Automated Microscopy and High Throughput Image Analysis in *Arabidopsis* and *Drosophila*

by

Daniel L. Mace

Computational Biology & Bioinformatics
Duke University

Date: ____________________________

Approved:

______________________________
Uwe Ohler, Advisor

______________________________
Philip N. Benfey

______________________________
Alexander Hartemink

______________________________
Carlo Tomasi

______________________________
Greg Wray

Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Computational Biology & Bioinformatics in the Graduate School of Duke University 2009
Abstract
(Image Analysis)

Automated Microscopy and High Throughput Image Analysis in Arabidopsis and Drosophila

by

Daniel L. Mace

Computational Biology & Bioinformatics
Duke University

Date: ____________________

Approved:

__________________________
Uwe Ohler, Advisor

__________________________
Philip N. Benfey

__________________________
Alexander Hartemink

__________________________
Carlo Tomasi

__________________________
Greg Wray

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Computational Biology & Bioinformatics in the Graduate School of Duke University 2009
Abstract

Development of a single cell into an adult organism is accomplished through an elaborate and complex cascade of spatiotemporal gene expression. While methods exist for capturing spatiotemporal expression patterns—in situ hybridization, reporter constructs, fluorescent tags—these methods have been highly laborious, and results are frequently assessed by subjective qualitative comparisons. To address these issues, methods must be developed for automating the capture of images, as well as for the normalization and quantification of the resulting data. In this thesis, I design computational approaches for high throughput image analysis which can be grouped into three main areas. First, I develop methods for the capture of high resolution images from high throughput platforms. In addition to the informatics aspect of this problem, I also devise a novel multiscale probabilistic model that allows us to identify and segment objects in an automated fashion. Second, high resolution images must be registered and normalized to a common frame of reference for cross image comparisons. To address these issues, I implement approaches for image registration using statistical shape models and non-rigid registration. Lastly, I validate the spatial expression data obtained from microscopy images to other known spatial expression methods, and develop methods for comparing and calculating the significance between spatial expression patterns. I demonstrate these methods on two model developmental organisms: *Arabidopsis* and *Drosophila*. 
Contents

Abstract iv

List of Tables viii

List of Figures ix

List of Abbreviations and Symbols xi

Acknowledgements xii

Introduction 1

1 Biological Background and Arabidopsis Development 6

1.1 The Arabidopsis Root as a Model Development Organism . . . . . . . . 9

1.1.1 The Biochemical and Genetic Components to Root Development 12

1.1.2 The Root in Response to Environmental Cues . . . . . . . . 15

2 Image Analysis in Computational Biology 17

2.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 17

2.2 Biological Problems . . . . . . . . . . . . . . . . . . . . . . . . . . 18

2.3 Biological Markers . . . . . . . . . . . . . . . . . . . . . . . . . . 19

2.4 Microscopy . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 24

2.5 Image Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 26

2.5.1 Segmentation and Registration . . . . . . . . . . . . . . . . . . . 26

2.5.2 Normalization and Quantification . . . . . . . . . . . . . . . . . . 33
5.2.3 Fluorescence Activated Cell Sorting (FACS) .......................... 69
5.3 Methods ............................................................................ 69
  5.3.1 Noise Removal and Contour Detection ............................. 70
  5.3.2 Registration ................................................................. 72
  5.3.3 Quantification .............................................................. 79
  5.3.4 Image Processing/Data Analysis ..................................... 80
  5.3.5 Results Scoring Metric .................................................. 80
5.4 Results and Discussion .......................................................... 81
5.5 Summary/Outlook ............................................................... 84

6 Automated High Throughput Extraction and Comparison of Gene Expression Patterns from Drosophila RNA In Situ Hybridization Images 87
  6.1 Introduction ...................................................................... 87
  6.2 Material and Methods ...................................................... 90
    6.2.1 Datasets .................................................................... 90
    6.2.2 Image Registration .................................................... 91
    6.2.3 Metrics and Evaluation Measures ................................. 95
    6.2.4 Biological Significance Testing ..................................... 98
  6.3 Results ............................................................................. 100
  6.4 Discussion ....................................................................... 105

7 Applications, Conclusions and Future Work 111
  7.1 Applications .................................................................... 111
  7.2 Conclusions .................................................................... 114
  7.3 Future Directions ............................................................. 114

Bibliography ................................................................. 116

Biography ................................................................. 127
List of Tables

6.1  Comparison of enrichment terms based on similarity metric . . . . . . 110
List of Figures

1.1 A tissue map of the *Arabidopsis* root .................................................. 9
2.1 Biological markers ....................................................................................... 21
2.2 Segmentation overview ............................................................................... 28
3.1 Schematic representation of the Rootarray .................................................. 42
3.2 Overview of controlling the microscope ...................................................... 45
4.1 Converting a Markov Random Field to a factor graph ............................... 48
4.2 Multiscale probabilistic model represented using a factor graph ............... 50
4.3 Loopy belief propagation updates for the model ......................................... 55
4.4 Two level multiscale model on color images .............................................. 59
4.5 Five level multiscale model on color images .............................................. 60
4.6 Comparison of two level and five level model .......................................... 61
4.7 *Arabidopsis* root segmentation ................................................................. 62
5.1 Three main regions of the *Arabidopsis* root ............................................ 68
5.2 Processing and straightening of the *Arabidopsis* root ............................. 70
5.3 Summary of the registration process of *Arabidopsis* roots .................... 76
5.4 Comparison of image quantification with FACS expression data for *Arabidopsis* .............................................................................................................. 78
5.5 Evaluation of a low correlated image .......................................................... 83
6.1 Training the shape model for *Drosophila* embryo registration ................ 92
6.2 Automated image registration of *Drosophila* embryos ............................ 102
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.3 Regulation of the even-skipped gene</td>
<td>105</td>
</tr>
<tr>
<td>6.4 Generation of background data sets and pairwise significance tests</td>
<td>106</td>
</tr>
<tr>
<td>7.1 Expression of SCR in SHR inducible mutants</td>
<td>112</td>
</tr>
<tr>
<td>7.2 Gene response to salt treatment of the <em>Arabidopsis</em> root</td>
<td>113</td>
</tr>
<tr>
<td>7.3 Expression along the longitude of the <em>Arabidopsis</em> root</td>
<td>113</td>
</tr>
</tbody>
</table>
List of Abbreviations and Symbols

Symbols

\( \mathbf{X} \) \ A Matrix \( X \)
\( \mathbf{x} \) \ A vector \( x \)
\( x_i \) \ The \( i \)th element in the vector \( x \)
\( \mathbf{x}_i \) \ The \( i \)th row of the matrix \( X \)
\( \mathbf{x}_j \) \ The \( j \)th column of the matrix \( X \)
\( y_i^{(k)} \) \ The \( i \)th element of the \( k \)th level (used in multiscale model)

Abbreviations

(G,Y,B,R)FP \ (Green,Yellow,Blue,Red) Fluorescent Protein
MRF \ Markov Random Field
ROI \ Region Of Interest
FG \ Factor Graph
BN \ Bayesian Network
MLE \ Maximum Likelihood Estimate
GUI \ Graphical User Interface
API \ Application Programming Interface
Acknowledgements

I’d like to thank everyone that has either helped me either directly by providing me data, or indirectly through intellectual conversations. In particular, I’d like to thank my advisor, Uwe Ohler, for his continued support and patience through my graduate work. To members of the Rootarray project, for their contributions of data and assistance. In addition, I would like to thank all the members of my lab, as well as the labs of Sayan Mukherjee, and Terry Furey for their intellectual support.

I’d also like to thank the following groups for providing me funding during my graduate work: the Duke graduate school and ISMB for providing travel money for various conferences. Additionally I have been generously supported by the NSF-MCB-06-18304 grant.

At lastly to Nicky, who has been extremely patient and caring while I was writing this thesis.
Introduction

Computational Biology has become so pervasive that it is difficult to imagine the state of the biological community today without its presence. From large scale projects such as the human genome project, ENCODE, and the HapMap project; to small scale studies such as microRNA predictions, gene knockout models, and personalized cancer treatments; computational biology has led to significant advances in biological understanding.

While sequence data has often been the data choice of interest for computational biologists, an equal, if not more important data source for computational biology, has been expression data. Often taking the form of microarray data, gene expression data provides a snapshot of the messenger RNA (mRNA) levels of a sample (cell, tissue, organism) at a specific time point. Traditionally, mRNAs have been seen as an acceptable marker for protein levels, and models representing protein interactions are often applied to expression data (e.g., transcription factor binding models).

Unfortunately, methods for quantifying expression data using microarrays rely on a large collection of cells to provide accurate measurements; additionally, the procedure is rather invasive: requiring the biologist to destroy the organism/cell. While cell sorting and cell synchrony methods have been successful in isolating individual tissues at specific time stages, they often require large amounts of biological materials, and do not achieve full high resolution measurements. As such, it is not possible to use microarray technologies to obtain reliable real-time measurements of
individual cells over a continuous time. This is quite problematic in the study of developmental biology where changes in gene expression in individual cells, at specific time stages, are necessary for proper organogenesis—the lack of which can result in abnormalities/birth defects.

Given the limitations of microarray data, a new data source must be used to obtain such high resolution data. Fortunately, biologists have already developed methods for capturing spatiotemporal gene expression patterns—from images obtained from biological microscopy. Microscopy has been a primary tool for many areas of biology; in particular it has been widely adopted for monitoring cellular dynamics in developmental biology. As evidenced by the recent Nobel prize winners, reporter constructs such as Green Fluorescent Proteins (GFP) have been widely successful in observing spatiotemporal gene expression; and have contributed to many new advancements: sex determination, genetic defects, cell cycle processes. Unfortunately, much of this advancement has been done in a highly manual fashion, with the analysis relying on qualitative comparisons.

Recently, advances in high-throughput imaging protocols have led to an explosion of data available for quantitative analysis—with tens of thousands of images for thousands of genes. The two main sources of these data come from work done in *Drosophila* by the Berkeley *Drosophila* Gene Project [105], and *C. elegans* [42]. While these databases often provide interfaces for browsing and obtaining the images, little work has been done on providing methods for high throughput comparison and analysis. The work presented in this thesis is designed to address this problem by developing methods for dealing with these data sources.

*Contribution and Goals* The majority of this thesis is aimed at developing quantitative methods for biological image microscopy data. Here I focus on developing fully automated and robust methods for quantifying and comparing spatial expres-
sion patterns of developing *Arabidopsis thaliana* roots and *Drosophila melanogaster* embryos.

It is notable that this work is not the only approach for dealing with automated analysis of microscopy images. In fact, many of the methods described for dealing with image data are derived from methods used in the computer vision community for biomedical images. However, the emphasis of this thesis is to obtain robust methods for normalizing, quantifying, and comparing spatial expression patterns from biological microscopy images. These methods differ from previously developed methods as they are more robust, applied, and validated across a much larger dataset. Additional emphasis is placed on devising biologically relevant methods for comparing expression patterns between images. These approaches provide a strong framework for dealing with biological image data obtained from microscopy images.

In addition to developing methods for existing image data, I also introduce the Rootarray: a novel engineering platform developed in a cross-discipline collaboration between the Clark engineering and Benfey biology labs. I present novel imaging methods that are designed for controlling and obtaining high throughput image data.

To this end, this thesis is focused on the following three areas:

*Segment and Register Image Data* In order to analyze biological data quantitatively, the data need to be segmented, and registered to a common frame of reference. This is often referred to as image registration. Differences in organisms/species often result in custom tailored registration methods for each problem. Here I develop multiple methods for registering and segmenting images for *Arabidopsis* as well as *Drosophila* images. In addition to registration methods, I also present real-time high-throughput methods for identifying *Arabidopsis* roots in the Rootarray.
Quantify and Validate Expression Data  With the images registered, it is necessary to provide methods for quantifying the expression data of the organism. Often one is faced with decisions on the level of detail of the obtained expression data and how to compare it to other images. While approached from the perspective as the next step in the pipeline, the quantification is tightly coupled with the registration—the accuracy of the quantification is only as accurate as the registration methods. Methods are developed for quantifying expression data from both *Arabidopsis* and *Drosophila* images, and are validated against known microarray data.

Develop Methods for Comparing/Modeling Spatiotemporal Gene Expression Data  One of the largest and most open-ended questions is how to deal with the quantified expression data. Unlike microarray data, spatiotemporal gene expression data are complex and of higher dimensionality—resulting in the 4D perspective (space, time, gene, interaction). To deal with these data, methods are developed that more properly take into account the structure of the data.

Outline  The structure of this thesis is outlined as follows.

*Chapter 1*  provides an introduction to developmental biology, with a particular emphasis on *Arabidopsis thaliana*. Here I discuss the importance of differential gene expression in the context of developmental biology

*Chapter 2*  introduces biological imaging and computational biology methods for image data. This chapter provides a thorough review of the types of imaging/reporter constructs available, as well as a review of recent advancements and approaches in image analysis
Chapter 3 presents the Rootarray—a platform for high throughput imaging of Arabidopsis roots. This chapter additionally describes the informatics work needed to integrate and manage the data generated from the Rootarray.

Chapter 4 describes a method for detecting and locating growing Arabidopsis roots in the Rootarray. This is accomplished by using a multiscale factor graph representation for image segmentation.

Chapter 5 details methods for registering, quantifying, and validating expression data obtained from Arabidopsis confocal images. Additionally we compare and validate the expression data obtained from images with comparable tissue specific microarray data.

Chapter 6 describes methods for registering, and quantifying Drosophila expression data. Methods are also developed for comparing the spatial expression patterns between images. Emphasis is placed on obtaining significance values when comparing spatial expression data.

Chapter 7 demonstrates possible experimental applications for the Rootarray, as well as future directions.
Biological Background and *Arabidopsis* Development

The development of biological systems is a complex process that is controlled by specific spatiotemporal interactions between genes which establish differential expression in an organism. In Yeast, the proper timing of cycling genes is important for transitioning the organism to different phases and initiating cellular division. Similarly, in *Drosophila*, the time dependent concentration gradients leads to increasingly complex spatial patterns that ultimately are responsible for specifying organs, tissues, and individual cell types. These space-time interactions between genes are critical for the proper organogenesis of developing organisms; whereas single perturbations of the underlying system can lead to undesired characteristics (e.g. birth defects). In order to understand these developing systems, it is important to understand the underlying mechanisms that control gene regulation.

*Transcriptional Regulation*  Transcriptional regulation is a key process for all living organisms, and is well studied in model developmental organisms. The primary mode of transcriptional regulation occurs through the use of a family of genes known
as transcription factors. Transcription factors are proteins that bind to upstream regions of genes, thereby altering the rate at which a gene is transcribed. The rate of transcription can be either increased (activators), or decreased (repressors) by transcription factors—in some cases a transcription factor can serve as both an activator for some genes as well as a repressor for others (in even rarer cases it may serve as both an activator and a repressor for the same gene).

The ability to alter the rate of gene transcription occurs from the unique structure of transcription factors. Both activators and repressors contain one or more DNA binding domains that allow the protein to bind to unique regions of DNA in the genome. Activators also contain a domain that assists in recruiting RNA polymerase to the promoter of the gene, and henceforth initiating transcription. Both activators and repressors possess other domains that further refine their transcriptional control. In addition to the DNA binding domain, transcription factors can contain trans-acting domains that allow additional proteins to bind—serving as coregulators. This can occur by binding to a transcription factor that is already bound to DNA—either modifying the rate at which a gene is transcribed, or enabling or disabling the state of a transcription factor—or by binding to unbound transcription factors and either enabling or disabling the ability to bind to DNA.

It is important to note that many other mechanisms exist to control gene regulation. Some genes interact with the structure of the chromosomes and alter the rate of gene transcription. The histones tend to be a prime target, and are often modified through methylases and kinases. Gene regulation can also be controlled after a gene is transcribed. The structure and folding of the mRNA of a gene can cause an inhibitory effect; for example, IRIS’s can form which either prevent ribosomes from binding—preventing translation, or stalling the ribosome polypeptide production. Additional regulation can also be controlled through microRNA’s, which can either degrade mRNA or prevent the binding of ribosomes to mRNA.
In developmental biology, the cascade of transcriptional regulation is a key component to proper development—and often follows a predefined set of transitions between transcriptional states. For example, in \textit{C. elegans}, cellular division and cell fate occur in a very determined pattern from early embryo to full adult. This precise transcriptional regulation is not only controlled by the transcriptional machinery of the organism, but also by maternal factors present in the egg. For example, in \textit{Drosophila}, nurse cells are responsible for establishing several of the initial gradients of expression, including the gene \textit{bicoid}. These maternal effects establish a predefined state for developing organisms, from which, all intermediate or final states of the organism will depend on.

\textit{Differential Expression}  The transcriptional control just described is responsible for establishing the differential expression in an organism, which occurs from three main possible scenarios. First, polarity within an individual cell that is undergoing cellular division can cause uneven distributions of mRNA levels within mother and daughter cells which results in different gene expression patterns across cell lineages [68]. Second, cell to cell receptors between neighboring cells can cause transcriptional regulation, leading to differentiation expression between cells [19]. Lastly, and most importantly, signaling molecular are transported through an organism, often causing differential expression not only between neighboring cells, but also across long distances [61].

This last method of differentiation—through the use of signaling molecules—is of particular importance. In some cases the signaling of a message occurs over short range, and can often result in complex patterning of gene expression between cells. However, in other instances, the signalling can be over a large distance and can affect the differentiation of many cells. In these cases, the signaling often takes the form of a gradient of expression emitted from a localized source in the embryo, and
is referred to as a morphogen. Morphogens achieve cellular differentiation through complex mechanisms in both the control and maintenance of the gradient, as well as the response and differentiation of cells in response to the gradient.

It is important to note that an organism will often use one or more of these approaches to achieve cellular differentiation not only within an organism, but also within a cell. Combined, these approaches allow both the proper development of an organism, as well as the ability to react and respond to its surrounding environment.

1.1 The Arabidopsis Root as a Model Development Organism

The root of the Arabidopsis plant is a well studied organism in developmental biology. Of particular importance is its well defined tissue arrangement and stem cell center (often referred to as a “stem cell niche”). In addition to its role as a study of stem cell development, the Arabidopsis root also serves in two highly important capacities: as the primary point of uptake of nutrients, and as the first layer of contact with pests/toxic environments. These capabilities are achieved through complex signaling pathways that allow plants to adapt to and survive to diverse environmental conditions. This section is organized as an introduction to these two roles for the Arabidopsis root.

The tissue patterning and stem cell center can be easily determined and identified through visual analysis using confocal microscopy—an illustrative diagram of the tis-
sue specification can be seen in Figure 1.1. In total, there are fifteen different tissue layers in the root, each having a rather unique role in the maintenance of the root. A subset of these tissues constitute the "stem cell niche" and are composed of two types of stem cells: functional stem cells, and structural stem cells. Functional stem cells (the initials), act akin to adult stem cells in mammals: they are responsible for extending the cell layers of differentiated cells. This occurs during cellular division, where one daughter cell differentiates and is added to the growing cell line that forms a tissue, while the other daughter cell retains its state as an undifferentiated stem cell. Structural stem cells (the quiescent center (QC)), act much differently than functional stem cells: they are primarily responsible for maintaining the cell fates of the surrounding initials, and spend the majority of their time in the G1 cell cycle phase [24]. As a result, structural stem cells rarely divide, and if they do, it is often to replace neighboring functional initials that are damaged or destroyed.

It was shown early on in roots that cell fate is controlled by positional information, and not by ancestral lineage. Experiments using cellular ablation demonstrated that, by destroying an internal cell in an organism, the dead cell would be removed, and the neighboring cell would invade the position of the dead cell [109]. While these neighboring cells were from a different tissue (and therefore, originally of a different fate), they took on the cell fate of the recently ablated cell. Thus, cell fate did not depend on the lineage from which it was derived, but rather on the interaction between neighboring cells. It was from these ablation experiments that the role of the QC as a regulator of the neighboring initials was determined. Like previous experiments demonstrated, ablating out a QC cell resulted in a neighboring (often stele) cell to invade the dead cells locations and take on the fate of a QC cell. However, in destroying the QC cells, some of the neighboring initial cells lose contact with the QC cell, and henceforth, differentiate into their non-replicating cell types. As such, the QC cells were demonstrated to be responsible for maintaining the stem cell fates
of their neighboring initials.

Although the majority of growth occurs at the initials of the root, additional development occurs further up along the root. After cells have initially divided, they progress in time, extending themselves (relatively) further up along the root. When they begin to distance themselves from the stem cell niche, they enter the elongation region of the root, which, as the name implies, is the region where the cells start to grow and become elongated along the lateral axis. While infrequent, cellular divisions continue to occur further up along the root.

