivin-borealin-incenp core complex reveals how chromosomal passengers travel together. *Cell*, 131(2):271–85, 2007. doi: 10.1016/j.cell.2007.07.045.

G. Jones, D. Jones, L. Zhou, H. Steller, and Y. Chu. Deterin, a new inhibitor of apoptosis from drosophila melanogaster. *J. Biol. Chem.*, 275(29):22157–65, 2000. doi: 10.1074/jbc.M000369200.

T. R. Jones, A. Carpenter, and P. Golland. Voronoi-based segmentation of cells on image manifolds. *Computer Vision for Biomedical Image Applications*, 3765:535, 2005. doi: 10.1007/11569541˙54.

T. R. Jones, I. H. Kang, D. B. Wheeler, R. A. Lindquist, A. Papallo, D. M. Sabatini, P. Golland, and A. E. Carpenter. Cellprofiler analyst: data exploration and analysis software for complex image-based screens. *BMC Bioinformatics*, 9:482, 2008. doi: 10.1186/1471-2105-9-482.

R. E. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME- Journal of Basic Engineering*, 82(Series D):35–45, 2002.

E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457, 1958. doi: 10.2307/2281868.

Y. V. Karpievitch, E. G. Hill, A. P. Leclerc, A. R. Dabney, and J. S. Almeida. An introspective comparison of random forest-based classifiers for the analysis of cluster-correlated data by way of rf++. *PLoS ONE*, 4(9):e7087, 2009. doi: 10.1371/journal.pone.0007087.

M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):321, 1988. doi: 10.1007/BF00133570.

L. Kaufman and P. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. Wiley series in probability and mathematical statistics. Applied probability and statistics. Wiley, 2005. ISBN 9780471735786.

L. N. Kazembe and P. M. G. Mpeketula. Quantifying spatial disparities in neonatal mortality using a structured additive regression model. *PLoS ONE*, 5(6):e11180, 2010. doi: 10.1371/journal.pone.0011180.

A. B. Keeton, K. D. Bortoff, W. L. Bennett, J. L. Franklin, D. Y. Venable, and J. L. Messina. Insulin-regulated expression of egr-1 and krox20: dependence on erk1/2 and interaction with p38 and pi3-kinase pathways. *Endocrinology*, 144(12):5402–10, 2003. doi: 10.1210/en.2003-0592.

A. E. Kelly and H. Funabiki. Correcting aberrant kinetochore microtubule attachments: an aurora b-centric view. *Curr. Opin. Cell Biol.*, 21(1):51–8, 2009. doi: 10.1016/j.ceb.2009.01.004.

A. E. Kelly, C. Ghenoiu, J. Z. Xue, C. Zierhut, H. Kimura, and H. Funabiki. Survivin reads phosphorylated histone h3 threonine 3 to activate the mitotic kinase aurora b. *Science*, 330(6001):235–9, 2010. doi: 10.1126/science.1189505.

Z. Khan, T. Balch, and F. Dellaert. Mcmc-based particle filtering for tracking a variable number of interacting targets. *IEEE transactions on pattern analysis and machine intelligence*, 27(11):1805–19, 2005. doi: 10.1109/TPAMI.2005.223.

H. C. Kim and A. Genovesio. Neuron branch detection and description using random walk. *Conf Proc IEEE Eng Med Biol Soc*, 2009:1020–3, 2009. doi: 10.1109/IEMBS.2009.5334083.

R. Kittler, L. Pelletier, A.-K. Heninger, M. Slabicki, M. Theis, L. Miroslaw, I. Poser, S. Lawo, H. Grabner, K. Kozak, J. Wagner, V. Surendranath, C. Richter, W. Bowen, A. L. Jackson, B. Habermann, A. A. Hyman, and F. Buchholz. Genome-scale rnai profiling of cell division in human tissue culture cells. *Nat. Cell Biol.*, 9(12):1401–12, 2007. doi: 10.1038/ncb1659.

U. R. Klein, E. A. Nigg, and U. Gruneberg. Centromere targeting of the chromosomal passenger complex requires a ternary subcomplex of borealin, survivin, and the n-terminal domain of incenp. *Mol. Biol. Cell*, 17(6):2547–58, 2006. doi: 10.1091/mbc.E05-12-1133.

R. Kohavi. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273, 1997. doi: 10.1016/S0004-3702(97)00043-X.

J. Kononen, L. Bubendorf, A. Kallionimeni, M. Bärlund, P. Schraml, S. Leighton, J. Torhorst, M. J. Mihatsch, G. Sauter, and O.-P. Kallionimeni. Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nature Medicine*, 4(7):844, 1998. doi: 10.1038/nm0798-844.

S. Kumar. A rule-based approach for robust clump splitting. *Pattern Recognition*, 39(6): 1088, 2006. doi: 10.1016/j.patcog.2005.11.014.

S. Lankton and A. Tannenbaum. Localizing region-based active contours. *IEEE Trans Image Process*, 17(11):2029–39, 2008. doi: 10.1109/TIP.2008.2004611.

M. A. Larkin, G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson, and D. G. Higgins. Clustal w and clustal x version 2.0. *Bioinformatics*, 23(21):2947–8, 2007. doi: 10.1093/bioinformatics/btm404.

S. M. A. Lens, J. A. Rodriguez, G. Vader, S. W. Span, G. Giaccone, and R. H. Medema. Uncoupling the central spindle-associated function of the chromosomal passenger complex from its role at centromeres. *Mol. Biol. Cell*, 17(4):1897–909, 2006. doi: 10.1091/mbc.E05-08-0727.

R. M. Levenson and J. R. Mansfield. Multispectral imaging in biology and medicine: slices of life. *Cytometry A*, 69(8):748–58, 2006. doi: 10.1002/cyto.a.20319.

M. Leventon, W. Grimson, and O. Faugeras. Statistical shape influence in geodesic active contours. In *5th IEEE EMBS International Summer School on Biomedical Imaging, 2002.*, pages I: 316–323, 2002. doi: 10.1109/SSBI.2002.1233989.

C. Li. Level set evolution without re-initialization: A new variational formulation. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 05)*, volume 1, page 430, 2005. doi: 10.1109/CVPR.2005.213.

G. Li, T. Liu, J. Nie, L. Guo, J. Chen, J. Zhu, W. Xia, A. Mara, S. Holley, and S. T. C. Wong. Segmentation of touching cell nuclei using gradient flow tracking. *Journal of microscopy*, 231(Pt 1):47–58, 2008a. doi: 10.1111/j.1365-2818.2008.02016.x.

K. Li, E. D. Miller, M. Chen, T. Kanade, L. E. Weiss, and P. G. Campbell. Cell population tracking and lineage construction with spatiotemporal context. *Medical image analysis*, 12(5):546–66, 2008b. doi: 10.1016/j.media.2008.06.001.

S. Li, C. Buehnemann, B. Hassan, and J. Noble. Segmentation of cell clumps for quantitative analysis. *Conf Proc IEEE Eng Med Biol Soc*, 2010:4813–6, 2010. doi: 10.1109/IEMBS.2010.5628032.

S. Li, J. Wakefield, and J. A. Noble. Automated segmentation and alignment of mitotic nuclei for kymograph visualisation. In *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, page 622, 2011. doi: 10.1109/ISBI.2011.5872484.

A. Liaw and M. Wiener. Classification and regression by randomforest. *R News*, 2(3): 18–22, 2002. URL `http://www.r-project.org/doc/Rnews/Rnews_2002-3.pdf`.

J. W. Lichtman and J.-A. Conchello. Fluorescence microscopy. *Nat. Methods*, 2(12):910–9, 2005. doi: 10.1038/nmeth817.

J. Lippincott-Schwartz and G. H. Patterson. Development and use of fluorescent protein markers in living cells. *Science*, 300(5616):87–91, 2003. doi: 10.1126/science.1082520.

C. L. Liu, W. Prapong, Y. Natkunam, A. Alizadeh, K. Montgomery, C. B. Gilks, and M. van de Rijn. Software tools for high-throughput analysis and archiving of immunohistochemistry staining data obtained with tissue microarrays. *The American Journal of Pathology*, 161(5):1557, 2002. doi: 10.1016/S0002-9440(10)64434-3.

D. Liu, G. Vader, M. J. M. Vromans, M. A. Lampson, and S. M. A. Lens. Sensing chromosome bi-orientation by spatial separation of aurora b kinase from kinetochore substrates. *Science*, 323(5919):1350–3, 2009. doi: 10.1126/science.1167000.