Beyond the elongation region, the root enters the differentiation zone. This zone is particularly important as it is responsible for two main events: the formation of root hairs, and the formation of lateral roots. Root hairs are a crucial component to the root, and are primarily responsible for the uptake of nutrients in the root, such as: $Ca^{2+}$, $NO_3^−$, $Mn^{2+}$ [35]. The root hairs accomplish this by creating a large surface area for the root to interact with its environment. This increase in surface area, combined with a strong electrochemical gradient—which results in high expression of $H^+$-ATPase’s in the roots—leads to a rapid increase in nutrient uptake in the roots.

Root hairs are formed from the epidermis of the root, which takes on the fate of either a trichoblast (root hairs) or an atrichoblast (non-root hairs) [96]. The specification of trichoblasts occurs when the epidermal cells undergo an asymmetrical division—daughter cells that divide and overlap with two cortex cells become trichoblasts, while cells that overlap with one cortex cell become atrichoblasts. The process of root hair formation occurs in four steps: specification of root hair cell types (trichoblasts), initiation of root hair growth, bulging of the external cell wall, and growth of root hair tips. After specification, the root hairs begin to bulge and initiation of the root tip takes place. The root hairs then continue to grow outwards from the bulge, where, once done growing, a suitable cytoskeleton is formed.
In addition to root hairs, lateral roots are also an important mechanism in root development [15]. Lateral roots are fully functioning roots that form off of the primary root—occurring further up along the root, often in the differentiation zones. The formation of lateral roots occurs in two stages: identifying the region for lateral root formation, and formation of a functional meristem. After the region for lateral root formation is identified, a new functional meristem is formed. This formation begins in the pericycle region of the root tip, along cells that are neighboring to the protoxylem pole. A series of precise and predictable cell divisions occur that lead to the establishment of the quiescent center and the columella initials. The lateral root then begins to break down the cell wall structure of neighboring cells and protrudes out of the root. The outgrowth of the lateral root is controlled through the gravitropic set-point angle (GSA), which initially causes the lateral roots to grow out horizontally, and then proceed in a downward fashion. Not all roots that undergo formation of a primordium immediately penetrate the cell walls and begin to grow—some lateral roots are initiated and then remain in stasis, waiting for an external stimuli to trigger the outgrowth through the outer cell wall. It is suggested that lateral roots, advantageous roots, and embryonic roots share many of the same developmental pathways—although small differences do exist [75].

1.1.1 The Biochemical and Genetic Components to Root Development

A combination of genes and molecular signals are what allows the root to develop into a fully functional organism. At the forefront of the molecular department, is the photohormone auxin. Auxin’s distribution along the root is controlled in a complex fashion: auxin is transported down from the cotyledon of the plant to the stem cell niche, where it accumulates in high concentration. Excess auxin is then transported back up the root through the outer cell tissues where it is slowly transported back into the stele. This whole process of auxin transport is facilitated by a series of
PIN family proteins. Auxin is responsible for a large majority of cellular process, including the proper establishment of gravitropism through the gene AUX1 [77]. Transcriptional response to auxin is controlled by a family of transcription factors called Auxin Response Factors (ARFs)—which are responsible for responding to auxin levels by indirectly binding to auxin; and further, binding to promoter regions of auxin-inducible genes and triggering transcription initiation [108].

The largest contribution of auxin in root development is in its role in the formation of the stem cell niche. As described above, auxin is transported downwards through the stele where it accumulates in high quantity below the QC cells. It is through this high concentration of auxin that the root is capable of forming the stem cell niche. This primarily occurs through the interaction of the gene PLETHORA (PLT), which is responsible for activating a large series of downstream genes that create the stem cell niche—including the overexpression of the gene CYCD3—a cyclin responsible for maintenance of the G1 cell cycle phase. High concentrations of auxin at the root tip are maintained through the interaction of PLT and PIN proteins, which inhibit the transport of auxin out of the newly formed primordium [9]. This mechanism of auxin transport and PLT induction is also responsible for the formation of lateral root meristems.

Elongation of the cell wall structure in the elongation region is also controlled by auxin levels. It is speculated that interactions between auxin levels, the PIN proteins, and DELLA proteins control the localized expansion of cell wall structures [26]. The expansion is mediated by genes from the expansin family, which are responsible for remodeling and reshaping the structure of the cell wall. Remodeling and directionality of expanding cells are controlled by the gene COBRA, which is a protein that localizes itself to the growing wall segments.

Many of the genes and pathways involved in cell wall expansion are also responsible for the elongation of root hairs. After initiation, root hairs begin a process of
elongation in which the root hair tip is extended through the help of expansin proteins. Like cells in the elongation region, it is speculated that a cell receptor, MRH4, is responsible for initiating the elongation of the root tip [94]. After the initialization of the root hair, the gene MRH3 helps establish a Ca2+ gradient at the tip of the root hair. It is suggested that the influx of Ca2+ at the root hair tip establishes a positive feedback system, resulting in more MRH3 receptors, and, as a result, a higher increase of calcium concentration at the tip of the root hairs. In addition, other genes are required for root hair elongation: plants with a mutant version of the TINY ROOT HAIR1 protein are able to develop, but do not elongate root hairs.

Lateral root formation is also affected by auxin concentration levels. Like the formation of the primary meristem, lateral root formation is triggered by an influx of auxin from the shoot of the plant. It is speculated that this influx causes local changes in auxin concentrations—inducing the formation of new quiescent centers and lateral roots. Wilmoth et al. showed that the genes NPH4/ARF7 and ARF19 are possible candidates for lateral root formation [112]. Here, ARF19, and ARF7 respond to local change in auxin level, and trigger downstream genes that initiate lateral root growth. It has been shown that roots respond to local nitrate levels in the surrounding environment, where, higher levels of nitrate supply result in a large increase in lateral root formation. Cell cycle genes are also known to affect the growth of lateral roots. Prior to lateral root formation, cells in the pericycle express the CYCD4 gene, known to be active in the G1 cell cycle phase. After initiation, mRNA profiling has shown a rapid induction of genes that are known to control G2/M cell cycle transitions—ultimately causing the cellular divisions and differentiations needed for new lateral root formation.
1.1.2 The Root in Response to Environmental Cues

Unlike animals, which are capable of changing/moving to a new environment if their current one becomes inhospitable, plants are fairly immobile, and must deal and interact with their local environment. The root of the plant must be capable of growing and adapting to the soil it is currently in—adjusting its growth path for maximal uptake of nutrients while at the same time avoiding areas that are hostile (low water, pests). The faster a root is capable of responding to its local environment, the higher chance it will survive and uptake beneficial nutrients from the environment. Plants that have highly complex adaptive root architecture, tend to be more evolutionarily competitive than plants with simpler mechanisms.

There are three main ways that a root can grow and adapt to its environment. First, the rate of cellular divisions in the stem cell niche of the meristem controls the overall speed at which the root explores its environment. Second, initiation of lateral root growth can increase the ”spanning” through the environment—increasing the number of meristems that are seeking nutrients. Lastly, root hairs increase the overall surface area and induction of nutrients into the root. In order for a root to respond to its environment, it must be capable of altering at least one of these mechanisms in response to environmental cues. As mentioned above, most of these processes can be altered by auxin and other hormones in the plant, and, as such, many of the environmental response pathways are overlapping and interact with the auxin pathway.

Nitrogen

Nitrogen is an essential nutrient to all major organisms of life, plants included. The ARABIDOPSIS NITRATE RESPONSE (ANR1) gene is capable of responding to the nitrate level in the soil. It is suggested that this gene is capable of triggering downstream genes in the auxin and ABA pathways that affect root growth.
Roots that grow into areas that are rich in nitrogen trigger a two step response: lateral root initiation and dormant lateral root primordium are initialized around the local area; and global initiation of lateral roots is suppressed along the entire root. This results in a large increase in lateral root formation around the nitrate supply, and a suppression of lateral root formation in other regions [118].

**Phosphates**  Plant species that are capable of responding to high levels of phosphates often incur a competitive advantage over plants that do not. It is suggested that the gene PI DEFICIENCY RESPONSE (PDR2), is responsible for sensing and activating downstream genes responsible for phosphate response [104]. Plants that are exposed to low phosphate levels undergo a rapid response in root hair growth: occurring more frequent, and growing longer root hairs along the root. In addition to the root hairs, the formation of lateral roots is also increased in response to low phosphate levels. At the same time that the architecture is adapting, internal molecular mechanisms also undergo change: low phosphate levels trigger the increase in production of phosphate uptake genes—allowing the root to increase the production and intake of phosphate under starvation conditions. Phosphates are also know to cause a change in the angle of root growth—favoring shallow/more dispersed root architectures rather than deep thin architectures.

**Sulphate**  Roots growing in sulfate deprived environments also end up having altered root architectures, with lateral roots occurring more frequent at the tip of the root and in higher numbers. This response occurs through the NITRILASE3 (NIT3) gene, which is known to be transcribed at high levels in lateral root primordium [70]. Roots responding to a low sulfate environment also induce higher levels of auxin concentration, suggesting an interaction with the sulphate response pathway and the auxin transport pathway.
2.1 Introduction

To capture and model the spatiotemporal dynamics of a developing organism, it is necessary to design not only methods for marking and visualizing the organism, but also methods for quantifying the data for modeling. To date, much emphasis has been put on the former—developing new markers for biological imaging and advancing technology relating to microscopy—and not much on the latter; leading to a large accumulation of biological data and academic papers, where the majority of the analysis relies on qualitative assessments. To further complicate the problem, much of the data resulting from these experiments is not useful for quantitative modeling postmortem; where, either the data lacks the experimental rigor for proper quantification, or, the amount of data are not sufficient in number to answer the proposed biological question. It is for this reason that careful consideration should be placed on both the experimental design and quantitative modelling approach of the data.

Quantitative data used for modeling spatiotemporal dynamics can largely be
grouped into one of two categories: shape/phenotype data, and molecular/protein data. In some cases shape models are often used to describe the phenotype of an organism (e.g. *Drosophila* mutants), however, in many cases the phenotype often takes the form of a much simpler representation (e.g. cell death in response to an environment stimulus). Molecular/protein data constitute perhaps the largest source of both qualitative and quantitative data in biological imaging. In this case, biological markers (e.g. Green Fluorescent Protein (GFP) constructs) are used to visualize and capture spatiotemporal dynamics of molecular and protein data.

These data sources often compromise the “observed data” for a quantitative model, and can ultimately be used to address a large host of biological questions. In shape and phenotype models, often the goal is to understand the variability or uncertainty in the shape across experiments, time points, or species. For example, in a high content screening experiment, the goal may be to classify cell types for the identification of genetic mutants (e.g. cell proliferation, division, shape); or to model the variability of a population in response to environmental stimuli (e.g. drug treatment). Classification can also be extended to multiple cellular organisms, where the goal is to model or describe the shape characteristics of an organism in response to an environment, or in intr(a/e)species comparisons.

### 2.2 Biological Problems

In contrast to microarray data, molecular/protein data obtained from images allows an individual to visualize and quantify the expression patterns of one or more genes over space and time, and in a much higher resolution than possible with microarray data. This increase in resolution has allowed researchers to address a host of additional problems and has become a popular data source for modeling the spatiotemporal dynamics of, and interactions between, transcription factors in an organisms. While this has traditionally been done using differential equations, recent work has
taken a machine learning approach with the goal of automatically classifying or annotating spatiotemporal patterns into their functional categories. A further interest in studying evolutionary development (evo-devo) problems can also be applied to spatiotemporal data. Experiments involving reporter constructs have been used to compare the expression domain of orthologous genes, which, when combined with sequence data, can help address more recent problems in the evo-devo domain, such as exadaptation and directed programming.

The remainder of this chapter is outlined as follows. The biology markers and microscopy section is intended to give a brief introduction to microscopy and imaging for non-biologists. This section is followed with an introduction to core concepts of methods for image analysis: image registration and image segmentation. We then discuss approaches for normalizing and quantifying the data after it has been segmented/registered. We conclude with some examples of ongoing research and discuss available tools/Application programming interfaces (API’s) and data sources for image analysis in computational biology.

2.3 Biological Markers

The basic necessities for obtaining data from biological imaging are: a technique for “marking” your specimen, and a method for capturing/visualizing these markers. Marking a specimen is accomplished by using a biological marker, and is used in two contexts: to mark the sub-components you are interested in visualizing, and to mark elements that provide a frame of reference for your measured sub-components. Examples of biological markers are shown in Figure 2.1, where various fluorescent markers are used to visualize the gene expression (sub-components of interest) as well as to identify reference frames (cell nuclei, cell walls). This delineation in roles is not always distinct and separate—a marker may be used to delineate cell wall structures, but your quantity of interest is the shape of the cells. In this context,
your cell wall marker is used both for the frame of reference as well as the measure of interest. However, this is often not the case, and most microscopy experiments involve one or more markers for each case. For a more thorough review on biological markers, see [111].

Choosing the most optimal markers for experiments is often a complex process and depends on many factors. The organism you are studying can have an impact on marker selection; many markers are specific to certain organisms, or in some cases, mark different subregions depending on the organism. The choice of experimental conditions can also have a dramatic effect on marker selection. Some markers are only effective in strict biological conditions—if experiments are designed to test the response of an organism to various levels of environmental stress (e.g. pH levels, temperatures), the markers may not be effective across all experiments. Problems can also arise when confronted with choosing multiple markers; the addition of a new marker may cause unwanted effects in other markers already used: e.g., using multiple fluorescent protein constructs with identical promoter regions might cause competitive transcription rates, leading to an overall decrease in quantity for both markers. However, the largest problem in choosing multiple markers is the result of overlapping emission spectra, requiring experimentalists to select suboptimal non-overlapping markers.

The most common markers used in fluorescent microscopy are organic fluorescent dyes. Over the years, a variety of dyes have appeared with a large range of emission properties and specific target bindings: cell walls, cell nuclei, spindle polls, cytoskeleton. Most of these markers are derived from fluorescein and rhodamine; new dyes are often derived from old dyes by replacing the reactive groups to mark their target of interest. While some of these dyes are highly invasive and require fixing the cells prior to dyeing, many of these dyes are minimally evasive and can be used to measure dynamic properties of a specimen; such as membrane potentials, pH concentrations
Figure 2.1: Biological markers come in many colors and can be used to visualize biological specimens. (A) An Arabidopsis root with a red dye used to stain the cell walls, and a green fluorescent protein used to monitor the expression. (B) A Strongylocentrotus droebachiensis sea urchin embryo image. A (blue) dye is used to stain the nucleus, and a cyan dye is used to stain the expression (Image courtesy of Jenifer Croce). (C) A Drosophila embryo with a red dye used to stain the cell nuclei, and a green dye to stain the gene expression (FISH).

and changes in calcium levels [49, 8, 79].

Fluorescent dyes were used as one of the first approaches for measuring and quantifying spatiotemporal gene expression through a technique called fluorescence in situ hybridization (FISH) [107]. FISH allows gene expression patterns to be visualized by using fluorescent probes that mark mRNA in an organism. These probes generally consist of two parts: a complementary oligonucleotide designed to bind to the sequence of interest, and a fluorescent marker for visualization. The FISH protocol consists of a multistep process: (1) the organism is fixed, (2) specimens are preprocessed for hybridization, (3) probes are hybridized to the specimen, (4) the specimen is washed to remove any unbound probes, (5) the expression is visualized using microscopy. In some cases the fluorescent marker is not attached directly to the probe, but rather indirectly by linking the probe to an antibody reporter with an additional step after the initial hybridization to insert the fluorescent antibody. Fluorescent antibodies can also be created such that they will directly bind to the actual protein of interest allowing direct visualization. However, this approach is often highly laborious, and is not considered as a viable high throughput approach for monitoring spatiotemporal
While FISH has been a largely successful method for visualizing gene expression, it has a rather large drawback: the organism needs to be fixed prior to hybridization, killing the organism in the process, and preventing continuous measurements over time. In order to overcome this limitation, researchers needed a reporter construct that was minimally invasive—allowing continuous visualization of gene expression while not disrupting the natural process of the organism. This was achieved when the Green Fluorescent protein (GFP) was cloned from the jellyfish *Aequorea victoria* and used as a gene expression marker [16]. GFP is a small 238 amino acid protein with a unique structure: a series of 11 beta sheets encase a Serine, Tyreine, Glycein triplet chromophore—providing a protective environment for the chromophore to fluoresce. While the fluorescent protein was a major advancement for biologists—a mechanism for fluorescence that can be produced/adapted using the cells own machinery—it did have its drawbacks. The protein is not suitable for instantaneous reporting, requiring a total time of 45 minutes before the protein folds and the fluorophore becomes active. Additionally, the highly stable structure of the protein leads to slow degradations rates, making it difficult to discern “new” GFP from “old” GFP in time series measurements.

Notwithstanding these drawbacks, GFP has become widely popular and has been used for many purposes—including protein-protein interactions, biosensors for cellular stimuli; as well as serving as an alternative for many organic fluorescent dyes: including markers for chromosomes, nuclei, and membranes [21]. One of the most important roles of GFP, however, has been as a marker for spatiotemporal gene expression—often referred to as reporter constructs. GFP reporter constructs come in two types: transcriptional reporter constructs, and translational reporter constructs. Transcriptional reporter constructs are genetic constructs, where the gene for GFP is attached to a promoter of the gene of interest. The transcriptional
reporter is then inserted into the genome, under the premise that, any transcription factor that would normally bind to the promoter region of the gene, will also bind to the promoter of the construct; and therefore, expression of the GFP construct should be representative of the endogenous gene. Translational reporters include both the promoter region as well as the gene of interest within the construct. Traditionally, the gene for GFP is attached to the end of the construct with a small interlinking region between the gene of interest and the GFP; however, other locations of insertion can be used, usually guided by structural information or other a priori knowledge of the gene of interest. Translational reporter constructs differ from transcriptional constructs as GFP itself is attached to the protein of interest; therefore, it can be used to monitor subcellular localization and/or protein-protein interactions using Fluorescence Resonance Energy Transfer (FRET). While translational reporter constructs are often designed to be minimally invasive to the protein of interest, their attached GFP often has a lower fluorescence.

Recent advancements have resulted in new fluorescent protein that have been adapted and optimized for many purposes. The most notable of these include differences in the emission spectrum for red, blue, and cyan variants [37]—allowing for multiple fluorescent proteins to be used in a single experiment. Advances in bioengineering have led to photoswitchable fluorescent proteins; which allow markers from the same spectrum to be imaged sequentially [1]. Further adaptations have addressed the issue of the degradation rate of GFP, including unstable fast degrading proteins [65], and ‘fluorescent timer’ proteins [101].

It is important to note that, in terms of measuring spatiotemporal gene expression, both FISH and fluorescent proteins have their drawbacks. The creation of fluorescent protein constructs, and respective insertion into an organisms genome, is often a lengthy process. In addition to the time necessary to perform the experiments, specimens often have to be reared and selected for over multiple generation
to ensure the insertion is homozygous. In some organisms/experiments, this time requirement can be burdensome and intractable to the problem at hand. Conversely, FISH is an invasive process, requiring the organism to be fixed, destroying it in the process. This can be problematic in experimental conditions that require continuous high resolution measurements. As such, the choice of marker is a difficult process and highly dependent on the overall experimental goals.

2.4 Microscopy

Fluorescence microscopy is used for nearly all forms of visualizing gene expression—even for measuring gene expression from microarrays. The basic form of fluorescence microscopy is widefield microscopy and follows a basic principle: a lamp is used to emit a light source and focused onto the specimen of interest. The emitted light from the lamp interacts with the fluorophore in the specimen, exciting them, and causing them to emit a light source of their own. This light source emitted from the fluorophores is filtered and then captured by the microscope, producing a two dimensional \((x,y)\) image of the specimen. The choice of the emitted light is entirely dependent on the choice of fluorescent marker, as different markers have different excitations ranges.

Due to the lack of specificity of basic fluorescence microscopy, it is often limited to single-cell organisms, or organisms with small optical thickness. For organisms that do not meet this criteria, or for organisms where it is important to discern differences along the third dimension \(z\), different methods must be applied. One approach for dealing with these criteria is to use confocal microscopy. Confocal microscopy extends standard fluorescence microscopy by allowing the microscope to focus its scanning region to small sub-regions of the specimen \((x,y,z)\). This is accomplished by using a series of mirrors, pinholes and objectives to focus the light onto specific subregions of the specimen (the \(z\) range is often controlled by a motorized tray/stage). The emitted
light is then passed through an additional pinhole to filter out-of-focus emissions and subsequently collected using a photodetector. The confocal microscope is capable of scanning an entire specimen by continually refocusing the light onto specific sub-areas of the image, scanning the region, and moving onto the next region. This creates a scanning process where a specimen is scanned by imaging every element in the image (e.g., a (512,512,20) region would require over 5 million scans). The individual scans are often short and whole specimen scanning can be accomplished in a reasonable time frame.