V. Ljosa and A. Singh. Probabilistic segmentation and analysis of horizontal cells. In *Sixth International Conference on Data Mining (ICDM'06)*, page 980, 2006. doi: 10.1109/ICDM.2006.129.

T. A. Longacre, M. Ennis, L. A. Quenneville, A. L. Bane, I. J. Bleiweiss, B. A. Carter, E. Catelano, M. R. Hendrickson, H. Hibshoosh, L. J. Layfield, L. Memeo, H. Wu, and F. P. O'malley. Interobserver agreement and reproducibility in classification of invasive breast carcinoma: an nci breast cancer family registry study. *Mod. Pathol.*, 19(2):195–207, 2006. doi: 10.1038/modpathol.3800496.

L.-H. Loo, L. F. Wu, and S. J. Altschuler. Image-based multivariate profiling of drug responses from single cells. *Nat. Methods*, 4(5):445–53, 2007. doi: 10.1038/nmeth1032.

L.-H. Loo, H.-J. Lin, R. J. Steininger, Y. Wang, L. F. Wu, and S. J. Altschuler. An approach for extensibly profiling the molecular states of cellular subpopulations. *Nat. Methods*, 6(10):759–65, 2009. doi: 10.1038/nmeth.1375.

A. Lupas, M. Van Dyke, and J. Stock. Predicting coiled coils from protein sequences. *Science*, 252(5009):1162–4, 1991. doi: 10.1126/science.252.5009.1162.

J. MacCormick. Probabilistic exclusion principle for tracking multiple objects. *International Journal of Computer Vision*, 39(1):57–71, 2000. doi: 10.1023/A:1008122218374.

A. M. Mackay. Molecular analysis of the incenps (inner centromere proteins): separate domains are required for association with microtubules during interphase and with the central spindle during anaphase. *The Journal of Cell Biology*, 123(2):373, 1993. doi: 10.1083/jcb.123.2.373.

A. M. Mackay, A. M. Ainsztein, D. M. Eckley, and W. C. Earnshaw. A dominant mutant of inner centromere protein (incenp), a chromosomal protein, disrupts prometaphase congression and cytokinesis. *J. Cell Biol.*, 140(5):991–1002, 1998. doi: 10.1083/jcb.140.5.991.

R. Malladi, J. Sethian, and B. Vemuri. Shape modeling with front propagation: a level set approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(2):158, 1995. doi: 10.1109/34.368173.

N. Malpica. Applying watershed algorithms to the segmentation of clustered nuclei. *Cytometry*, 28(4):289, 1997. doi: 10.1002/(SICI)1097-0320(19970801)28:4¡289::AID-CYTO3¿3.0.CO;2-7.

A. Mansouri, A. Mitiche, and C. Vazquez. Multiregion competition: A level set extension of region competition to multiple region image partitioning. *Computer Vision and Image Understanding*, 101(3):137, 2006. doi: 10.1016/j.cviu.2005.07.008.

T. J. Maresca, A. C. Groen, J. C. Gatlin, R. Ohi, T. J. Mitchison, and E. D. Salmon. Spindle assembly in the absence of a rangtp gradient requires localized cpc activity. *Curr. Biol.*, 19(14):1210–5, 2009. doi: 10.1016/j.cub.2009.05.061.

K. Masood and N. Rajpoot. Texture based classification of hyperspectral colon biopsy samples using clbp. In *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, page 1011, 2009. doi: 10.1109/ISBI.2009.5193226.

A. Matov, M. M. Edvall, G. Yang, and G. Danuser. Optimal-flow minimum-cost correspondence assignment in particle flow tracking. *Computer vision and image understanding : CVIU*, 115(4):531–540, 2011. doi: 10.1016/j.cviu.2011.01.001.

S. A. McKinney, C. S. Murphy, K. L. Hazelwood, M. W. Davidson, and L. L. Looger. A bright and photostable photoconvertible fluorescent protein. *Nat. Methods*, 6(2):131–3, 2009. doi: 10.1038/nmeth.1296.

J. G. McNally, T. Karpova, J. Cooper, and J. A. Conchello. Three-dimensional imaging by deconvolution microscopy. *Methods*, 19(3):373–85, 1999. doi: 10.1006/meth.1999.0873.

E. Meijering. Tracking in molecular bioimaging. *IEEE Signal Processing Magazine*, 23(3): 46, 2006. doi: 10.1109/MSP.2006.1628877.