While confocal microscopy has been extremely beneficial to biological imaging, it does have drawbacks. The confocal microscope has a limited range, with deep scans into specimens resulting in the attenuation of light and overall decrease in signal-to-noise ratio of the sample. This is attributed to the fact that the light diffracts off of the elements in the specimen, where deeper tissues of a specimen result in more diffraction of the light. Recent advances in two-photon microscopy have helped alleviate this problem [23] by using a pulsed laser to emit two photons onto a location. In this case, the fluorophores emit light only when they are simultaneously excited from both photons (an event that only takes places in the intersection of the two photons). This results in a significant decrease in regions surrounding the scanning area, as the likelihood of a fluorophore being excited outside this region remains extremely low. The largest downside of two-photon microscopy is in the overall cost of using a pulsed laser setup. Additionally, determining and measuring the absorption spectrum of new two-photon fluorophores is a slow and lengthy process.

A more recent advance for full three dimensional imaging is Single Plane Illumination Microscopy (SPIM) [41]. SPIM works by projecting a plane of light through a sample with the detector being orthogonal to the projected light. Further improvements can be done by rotating and imaging the sample from multiple angles. Image reconstruction methods can then be used to correct for attenuation and depth from
scanning issues. As a result, SPIM can be used to image deeper into specimens than
standard confocal microscopy.

2.5 Image Analysis

Designing the experiments and obtaining the images only comprises half of the prob-
lem when dealing with image data. An equally, if not more difficult problem, is
extracting information from the images—often termed image analysis. Here we split
the problem of image analysis into two parts: methods for identifying and normal-
izing specimens (Registration and Segmentation), and methods for quantifying and
normalizing the data (Data Normalization).

2.5.1 Segmentation and Registration

Image segmentation is the process of partitioning an image into several regions
[92], while image registration [120] deals with establishing a mapping between im-
ages/models so that they are comparable. In many cases, image registration and
segmentation are a coupled process and are carried out simultaneously. For others
the process may be quite separate, with distinct phases needed for the segmentation
and registration; or it may be that the only the segmentation of an object is neces-
sary and no registration is needed to be done. In a further extreme, it is possible
that no registration and segmentation is necessary at all, with the image being the
quantified data (e.g. raw classification of whole images, aggregate measures). For
the purpose of this section, the concepts of registration and segmentation are treated
as the two basic groups of methods necessary for biological image processing.

Image segmentation is the process of identifying and differentiating one or more
objects in an image from each other. In single celled organism studies, this may
be the identification and labeling of each individual cell in an image. For whole
organisms, it may be the identification of an organism from the background image,
or the identification of individual cells/tissues within the organism. Regardless of application, the goals of segmentation are consistent: to accurately and robustly identify and label subregions of an image into their respective categories.

While segmentation is more focused on identifying and differentiating individual objects within an images, image registration deals with the higher level problems of defining a mapping between images/objects. Defining a mapping between images/objects allows images/objects to be directly comparable across different experiments/time. It is also possible that image registration can be used to map different specimens within a single image.

**Segmentation**

There are a variety of methods for image segmentation of biological images and can be broadly grouped into two classes: model based, and non-model based approaches (Figure 2.2). Both of these attempt to solve one of the following segmentation problems: feature/point segmentation, or region segmentation. Feature/point based segmentation is used to identify elements of an image that occur on a sub-manifold of the image (points, lines (2D images), surfaces (3D images)); examples of feature/point segmentation include the identification of cell walls or cell nuclei. Region based segmentation is used to identify specific regions (areas, volumes) of an image, such as the detection of cells or whole organisms in an image.

Non-model based approaches to image segmentation can be effective when the objects of interest are readily discernable, and variability between images is kept to a minimum. Thresholding methods, such as adaptive thresholding, were used by Kumar et al. to identify and segment gene expression patterns from *Drosophila* embryos [59]. Region growing—methods that often begin with an initialization of a region and grow until it encapsulates the object of interest—are also frequently used. In *C. elegans*, region growing methods were used to identify growing eggs [34].

27
Figure 2.2: A comparison of image segmentation methods. (A) Initial image and final segmented image (left, right). Segmentation can be done using non-model based methods (B), which often take the form of a pipeline of filters, where the final output is the segmented image. Alternatively, model based approaches such as the level set framework (C), can be used to segment the image. In the level set framework, the segmentation is specified by a model, which evolves over time to fit the natural segmentation of the object.

particularly popular region growing method is the watershed algorithm, which has been used as an initial step in identifying cellular nuclei for developing drosophila embryos [40]. Non-model based approaches are often tied together in the form of a pipeline, where earlier steps are then used as inputs for other segmentation methods.
While non-model based registration can often produce good results, in general, model based methods for image segmentation are more tolerable to noise/distortions than non-model based segmentation methods. One of the more popular approaches for model based segmentation has been the active contours/level sets framework. Level sets [76] and geodesic active contours [14] have been largely successful in segmenting objects from many different disciplines. The level set methodology represents the segmentation problem in the context of an evolving higher dimensional hypersurface over the image; segmentation occurs when the level set is “cut” at a specific level. The level set model is initialized in either a random fashion, or using some prior information for the image (e.g. output from lower level segmentations (watershed, region growing)). Using a series of update equations, the curve is evolved over a series of iterations until convergence. The result of this process is a smooth segmentation that is more tolerant of noise and missing information than lower level approaches. Level sets have been used in a large variety of image segmentation problems, including detecting *Drosophila* cells from RNAi fluorescence experiments [115].

A large reason behind the adoption of level set methods for image segmentation is that they can be easily extended to a large variety of problem types [20]: extensions for time series tracking have been developed [84], and applied to biological microscopy images [81]. An additional benefit of the level set methodology is the ability to incorporate prior information into the model. In many cases this can be in the form of constraints on the shape of the segmented object [64], or on the overall texture of the background/object [85]. This ease of extensibility makes this a solid and attractive general framework for image segmentation problems.

Another prominent approach for image segmentation is in the context of probabilistic graphical models. Graphical models, such as Markov Random Fields (MRFs) [33], represent the problem of image segmentation in a probabilistic framework. This
is done by representing the segmentation of the image as an unobserved process, with the pixel intensities being the observed variables, and the segmented labels being the unobserved hidden variables. MRFs can also be incorporated into the level set methodology by representing the problem of an evolving curve in a probabilistic framework, with the labeling from the MRF used as a prior to guide the level set updates [95].

Registration

When comparing or modeling data obtained from multiple images, it may often be the case that a mapping between the two images must be defined. This is apparent when comparing gene expression patterns from microscopy images, where small changes in imaging conditions (pinhole, laser intensity), location (objects move, or separate scans), and inter/intra species variations require the images to be mapped to a common frame of reference before the data can be modelled. The process of defining this mapping is called image registration, and can be used for a large variety of image problems.

Image registration can be largely grouped into one of four categories: sensor registration, viewpoint registration, temporal registration, and model registration. Sensor registration is primarily concerned with problems relating to the capture of the data; markers may be imaged separately and specimens may move—disrupting the imaging process. Viewpoint registration occurs when multiple images of an object/scene are taken with the goal of merging the multiple viewpoints into one representation. As discussed above, the merging of multiple viewpoints from SPIM microscopy is an example of viewpoint registration. Temporal registration deals with the problem of registering images taken from multiple time points to one another—in many cases the goal is to track multiple objects as they move through time, or, perhaps, identify cellular divisions. As discussed in Section 2.5.1, methods for segmentation applied to
time series images (e.g. level sets, MRF’s), can be extended to temporal registration problems. Finally, model based registration involves registering objects to a common frame of reference so that they are comparable. While sensor, viewpoint and temporal registration problems occur in biological imaging, this section is focused on model based registration. For more thorough review of the other forms of image registration see Zitova et al. [120].

Model based image registration is a large area of research in the computer vision community. Originally driven by the registration and identification of regions of the brain from MRI images [113], model registration has become an increasingly pervasive method for biological image analysis. The goal of model based registration is to define a mapping between the observed image and a model image (often referred to as an atlas image), where every location in the observed image is mapped to a corresponding location in the model image. This can take on many forms in biological imaging: for example, we may be interested in determining the mapping between markers denoting nuclei in an observed image to a model describing the average nuclei locations; or we may be interested in mapping all the voxels in an observed image to all the voxels in a model image.

In general, the registration between two images takes one of the following two forms: rigid registration and non-rigid registration. Rigid registration is the process of mapping locations between an image and a model by using transformations that are global across the whole image. These transformations are consistent across all spatial locations; most often the parameters of rigid transformations are: rotation, scale, and translation (shifts in x,y,z). Non-rigid transformation consists of a non-global transformation from the image to the model, where the displacement between locations can not be described using simple rigid parameters (e.g. transformation, rotation, scale). Examples of methods for non-rigid transformations including methods based on elastic or diffusion based methods, such as thin plate splines [10] or demons
registration [103]. A more thorough discussion of non-rigid registration methods can be found in [80]. In many cases, both rigid and non-rigid registration methods are used to define a mapping.

The choice of mapping comprises only part of the registration problem. To define the mapping between an image and a model, it is necessary to determine the parameters required to define the mapping. It may be possible to mathematically define the mapping; as was the case for mapping point sets for straightening *Arabidopsis* roots [73] and *C. elegans* [88]. However, in most cases there is no closed form solution to determine the mapping; instead, the parameters for the mapping must be estimated. In general, estimating the parameters for registration requires two components: a method for evaluating the “goodness of fit” between the original image and the model image (evaluation function), and a method for estimating the parameters given the evaluation function (optimization function). There are many methods for parameter estimation, ranging from greedy search optimization to Bayesian approaches and Gibbs samplers. The evaluation function can be broadly grouped into one of two categories: feature space/point set methods and intensity/whole image methods. In point set methods, the approach is to minimize the overall distance between the two sets of points, and can take the form of simple mean squared error, Hausdorff distance, or more intricate iterative closest point methods. Intensity based methods differ slightly with point set methods, but have a similar underlying goal of minimizing the distance between two images: mean squared error, mutual information, wavelet pyramid matching. The optimization function is then used in an iterative fashion to estimate the parameters that either minimize or maximize the evaluation function.
Segmentation and Registration: Joint Models

Recently, approaches have begun to be developed for combining registration and segmentation problems into a unified framework [2]. This is often done by incorporating the segmentation, shape, and transformation parameters into a hierarchical statistical model. The benefit of this approach is that it is more robust and accurate compared to standard pipeline approaches; and the statistical representation provides a strong framework for incorporating a priori knowledge. Unfortunately, these approaches often induce longer computation times and larger memory requirements; additionally the methods are often more complex and requiring more expertise to implement/design. Nevertheless, it is an emerging area of research that shows great promise for image registration problems.

2.5.2 Normalization and Quantification

The previous two sections provided an introduction to both the biological foundations of microscopy, as well as the computational methods needed to analyze the images. In this section, we cover the methods used for normalizing and quantifying the data after image segmentation and registration have been done. While modeling approaches for the quantified data are discussed, it is important to note that they are not catch-all solutions for data analysis. As we mention at the end of this section, many of the approaches for modeling the data do not adequately reflect the biological properties of the system, and new approaches for modeling the data are needed (Section 6).

Depending on the choice of markers and microscopy methods used, methods for modelling biological data can be split into two groups: aggregate methods across an entire cell/organism, and methods for quantifying down to the sub-organism, tissue, cellular level. Aggregate methods are often the choice for large scale screening methods, examples of which include methods designed to detect and measure the effect of drug/environmental responses across a population of cells/organisms. In most of
these cases, cells are identified using image segmentation, and the data are quantified on a population level by measuring differences between cells, or distributions of response across cells. Since there are already many papers devoted to the subject of high throughput single celled organisms, the focus of this section will be on data quantification and normalization after image registration.

Often, the goal is to quantify the measurement of the reporter constructs, where, the reporter construct is representative of gene expression. If the data are normalized to a standard atlas image, the quantification and comparison of the data are fairly straightforward. In *Arabidopsis*, we measured the correlation between quantified tissue-specific expression obtained from confocal microscopy images, and comparable fluorescence activated cell sorting microarray data [73]. In this case, an atlas image was used to define the tissue locations in an *Arabidopsis* root for every location in the observed image, the data were summed and binned based on their respective tissue as denoted by the atlas image. In *Drosophila*, we used statistical shape models to register and quantify the expression patterns of the images by projecting the data along the anterior-posterior axis (Section 6). Other approaches have obtained expression at higher resolution through the use of full 3D imaging of drosophila embryos [55].

Unfortunately, raw expression data from images are not directly comparable. If the data were obtained under varying microscopy conditions (e.g., different labs, different microscopes/objectives, different microscope settings), then the quantified data may not be consistent across all images—a measure of 100 in one image might be equivalent to a measure of 50 in another image. Additionally, when using FP reporter constructs the efficiency of the reporter can be dependent on the location of the insertion into the genome. An additional problem also arises when using confocal microscopy, as the natural attenuation of the light will often lead to varied measurements based on the depth of the scanning. One natural way to overcome
this issue is to employ an additional reporter to standardize the intensity of light. Damle et al. used known measurable microinjections to normalize the data; in this way correcting for both varying microscopy conditions as well as attenuation due to light [97]. In some cases, however, the choice of modeling approach can make normalization of the data an unnecessary step (with the exception of light attenuation). Pairwise comparison studies can often avoid normalization by either comparing relative expression levels across tissues [73], whitening the data to be zero mean and one standard deviation, or by using similarity metrics that are invariant to relative level shifts such as mutual information [100].

2.6 Data Modelling

With the data obtained, registered, and normalized, we can begin to use the data to address biological problems. One of the more prominent interests is in understanding and comparing spatial expression patterns of genes. In *Drosophila*, much work has been done in comparing spatial expression patterns of genes. Pairwise similarity values have been applied and tools have been developed for matching and extracting expression patterns [89, 38, 59]. Recently, this pairwise comparison has been extended for predicting regulatory interactions using regression analysis [30]. To understand the relationships between groups of genes, data clustering and component analysis have been used to extract commonalities between genes [87].

Extending on the previous approaches, more recent work has focused on using image datasets to develop tools for biologists. To address some of the uncertainty in manual annotation of genes, machine learning classification methods were used to identify the time stage of *Drosophila* images [119, 116], as well as the orientation [31]. Classification methods have also been used on the gene expression data from images—classifying expression patterns into previously defined Controlled Vocabulary (CV) Annotation Terms [47, 119].
In addition to machine learning styled approaches, more in-depth work has been done modelling the expression patterns using differential equations. Jaeger et al. used differential equations to model the gene circuitry of a small subset of genes [44]. Additional work extended this work by incorporating binding site information into the model [45]. In Arabidopsis, differential equations have been used to describe spatial patterning of gene expression in the shoot of the plant [51].

Understanding and comparing the morphology either within or between species can often provide insight into the development of an organism. In Arabidopsis, statistical models have been used for describing the variability in the shape of leaves of plants under varying mutations [6]. Similar approaches have also been used in Zebrafish to describe phenotypical relationships of neuron structures [66]. In C. elegans, the effect of aging on morphology was studied, and models were used to predict the age of an organism [50].

2.7 Availability of Software

Here we list available software packages for dealing with biological data and publically available image datasets.

2.7.1 Image Analysis

<table>
<thead>
<tr>
<th>Name</th>
<th>I/V</th>
<th>G/T</th>
<th>Image Dim.</th>
<th>Lang.</th>
<th>Website</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ITK</strong></td>
<td>I</td>
<td>T</td>
<td>3D+</td>
<td>C++</td>
<td><a href="http://www.itk.org/">http://www.itk.org/</a></td>
</tr>
<tr>
<td>Notes: The Image Registration and Segmentation toolbox (ITK) is an open source project with a BSD style license. In addition to the framework, a large community driven application/extensions are available from the Insight Journal <a href="http://www.insight-journal.org/">http://www.insight-journal.org/</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| **VTK** | V   | T   | 2D-3D      | C++   | [http://www.vtk.org/](http://www.vtk.org/) |
| Notes: The Visualization Toolkit is the visualization counterpart of ITK that is also open source with a BSD style license. Libraries exist that connect the data formats of ITK and VTK, allowing image processing algorithms to be developed in ITK, and displayed in VTK |

| **Amira** | IV | GT | 3D+ | C++ | [http://www.amiravis.com/](http://www.amiravis.com/) |
| Notes: Amira supports both image processing and visualization of images. While the main interaction is with the GUI, a development version allows customized C++ code to be written |

| Notes: The Matlab Image Processing Toolbox (IPT) is a toolbox that can be added to the standard matlab version |

| Notes: ImageJ is an open source community project that supports both image processing and visualization. In addition to the base software, a large amount of plugins have been developed for custom applications. |
### 2.7.2 Available Data

<table>
<thead>
<tr>
<th>Name</th>
<th>Organism</th>
<th>Data Type</th>
<th>Current Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Berkeley Drosophila Genome</td>
<td><em>Drosophila</em></td>
<td>2D ISH</td>
<td>80,000 images 80,000 images</td>
</tr>
<tr>
<td>University of British Colombia C. elegans Expression Project</td>
<td><em>C. elegans</em></td>
<td>4D GFP</td>
<td>10,000 images 1,800 genes</td>
</tr>
<tr>
<td>University of Toronto Fly FISH</td>
<td><em>Drosophila</em></td>
<td>2D FISH</td>
<td>25,000 images 2,300 genes</td>
</tr>
<tr>
<td>Allen Brain Atlas</td>
<td><em>Mouse Brain</em></td>
<td>4D ISH</td>
<td>10,000 images 160 genes</td>
</tr>
<tr>
<td>Arex</td>
<td><em>Arabidopsis</em></td>
<td>2D GFP</td>
<td>200 images 100 genes</td>
</tr>
</tbody>
</table>

*URL: [http://www.fruitfly.org/cgi-bin/ex/insitu.pl](http://www.fruitfly.org/cgi-bin/ex/insitu.pl)*

*URL: [http://gfpweb.aecom.yu.edu/index](http://gfpweb.aecom.yu.edu/index)*

*URL: [http://fly-fish.ccbr.utoronto.ca/](http://fly-fish.ccbr.utoronto.ca/)*


*URL: [http://www.arexdb.org](http://www.arexdb.org)*
3 High Throughput Microscopy and the Rootarray

3.1 High Throughput Microscopy

The genomics age has led us to era of high throughputs: high throughput sequencing, high throughput expression analysis, high throughput genotyping, and, more recently, high throughput microscopy. Much like the previous high throughput breakthroughs, high throughput microscopy brings the possibility of greatly increasing our understanding of biology that previously were unattainable, or intractable. One of the largest benefits of high throughput microscopy is that we can simultaneously image hundreds of images during a single session. This is largely important for screening experiments, where, either different reporter constructs, or different phenotypes, can be exposed to stimuli and captured simultaneously. Another large benefit of high throughput microscopy is the ability to monitor the changes of a dynamic system over time (again, either between different phenotypes or reporter constructs). Much like how high throughput expression analysis greatly increases our ability to collect gene expression data, high throughput microscopy stands to greatly increase our ability to look at image data (e.g. spatial expression, shape).
Platforms have already been designed for single cell organisms/cell cultures, and are in use today. Unfortunately, for more complex organisms, it is difficult to envision a universal platform that would be suitable across all organisms. To this regard, high throughput microscopy imaging in developing organisms is in its infancy, and significant challenges still remain before it achieves widespread adoption.

3.2 Designing a High Throughput Platform

In order to understand the challenges that exist with designing high throughput microscopy platforms, it is important to recognize the areas that high throughput microscopy needs to address. The largest area high throughput microscopy seeks to improve on is in reducing both the time, and the variability, that exist in manual imaging. There are many reasons imaging requires a large amount of time: in traditional imaging, samples have to be prepared/grown in a separate medium. The samples then have to be transferred into a medium suitable for imaging (e.g. *Arabidopsis* plants have to be moved from a dish surrounded with agar to a glass slide immersed in water). Also in many cases, the goal is not to take a single image of the specimen at one time point, but rather to monitor the change of the specimen over its natural growth, or in response to an external stimuli (e.g. environmental changes, RNAi experiments). In order to accomplish the latter, the specimen would either need to be moved to a separate medium that allows the external stimuli to occur, or be imaged in an apparatus that allows imaging while administering an external stimuli. In all, these steps require a large amount of time and preparation for biologists.

In addition to the large time commitment required by biologists, the process involves multiple steps that open the opportunity for errors/variability between experiments. Adding to experimental issues arising from user mistakes (e.g. measurement errors), an even larger source of error arises from performing experiments across mul-
tiple specimens. This presents a problem with synchronization, where, depending on
the time stage of the experiment, multiple experiments may not be receiving the
stimuli at the same time, or, even worse, may not be in the correct developmental
stage when the stimuli is administered.

In order to overcome these limitations, platforms must be designed to allow biol-
ogists to perform all of the steps for imaging with minimal involvement. This design
must take into account several important aspects; first, the specimens must be able
to be placed and grown in the platform in a non-obtrusive fashion. While it may
not be necessary to develop a platform that is capable of supporting the specimen
from its initial growth, it should, however, be capable of supporting the organism
throughout its imaging lifetime. Second, the platform must be capable of handling
and distributing the external stimuli with little to no user involvement; computer-
ized controls are a very practical approach. Lastly, and most importantly, specimens
must be able to be imaged while maintaining the above two criteria.

3.3 The Root Array as a High throughput Imaging Platform

3.3.1 The Design

The Rootarray is a engineering project that is designed to meet all of the above
described requirements. Designed in collaboration with the Clark and the Benfey
lab, the Rootarray is a high throughput platform that allows up to sixty-four roots
to be simultaneously imaged under identical conditions. The Rootarray consists of
three distinct layers that allow roots to be planted, grown, and imaged with minimal
involvement from biologists. The top-most layer is the air environment, and is where
the shoots and leaves of the plants will grow into. The middle layer is the growth
chamber, and consists of a series of wells where seeds are initially planted. The
bottom layer is the imaging chamber, and is where the roots of the plant will grow
into (Figure 3.1).
Figure 3.1: Schematic representation of the Rootarray. Seeds are planted in the middle layer into individual wells. The shoots of the plant grow into the top layer, while the roots will grow into the bottom layer where they are automatically imaged.

The middle layer is an important layer as it provides the structural support for the growing plants. Seeds are placed into individual wells, where they are suspended with agar. Over time, the seeds with germinate, where, due to their natural tendencies, the stem and leaves will grow into the top layer, while the roots will grow into the bottom layer. To allow for the control of external stimuli, both the top and bottom layers are sealed off using glass slides—separate ports are used to feed in air and liquid environments respectively. These environments can be changed by altering the air and liquid flow going into the ports through a series of valves and peristaltic pumps.

In addition to providing environmental stimuli to the roots, the bottom layer also serves as the imaging chamber. Sealed off with a glass cover slide, the bottom layer is imaged using an inverted confocal microscope: the sample is placed in the upright position, where a laser will progressively scan and capture the growing roots from the bottom of the array. Since the depth of the imaging chamber is small (less than
300um), the microscope is capable of imaging the entire imaging chamber.

3.3.2 Automation and Informatics

While the bottom layer allows the whole array to be imaged for each individual root, the volume of the imaging chamber is exceptional large; high resolution imaging of the whole chamber would take many hours—eliminating any reasonable time interval. This large time commitment on imaging however is not a direct result of the number of roots, but rather an artificial constraint imposed by allowing the imaging chamber to serve as the growth chamber, which leads to a much larger volume than desired for imaging. Since the roots only occupy 2% of the volume of the imaging chamber, fast methods for detecting the images a priori to high resolution scanning would significantly speed up image scanning.

To deal with this issue, we have developed automated methods that allow us to localize the high resolution scanning only to regions of the array where roots are growing. This is accomplished by using a two step scanning process: an initial fast low resolution scanning is performed which results in a low magnification tiling image of the entire Rootarray. Using the low magnification image, we use a multiscale probabilistic model (Section 4) to identify the locations of the roots, subsequently referred to as the Regions of Interest (ROI). The scanning then enters a high resolution phase where the much smaller ROI are sequentially scanned with a higher magnification objective and subsequently transferred to an external server for permanent storage.

The microscope can be controlled using an Application Programming Interface (API) provided by Zeiss. This API consists of a series of files that can be used to update the locations, and notify the microscope that new locations need to be imaged. Additionally, the state of the scanning process can be inferred by monitoring the directory where images are written to. Using these two interfaces, we can construct a semi-automated method for identifying and scanning roots.
A Graphical User Interface (GUI) is used to allow users to initialize the settings for the experiment (experiment name, directories, imaging parameters, etc.). The basic control of the program is as follows and is illustrated in Figure 3.2.

- **Listening:** A series of threads are responsible for listening to both the image directory and the directory of the multitime file (for the initialization). These threads allow the program to monitor the state of the microscope.

- **Low Resolution Merging:** As the images are written to the directory, they are read in and merged into a tiling of the Rootarray image. The final tiling is then used for the ROI detection.

- **ROI Detection:** Due to the speed and memory issues of 3D detection, the ROI program is run remotely. The images are transferred to a run on a remote server, where the results are returned and used to create the multitime file to be used for high resolution scanning.

- **High Resolution Merging:** The high resolution merging is responsible for merging the tiles of a high resolution scan.

- **SSH Transfer:** All files resulting from an experiment are moved over to the server for permanent storage.

All tasks (Merging (High and Low), SSH, ROI), are done in an asynchronous fashion to minimize downtime of the automation.
Figure 3.2: Multitime Control: Overview of controlling the microscope. The scanning consists of many individual tasks that are responsible for controlling and identifying regions of the Rootarray. A more thorough overview of the individual tasks is described in Section 3.3.2
We have discussed methods for automating the microscope and the respectiveinformatics necessary to go with it. In this chapter, we discuss a novel multiscalefactor graph segmentation method that is used to identify the growing roots in theRootarray.

4.1 Introduction

Image segmentation is the process of partitioning an image into sets of non-overlappingcategorical labels. In many cases, these sets of pixels are representative of true worldobjects: e.g. a segmentation of a natural scene image might partition the image intobackground objects (e.g. the sky), and the foreground objects (e.g. cards, roads,buildings). Approaching the problem of image segmentation using a probabilisticframework can be difficult—images are often complex and the true segmentationof an image is often dependent on many different variables (e.g. neighboring pix-els, shading, context of image). As a result of their simplifying modular approach,
probabilistic generative models have increasingly gained recognition as powerful and practical tools to address image segmentation.

One of the most widely known approaches for image segmentation using a probabilistic framework is the Ising Model—a specific representation of Markov Random Fields (MRFs) \[33\]. The Ising model is a graphical representation where the true partitioning of an image is obtained by minimizing an energy function that incorporates both information from the image as well as the labeling of neighboring pixels. One of the main benefits of MRFs is the simplicity in its probabilistic structure, which allows easy customization to specific applications. Unfortunately, this simplicity comes at a cost, and many of the initial approaches are unable to deal with more complex segmentation problems. To combat this issue, extensions to MRFs were developed—the most important being the multiscale model. The multiscale model extends MRFs by introducing multiple dyadic scaling levels and defining dependencies between scales—allowing larger scale dependencies to be efficiently modeled.

In this paper we develop a generative probabilistic model for image segmentation trained on \textit{a priori} segmented images. Our generative model extends previous work by representing the segmentation in the framework of a hierarchical mixture model with latent variable levels. However, unlike previous methodologies, our representation incorporates multiple levels of dyadic scaling structure, without requiring the prelabeling of individual sublevels.

To present our model, we incorporate it in a factor graph representation. Factor graphs generalize previous graphical models (MRFs, Bayesian networks, etc.) by adding additional ”factor” nodes to the graph to describe dependencies between variables. Edges between variables are then replaced by an intermediate edge from a variable to a factor, and then from the factor to the variable. This representation creates a bipartite graph where edges exist only between factor nodes and variable nodes. The factor graph representation provides a natural intuitive representation for
Figure 4.1: **Factor Graphs:** (a) A simple Markov random field model for image processing. Here, neighboring pixels \((x_i, x_j)\) share a dependency and are illustrated by drawing an edge between them \(e_{ij}\)—often described through potential functions \(\phi_{ij}\). (b) This graphical structure can easily be converted to a factor graph representation by replacing the edges between neighboring pixels with an edge between the pixel \(x_i\) and a factor \(f_{ij}\), and from the factor \(f_{ij}\) to the other pixel \(x_j\). This factor graph is identical to the Markov random field representation.

belief propagation—which, as we will show, is used for both the decoding of images as well as the training of the model. Since factor graphs are fully interchangeable with current graphical models, we lose no expressive power in using this representation. Recently, factor graph models have begun to be used in the computer vision community [60]. A more thorough explanation of factor graphs can be found in [58].

The rest of the chapter is organized as follows: Factor graphs and belief propagation for single scaled models are introduced in Section 4.2. The concept of a latent variable level and a hierarchical approach are illustrated in Section 4.2.2. Our multiscale generative model is described in Section 4.2.3, and methods for training the model are detailed in Section 4.2.4. Section 4.2.5 describes several approaches that considerably increase the computational and memory performance of the factor graph representation. In Section 4.3 we demonstrate the effectiveness of the model by applying it to several image segmentation problems.
4.2 Methods

We begin by introducing the basic formalization for image segmentation with probabilistic generative models. Let $X = \{x_1, ..., x_n\}$ be the observed pixels of an image, and $Y = \{y_1, ..., y_n\}$ be the unobserved labels of an image (e.g., sky, vs ground). Our goal is to define a function $g$, that describes the dependency between the observed pixels and the unobserved labels $g(x_1, ..., x_n, y_1, ... y_n)$. A simple assumption is that the function $g$ can be factored into a product of local functions $g(x_1, ..., x_n, y_1, ... y_n) = \prod_i f_i(x_i, y_i)$, or, stated otherwise, the unobserved label of an image is dependent only on the observed pixel value for that location $i$.

While the function, $f_i(x_i, y_i)$ can take on many forms, one approach is to assume that each pixel is generated by a mixture of gaussians. In this case $f$ is a generative probabilistic model: $f(x_i, y_i) = p(x_i|y_i = j) p(y_i = j)$ where $p(x_i|y_i = j) = \mathcal{N}(\mu_j, \sigma_j)$ is the standard gaussian distribution for mixture $j$, and $p(y_i = j) = \pi_j$ is the mixing ratio for mixture $j$ with constraints $\pi_j \geq 0, \sum \pi_j = 1$. We can calculate the probability that a pixel was generated by class $k$ by calculating the posterior probability of the mixture

$$p(y_i = k|x_i) = \frac{p(x_i|y_i = k) p(y_i = k)}{\sum_j p(x_i|y_i = j) p(y_i = j)} \quad (4.1)$$

### 4.2.1 Factor Graphs and Belief Propagation

An alternative approach to the above is to consider the calculation of the posterior in the context of belief propagation on factor graphs [58]. Belief propagation is a method for computing marginal probabilities by using the factorization of a model to efficiently propagate uncertainties about variables. This is accomplished by constructing a message passing scheme on the factor graph, where uncertainties about variables are passed as messages between factor nodes and variable nodes. To fully make use of the factorization, we split our mixture model into two factors:
Figure 4.2: Factor Graph Representation: (a) A single scale factor graph representation, the observed pixels \( x \), and unobserved labels \( y \) are described through the factor nodes \( f^1 \) (multivariate gaussian) and \( f^2 \) (prior class mixing ratios). The posterior probability can be calculated by propagating the messages along the graph (blue arrows) (b) A hierarchical factor graph representation. An intermediate latent level is added \( y^{(1)} \), a mixture of multinomials to describe them \( f^2 \), and the multinomial mixture ratios \( f^3 \). The beliefs about the final label are calculated by propagating messages along the graph. (c) Graph structure is identical to [b], however, we calculate the beliefs about the latent variable level. (d) Partial illustration of the full multiscale model with neighborhood structure. The underlying levels are superimposed on higher levels to demonstrate the scaling effects of the model. While the overall neighborhood structure remains the same, the scaling effect incorporates more observed pixels at each level (5 → 13 → 41).

\[ f^1_i = p(x_i | y_i) \] and \( f^2_i = p(y_i) \) as shown in Figure 4.2. Our message passing scheme now consists of three messages: the observed pixels to the factor node (variable to factor), \( m_{x_i \rightarrow f^1_i}(x) \), the mixture factor node to the unobserved label \( m_{f^1_i \rightarrow y_i}(y) \), and the mixing ratios to the unobserved label \( m_{f^2_i \rightarrow y_i}(y) \).

The message from the observed pixel to the factor node is the identity message:

\[
m_{x_i \rightarrow f^1_i}(x) = \begin{cases} 1 & x = x_i \\ 0 & x \neq x_i \end{cases}
\]

The message from the factor node to the unobserved label is a little more complex:

\[
m_{f^1_i \rightarrow y_i}(y_i) = \sum_x p(x_i | y_i) m_{x_i \rightarrow f^1_i}(x)
\]
Since $m_{x_i \rightarrow f_i}^1(x)$ is an identity function, equation 4.2.1 reduces to the simple conditional probability shown in equation 4.1. The message from the mixing ratios to the unobserved label is trivial:

$$m_{f_i^2 \rightarrow y_i}(y) = \pi$$

To calculate the beliefs about the unobserved labels, the messages are updated in an iterative fashion: (1) the observed messages $m_{x_i \rightarrow f_i^1}(x)$; (2) followed by the messages from the gaussian mixture factor to the unobserved factor $m_{f_i^1 \rightarrow y_i}(y_i)$; (3) and finally, the prior factor messages $m_{f_i^2 \rightarrow y_i}(y_i)$ as shown in Figure 4.2a. With the messages updated, the beliefs $b(y_i)$, can be calculated.

$$b(y_i) \propto \prod_{h \in n(y_i)} m_{f_h^i \rightarrow y_i}$$ (4.2)

where $n(y_i)$ describes all factor nodes that share an edge with variable node $y_i$. After normalization, equation 4.2 is equivalent to the already seen posterior probability shown in equation 4.1.

4.2.2 Hierarchical Factor Graphs

In many cases, a single gaussian mixture would not be sufficient to accurately label an image pixel. For example, a generic label of “sky” would need to describe pixels that come from both clear as well as cloudy sky regions. To account for this, we add an additional latent variable level $y_i^{(1)}$, where $y_i^{(1)}$ can take on an intermediate label (e.g. clouds, or clear sky). The final true label $y_i^{(2)}$, can then be described as a mixture of $y_i^{(1)}$ in a hierarchical fashion.

$$f\left(x_i, y_i^{(1)}, y_i^{(2)}\right) = p\left(x_i|y_i^{(1)}\right) p\left(y_i^{(1)}|y_i^{(2)}\right) p\left(y_i^{(2)}\right)$$ (4.3)
We define \( p\left(y_i^{(1)} \mid y_i^{(2)}\right) \) to be the multinomial distribution with parameters \( p_j \) for the \( j^{th} \) mixture, and \( p\left(y_i^{(1)}\right) = \pi_y^{(2)} \) are the familiar mixing proportions.

Equation 4.3 can be factorized: \( f^1 (x_i, y_i^{(1)}) \), \( f^2 (y_i^{(1)}, y_i^{(2)}) \), and \( f^3 (y^{(2)}) \). This factorization extends our model as shown in Figure 4.2b. Our new model now has three additional messages being passed leading to a total of six messages: the observed pixels to the gaussian mixture factor, \( m_{x_i \rightarrow f_1^i} \); two messages between the gaussian mixture factor and the intermediate layer, \( m_{f_1^i \rightarrow y_i^{(1)}} \) and \( m_{y_i^{(1)} \rightarrow f_1^i} \); two messages between the multinomial factor and the final layer, \( m_{f_2^i \rightarrow y_i^{(2)}} \) and \( m_{y_i^{(2)} \rightarrow f_2^i} \); and the final mixing proportions for the multinomial, \( m_{f_3^i \rightarrow y_i^{(2)}} \). The new messages are as follows (ignoring the identity, gaussians, and final mixing proportions):

\[
\begin{align*}
m_{y_i^{(1)} \rightarrow f_1^i} (y^{(1)}) &= m_{f_2^i \rightarrow y_i^{(1)}} (y^{(1)}) \\
m_{y_i^{(2)} \rightarrow f_2^i} (y^{(2)}) &= m_{f_3^i \rightarrow y_i^{(2)}} (y^{(2)}) \\
m_{f_2^i \rightarrow y_i^{(2)}} (y^{(2)}) &= \sum_{y^{(2)}} p\left(y_i^{(1)} \mid y_i^{(2)}\right) m_{y_i^{(2)} \rightarrow f_2^i} (y^{(2)})
\end{align*}
\]

The largest difference is \( m_{f_2^i \rightarrow y_i^{(2)}} (y^{(2)}) \), which is equivalent to the mixing ratios of the intermediate classes \( p\left(y^{(1)}\right) \). While previously we defined these as explicit mixing ratios \( \pi_{y^{(1)}} \), here they take on a hierarchical formulation and are equivalent to:

\[
p\left(y^{(1)}\right) = \sum_{y^{(2)}} p\left(y_i^{(1)} \mid y_i^{(2)}\right) m_{y_i^{(2)} \rightarrow f_2^i}
= \sum_j \pi_{y^{(2)}y^{(1)}} p_j^{(1)} \cdots p_j^{(m)}
\]

52
The belief about the true label can be calculated similarly to the single scaled model, by propagating messages up the model until the top layer, and propagating the mixing ratios downwards. Additionally, we can also calculate the beliefs about the intermediate latent variable by propagating messages upwards from the observed pixels to the latent variable, and all messages downwards from the final mixing ratios to the latent variable as shown in Figure 4.2c. The beliefs about both levels are:

$$b \left( y_i^{(2)} \right) \propto \prod_{h \in n} m_{f_h \rightarrow y_i^{(2)}} \left( y_i^{(2)} \right) \quad \text{and} \quad b \left( y_i^{(1)} \right) \propto \prod_{h \in n} m_{f_h \rightarrow y_i^{(1)}} \left( y_i^{(1)} \right)$$

### 4.2.3 Multiscale Factor Graphs and Loopy Belief Propagation

We have introduced our generative probabilistic model, and shown the equivalence of using belief propagation to calculate the posterior labeling of image in both simple and hierarchical mixture models. However, the previous models suffer a major limitation by assuming independence between pixels. It is often the case that this assumption is an oversimplification—the labeling of a pixel is dependent on both the observed as well as unobserved neighboring pixels. Ignoring these relationships tends to result in less successful models.

To overcome these limitations, we expand the previous hierarchical model to incorporate both pixel-label as well as label-label dependencies. We begin by assuming that the labeling of a pixel on the intermediate layer is not dependent on its direct pixel value, but rather on a set of neighboring pixels. Let $X_{N_i}$ be the set of all neighboring pixels of $x_i$ at scale level $s$, where $x_j \in X_{N_i}$ if $x_j \in x_i + 2^s \{(0,0), (-1,0), (0,-1), (1,0), (0,1)\}$, and $s = 0$ is the scaling level for the first level (in this case, all immediate neighbors of distance $\leq 1$). The gaussian then becomes a multivariate gaussian, and equation 4.2.1 extends to the following:
$$m_{f^i_1} (y^{(1)}_i) = \sum_{\sim \{y^{(1)}_i\}} p (X_{N_1^i} | y^{(1)}_i) \prod_{h \in n(f^1) \setminus y^{(1)}_i} m_{h \rightarrow f^i_1} (h)$$

Here we introduce the notation $\sim \{y^{(1)}_i\}$ to represent the summation over all edges with the exception of $y^{(1)}_i$, and $h \in n(f^1) \setminus y^{(1)}_i$ the product over all edges with the exception of $y^{(1)}_i$. The parameters for mixture $j$ of the multivariate gaussian become a vector of means $\mu_j$, and a covariance matrix $\Sigma_j$. For simplicity, we assume a diagonal covariance matrix. To allow for rotational invariance, we split our neighboring pixels into two groups: the observed pixel directly underneath the label, $x_i$, and the neighboring pixels, $K = x_j \in x_{N_1^i} \sim \{x_i\}$. Let $e : K \leftarrow K$ be a permutation of the locations with $e = e_1, ..., e_n$ being the set of all rotational permutations. The probability under the multivariate gaussian then becomes the maximal probable rotation:

$$m_{f^i_1} (y_i) = \arg \max (e_j) \sum_{\sim \{y\}} p (X_{N_1^i} | y^{(1)}_i) \prod_{h \in n(f^1) \setminus y^{(1)}_i} m_{h \rightarrow f^i_1} (h)$$

Likewise, we extend the multinomial model to a multivariate multinomial. Similar to our multivariate gaussian, we split our inputs into two sets: the label directly underneath $y^{(1)}_i$, and the neighboring labels $y^{(1)}_{N_1^i} \sim \{y^{(1)}_i\}$. The individual mixtures are then represented as:

$$p (y^{(1)}_i | y^{(2)}_i) = p_{y^{(1)}_i} \prod_{j=1}^{m} \left[ \frac{n!}{y^{(1)}_i! \cdots y^{(m)}_i!} q_{j_1}^{y^{(1)}_i} \cdots q_{j_m}^{y^{(m)}_i} \right]$$

The scaling and determination of the neighborhood regions are identical to that of the multivariate gaussian, starting at $s = 0$. As we add additional intermediate levels to the model, the neighborhood set scales in a dyadic fashion, increasing in distance by powers of 2. This increase in scaling allows increasingly large pixels regions to be used in determining the final label as shown in Figure 4.2d.