B. H. Menze, B. M. Kelm, R. Masuch, U. Himmelreich, P. Bachert, W. Petrich, and F. A. Hamprecht. A comparison of random forest and its gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics*, 10:213, 2009. doi: 10.1186/1471-2105-10-213.

C. Mollinari. The mammalian passenger protein td-60 is an rcc1 family member with an essential role in prometaphase to metaphase progression. *Developmental Cell*, 5(2):295, 2003. doi: 10.1016/S1534-5807(03)00205-3.

M. Momiyama, Y. Omori, Y. Ishizaki, Y. Nishikawa, T. Tokairin, J. Ogawa, and K. Enomoto. Connexin26-mediated gap junctional communication reverses the malignant phenotype of mcf-7 breast cancer cells. *Cancer Science*, 94(6):501, 2003. doi: 10.1111/j.1349-7006.2003.tb01473.x.

P. A. P. Moran. Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17, 1950. doi: 10.2307/2332142.

D. Muzzey and A. van Oudenaarden. Quantitative time-lapse fluorescence microscopy in single cells. *Annu. Rev. Cell Dev. Biol.*, 25:301–27, 2009. doi: 10.1146/annurev.cellbio.042308.113408.

Y. Nakajima, A. Cormier, R. G. Tyers, A. Pigula, Y. Peng, D. G. Drubin, and G. Barnes. Ipl1/aurora-dependent phosphorylation of sli15/incenp regulates cpc-spindle interaction to ensure proper microtubule dynamics. *J. Cell Biol.*, 194(1):137–53, 2011. doi: 10.1083/jcb.201009137.

S. Nath, F. Bunyak, and K. Palaniappan. Robust tracking of migrating cells using four-color level set segmentation. *Lecture notes in computer science*, 4179(LNCS):920–932, 2006a. doi: 10.1007/11864349˙84.

S. K. Nath, K. Palaniappan, and F. Bunyak. Cell segmentation using coupled level sets and graph-vertex coloring. *Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention*, 9(Pt 1):101–8, 2006b. doi: 10.1007/11866565˙13.

C. Nelson and C. Chen. Cell-cell signaling by direct contact increases cell proliferation via a pi3k-dependent signal. *FEBS Letters*, 514(2-3):238, 2002. doi: 10.1016/S0014-5793(02)02370-0.

M. Niepel, S. L. Spencer, and P. K. Sorger. Non-genetic cell-to-cell variability and the consequences for pharmacology. *Curr Opin Chem Biol*, 13(5-6):556–61, 2009. doi: 10.1016/j.cbpa.2009.09.015.

J. K. Ord and A. Getis. Local spatial autocorrelation statistics: Distributional issues and an application. *Geographical Analysis*, 27(4):286, 1995. doi: 10.1111/j.1538-4632.1995.tb00912.x.

N. Otsu. Threshold selection method from gray-level histograms. *IEEE Trans Syst Man Cybern*, SMC-9(1):62–66, 1979. doi: 10.1109/TSMC.1979.4310076.

D. Padfield, J. Rittscher, N. Thomas, and B. Roysam. Spatio-temporal cell cycle phase analysis using level sets and fast marching methods. *Medical image analysis*, 13(1): 143–55, 2009. doi: 10.1016/j.media.2008.06.018.

D. Padfield, J. Rittscher, and B. Roysam. Coupled minimum-cost flow cell tracking for high-throughput quantitative analysis. *Med Image Anal*, 2010. doi: 10.1016/j.media.2010.07.006.

H. Peng. Bioimage informatics: a new area of engineering biology. *Bioinformatics*, 24(17): 1827–36, 2008. doi: 10.1093/bioinformatics/btn346.

T. Peng, G. M. C. Bonamy, E. Glory-Afshar, D. R. Rines, S. K. Chanda, and R. F. Murphy. Determining the distribution of probes between different subcellular locations through automated unmixing of subcellular patterns. *Proc. Natl. Acad. Sci. U.S.A.*, 107 (7):2944–9, 2010. doi: 10.1073/pnas.0912090107.

R. Pepperkok and J. Ellenberg. High-throughput fluorescence microscopy for systems biology. *Nat. Rev. Mol. Cell Biol.*, 7(9):690–6, 2006. doi: 10.1038/nrm1979.