Previously, the calculation of the beliefs for the labels was relatively simple, mes-
sages would be passed in a chain-like fashion until reaching the node of interest. By adding neighboring edges we have created a graph with many cycles and loops—requiring us to use loopy belief propagation \[82\]. Unlike standard belief propagation, loopy belief propagation has no guarantee of convergence—it is possible that the graph can get stuck in a suboptimal state, or alternate between multiple states. It is therefore important to choose a belief propagation update scheme that converges more consistently.

Our loopy belief propagation update scheme consists of two stages: an initialization stage, and an iterative stage. The initialization begins by propagating all messages from observed nodes \((m_{x_i \rightarrow f_1}(x))\), and all factors with a single “varying” edge. We define a varying edge as an edge between a factor and an unobservable (the messages from the prior term to the final label level \(m_{f_3 \rightarrow y_i^{(2)}}(y_i^{(2)})\), and from the multivariate gaussian factor to the intermediate level \(m_{f_1 \rightarrow y_i^{(1)}}(y_i^{(1)})\)). The propagation then becomes iterative; at each iteration, messages are passed in a bottom-up→top-down fashion: the messages from the observables are propagated upwards until they reach the top unobservable, at which, the process is repeated in a downward fashion as shown in Figure 4.3. This process is repeated until convergence.

**Figure 4.3: Belief Propagation Update:** A graphical representation demonstrating the belief propagation update scheme of a two level model. (1) The propagation is initialized by updating all messages from the observables \((X)\) to the multivariate gaussian \(f_1\), (2) and from the multivariate gaussian to the intermediate level \(Y^{(1)}\). (3) The prior class probabilities are then propagated from the prior factor \(f_3\), to the final class level \(Y^{(2)}\). The update scheme then enters an iterative process: (4,5) messages are propagated up the model from the unobservable levels \(Y^{(1−2)}\) and the multinomial factor \(f^2\); (6,7) when the messages reach the top of the model, they are propagated downward to the bottom level. This process is continued until convergence.
We deviate slightly from the standard updates on the variable level. Traditionally, the update at the variable level would take on the product of all incoming edges with the exception of the update edge \( m_{y(1) \rightarrow f^2} = \prod_{h \in n(y(1)) \setminus \{f^2\}} m_{h \rightarrow y(1)} (y^{(1)}) \). Instead we normalize the messages that are sent out between each level \( m_{f^2 \rightarrow y(1)} (y) \) and \( \bar{m}_{f^3 \rightarrow y(1)} (y) \), where \( \bar{m}_{f^3 \rightarrow y(1)} (y) = \frac{1}{|E|} \sum_{h(f^3) \in n(y(1)) \setminus \{f\}} e_h m_{f^3 h \rightarrow y(1)} (y) \) are the normalized messages coming from \( f^3 \) to \( y^{(1)} \) with \( e_h \) being the normalized inverse entropy of message \( h \). Although it is possible to describe this graphically by adding an additional level and factor, for brevity, it is assumed that all variable updates will take this form.

4.2.4 Training

We approach the training from an expectation maximization (EM) perspective. For each proposed set of parameters, the expectation of the final labeling is calculated by propagating the beliefs through the model as described in 4.2.3. The maximization is then calculated on the likelihood of the data, based on a classification likelihood of training data, \( T \). For the likelihood, we use a mixture approach that incorporates both a mixture likelihood on the individual layers, and a classification likelihood on the labeled data:

\[
\arg \max (\theta) = \pi_1 p(y) + \pi_2 p (y^{(m)} = T)
\]

where \( \pi_1 \) and \( \pi_2 \) are mixing ratios. For simplicity, we describe the individual likelihoods in the log likelihood equivalent. For the classification log likelihood, we use an averaged classification log likelihood:

\[
\log p (y^{(m)} = t_i) = \sum_{j=1}^{k} \frac{1}{|T_j|} \sum_{s \in T_j} \log p (y_{s}^{(m)} = j)
\]

Here, \( T_j \) is the set of all labels in the training data set with label \( j \), and \( |T_j| \)
is the number of pixels labeled $j$. The summations for each class are reweighted to its geometric mean. For the mixture likelihood, we use an averaged mixture log likelihood:

$$
\log p(y) = \frac{1}{|Y|} \sum_{j=1}^{m} \sum_{s \in y^{(j)}} \log \sum_{i} z_{i,s}^{(j)} p(y_{s}^{(j-1)}|y_{s}^{(j)} = i) p(y_{s}^{(j)} = i)
$$

This averaged mixture log likelihood is based on a standard likelihood on mixture models, but re-weights the model to the geometric mean. Additionally, we add an indicator variable $z_{i,s}^{(j)}$, where $z_{i,s}^{(j)} = 1$ for all intermediate levels and $z_{i,s}^{(j)} = 1$ iff $t_{s} = i$ for the final level. Using the log likelihood we find the parameters using an house version of the particle swarm optimizer [54].

4.2.5 Performance

Despite its increase in consistency and convergence, belief propagation on factor graphs has not received a lot of recognition for image segmentation. This can largely be attributed to the increased demands in computational and memory requirements as compared to standard search methods on MRFs. Here we describe methods that considerably improve the computational and memory performance of factor graphs.

**Computational** In many cases, the summation over all possible states can be very costly. For example, the sum product for the mixture of multinomials requires a summation over all possible state combinations of the previous level for each mixture component

$$
m_{f_{2} \rightarrow y_{1}}(y^{(1)}) = \sum_{\sim \{y^{(1)}_{j}\}} p\left(y_{1}^{(1)}|y_{i}^{(2)}\right) \prod_{h \in n(f_{2}^{(2)}) \setminus \{y^{(1)}_{j}\}} m_{h \rightarrow f_{2}}(h)
$$
An alternative approach is to use particle methods to approximate this summation [22]. Let

$$L = \{(w_1, \hat{y}_1), \ldots, (w_n, \hat{y}_n)\}$$

be a list of weights and samples, where \(\hat{y}_i \in \{y^{(2)}_1, y^{(2)}_2, \ldots, y^{(1)}_k\}\) is a sample, and \(w_i = (w^1_i, \ldots, w^k_i)\) is a vector of update weights for each edge. Using this list of samples and weights, a message can be updated along edge \(j\) by the following:

$$m_{f^{(2)}_i \rightarrow y^{(1)}_j}(y^{(1)}) = \frac{1}{n} \sum_{l=1}^{n} w^j_{jl} p(\hat{y}_l)$$

For this summation to be consistent, the samples and the weights must be drawn in a manner that is representative of their true distributions. In a single edge update, we would sample directly from the incoming edges \(u_j(\hat{y}) = \prod_{h \in n(f^j_i)} m_{h \rightarrow f^j_i}(h)\). This is a Gibbs sampling process, where \(\hat{p}_i\) would be sampled directly from the observed distribution, with weights \(w^j_i = 1\). To account for multiple edge updates, we extend the sampler to an importance sampler by using the following proposal distribution:

$$v(\hat{y}) = \pi_1 U(\hat{y}) + \pi_2 \left[ \prod_{h \in n(f^j_i)} m_{h \rightarrow f^j_i}(h) \right]$$

\(\pi_1\) and \(\pi_2\) are user defined mixing ratios, and \(U(\hat{y})\) is the uniform distribution. The proposal distribution balances the need to draw accurate samples that are representative of the observed distribution across all incoming edges with the need to visit configurations that might otherwise be inaccessible. The weight for an individual edge can then be calculated from the proposal and true distributions of the data, \(w^j_i = u_j(\hat{y}_i) / v(\hat{y}_i)\). In our experience, the number of samples is often in the
Figure 4.4: Performance on a testing image of a two level model on a color image. (A) The original image (top) and final segmentation (bottom). The final segmentation is obtained from the second level, the first, level is a latent variable level. (B) The beliefs (posterior probabilities) for each label at each level. 1-2 are the individual levels \(Y^{(1-5)}\), while I-IV are the subcomponents per level \(I = b(y_i = 1)\).

hundreds compared to thousands with the full sum product, resulting in a 25-50 fold reduction in computational time.

Memory  Since our update scheme follows a directional process, it is not necessary to store all the messages in the model. Instead, we only need to store the direction of one of the messages during propagation [28]. Overall this allows a reduction in memory requirements by 50%.

4.3 Results

In this section we present several applications of our model to segment images. For our first application, we demonstrate use of the model on a sample set of color images with the goal of labeling images into two groups: the sky (including clouds), and land. We compare the performance of two models: a two level model, and a five level model, where the final labelling is third level, with the fourth and fifth levels
being latent levels.

The model segments pixels into hierarchical structures through two events: label merging and label splitting. Merging occurs when multiple labels at a lower level are merged into one label at a higher level. This is apparent in the second level of the model where various forms of land pixels are merged into one final land component. Conversely, splitting occurs when labels grouped at lower levels appear ungrouped at higher levels. At lower levels, part of the river is included in the labeling as sky pixels; at higher levels these components are split and correctly assigned.
Figure 4.6: Comparing the segmentations using a two level and five level model. (top row), input images (middle row) the two final segmentations from the two level model (bottom row) the final five level segmentation’s. The two level model is more susceptible to small fluctuations in color, which occasionally results in mis-segmentations (first column, second column). The five level model is more robust to small fluctuations in color, although can occasionally oversmooth the results (second column, fifth column).

In comparing the models, the five level model generally performs better than the two level model. For the first and second images, the two level model incorrectly labels the river and part of the street/cars as sky. In contrast, the five level model is susceptible to oversmoothing of the label field. For example, in the second image, it rounds off some of the pillars on the chapel, and in the fifth image, it removes some of the structure of the tree.

We are also interested in using the multiscale model for identifying and segmenting Arabidopsis roots growing in a high-throughput array. The results from the segmentation will be used to for microscope automation; therefore, it is important that the segmentation is smooth and low in error. After identification, the segmented roots will be used for further high resolution analysis [73]. We show the results of a trained three level model in Figure 4.7.

More extensive comparisons of this approach are needed to evaluate this model with respect to other models. In the context of segmenting Arabidopsis roots, our
method balances the need for having an accurate model that is capable of handling three dimensional data, with the necessity of calculating the segmentation in a time dependent manner. Earlier attempts using either simple Markov Random Fields, watershed segmentation, or level set approaches were either not accurate enough, or did not finish in a reasonable time frame.

4.4 Discussion

We have presented a multiscale factor graph model for image segmentation. Our method provides a smooth result, does not suffer from the above described blocking issue, and is capable of accurately segmenting objects. In addition to the theoretical advances, we have applied many practical advances for factor graphs including: efficient sampling methods, and an overall reduction in memory requirements. These advances provide a practical approaches for performing belief propagation on factor graphs.

Currently, the number of layers and mixtures is a user defined parameter and is determined via trial and error. Determining the number of layers and mixtures
is a difficult matter, as can be seen from the large variability of approaches for simple mixture models. Our multiscale approach simplifies the process by introducing the effect of dyadic scaling—as the scaling increases in distance, the likelihood of describing the data decreases (similar to a Kriging correlation function). Extending our likelihood to incorporate intermediate layers and the number of mixtures will allow us to extend our model to semi-supervised and unsupervised approaches.

While we have chosen to model the labels through the use of their direct pixel values, our model is not limited to this type of input. An extension of our model to wavelets could done by replacing the pixel inputs of the multivariate gaussian with the coefficients from a wavelet filter bank, or by creating multiple layers of gaussian mixtures, with additional factor nodes linking the intermediate layers. An extension of our model to wavelets, combined with the previously described likelihood on the individual layers, will provide an effective framework for filter bank selection.
Automatic Image Registration in *Arabidopsis*

So far, we have discussed methods for identifying and controlling the microscope during the scanning process for automated microscopy. Once obtained, methods must be developed that Once the images have been obtained, we need methods for extracting and normalizing and registering images to a common frame of reference. Once registered, the tissue specific expression data can be quantified and compared across images. In this chapter, I present such a method for segmenting and registering *Arabidopsis* images from confocal microscopy experiments.

5.1 Introduction

While traditional biology has analyzed connections in these networks using a bottom-up approach (e.g. gene knockouts or knockdowns), technologies such as microarrays provide data for the inference of regulatory connections through the analysis of expression levels—often referred to as a top-down methods. However, the established way of measuring gene expression by DNA microarrays frequently averages over areas with different expression signatures and does not provide cues as to preferred spatial expression. To obtain a thorough understanding of gene regulation, we must move
beyond these limits towards an accurate and detailed description of spatiotemporal (4-D) gene activity and regulatory interactions. High throughput digital microscopy has begun to deliver large datasets describing where a gene is expressed at a particular stage in living organisms. We are now faced with the task of how to use this rich information resource in combination with computational approaches with the aim of elucidating regulatory interactions in the development of multicellular organisms.

The process of extracting information from images is not new and has been particularly established for biomedical problems; examples include the mapping of brain scans and the automatic identification of breast cancer tumors \[74, 113\]. Recently, these techniques have begun to be adapted to molecular biology. In \textit{Drosophila}, analysis of RNA in situ hybridization images has been used to identify expression patterns \[59, 90\]. Due to the variability in staining, in situ patterns are not capable of providing accurate expression values and are more of a qualitative nature. As discussed in Section 2.3, fluorescent proteins, such as Green Fluorescent Protein (GFP), can be used to quantitatively visualize the expression of a gene \[18\]. It has been demonstrated that the intensity values from these fluorescent protein fusions are capable of recapitulating the underlying molecular biology of yeast with high confidence \[114\]. Quantification of fluorescent proteins in yeast does not need to address issues of attenuation due to depth or multiple tissue regions that are present in multi-cellular organisms. Additional work in sea urchins has shown that, by using a known injected fluorescent standard, one can correct for this attenuation and provide accurate measurements \[25\]. GFP reporter constructs have also been used to derive precise quantitative models of a small regulatory cascade in early \textit{Drosophila} development \[44\]. This work has demonstrated the potential for extracting expression profiles from confocal images.

Here we present an approach to automatically obtain transcription factor expression levels from GFP confocal images in \textit{Arabidopsis}. In particular, we will consider
longitudinal images of the root region. We have chosen the Arabidopsis root as a model because it provides a distinctive spatial patterning of cell differentiation, allowing us to restrict the current analysis to 2-D cross-sections [5]. In addition, we also have a unique resource: tissue enriched microarray data at our disposition, which will provide a standard to validate our method [7].

In order to identify and correctly map root tissues, we employ image registration algorithms. Image registration is a very broad subject with many applications to biological and biomedical data [74, 120]. After a brief overview of the data available to us (Section 5.2), we describe the details of our registration process (Section 5.3) to map an image onto a representative model—in our case, a labeled tissue map of a model Arabidopsis root. In Section 5.3.3, we show that this method is capable of identifying and quantifying the expression profiles of 13 tissues in the Arabidopsis root, and we evaluate how well microarray and image-derived expression values correlate with each other. Section 5.4 addresses future developments and the implications of our results for the inference of regulatory mechanisms and pathways in multi-cellular organisms. An earlier version of this method was used in a large-scale study to assess the influence of post-transcriptional gene regulation on the expression of transcription factors [62]. This work differs by utilizing new methods that expand our identification from 4 to 13 tissues, and also allow for analysis of images taken from all regions of the root.

5.2 Data

5.2.1 GFP Promoter Fusion

Using the coding sequence for GFP, transcriptional fusion constructs were created by attaching the promoter region of the gene of interest (3kb upstream of the translation start site or the intergenic region—whichever is shorter) to the coding region of the GFP gene. In contrast to translational fusions—which incorporate the GFP as
a domain into the protein—transcriptional fusions function as a marker for mRNA expression. The constructs were inserted into the genome with the assumption that its transcriptional regulation will be similar to that of the endogenous gene. While this concept ignores some of the regulatory steps of gene expression (e.g. translational/transcription inhibitions, chromatin modification, etc.), previous work has shown that, in Arabidopsis, this type of construct recapitulates tissue specific gene expression with high fidelity [62].

5.2.2 Images and Image Selection

The Arabidopsis images depict optical longitudinal sections of transcription factor GFP constructs taken from all three main zones of the root: meristematic (primary root growth and location of initials), elongation (elongation of cell size), and maturation (root hair growth) region (Figure 5.1). The use of longitudinal images allows us to identify 13 tissue regions (Figure 1.1): columella root cap, columella initials, cortex, cortex initials, endodermis, epidermis, lateral root cap, lateral root cap initials, pericycle, pericycle initials, stele, quiescent center (QC), and vascular bundle (VB) initials. Tissues that cannot be distinguished in longitudinal images are the atrichoblast and trichoblast (epidermis), and the xylem and phloem (stele).

Images are composed of three channels: a red channel highlighting cell wall boundaries stained using a dye, a green channel containing the GFP expression, and a blank blue channel. Selection of images for comparison were based on the following criteria:

- The cell wall stain was strong on the external boundary and at least partially visible in the interior of the root.

- Roots were centered and not heavily skewed to one side of the image.

- Images were chosen from transgenic lines known to harbor detectable tran-
Figure 5.1: Three main regions of the *Arabidopsis* root: meristematic, elongation, and maturation.

122 images representing 30 transcription factors met these criteria. 64 of them expressed in the elongation/maturation region, and the remaining 58 expressed in the meristematic region. To segment roots into the 13 tissue regions, we used an atlas image which contains a tissue map for a representative *Arabidopsis* root. We created this atlas by fusing two high resolution images: one of the meristematic zone up to the elongation zone, and one from the elongation zone to the maturation zone. Within this composite image, we marked the tissue regions as depicted by a standard template (Figure 1.1).
5.2.3 Fluorescence Activated Cell Sorting (FACS)

The tissue-specific microarray data is collected using a Fluorescence Activated Cell Sorting technique [7]. *Arabidopsis* roots with GFP expression enriched for a particular tissue are run through a fluorescent activated cell sorter. FACS separates cells expressing GFP from non-GFP expressing cells, obtaining the enrichment of cells from an individual tissue. RNA from the sorted cells is then analyzed on a microarray, providing tissue enriched gene expression data. Eight tissue regions were common to both image registration and the tissue enriched gene expression data: columella root cap, cortex, endodermis, epidermis, lateral root cap, pericycle, quiescent center (QC), and stele. We will refer to these expression values as $T_{GFP}$ and $T_{FACS}$. The five tissues not present in the microarray data are the initials: columella initials, cortex/endodermis initials, lateral initials, pericycle initials, and vascular bundle initials. Due to lack of promoters specific to each type of initials, it is not currently possible to use FACS to isolate initials. Differences on what constitutes a specific tissue exist between the FACS and GFP data. First, the overall area that expression is averaged over differs: not all GFP lines used for FACS are expressed ubiquitously across all regions of the root, and the region the image is taken from might not encompass the full range that the FACS data is obtained from. Second, some GFP lines used for sorting have a partial inclusion of additional tissues that leads to a slightly convoluted FACS measurement for that tissue. Despite these minor differences (see supplementary data for full details), the regions are treated as homologous.

5.3 Methods

The system to quantify tissue-specific expression from images consists of three main steps. First, noise from the imaging process or normal root growth must be removed.
Second, roots are registered to a master atlas image. Third, using the registered image and the atlas image, GFP levels are quantified, and tissue-specific expression values are obtained.

5.3.1 Noise Removal and Contour Detection

Noise can result from the imaging process itself (such as blurring, the addition of speckle noise, etc.), or due to variability which naturally occurs in the root and which would present a difficulty during the later image registration stages. Morphological
operators such as image closing, restoration, and thresholding are applied on all channels to eliminate the general noise present from imaging.

Two types of root variability exist that provide difficulties with registering the images: boundary cells and root hairs. Boundary cells are lateral root cap cells that have detached themselves from the rest of the root. This is a natural process as the boundary cells provide the lubrication needed for the root to penetrate the soil. Once removed from the tissue layer, these boundary cells often adhere to, or reside in close proximity to, the outer cell walls—making it difficult to accurately detect the shape of the actual root. Unlike boundary cells (which have lost the connectivity and henceforth any cell signaling pathways), root hairs are a viable part of the root. Their occurrence along the longitudinal axis is not very predictable, which impedes any type of accurate registration process.