A. J. Pereira and H. Maiato. Improved kymography tools and its applications to mitosis. *Methods*, 51(2):214–9, 2010. doi: 10.1016/j.ymeth.2010.01.016.

Z. E. Perlman, M. D. Slack, Y. Feng, T. J. Mitchison, L. F. Wu, and S. J. Altschuler. Multidimensional drug profiling by automated microscopy. *Science*, 306(5699):1194–8, 2004. doi: 10.1126/science.1100709.

M. Petronczki, P. Lénárt, and J.-M. Peters. Polo on the rise-from mitotic entry to cytokinesis with plk1. *Dev. Cell*, 14(5):646–59, 2008. doi: 10.1016/j.devcel.2008.04.014.

J. P. Peyrat and J. Bonneterre. Type 1 igf receptor in human breast diseases. *Breast Cancer Research and Treatment*, 22(1):59, 1992. doi: 10.1007/BF01833334.

M. N. Pollak, E. S. Schernhammer, and S. E. Hankinson. Insulin-like growth factors and neoplasia. *Nat. Rev. Cancer*, 4(7):505–18, 2004. doi: 10.1038/nrc1387.

A. Raj and A. van Oudenaarden. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell*, 135(2):216–26, 2008. doi: 10.1016/j.cell.2008.09.050.

P. Rämö, R. Sacher, B. Snijder, B. Begemann, and L. Pelkmans. Cellclassifier: supervised learning of cellular phenotypes. *Bioinformatics*, 25(22):3028–30, 2009. doi: 10.1093/bioinformatics/btp524.

D. Reid. An algorithm for tracking multiple targets. *IEEE Transactions on Automatic Control*, 24(6):843–854, 1979. doi: 10.1109/TAC.1979.1102177.

Ren and Malik. Learning a classification model for segmentation. In *Proceedings Ninth IEEE International Conference on Computer Vision*, page 10, 2003. doi: 10.1109/ICCV.2003.1238308.

J. Rittscher. Characterization of biological processes through automated image analysis. *Annu Rev Biomed Eng*, 12:315–44, 2010. doi: 10.1146/annurev-bioeng-070909-105235.

P. Rørth. Gal4 in the drosophila female germline. *Mechanisms of Development*, 78(1-2): 113, 1998. doi: 10.1016/S0925-4773(98)00157-9.

S. E. Rosasco-Nitcher, W. Lan, S. Khorasanizadeh, and P. T. Stukenberg. Centromeric aurora-b activation requires td-60, microtubules, and substrate priming phosphorylation. *Science*, 319(5862):469–72, 2008. doi: 10.1126/science.1148980.

P. Royston. The use of fractional polynomials to model continuous risk variables in epidemiology. *International Journal of Epidemiology*, 28(5):964, 1999. doi: 10.1093/ije/28.5.964.

S. Ruchaud, M. Carmena, and W. C. Earnshaw. Chromosomal passengers: conducting cell division. *Nat. Rev. Mol. Cell Biol.*, 8(10):798–812, 2007. doi: 10.1038/nrm2257.

D. Rueckert, L. I. Sonoda, C. Hayes, D. L. Hill, M. O. Leach, and D. J. Hawkes. Nonrigid registration using free-form deformations: application to breast mr images. *IEEE Trans Med Imaging*, 18(8):712–21, 1999. doi: 10.1109/42.796284.

Y. Sato, J. Chen, R. A. Zoroofi, N. Harada, S. Tamura, and T. Shiga. Automatic extraction and measurement of leukocyte motion in microvessels using spatiotemporal image analysis. *IEEE Trans Biomed Eng*, 44(4):225–36, 1997. doi: 10.1109/10.563292.

I. F. Sbalzarini and P. Koumoutsakos. Feature point tracking and trajectory analysis for video imaging in cell biology. *J. Struct. Biol.*, 151(2):182–95, 2005. doi: 10.1016/j.jsb.2005.06.002.

O. Schmitt. On the decomposition of cell clusters. *Journal of Mathematical Imaging and Vision*, 33(1):85, 2009a. doi: 10.1007/s10851-008-0110-5.

O. Schmitt. Morphological multiscale decomposition of connected regions with emphasis on cell clusters. *Computer Vision and Image Understanding*, 113(2):188, 2009b. doi: 10.1016/j.cviu.2008.08.011.

D. B. Seligson, S. Horvath, T. Shi, H. Yu, S. Tze, M. Grunstein, and S. K. Kurdistani. Global histone modification patterns predict risk of prostate cancer recurrence. *Nature*, 435(7046):1262–6, 2005. doi: 10.1038/nature03672.

F. Sessa, M. Mapelli, C. Ciferri, C. Tarricone, L. B. Areces, T. R. Schneider, P. T. Stukenberg, and A. Musacchio. Mechanism of aurora b activation by incenp and inhibition by hesperadin. *Mol. Cell*, 18(3):379–91, 2005. doi: 10.1016/j.molcel.2005.03.031.

J. A. Sethian. Level set methods: An act of violence - evolving interfaces in geometry, fluid mechanics, computer vision and materials sciences. *American Scientific*, 1997.

J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888, 2000. doi: 10.1109/34.868688.

T. Shi and S. Horvath. Unsupervised learning with random forest predictors. *Journal of Computational and Graphical Statistics*, 15(1):118, 2006. doi: 10.1198/106186006X94072.

Y. Shi. Real-time tracking using level sets. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 05)*, volume 2, page 34, 2005. doi: 10.1109/CVPR.2005.294.

Y. Shi and W. C. Karl. A real-time algorithm for the approximation of level-set-based curve evolution. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, 17(5):645–56, 2008. doi: 10.1109/TIP.2008.920737.

D. Shutin and O. Zlobinskaya. Application of information-theoretic measures to quantitative analysis of immunofluorescent microscope imaging. *Comput Methods Programs Biomed*, 97(2):114–29, 2010. doi: 10.1016/j.cmpb.2009.05.009.

A. C. Silva, P. C. P. Carvalho, and M. Gattass. Analysis of spatial variability using geostatistical functions for diagnosis of lung nodule in computerized tomography images. *Pattern Analysis and Applications*, 7(3):227, 2004. doi: 10.1007/s10044-004-0219-0.

D. K. Singh, C.-J. Ku, C. Wichaidit, R. J. Steininger, L. F. Wu, and S. J. Altschuler. Patterns of basal signaling heterogeneity can distinguish cellular populations with different drug sensitivities. *Mol. Syst. Biol.*, 6:369, 2010. doi: 10.1038/msb.2010.22.

M. D. Slack, E. D. Martinez, L. F. Wu, and S. J. Altschuler. Characterizing heterogeneous cellular responses to perturbations. *Proc. Natl. Acad. Sci. U.S.A.*, 105(49):19306–11, 2008. doi: 10.1073/pnas.0807038105.

I. Smal, K. Draegestein, N. Galjart, W. Niessen, and E. Meijering. Particle filtering for multiple object tracking in dynamic fluorescence microscopy images: application to microtubule growth analysis. *IEEE transactions on medical imaging*, 27(6):789–804, 2008. doi: 10.1109/TMI.2008.916964.

M. A. Smith, N. L. Seibel, S. F. Altekruse, L. A. G. Ries, D. L. Melbert, M. O'Leary, F. O. Smith, and G. H. Reaman. Outcomes for children and adolescents with cancer: challenges for the twenty-first century. *J. Clin. Oncol.*, 28(15):2625–34, 2010. doi: 10.1200/JCO.2009.27.0421.

B. L. Sprague and J. G. McNally. Frap analysis of binding: proper and fitting. *Trends Cell Biol.*, 15(2):84–91, 2005. doi: 10.1016/j.tcb.2004.12.001.

A. Statnikov, L. Wang, and C. F. Aliferis. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Bioinformatics*, 9:319, 2008. doi: 10.1186/1471-2105-9-319.

C. Steger. An unbiased detector of curvilinear structures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(2):113, 1998. doi: 10.1109/34.659930.

C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn. Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics*, 8:25, 2007. doi: 10.1186/1471-2105-8-25.

C. Strobl, J. Malley, and G. Tutz. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychol Methods*, 14(4):323–48, 2009. doi: 10.1037/a0016973.

J. Su, P. J. Zapata, C.-C. Chen, and J. C. Meredith. Local cell metrics: a novel method for analysis of cell-cell interactions. *BMC Bioinformatics*, 10:350, 2009. doi: 10.1186/1471-2105-10-350.

V. T. Ta, O. Lezoray, and A. Elmoataz. Graph based semi and unsupervised classification and segmentation of microscopic images. In *2007 IEEE International Symposium on Signal Processing and Information Technology*, page 1160, 2007. doi: 10.1109/ISSPIT.2007.4458172.