To remove these variabilities, we adapt a snake/active contour model. A snake is an iterative contour detection algorithms that can grow and shrink based on a set of force balancing equations [53]. We use an improved active contour model called a Gradient Vector Field Snake, or GVF [17]. Expanding on previous snake algorithms, a GVF snake is governed by two sets of forces: internal forces (such as elasticity/rigidness of the growing contour) and external forces (an external constant pressure force, viscosity, and a gradient vector field). With the exception of the gradient vector field, all parameters for these forces are user defined but kept constant for all images in the data set. A gradient vector field is a modification of a standard first order gradient edge map, in which the radius of the force field is increased. This causes it to extend its influence on the snake algorithm to areas outside those in close proximity to an edge.

For the gradient edge map, we use the external contour of the root. This contour is determined by performing a watershed segmentation [71] on the image—segmenting the individual cells and background into different regions. For our current image set,
we can safely assume that the region with the largest area can be labeled as background. In addition, regions with an area greater than one sixth and mean red and green intensity less than twice of the largest one are also labeled as background, to take cases into account in which the root partitions the background into two or more regions. Regions not fitting this criterion are considered part of the root. This creates a contour that contains all the root hairs/boundary cells which we seek to remove (Figure 5.2B, black outline). To initialize the starting contour for the snake algorithm, we perform a morphological erosion on the filled object. Due to differences in magnification/image size, this erosion is performed using a disk structuring element with variable size, which is set to 0.35 times the diameter of the root. This internal contour still contains some of the noise from the root hairs/boundary cells, but provides a smoother initial contour (Figure 5.2B, red outline). The algorithm is then run for a number of iterations adjusted to the size of root diameter (Figure 5.2C). By tuning the parameters of the snake algorithm (in particular, the elasticity/rigidness) once for our application, the final active contour can be adjusted so that it minimizes the amount of noise from boundary cells/root hairs. The resulting image is a clean smoothed external contour that does not contain root hairs/boundary cells (Figure 5.2D).

5.3.2 Registration

After removal of general noise and standardization with respect to boundary cells and root hairs, we can proceed with registering the images to an atlas image. Two major issues must be considered for this registration process. First, the roots grow in a curved fashion, bending and twisting in response to the environment that they are growing in. This poses a non-rigid registration problem. Second, the images in the data set show different regions of the root, corresponding to a partial mapping problem. Non-rigid registration requires the calculation of complex transformation
fields and cannot be solved by simple affine movements (e.g. scale, rotation, translation, etc.). Partial mapping requires the registration of an image under missing information—i.e. with occlusion of the object, or images showing only parts of a complete object (which is the case here). These two problems of non-rigid transformation and partial mapping are often mutually exclusive, and as such, we deal with them separately.

Non-rigid Registration

To allow roots to be aligned to a master atlas image, they need to be straightened. In order to do this, we will use a non-rigid transformation algorithm called Thin PlateSplines (TPS) [10]. TPS is a transformation function which is derived from the physical bending energy of thin plates. TPS require a set of homologous points between a standard image and a reference image. A deformation field can be created based on the distance between the homologous points. This deformation field is then applied to every pixel in the entire image.

Successful straightening and registration using TPS is highly dependent on the set of homologous points that one chooses. Determining which, and how many, points to use for this mapping is referred to as feature detection. Successful feature detection for image registration requires that features are easily identifiable and abundant. This is often a problem with biomedical data, as the images do not contain features that fit these criteria [120]. In our case, however, after the removal of boundary cells and root hairs, the contour of the root provides a source of features that fit both of these criteria.

Given the situation of partial mapping, as is the case for the root images, we encounter the problem of how to define a set of homologous points for the TPS. We address this by choosing a set of unique points for every image. This total feature set is then used to automatically derive a new set of points describing a straight root.
Our feature set will contain two groups of features: a modification of the major axis endpoints (which we will refer to herein as the extreme points), and pairs of cross-sectional cuts, which are defined as the locations on the contour that result from the orthogonal bisection of the medial axis of the root.

The medial axis transformation (MAT) algorithm can provide us with the knowledge to derive these features [83]. The MAT function determines the medial axis by calculating the distance between every point in the interior of the object to the contour elements. The minimum distance of an internal point to its closest contour element is defined as the Voronoi distance. When the set of Voronoi elements is greater than one, i.e. the shortest distance to the contour is shared by two or more contour elements, it is considered part of the medial axis. To adapt for the partial occlusion of the images, the end points of the external contour are determined and the image is extended to create a new image three times the size of the original image, with the original being centered in the middle. Starting at the end points of the contour, we extend it into the added regions. Unlike traditional replicate padding of images, the extensions of the contour provide a better estimate of our expected shape of the root, and henceforth, a more accurate MAT estimate. While our initial noise removal algorithm is efficient in removing contour noise, the MAT is very sensitive to perturbations in the boundary of the object (e.g. natural distortions, small noise), and an additional step of pruning the MAT is required for removing any small branches. The extended regions are then removed.

Using the MAT, we can calculate a set of cross-section pairs. The MAT is treated as a continuous contour, and its curvature angle is calculated by using a standard first order derivative. Orthogonal lines are then drawn from the MAT, intersecting with the smoothed contour of the root. The intersection occurs on both sides of the root providing us with a pair of points where each point consists of an x and y component denoted as \([x_i, y_i]\). We will refer to this as a cross section pair: \(p_i = \{l_i, r_i\}\).
where \(l_i, r_i\) are the points on the left and right side of the contour respectively and \(P = \{p_1, ..., p_n\}\) is the set of all cross section pairs (Figure 5.2E). The curvature of the root contour, in combination with the partial imaging of the root, leads to a subset of the cross-sectional pairs being incomplete as one of the pairs is occluded by the imaging process. To eliminate this abnormality, the location where pairs become occluded is determined for both ends of the roots. All regions of the root beyond this threshold are removed, resulting in an adjusted contour with blunt cut ends for those sides where occlusion was present.

This process leads to the set of cross-sectional cut points which we subsequently use to determine the remaining features—the extreme points of the root. Most major-axis algorithms for determining extreme points of objects are not appropriate here due to both the nonrigidness as well as the partial occlusion of the roots. The medial axis is trimmed to 70% of its normal size to eliminate small perturbations occurring at the ends. Using both end points of the trimmed medial axis, the extreme points are extrapolated to intersect with the new adjusted contour of the root labeled as \(E = \{e_t, e_b\}\)—the top and bottom end points respectively. These intersection points represent the extreme points of the root, and by definition, partition the root into two separate sides (Figure 5.2F).

Given this feature set selected from a given image, we can proceed to create a homologous mapping to an approximately straight root. The original set of points can be separated into two groups: the extreme points of the image \(E = \{e_t, e_b\}\) (teal points Figure 5.2F), and the pairs of cross-section pairs \(P = \{p_1, ..., p_n\}\) (green dots Figure 5.2E). We define a new set of points \(M\) based on their location along the root, as the ordered set of the middle point between each cross-section pair \(c_i = \left[\frac{1}{2} (l_{ix} + r_{ix}), \frac{1}{2} (l_{iy} + r_{iy})\right]\), and the two extreme points defined above, leading to \(m \in M = \{e_t, c_1, ..., c_n, e_b\}\). In our first step in deriving a new set of points, we define two distance functions \(D_1\) and \(D_2\). The first function, \(D_1\), determines the distance
between the ordered middle points and the first extreme point $D_1 (i) = |m_i - e_t|$. 

The second distance function, $D_2$, measures the distance between the middle points and their respective cross-sectional pairs: 

$$D_2 (i) = \frac{1}{2} |l_i - m_i| + \frac{1}{2} |r_i - m_i|.$$  

Using these middle points and distance functions, we can now define a new set of straightened points $E^*, P^*, M^*$. We additionally use a parameter $a_x$, defined as the medial x-axis location of the image. Starting from the point furthest away from the tip of the root, we set the x-coordinates of our homologous middle points to this medial axis $m_{ix}^* = a_x \forall m^* \in M^*$. The y-coordinates for these middle points is
determined from the first distance function \( m_{iy}^* = D_1(i) \). This maps the new middle points along the medial x-axis, separating them by the same distance between their original middle points. The location of each pair of cross-sectional points is then determined by translating each point in the pair by an equal distance in opposite directions from the medial x-axis; \( l_{ix}^* = a_x - D_2(i) \) and \( r_{ix}^* = a_x + D_2(i) \), and the y-coordinate is set to its middle point \( l_{iy}^* = r_{iy}^* = m_{iy}^* \), creating a new set of pairs \( P^* = \{p_1^*, ..., p_n^*\} \). The resulting set of points \( F^* = \{e_t^*, p_1^*, ..., p_n^*, e_b^*\} \) provides a mapping from our original image \( F = \{e_t, p_1, ..., p_n, e_b\} \) onto a straight root (Figure 5.2G).

This homologous mapping of points finally provides us with the information needed to perform the TPS transformation. A transformation field is calculated from these sets of points and is applied to every pixel in the image. Due to the complexity and memory requirements of the TPS function, the two extreme points and a subset of the cross-sectional pairs (15 equidistantly spaced pairs) are used. The resulting image has eliminated most of the curvature and non-rigid abnormalities that exist on the contour of the root (Figure 5.2).

**Partial Mapping**

The final registration process addresses an affine registration between two images: the straightened root we have just obtained from the TPS registration, and the master atlas image which provides the tissue label information. An affine registration consists of minimizing a scoring metric. Traditional affine registration parameters encompasses rotation, skew, scale (both in x and y), and translation (both in x and y). In the process of root straightening, the TPS has already restricted the transformations required to register the root. Fixed along the center of the image with the root tip pointing to the bottom of the image, rotation, skew, and translation in x have already been determined. We assume that the scale is the same in both the
Figure 5.4: Comparison of image registration (x-axis) and FACS correlation (y-axis) scores. Arabidopsis Gene Index (AGI) identifiers are given. The figure shows registration scores between 0.65 and 1.0 and correlation scores between -0.6 and 1.0. No other data points fell outside this region.

x and y coordinates and will be treated as one parameter. The remaining degrees of freedom are then the translation along the y-coordinate and the scale of the image.

The affine scoring metric is motivated by the Hausdorff score for partial mapping [43]. The Hausdorff scoring metric is formally defined as:

\[ H(A, B) = \max (h(A, B), h(B, A)) \]

where

\[ h(A, B) = \max_{a \in A} \min_{b \in B} (|a - b|) \]

with \(|a - b|\) being the distance between a and b and \(A = \{a_1, \ldots, a_p\}\) and \(B = \{b_1, \ldots, b_q\}\) are two sets of points. Here, we modify this Hausdorff scoring metric to be

\[ H(A, B) = \sum_{i} h(A_i, B_i) \]
where $i$ denotes a subgroup of points and $A$ and $B$ are points in our image and atlas respectively. The first group is the subset of points we used for the TPS straightening, i.e. the contour of the root $A_1 = P^*$. For images taken from the meristematic region, the second group is a single point denoting the center of QC that is placed by an expert after the TPS process ($A_2 = Q^* = \{q^*\}$). This placement of a marker is currently necessary, as the internal cell staining is not robust enough to automatically determine its location. This leads to the full set of points: $A = \{A_1, A_2\} = \{p_i^*, ..., p_n^*, q^*\}$. Identical contour and QC markers are pre-determined and marked in the atlas image $B = \{B_1, B_2\} = \{p_i, ..., p_n, q\}$. A number of numerical optimization algorithms are appropriate; here we use a Particle Swarm Optimizer (PSO) [54]. We limit the range of scale values from 0.1 to 10, and translation values from 0 to 2500. The optimization converges in less than 400 iterations (Figure 5.3D).

5.3.3 Quantification

We now have two images with identical dimensions, from which we proceed to extract expression values. The first image is the expert created atlas image describing the tissue map of the root (Figure 5.3C). The second is the result of the affine registration, with the green channel detailing our gene expression values (Figure 5.3D). For every pixel in the registered image, the intensity value is summed and binned according to the tissue map as defined by the homologous point in the atlas image. The thirteen tissue expression values are normalized by dividing the total expression values by the total area that each tissue region occupies (Figure 5.3E). This helps to normalize against the occlusion of certain tissue layers due to imaging and provides standardization similar to microarray data.
5.3.4 Image Processing/Data Analysis

Most of the image analysis was carried out using the Matlab Image Processing Toolbox (IPT), with the exception of the Thin Plate Splines algorithm [27] as well as the GVF Snake algorithm. The Hausdorff partial mapping and Particle Swarm Optimization are implemented in C#. Statistical analysis was performed using the R Statistical Language [93].

5.3.5 Results Scoring Metric

Registration Scoring Metric

To determine the accuracy of the registration process, we modify a commonly used accuracy measure called Test Point Error (TPE) [120]. TPE measures the accuracy of the registration process by creating a set of homologous points with the atlas image that are not used in the registration process itself, but are used as an accuracy measure for the registration process. Our modification to the TPE does not use fixed points themselves, but instead marked regions. These marked regions are a subset of the total cells in the image, manually chosen based on their clearly distinguishable cell boundaries by an expert. As a result, not all tissue regions may be marked due to difference in quality of staining and localization of images (8 tissues are specific to the meristematic region and are not present in elongation/maturation images). Our scoring method is formally defined as \( \frac{\text{total}}{\text{matched}} \) where

\[
\text{matched} = \sum_{i=1}^{S} I(a_i = b_i)
\]

allowing I to be the indicator function equaling 1 if \( a_i, b_i \) are equal and 0 otherwise and \( \text{total} = \text{size}(s) \) with \( i \in s \) if \( a_i, b_i \in [1, 13] \). The numbers 1-13 refer to a unique tissue, and 0 to no tissue mapping available (either a cell wall or a region outside the root).
FACS Scoring Metric

The quality of GFP derived expression values will be assessed by comparison to respective FACS microarray data. By treating each data source as a random variable where each tissue is a sample (i.e. $X = T_{GFP}$ and $Y = T_{FACS}$), we can calculate the Pearson correlation value between the two data sets for every image. Comparing the correlations on the level of each image is required, as variations in gain, laser power, and pinhole settings between images (used to obtain maximum visual contrast) prohibit large scale correlation calculations. Images from the elongation and maturation zones did not contain GFP measurements for three of the tissue regions (QC, columella root cap, and lateral root cap), and both FACS and GFP measurements for these tissues were not included in the Pearson Correlation calculation.

5.4 Results and Discussion

We applied this prototype to a data set of 122 GFP images depicting the expression of 30 transcription factors in different regions of the root. For 7 out of the total data set of 122 images, the system was unable to eliminate noise and perform the straightening. This was due in most part to boundary cells/root hairs being present at the edges of the images—a known limitation of our noise removal algorithm. For images requiring QC marking, we were unable to unambiguously locate the QC region in 5 of the images. The remaining 110 images were passed on to the second phase of registration and quantification.

Figure 5.4 shows the scatterplot of registration score on the x-axis and FACS correlation score on the y-axis. The majority of the images were successfully registered to the master atlas image: The mean registration score was 0.93, i.e. only 7% of the root is mapped to the wrong tissue type. The FACS correlation scores had a mean of 0.64. Considering that with exception of the QC marker, no root
feature was manually marked to help register the images, the high accuracy of the registration process is very encouraging. It is notable that low registration scores do not necessarily lead to bad FACS correlation values: a mis-registration of a tissue layers may occur in a location where there is no GFP expression and have no effect on the correlation score.

While the FACS correlation scores had an overall good average of 0.64, a portion of these were rather poor. It is apparent from Figure 5.4 that several of these lowly correlated values are clustered within image groups of the same gene or line, suggesting potential issues with the promoter fusion of the GFP reporter constructs, or with the probe for the FACS data. The mean correlation score of 0.64 is reminiscent of previous studies for expression analysis, where it was found that correlations between platforms varied from 0.46 to 0.83 [56, 86, 110].

**Lowly expressed genes.** Poor correlation scores between platforms are frequently attributed to various sources of noise in different expression analysis platforms, and are increasingly observed for low expression values. Limiting our correlation calculation to FACS data where the median tissue expression is greater than 150 (used as noise threshold for *Arabidopsis* microarray data [62]) we increase our mean correlation to 0.70, thus reaffirming that lowly expressed values are more likely to have a negative effect on our correlation scores. In addition to the standard noise conditions present in microarray experiments (e.g. hybridization, background fluorescence, probe ambiguity etc.), a potential source arises from the FACS sorting of the data. While traditional microarray experiments use one sample per experiment, our data set requires 8 different sorting and microarray experiments, increasing the likelihood for biological variability.

Part of the noise in the FACS comparison originates from our approach to quantifying expression levels. Autofluorescence in the plant results in a mean background noise of 5 – 20% of the maximum possible intensity value in the green channel. On
examining the images with the lowest correlation values, we noticed that a majority of these were from images with GFP expression levels barely above this autofluorescence level, suggesting that the correlation values were skewed by the background in the green channel. In addition to this background noise, our method did not take into account the attenuation due to depth which affects the inner tissues of the root, such as the endodermis, pericycle and stele. An initial simple approach to normalizing these regions by multiplying the tissue types by 1.3, 1.4 and 1.5 respectively increased our mean correlation score from 0.64 to 0.7. A comprehensive study of this problem may therefore lead to a more systematic correction of GFP derived expression values.

Taken together, this suggests that improvements in increasing the signal to noise ratio in the green channel are paramount for the GFP quantification of lowly expressed genes.

**Tissue-specific expression differences.** The remainder of the lowly correlated genes suggests some inconsistencies in our images as well as in our approach in normalization of the data. Failure to differentiate between expression in the pericycle
and endodermis can lead to low correlation values between the data sets. This occurs when large pinhole settings during the imaging lead to longitudinal images showing expression in both tissue regions, but where radial images show expression in only one. This problem tends to occur in a subset of the images for a given gene. Visual inspection confirmed that low correlation values for some genes expressed above background were not caused by problems with the image analysis, but by actual differences in expression values as reported by microarrays and GFP. Such differences can originate from reporter constructs which do not fully recapitulate the expression of the native gene, or due to discrepancies in the tissue-specific expression data. In either case, our system can serve as helpful resource to point out and quantify such problems.

**Cells versus tissues.** Finally, the current normalization of expression by total tissue area can blur expression which only occurs in a subset of the tissue. A representative example was the gene AT2G37590, where GFP expression resided in a subset of the stele region as compared to uniform expression across the whole tissue (Figure 5.5). When its expression was averaged over the whole stele, it barely exceeded the background noise level. It should be noted that microarray data is prone to the same issue—in the example, it showed expression in the stele at a value of 179, again barely above the background threshold of 150.

### 5.5 Summary/Outlook

In this chapter, we have presented a system for the automated quantification of gene expression levels from digital images of GFP reporter constructs. As a proof of concept, we successfully performed an automated registration of *Arabidopsis* roots, derived tissue-specific expression values of transcription factors, and demonstrated that these values correlate well with microarray expression data. The data set used for this evaluation was only of modest size. However, the number of images in
gene expression databases of other model organisms [105] is easily on the order of
tens of thousands, which demonstrates the growing need to adapt image analysis to
problems in computational biology. In addition to the biological significance of our
methodology, we have presented a unique approach to both a partial mapping as well
as a non-rigid registration problem. The combination of these two problems often
requires one to manually annotate images prior to registration.

Developing a universal method for image registration across all types of images
is considered an intractable problem [120]. Image registration often utilizes domain
specific information—incorporating unique modifications in the image registration
process to adapt for differences that are inherit to a specific set of images. In our
model, we have adapted methods in noise removal, feature detection, and feature
mapping that are specific to the elongated, symmetric shape of Arabidopsis roots.
However, it is expected that the series of algorithms used in our approach will be
useful for other confocal images, particularly for approximately symmetric objects.

Many of the difficulties with quantification of GFP (attenuation to depth, large
pinhole settings) can be addressed by expanding our work to 3D. Efforts are under
way to scale up microscopy and imaging from one 2-D cross-section to a stack of
images. All image processing algorithms used in our system were chosen because they
offer adaptations to 3D image processing. In addition to the increase in precision for
our quantification, the expansion to 3D will also allow us to differentiate our stele
measurements into xylem and phloem, as well as our epidermis measurements into
atrichoblast and trichoblast.

We note that our current method requires a step of manually marking the QC
region in images taken from the meristematic region. Improvements in cell wall
staining will likely allow for the automatic detection of this region using image seg-
mentation algorithms, such as the watershed algorithm, to identify individual cells.
Current attempts to automatically identify this region are not robust enough given
the present staining technology. This is exemplified by the fact that several of the QC regions could not even be manually annotated by an expert and were subsequently removed from the analysis. An alternative here is to use a second fluorescent marker which is constitutively expressed in the QC cells. Adequately registering images on cellular resolution will also allow us to identify differential expression of genes within tissues (cf. Figure 5.3 (yellow and green framed images) and Figure 5.5).