V.-T. Ta, O. Lézoray, A. Elmoataz, and S. Schüpp. Graph-based tools for microscopic cellular image segmentation. *Pattern Recognition*, 42(6):1113, 2009. doi: 10.1016/j.patcog.2008.10.029.

E. C. Toomey, J. D. Schiffman, and S. L. Lessnick. Recent advances in the molecular pathogenesis of ewing's sarcoma. *Oncogene*, 29(32):4504–16, 2010. doi: 10.1038/onc.2010.205.

Z. Tothova and D. G. Gilliland. Foxo transcription factors and stem cell homeostasis: insights from the hematopoietic system. *Cell Stem Cell*, 1(2):140–52, 2007. doi: 10.1016/j.stem.2007.07.017.

B. S. Tseng, L. Tan, T. M. Kapoor, and H. Funabiki. Dual detection of chromosomes and microtubules by the chromosomal passenger complex drives spindle assembly. *Dev. Cell*, 18(6):903–12, 2010. doi: 10.1016/j.devcel.2010.05.018.

R. Y. Tsien. The green fluorescent protein. *Annu. Rev. Biochem.*, 67:509–44, 1998. doi: 10.1146/annurev.biochem.67.1.509.

D. J. Unwin. Gis, spatial analysis and spatial statistics. *Progress in Human Geography*, 20 (4):540, 1996. doi: 10.1177/030913259602000408.

G. Vader, C. W. A. Cruijsen, T. van Harn, M. J. M. Vromans, R. H. Medema, and S. M. A. Lens. The chromosomal passenger complex controls spindle checkpoint function independent from its role in correcting microtubule kinetochore interactions. *Mol. Biol. Cell*, 18(11):4553–64, 2007. doi: 10.1091/mbc.E07-04-0328.

M. J. van de Vijver, Y. D. He, L. J. van't Veer, H. Dai, A. A. M. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E. T. Rutgers, S. H. Friend, and R. Bernards. A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.*, 347(25):1999–2009, 2002. doi: 10.1056/NEJMoa021967.

V. N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995. ISBN 0-387-94559-8.

M. Veronika, J. Evans, P. Matsudaira, R. Welsch, and J. Rajapakse. Sub-population analysis based on temporal features of high content images. *BMC Bioinformatics*, 10 Suppl 15:S4, 2009. doi: 10.1186/1471-2105-10-S15-S4.

L. A. Vese. A multiphase level set framework for image segmentation. *International Journal of Computer Vision*, 50(3):271, 2002. doi: 10.1023/A:1020874308076.

C. Wählby, J. Lindblad, M. Vondrus, E. Bengtsson, and L. Björkesten. Algorithms for cytoplasm segmentation of fluorescence labelled cells. *Analytical cellular pathology : the journal of the European Society for Analytical Cellular Pathology*, 24(2-3):101–11, 2002.

L. C. C. Wai, M. Mellor, and M. Brady. A multi-resolution cls detection algorithm for mammographic image analysis. *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2004*, 3217:865, 2004. doi: 10.1007/978-3-540-30136-3˙105.

Y. Wakamoto and K. Yasuda. Quantitative evaluation of cell-to-cell communication effects in cell group class using on-chip individual-cell-based cultivation system. *Biochem. Biophys. Res. Commun.*, 349(3):1130–8, 2006. doi: 10.1016/j.bbrc.2006.08.149.

T. Walter. A genome wide rnai screen by time lapse microscopy in order to identify mitotic genes— computational aspects and challenges. In *2008 5th IEEE International Symposium on Biomedical Imaging From Nano to Macro*, page 328, 2008. doi: 10.1109/ISBI.2008.4540999.

E. Wang, E. R. Ballister, and M. A. Lampson. Aurora b dynamics at centromeres create a diffusion-based phosphorylation gradient. *J. Cell Biol.*, 194(4):539–49, 2011. doi: 10.1083/jcb.201103044.

M. Wang, X. Zhou, F. Li, J. Huckins, R. W. King, and S. T. C. Wong. Novel cell segmentation and online svm for cell cycle phase identification in automated microscopy. *Bioinformatics*, 24(1):94–101, 2008. doi: 10.1093/bioinformatics/btm530.