Our system was evaluated using the 8 tissues common between the image and microarray data sets. In total, our image analysis identifies 13 unique tissues. As mentioned, the five tissues not common are the initials of the root, for which it is currently not possible to obtain specific microarray data. This is another example where our method can truly complement available expression data for the understanding of *Arabidopsis* development.

We have chosen to validate our model with a series of images taken at arbitrary time points. However, the largest benefit of using GFP reporters and automated image processing for expression analysis is that expression can be monitored in a living organism. In the long term, we plan to further develop our system to be part of an anticipated large scale effort to study transcription factor expression in root development under a variety of environmental conditions. As such, a system to quantify GFP expression values will provide the basis for the computational biology of spatiotemporal gene expression [4] and the high resolution elucidation of transcriptional regulatory networks.
Automated High Throughput Extraction and Comparison of Gene Expression Patterns from \textit{Drosophila} RNA In Situ Hybridization Images

Since the Rootarray is an ongoing project that has only recently begun to generate data, I will use \textit{Drosophila} images to address some of the final problems associated with image data: developing fully automated methods for high throughput image data. To deal with these types of data robust methods must be developed for accurately registering and quantifying the expression from a large database of images (\textasciitilde 70,000 images). In addition, it is important to develop methods for comparing expression patterns between genes that account for the spatial dependencies. I conclude this chapter with methods for calculating the significance between two spatial gene expression patterns.

6.1 Introduction

Advances in high-throughput microscopy have lead to a rapid increase of digital image data related to molecular biology. This large volume of images is a new data
source for computational and systems biology for which the development of computational approaches has only just begun. Methods for the automatic processing and analysis of single-cell data are probably best developed so far and have allowed the efficient detection of phenotypic changes in response to the knockdown of genes or exposure to chemical compounds [13, 69].

Recent developments in high throughput microscopy techniques have now provided data sets with spatiotemporal expression patterns for thousands of genes [42, 63, 105]. For instance, images of tens of thousands of RNA in situ hybridization experiments from multiple time stages of developing *Drosophila* embryos are now available. In the past, these images have been analyzed in a manual fashion, through qualitative visual inspection methods. In order to deal with such data sets more appropriately, it is necessary to develop fully automated methods for extracting and analyzing these images. Methods to quantitatively describe spatiotemporal expression patterns is a relatively new area of research that has just begun to be explored in model organisms [78]. In previous work applied on *Arabidopsis* root images, we used image registration techniques for the identification and quantification of tissue specific expression from GFP reporter lines [73, 62]. In sea urchins, a comprehensive analysis of normalization methods was used for determining and quantifying spatial expression data [97]. Recently, large databases of images for *C. elegans* have become available [42], accompanied by methods to analyze the data [3].

Traditional approaches to measure gene expression, e.g. based on microarrays, do not allow to measure expression at high spatial resolution, a critical feature of genes in a developing organism, where spatial patterning can be complex. A particularly interesting problem which we can address with the availability of image data is thus to compare genes on the level of their spatiotemporal expression profiles at high resolution, e.g. of the same gene under different conditions or across different species, or to cluster different genes akin to approaches developed for microarray data.
Several general questions arise when comparing spatiotemporal image expression data: (1) We need to develop methods to process the raw input images, to eliminate noise under a typical large range of imaging conditions (e.g. different viewpoints, different locations, multiple specimens per image) and to perform normalizations to decouple the variability in morphology from the variability in expression; (2) we need to represent the expression patterns and to specify appropriate similarity metrics capable of assessing spatiotemporal similarity; (3) and we need to assess the significance of observed similarities.

Computational methods for analysis of biological data have often been applied to static data sets that lack space-time dependencies. These methods are unsuitable for a proper analysis of image data, where the dependency between neighboring regions within an image require the use of appropriate spatial models. Similarly to models for time series data such as hidden Markov models, spatial models seek to expand traditional single- and multivariate models by incorporating dependencies between neighboring spatial regions (to capture qualitative topological dependencies such as coherent observations in discrete subregions, as well as quantitative dependencies such as gradients). Spatial statistical models have become an increasingly popular field within the past 10 years, partially due to an increase in computational power to handle their complexity, but also driven by the increased awareness in application areas such as environmental issues (global warming, climate prediction).

The majority of work in image analysis work in the context of development of model organisms has been carried out for *Drosophila*, and can be broadly grouped into two categories: fine-grained high resolution analysis of a relatively small selected subset of genes [46, 55, 30], and higher-throughput pattern analysis of thousands of genes ([89, 59, 39]). This study is set in the category of high-throughput analysis of expression images, and we use a *Drosophila* embryonic dataset to illustrate how to deal with the three questions outlined above. Our work distinguishes itself from
previous work on *Drosophila* data sets by three main contributions. First, robust
and fully automated image analysis techniques are used to process the raw images as
described in section 6.2.2. Through the use of statistical shape models and partial
mapping methods, these techniques are capable of handling sources of phenotypic
and imaging variability that limited previous approaches. Second, we comprehen-
sively compare different similarity metrics as described in section 6.2.3. We propose
to use appropriate similarity measures that incorporate spatial dependencies to dis-
tinguish complex spatial patterns, and validate the different measures against visual
annotation terms provided by experts. Third, we develop a new significance test-
ing framework for spatial similarity scores through constrained realization Monte
Carlo simulations in section 6.2.4. We validate this method of significance testing
on known biological examples, emphasizing the usefulness for fully automated image
registration/comparison models in the context of inference of regulatory interactions.
While we use a fly embryo dataset as example, we expect that the general approach
is adaptable and thus of interest to the study of expression patterns in a wide range
of model organisms.

6.2 Material and Methods

6.2.1 Datasets

*Berkeley Drosophila In Situ Database and Visual Annotation Terms*  The Berkeley *Drosophila*
in situ database consists of 78,621 images of 3,724 genes expressed in *Drosophila* em-

bryos across 6 groups of time windows (covering the developmental stages 1-3, 4-6,
7-8, 9-10, 11-12, 13-15). A high throughput RNA in situ hybridization staining
protocol was used to visualize spatial expression patterns as described in [105]. Vi-

sual annotations are based on an ontology describing embryonic expression patterns,

consisting of 314 annotation terms, and were obtained from the Berkeley *Drosophila*
Genome Project group [106]. The annotation set was curated by the BDGP group
by manually inspecting the in situ images and providing ontology terms for each image at every time stage.

6.2.2 Image Registration

The approach for segmenting and registering images is based on statistical shape models using signed distance maps to describe object contours [64]. Signed distance maps are a representation of contours which contain the distance from the contour of the object for every point in the image: negative distance values depict regions that are inside the object, positive distances are outside, and the magnitude represents the actual distance 6.2. Signed distance maps are an attractive choice for shape modeling as they provide a continuous representation of a discrete space which is easily interchangeable (the signed distance map can be directly converted from the contour, and the contour can be determined from the signed distance map by calculating the zero crossing of the distance map).

A Drosophila shape model was semi-automatically created from 120 embryo images. First, the contours of the embryo were manually segmented and transformed into signed distances maps. The objects were then automatically normalized in size by minimizing the distance of each individual signed distance map to the mean signed distance map. The resulting normalized maps were analyzed using a hierarchical Principal Component Analysis (PCA) decomposition [48]. Let $X$ be the set of all training images, and $X^b \subset X$ be the subset of training images belong to time stage $b$, where $b \in B, B = [1, 6]$, and $X^i \cap X^j = \emptyset$ if $i, j \in B, i \neq j$. In standard PCA, a new set of bases $w_t \in W$ is selected such that $w_t = \arg \max \var{w_t x_{t-1}}$ where $x_{t-1}$ is the deflated matrix from the previous iteration. Each vector $w_t$ can then be converted into two dimensional signed distance principal component image. Hierarchical PCA extends upon PCA by normalizing the contribution of each basis to each individual block. This is accomplished in an iterative fashion by calculating
Figure 6.1: **Shape Model** (a) Example image of the creation of the training set. Prior image data sets are curated into two components: the filtered pixel valued images, and the manually segmented contour of the object. The signed distance map can be calculated directly from the external contour. Areas in red denote increasingly negative values, while those in blue depict increasingly positive values. (b) A subset of the 120 images used for the training of the shape model. The external contour is projected on top of the original image. Regions in red depict areas internal to the object, while those in blue are external. (c) The training set is then normalized in size (d) The resulting contours are converted into signed distance maps and processed using a hierarchical principal component analysis. Four of the main principal shapes of the embryo are shown. These images depict 2 standard deviations of the principal component from the mean of the signed distance map.

and normalizing the $w_t$ by each individual block basis $w_t^b$ leading to matrix of basis $W$ as described in [48].

In addition to providing characteristic priors on the shape through hierarchical PCA, we also model the filtered intensity values for the *Drosophila* embryos. We create a set of histograms of the intensity values by binning observed intensity values by their respective signed distance. We concentrate on the intensity values close to the contour, i.e. we bin the intensities observed in distances from -25 to 25 in 1 unit increments, while remaining bins are not included.

The task of image registration is to find the optimal set of parameters, $\theta$, such that the images are accurately aligned to a common frame of reference. The parameters
we choose to optimize over can be separated into two categories: rigid transformation parameters of the image $\theta_r$, and the principle shape components of our shape prior $\theta_s$. The principle component parameters allow the underlying shape model to take on a variety of possible shapes, which is defined by the signed distance maps. The rigid transformation parameters are used to rotate, translate, scale, and flip (horizontally and vertically) each image so that it maps onto the evolving shape model. These two sets of parameters are simultaneously optimized using an in house version of a Particle Swarm Optimizer \[54\]. By using the shape parameters as well as the empirical histogram values, we can assess how well the image is registered with the following metric:

$$f(\theta_r, \theta_s) = \alpha_1 g(s, g) + \alpha_2 h(s, i)$$  \hspace{1cm} (6.1)

$s = \theta_s W$ is the signed distance map created from the linear combination of signed distance principal component images. $g(s, g) = \sum_{j=1}^{n} \left( \frac{1}{1 + s_j^2} (s_j - g_j) \right)$ is an extension to Leventon's original scoring function, and $h(s, i) = \sum_s p(s_j, i_j)$ is the empirical probability of observing a pixel value of $i_j$ at the signed distance location $s_j$ over all $s_j \in (-25, 25)$.

Our method differs from previous methods with three main distinctions. First, the effect of the intensity distribution is limited to the area around the zero crossing of the evolving shape. This is necessary to limit the contribution of internal staining and multiple/impeding embryos to the score. It also allows a narrow band approach to be used with minimal effect on the overall score, significantly increasing the runtime speed. Second, the variability in size of the objects is normalized prior to creating the shape model. This is required to provide accurate representations of the actual differences in the shape of the \textit{Drosophila} embryo not simply related to size. The size of the embryo then becomes an additional parameter in our optimization step.
Lastly, we provide an additional term that is based on the direct values of the filtered image $h(s, i)$. This term is not decomposed into individual components, but rather describes the overall distribution of the filtered image based on the signed distance location. This allows the registration to accurately detect and align the shape model to the correct gradient of the image.

**Column Vectors** After registration, the transformed image and shape model are used to calculate column vectors of expression data. The bounding box of the object is determined from the optimized shape model, and the minimum and maximum $x$ and $y$ locations are calculated. The row and column vectors are then created by dividing the image into equally spaced rows and columns, and computing the mean pixel intensity value of the masked image for each column entry, resulting in a vector denoted as $c$. To allow a dyadic decomposition for the Haar wavelets, 64 columns are used throughout.

**Evaluation Score** We evaluate the accuracy of the registration and segmentation process by comparing the automated process to a manually segmented subset. Let $E$ define the set of all images, and $A \subset E$ be a randomly selected subset of 200 images. For each element $a_k \in A$, the column entries were calculated from both the manual segmentation and the automated segmentation as described in section 6.2.2. The accuracy of the segmentation was then assessed by calculating the mean square error between the manual segmentation and the automated segmentation (equivalent to the Mean Square Error defined in section 6.2.3). The accuracy score for an individual image $k$ is given as $u_k$; the vector of scores as $u$, and the median value of these scores as $\bar{u}$. To provide a frame of reference for these scores, we compare the accuracy scores of the automated and manual segmentations to randomly drawn images. For each randomly drawn image $a_k$ an additional random image was drawn $r_k$, where
$r_k \in R$ and $R \subset E, R \cap A = \emptyset$. The mean squared error between the automated segmentations of the random images was then calculated as $t_k = d_{MSE}^{a_k,r_k}$, where $t$ is the vector of scores. An accuracy score of the segmentation can then be calculated from these two sets of errors by calculating the ratio of the median values $S_{\text{ratio}} = \overline{u}/\overline{t}$.

6.2.3 Metrics and Evaluation Measures

Four metrics were evaluated to determine the similarity between images: standard mean squared error (MSE), mutual information (MI), Haar wavelets (HW), and spatial mutual information (SMI). MSE and MI are two frequently used measures for comparing vectors of data. Unfortunately, both make the underlying assumption that the samples within each vector are independent, which is clearly violated for spatial and time series data. On the other hand, interaction terms between individual elements of a sample are capable of describing higher order structures, such as the formations of gradients, or the alternating striped pattern of odd and even. These dependencies are relevant not only from the perspective of kinetics and molecular diffusion, but also in the interactions between genes. It is for this reason that similarity metrics that account for spatial dependency can be expected to provide more biologically relevant measures for comparing images. We implement two such measures, Haar wavelets and spatial mutual information, which are the spatial counterparts of mean squared error and mutual information, respectively.

Mean Squared Error We implement a mean squared error scoring metric between two genes, $a$ and $b$ (used e.g. in [67] to compare RNA in situ brain images). For each column we sum up the difference between the elements as defined:

$$d_{ab}^{MSE} = \frac{1}{n} \sum_j^n (C_{j}^a - C_{j}^b)^2$$

(6.2)
**Mutual Information**  We use the standard mutual information, e.g. as defined in [100]:

\[
d_{ab}^{MI} = H(A) + H(B) - H(A, B),
\]

(6.3)

where \( H(A) \) and \( H(B) \) define the entropy of each variable (or in this case, the column vectors for each gene) defined as:

\[
H(A) = - \sum_{i=1}^{M_A} p(a_i) \log p(a_i)
\]

(6.4)

And \( H(A, B) \) is the joint entropy:

\[
H(A, B) = - \sum_{i=1}^{M_A} \sum_{j=1}^{M_B} p(a_i, b_j) \log p(a_i, b_j)
\]

(6.5)

Additionally, we used a Parzen window kernel density estimate for the mutual information calculation.

\[
p(x) = \frac{1}{N h} \cdot \frac{1}{\sqrt{2\pi}} \cdot \sum_{i=1}^{N} \exp \left( -\frac{(x - x_i)^2}{2h^2} \right)
\]

(6.6)

Where \( h \) the bandwidth parameter and controls the smoothness of the estimate. This window allows us to provide a smoother and more robust mutual information calculation when dealing with sparse data.

**Haar Wavelets**  Wavelet analysis allows one to simultaneously examine the frequency and resolution components of a signal. This is accomplished by iteratively decomposing the signal with high and low bandpass filters. The low frequency filter serves the purpose of downscaling the image into progressively smoother dyadic scales. Let \( s \in (1, 6) \) denotes the dyadic scaling factor of the low frequency filter. For the Haar wavelets, this low pass filter can then be written as: 

\[
L_i^s = 0.5 \cdot \left( L_{2i-1}^{s+1} + L_{2i}^{s+1} \right),
\]
where $s$ denotes the scale level of the image, and $i$ represents a specific spatial location. The high pass filter is responsible for creating the coefficients of the wavelet which ultimately will be used to represent the similarity between images. The high pass filter is written as: $H^s_i = 0.5 * (L^{s+1}_{2i-1} - L^{s+1}_{2i+1})$. The Haar distance can then be calculated using the Mean Squared Error between two images $A$ and $B$: 
$$d_{A,B}^{HW} = \sum_s \sum_{i \in s} (H^s_i - H^s_i).$$

**Spatial Mutual Information** Spatial mutual information is an expansion of mutual information as defined above to include neighboring relations. Instead of defining the entropy as the one dimensional probability of observing an event (or in this case, an expression value), we instead define it as the joint probability of observing an event $i$ while its neighboring expression value is $j$. The entropy values for each variable (gene) then become the two-dimensional entropy we have already seen in equation 6.5:

$$H(A, \hat{A}) = - \sum_{i=1}^{M_A} \sum_{j \in N_i} p(a_i, \hat{a}_j) \log p(a_i, \hat{a}_j)$$ (6.7)

where $N_i$ are the neighbors of $i$.

Describing the range and extent of a neighborhood relation is a difficult problem and is often refereed to as the Change of Support Problem (COSP) [32]. Choosing a very small spatial range will result in little change between neighboring elements (all elements of the two dimensional histogram would lie on the diagonal); with too large a spatial range, sharp gradients and pairwise patterns will be lost. We chose to use neighboring elements of distance 2. These values were chosen as they preserve biological significant spatial patterns (local gradients; patterns such as observed for pair-rule or gap genes), but are still computationally tractable.

The cross gene comparison then becomes a 4-dimensional joint entropy:
\[
H(A, \hat{A}, B, \hat{B}) = - \sum_{i=1}^{M_A} \sum_{j \in N_i} \sum_{k=1}^{M_B} \sum_{l \in N_k} p(a_i, \hat{a}_j, b_k, \hat{b}_l) \log p(a_i, \hat{a}_j, b_k, \hat{b}_l)
\] (6.8)

With the spatial mutual information being:

\[
d_{a,b}^{SMI} = H(A, \hat{A}) + H(B, \hat{B}) - H(A, \hat{A}, B, \hat{B})
\] (6.9)

**Visual Annotation Term Assessment** For each scoring metric we evaluated the significance of the annotation terms as follows. A pairwise blocked matrix was created to represent the similarities between genes where \(d_{s,t}^{i,j}\) represents the distance from the \(s\)th image of the \(i\)th gene to the \(t\)th image of the \(j\)th gene. Let \(D_{ij}\) represent the distance between gene \(i\) and \(j\) where \(D_{ij} = \min(d_{s,t}^{i,j})\) for the MSE and Haar metrics and \(D_{ij} = \max(d_{s,t}^{i,j})\) for MI and SMI. For each annotation term \(k\), in each gene \(i\), we perform a Mann-Whitney-Wilcoxon test. Let \(U_{i}^{k}\) represent the Wilcoxon signed rank value for gene \(i\) and annotation term \(k\). The significance for each annotation term, \(k\), was calculated by taking the expectation of the U statistic \(E(U_{i}^{k}) = \frac{1}{n} \sum_{i} U_{i}^{k}\), and calculating its resulting zscore \(z^{k} = (E(U_{i}^{k}) - m_{U^{k}}) / \sigma_{U^{k}}\). The p values for each annotation were then calculated directly from the z scores.

**6.2.4 Biological Significance Testing**

Evaluating the significance of a scoring metric requires the generation of a series of surrogate data sets to be used for hypothesis testing. Here, we create appropriate surrogate data sets by drawing constrained realizations of the data [102] which account for both the marginal distribution of the intensity values as well as the joint spatial dependencies between neighboring images. By incorporating spatial dependencies, we are capable of generating surrogate data sets whose spatial complexity is representative of the underlying spatial processes; genes with high spatial dependencies (e.g.
smooth gradients), will result in a surrogate data set with similar gradient patterns, while those having low spatial dependencies (e.g. hard gradients, even-skipped), will have dissimilar patterns.

**Constrained Realizations**  Sampling surrogate data sets requires the use of a sampler capable of drawing values from both the marginal as well as the spatial dependencies. In order to meet these requirements, we implemented a swap sampler. For each column vector, we initialize every element by drawing a sample from the marginal distribution. This initialization provides us with column vectors that have no spatial dependency. To account for the spatial dependencies, an iterative swap sampler is used on this random initialization. Let \( p^1(a_i, a_j), p^2(a_i, a_k), p^3(a_i, a_l) \) be the probability of observing an intensity \( a_i \) while its first, second, and third neighbor elements are \( a_j, a_k, a_l \), where \( j \in N^1_i, k \in N^2_i, l \in N^3_i \) are neighborhood association sets. Let \( p(a_i, .) \) be the cumulative probability of observing \( a_i \). For each iteration, four random locations are chosen: \( K = k_c, k_d, k_e, k_f \) with observed values \( L = o(k_c), o(k_d), o(k_e), o(k_f) \) and \( h : K \leftarrow K \) be a permutation of the locations with \( h = h_1, ..., h_n \) being the set of all permutations of \( K \) with \( h_i(k_c, k_d, k_e, k_f) = H^i = \hat{k}_c, \hat{k}_d, \hat{k}_e, \hat{k}_f \) being the permuted locations with observed values \( M_i = o(\hat{k}_c), o(\hat{k}_d), o(\hat{k}_e), o(\hat{k}_f) \). The probability of a swap is calculated for all permutations (including the identity permutation) and the most probable swap is chosen. By iteratively resampling from this distribution, surrogate data sets are created that account for both the marginal as well as the spatial dependency structure as shown in figure 6.4b.