W. Wang. Binary image segmentation of aggregates based on polygonal approximation and classification of concavities. *Pattern Recognition*, 31(10):1503, 1998. doi: 10.1016/S0031-3203(97)00145-3.

B. Weyn, G. van de Wouwer, S. Kumar-Singh, A. van Daele, P. Scheunders, E. van Marck, and W. Jacob. Computer-assisted differential diagnosis of malignant mesothelioma based on syntactic structure analysis. *Cytometry*, 35(1):23, 1999. doi: 10.1002/(SICI)1097-0320(19990101)35:1¡23::AID-CYTO4¿3.0.CO;2-P.

L. Wordeman, M. Wagenbach, and G. von Dassow. Mcak facilitates chromosome movement by promoting kinetochore microtubule turnover. *J. Cell Biol.*, 179(5):869–79, 2007. doi: 10.1083/jcb.200707120.

Y. I. Wu, D. Frey, O. I. Lungu, A. Jaehrig, I. Schlichting, B. Kuhlman, and K. M. Hahn. A genetically encoded photoactivatable rac controls the motility of living cells. *Nature*, 461(7260):104–8, 2009. doi: 10.1038/nature08241.

P. Yan, X. Zhou, M. Shah, and S. T. C. Wong. Automatic segmentation of high-throughput rnai fluorescent cellular images. *IEEE transactions on information technology in biomedicine : a publication of the IEEE Engineering in Medicine and Biology Society*, 12(1):109–17, 2008. doi: 10.1109/TITB.2007.898006.

C. Yau. Quantitative image analysis of chromosome dynamics in early drosophila embryos. In *2007 4th IEEE International Symposium on Biomedical Imaging From Nano to Macro*, page 264, 2007. doi: 10.1109/ISBI.2007.356839.

A. Yezzi, A. Tsai, and A. Willsky. A fully global approach to image segmentation via coupled curve evolution equations. *Journal of Visual Communication and Image Representation*, 13(1):195–216, 2002. doi: 10.1006/jvci.2001.0500.

Z. Yin, X. Zhou, C. Bakal, F. Li, Y. Sun, N. Perrimon, and S. T. C. Wong. Using iterative cluster merging with improved gap statistics to perform online phenotype discovery in the context of high-throughput rnai screens. *BMC Bioinformatics*, 9:264, 2008. doi: 10.1186/1471-2105-9-264.

W. Yu, H. K. Lee, S. Hariharan, W. Bu, and S. Ahmed. Level set segmentation of cellular images based on topological dependence. *Advances in Visual Computing*, 5358:540, 2008. doi: 10.1007/978-3-540-89639-5˙52.

Z. Yue, A. Carvalho, Z. Xu, X. Yuan, S. Cardinale, S. Ribeiro, F. Lai, H. Ogawa, E. Gudmundsdottir, R. Gassmann, C. G. Morrison, S. Ruchaud, and W. C. Earnshaw. Deconstructing survivin: comprehensive genetic analysis of survivin function by conditional knockout in a vertebrate cell line. *J. Cell Biol.*, 183(2):279–96, 2008. doi: 10.1083/jcb.200806118.

X. Zhou, F. Li, J. Yan, and S. T. C. Wong. A novel cell segmentation method and cell phase identification using markov model. *IEEE transactions on information technology in biomedicine : a publication of the IEEE Engineering in Medicine and Biology Society*, 13(2):152–7, 2009. doi: 10.1109/TITB.2008.2007098.

C. Zhu, E. Bossy-Wetzel, and W. Jiang. Recruitment of mklp1 to the spindle midzone/midbody by incenp is essential for midbody formation and completion of cytokinesis in human cells. *Biochem. J.*, 389(Pt 2):373–81, 2005. doi: 10.1042/BJ20050097.

C. Zimmer, E. Labruyère, V. Meas-Yedid, N. Guillén, and J.-C. Olivo-Marin. Segmentation and tracking of migrating cells in videomicroscopy with parametric active contours: a tool for cell-based drug testing. *IEEE Trans Med Imaging*, 21(10):1212–21, 2002. doi: 10.1109/TMI.2002.806292.

R. Zwiggelaar, S. M. Astley, C. R. M. Boggis, and C. J. Taylor. Linear structures in mammographic images: detection and classification. *IEEE Trans Med Imaging*, 23(9): 1077–86, 2004. doi: 10.1109/TMI.2004.828675.