**Significance Testing**  40 constrained realizations are created for each gene. The background similarity value is then constructed by calculating all 1,600 pairwise similarity values between the constrained realizations for each gene denoted \( b^*_{i,j} \) for
$s \in (1, 1600)$. The empirical p-value of the expression between gene $i$ and $j$, $p_{i,j}$, is then calculated by counting the number of background scores greater than the observed score $|b_{i,j}^s > d^S MI_{i,j}|$, and dividing it by the number of elements $p_{i,j} = |b_{i,j}^s > d^S MI_{i,j}|/1600$.

6.3 Results

*Image Registration*  Prior to any quantitative analysis of expression image data, it is necessary to normalize and register the images to a common frame of reference. The identification of specific regions within an image consists of image segmentation and registration, and is well established in the bio-medical imaging community. By first mapping different specimens such as fly embryos to a common reference, we can subsequently apply a large variety of methods for univariate and multivariate data analysis to perform cross-subject comparisons (between replicates, genes, time stages, or different strains or species). Image registration is often performed by creating an atlas image describing the prototypical locations of specific regions, and then aligning different subjects directly to the atlas image, or by a model-based approach which represents the typical variability as learned from labeled examples.

Using statistical shape models and improved numerical optimizers, we developed a fully automated registration method which we applied to the complete set of 78,621 images in the latest release of the Berkeley *Drosophila* in situ database. Overall, the approach is able to successfully address problems that limited the successful application of previous methods on the whole database, highlighting its ability of extracting embryos under a large variety of imaging conditions: multiple embryos or impeding boundaries, changes in lighting/microscope settings, and out of focus boundary regions. To more formally analyze the accuracy of this registration method, we randomly selected 200 images from the dataset for a more quantitative assessment. The method has a very low absolute failure rate of 1.5% (3/200). The most common
inconsistencies in registration resulted from incorrectly orienting the axis, and were observed in seven cases (3.5%) for the anterior-posterior axis, and in 13 cases (6.5%) for the dorsal/ventral axis.

After normalization and registration of the images, the expression patterns can be quantified. Here, a choice of abstraction level has to be made; instead of working with the complete pixel-based 2D patterns, it is common to map them to a smaller set of features representing subregions of the specimen. Like other approaches studying embryonic expression patterns [46], we chose to project the staining intensities along the anterior-posterior axis as described in section 6.2.3. We then addressed how accurate the registration process was with respect to this particular quantification. Each of the 200 images processed above was manually registered and segmented. The average mean squared error of the quantified expression between the automated and manual data was then calculated, resulting in a low error rate of 0.0015. To assess the significance of this error rate, the observed error was compared to a randomized dataset, leading to a low error ratio of observed versus randomized of 0.077 (cf. section 6.2.2).

**Correspondence of Images to Visual Annotations** Using the extracted expression patterns, we demonstrated the importance of using spatial metrics (Haar wavelets and spatial mutual information) by comparing their performance to two previously used non-spatial metrics (mean squared error and mutual information). To assess these metrics, we determined the correspondence of the similarity values computed by each metric to the visual annotation terms in the in situ database. We selected the 3,127 genes and 27,157 images contained in the window of developmental stages 4-6—this window was chosen to balance the frequency of spatially diverse expression patterns with the limitation that the annotation terms in later stages are derived from earlier time stage annotations. The window of stages 4-6 contained genes annotated with 38
unique annotation terms describing the spatial expression patterns. We removed all
genes within this time stage that had annotation terms describing no spatial variabili-
ity (e.g. ubiquitous, maternal, etc.), leaving us with 214 genes, 2,156 images, and 31
annotation terms. For each scoring metric, we calculate an enrichment significance
for each annotation term describing how often genes annotated with a particular
ontology term show the strongest similarity to genes annotated with the same term
(section 6.2.3). Using a p-value cutoff of 0.05, spatial mutual information performed
the best, with 21 of the 31 annotation terms being significantly enriched, while mu-
tual information lead to the second best result with 19 of the 31 terms. Both mean
squared error and Haar metrics lead to only 8 enriched terms. Most terms which
are not found to be enriched under any measure are to be expected: gap, pair rule,

![Figure 6.2: Automated image registration](image)

**Figure 6.2: Automated image registration:** (a) Examples of input RNA in situ
images. (b) Registration Flowchart: During an initialization step, gradient and filtered
images are calculated, and the parameters of the shape model are randomly chosen. The
parameters for the registration are then re-iteratively evaluated and optimized over for a
fixed number of iterations, using a particle swarm optimizer. The final parameters are then
used to transform the image and create the shape model as shown in (c). The optimized
shape models for the embryos in the example images are shown in red, superimposed on
the normalized input images.
and segmentation refer to a type pattern rather than a precise spatial localization; in addition, cellular blastoderm and yolk nuclei are describing subcellular patterning. Many of the remaining non-enriched categories are those denoted as *in statu nascendi*—or in the beginning phases of a differentiation process. Most of the genes annotated with these terms have very broad expression profiles at this stage and form consistent localized patterns only in later stages of development.

**Significance Testing** We have so far demonstrated the accuracy of our approach for high throughput image expression analysis, and illustrated how including spatial dependencies leads to an increased enrichment of annotation terms documenting expression patterns. However, the question of the significance of the similarity between any two given patterns remains and is a subject which has received little attention. Clearly, the significance of pattern similarity is highly dependent on the pair of genes under consideration: A gene with ubiquitous maternal expression will show overlap to any other pattern, whereas more restricted patterns such as those of gap genes have no overlap at all to many other genes.

We explore the issue of pattern similarity significance in the context of the segmentation network, using a similar set of the gap, pair, and segmentation genes used in [99]. The pair-rule, gap, and segmentation network consists of a set of genes with many direct regulatory interactions and shared functional roles; in many cases, these similarities in function/interaction are reflected in noticeable similarities of spatial expression patterns. We provide a quantitative measure of this similarity by computing significance values for the spatial mutual information metric which proved to most accurate (see above). The observed similarity value is compared to a null distribution that preserves spatial dependencies of the two gene expression patterns, as described in section 6.2.4 and shown in Figure 6.4a.

The significance values for the similarity scores from all pairwise comparisons are
shows in Figure 6.4b. Many of the genes who share functional roles are identified as being significant; for instance, *pdm2* and *nubbin* (also known as *pdm1*) are paralogs with highly similar functional roles and interactions [117]. Many significant similarity scores reflect regulatory relationships between genes with overlapping expression domains in the transcriptional network: *ocelliless* (also known as *orthodenticle*) is positively regulated by *bicoid* [29]; *ubx* indirectly regulates *dichaete* through the intermediate activation of *dpp* [11, 98]. In addition to activators, we observe significant similarities for repressors, a property of the mutual information score which scores correlation and anti-correlation equally: *hunchback* represses the expression of *nubbin*, *pdm1* [52] and *ubx* [91]; *giant* and *krueppel* mutually repress each other[57].

Additional significant spatial localization can be a result of conditions in which genes act either in concert, or independently, to regulate other downstream genes. Both *pdm2* and *nubbin* show significance with *dichaete*, and double knockout studies have shown that these three genes are essential for proper formation of eve stripes 1, 4, 5 and 6 [72].

Not all known interactions are detected as significant, nor is this to be expected given the data. Often, a spatial expression pattern of a gene is a result of complex interactions between many genes across several time stages. For example, the proper development of all *even-skipped* stripes requires the interaction of many genes; the stripes are encoded by distinct cis-regulatory regions; and some factors function to control only a subset of the stripes as shown in Figure 6.3.

In such cases, assessing similarity on the level of global gene expression pattern, as pursued here as illustrative example, may therefore be modified to deal with parts of patterns, such as disjoint expression domains or even the boundary of expression domains. We anticipate that the integration of quantitative spatial expression information with other “traditional” high throughput data (e.g. binding, expression), will be highly useful to infer complex interactions in the regulatory networks.
Figure 6.3: Regulation of even-skipped: Example of the complexity of regulation in *Drosophila*. The proper establishment of the even-skipped regulation requires a complex interaction between many different genes. A subset of genes are shown here for stripes 2-5, activators and repressors work in concert to establish well defined stripes of even-skipped expression.

6.4 Discussion

As high-throughput image datasets become more prevalent, the need to develop models for extraction, modeling, and analysis of this data becomes equally important. Given the anticipated volume of image data sets, this provides a new challenge to develop robust fully automated image analysis methods to normalize and register the data, as well as new approaches to model spatiotemporal data. In the context of gene expression, models for time series data have been proposed, but appropriate models for spatial expression patterns have not been extensively explored thus far.

To address questions that arise with the analysis of spatial gene expression patterns, we have implemented a complete pipeline for 2D *Drosophila* RNA in situ staining images that is fully automated and highly robust to variable conditions in...
Figure 6.4: Generation of background data sets and pairwise significance tests for 2D expression patterns. **Upper Right Panel:** Computation of significance values, as exemplified on in situ images for the genes *Nubbin* (a) and *Dichaete* (c). The extracted expression vectors (bottom) of *Nubbin* and *Dichaete* are used to calculate the background distribution specific for the comparison of these two expression patterns. For each gene, a set of random realizations with constraints on the correlation between spatially adjacent expression values is created. To create a constrained pattern, an initial column vector with values determined from the marginal distribution of the expression vector is created. The constrained realization is then iteratively resampled by swapping elements of the expression vector that result in a more probable spatial structure. The resulting set of constrained realizations (examples are shown in (b) and (d)) are used to compute a background distribution of similarity values (histogram in (e)). The observed similarity on the expression patterns is indicated with an orange line, and is used to calculate an empirical p value for each comparison.

**Lower Left Panel:** A modified set of previously described gap, pair-rule, and segmentation genes [99] was used to evaluate the significance testing. Images for each gene were registered, and their respective column expression vectors were calculated. Using the constrained realizations (histograms) and observed similarity values (orange lines), significance values were calculated for each similarity score. The color of the histogram represents the significance of the pairwise score (blue: > 0.1, green: (0.1, 0.05], yellow: (0.05, 0.01], red: < 0.01).

In addition to new methods for identifying and quantifying expression patterns, we utilized similarity metrics appropriate for spatial patterns. Such methods are critical for a proper interpretation of similarity between genes on the level of their spatial expression patterns with dependencies among neighboring areas. We also developed a significance test which takes these spatial dependencies properly into account, by sampling random patterns that preserve spatial structure.
methods allow us to calculate significance values that reflect the biological knowledge of the relationships between the underlying genes.

Our results clearly demonstrate the importance of including spatial dependencies in our models. However, they are neither complete, nor should they be considered a catch-all solution for spatial data analysis. By including dependencies between neighboring data points, we have limited our ability to pick up larger scale dependencies, where our choice of model choice and the sparseness of the data limited us to a fixed scale neighbor interaction, and to the projection of the expression patterns onto x/y axes. The effect of scaling and neighbor choice can have a large effect on the spatial model and is known as the modifiable areal unit problem (MAUP) or change of support (COS) [32, 36], and has long been an active area of research. Incorporating the effects of scaling through the use of scale invariant measures or multistage models not only provides more realistic and powerful models, but also provides a realistic and natural method of combining spatial expression patterns with more aggregate types of data (e.g., microarrays).

Imaging in situ hybridizations under bright field microscopy often introduces artifacts into the data. This can be quite problematic when making quantifiable comparisons: depending on the overall position, lighting conditions, and focal plane of the image, genes with near identical spatial expression patterns will result in low similarity values. While mutual information and spatial mutual information are more noise tolerant than the mean squared error and Haar distance, they still limited our ability to fully capture the 2D spatial expression data. Many of these quantification issues may be addressed by proper prospective experimental planning, including standardization of microscopy settings, number of replicates, time stages/conditions, and staining protocols. Unfortunately, this type of planning is not available in a retrospective study as this is; as a result, quantification of microscopy images will continue to be affected by these issues until methods are developed that reduce
and/or adequately model these sources of noise. A critical limitation which may also be alleviated in future studies is the current lack of biological replicates, to be able to filter input noise and model the variance of expression patterns in a way comparable to microarray data.

While our method for the automated registration of images is more robust and complete than many previous approaches, it is important to note that it cannot be a complete and ultimate solution for all image expression analysis problems. Differences in sample preparation and imaging technology, as well as differences in the morphology of specimens which are imaged, pose restrictions on the possible choices of analysis methods. For instance, Our previous work on Arabidopsis root images [73] deals with confocal images of GFP reporter constructs. In addition to the different morphology, the fly dataset poses additional challenges of dealing with multiple objects, and a high variability in imaging conditions; conversely, we needed to apply non-rigid registration for the plant roots. While general platforms for cell-based image analysis have been published [12], a universal framework for complete multicellular organs or organisms has to be placed at a higher level and poses a hurdle for the automated analysis of image expression datasets pertaining to molecular biology.

Even when a registration pipeline can be adapted quickly, the underlying research questions can strongly influence the way the image analysis is approached. The goal of our work at the current stage was to extract information from high-throughput image data, which can then be utilized as or combined with other sources of genomic expression data. For instance, it can provide quantitative similarity measures to cluster expression profiles (a task which is currently based on qualitative descriptors obtained by manual annotation [106]), or as a high-resolution complement to high-throughput data such as ChIP-chip, or microarrays. Clearly, other fly datasets [46, 55] are obtained with a different purpose, namely to study the expression of few genes but at a much higher quality in terms of resolution and quantification of expression
values.

Whatever the specific context, appropriate quantitative methods along the lines proposed in this study will allow us to move image expression data from qualitative descriptions to the quantification of gene expression, and to its use as phenotype for which we can assess the significance of changes under perturbations of the genotype or the environment.
<table>
<thead>
<tr>
<th>Annotation Term</th>
<th>Genes</th>
<th>MSE</th>
<th>Haar</th>
<th>MI</th>
<th>SMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>amnioserosa anlage *</td>
<td>20</td>
<td>0.0532</td>
<td>0.0329</td>
<td>0.0076</td>
<td>0.0255</td>
</tr>
<tr>
<td>anlage *</td>
<td>31</td>
<td>0.4590</td>
<td>0.4923</td>
<td>0.1413</td>
<td>0.0197</td>
</tr>
<tr>
<td>anterior endoderm anlage *</td>
<td>31</td>
<td>0.0191</td>
<td>0.0164</td>
<td>0.0084</td>
<td>0.0048</td>
</tr>
<tr>
<td>cellular blastoderm</td>
<td>32</td>
<td>0.2671</td>
<td>0.2064</td>
<td>0.6811</td>
<td>0.7571</td>
</tr>
<tr>
<td>clypeolabrum anlage *</td>
<td>10</td>
<td>0.6958</td>
<td>0.6461</td>
<td>0.0546</td>
<td>0.0252</td>
</tr>
<tr>
<td>dorsal ectoderm anlage</td>
<td>15</td>
<td>0.0916</td>
<td>0.0680</td>
<td>0.0231</td>
<td>0.0052</td>
</tr>
<tr>
<td>dorsal ectoderm anlage *</td>
<td>74</td>
<td>0.0004</td>
<td>0.0012</td>
<td>0.0003</td>
<td>0.0001</td>
</tr>
<tr>
<td>ectoderm anlage *</td>
<td>15</td>
<td>0.0097</td>
<td>0.0132</td>
<td>0.0007</td>
<td>0.0005</td>
</tr>
<tr>
<td>endoderm anlage *</td>
<td>10</td>
<td>0.4230</td>
<td>0.4065</td>
<td>0.0126</td>
<td>0.0113</td>
</tr>
<tr>
<td>foregut anlage *</td>
<td>26</td>
<td>0.0823</td>
<td>0.0839</td>
<td>0.0135</td>
<td>0.0016</td>
</tr>
<tr>
<td>gap</td>
<td>27</td>
<td>0.4067</td>
<td>0.3357</td>
<td>0.2997</td>
<td>0.2787</td>
</tr>
<tr>
<td>head epidermis anlage *</td>
<td>4</td>
<td>0.2205</td>
<td>0.0611</td>
<td>0.0001</td>
<td>0.0001</td>
</tr>
<tr>
<td>head epidermis dorsal anlage *</td>
<td>10</td>
<td>0.0421</td>
<td>0.0451</td>
<td>0.0001</td>
<td>0.0001</td>
</tr>
<tr>
<td>head mesoderm anlage</td>
<td>5</td>
<td>0.6142</td>
<td>0.5348</td>
<td>0.0057</td>
<td>0.0016</td>
</tr>
<tr>
<td>head mesoderm anlage *</td>
<td>17</td>
<td>0.2547</td>
<td>0.2212</td>
<td>0.0930</td>
<td>0.1507</td>
</tr>
<tr>
<td>hindgut anlage *</td>
<td>22</td>
<td>0.0923</td>
<td>0.1773</td>
<td>0.2040</td>
<td>0.0674</td>
</tr>
<tr>
<td>hypopharynx anlage *</td>
<td>2</td>
<td>0.0479</td>
<td>0.0648</td>
<td>0.0001</td>
<td>0.0001</td>
</tr>
<tr>
<td>mesectoderm anlage *</td>
<td>14</td>
<td>0.0205</td>
<td>0.0509</td>
<td>0.0189</td>
<td>0.0211</td>
</tr>
<tr>
<td>mesoderm anlage *</td>
<td>20</td>
<td>0.2579</td>
<td>0.3067</td>
<td>0.2095</td>
<td>0.2481</td>
</tr>
<tr>
<td>pair rule</td>
<td>5</td>
<td>0.4926</td>
<td>0.3871</td>
<td>0.1235</td>
<td>0.1725</td>
</tr>
<tr>
<td>pole cell</td>
<td>14</td>
<td>0.2836</td>
<td>0.1313</td>
<td>0.5479</td>
<td>0.6829</td>
</tr>
<tr>
<td>posterior endoderm anlage *</td>
<td>39</td>
<td>0.0585</td>
<td>0.0334</td>
<td>0.0290</td>
<td>0.0087</td>
</tr>
<tr>
<td>procephalic ectoderm anlage *</td>
<td>70</td>
<td>0.0068</td>
<td>0.0058</td>
<td>0.0011</td>
<td>0.0002</td>
</tr>
<tr>
<td>segmentally repeated</td>
<td>10</td>
<td>0.2463</td>
<td>0.3711</td>
<td>0.2220</td>
<td>0.1386</td>
</tr>
<tr>
<td>trunk mesoderm anlage</td>
<td>6</td>
<td>0.4244</td>
<td>0.3160</td>
<td>0.0350</td>
<td>0.0276</td>
</tr>
<tr>
<td>trunk mesoderm anlage *</td>
<td>19</td>
<td>0.4159</td>
<td>0.4798</td>
<td>0.3079</td>
<td>0.4616</td>
</tr>
<tr>
<td>ventral ectoderm anlage</td>
<td>11</td>
<td>0.1433</td>
<td>0.0728</td>
<td>0.0205</td>
<td>0.0043</td>
</tr>
<tr>
<td>ventral ectoderm anlage *</td>
<td>18</td>
<td>0.0045</td>
<td>0.0046</td>
<td>0.0120</td>
<td>0.0243</td>
</tr>
<tr>
<td>visual anlage *</td>
<td>18</td>
<td>0.0748</td>
<td>0.1046</td>
<td>0.0209</td>
<td>0.0019</td>
</tr>
<tr>
<td>yolk</td>
<td>2</td>
<td>0.5000</td>
<td>0.5232</td>
<td>0.0015</td>
<td>0.0007</td>
</tr>
<tr>
<td>yolk nuclei</td>
<td>37</td>
<td>0.6842</td>
<td>0.6263</td>
<td>0.4118</td>
<td>0.4272</td>
</tr>
</tbody>
</table>

Table 6.1: Table showing the enrichment terms for all genes within the second time grouping (time stages 4-6) for all four similarity values: Mean Squared Error (MSE), Haar wavelets (Haar), Mutual Information (MI), and Spatial Mutual Information (SMI). Each row represents an individual annotation term, light gray shading represents 0.05 significance, while dark gray represents 0.01 significance values. (* - GO term is listed as in statu nascendi)
Applications, Conclusions and Future Work

7.1 Applications

The Rootarray—and imaging in general—provide an emerging field for dealing with biological problems. Here I present several approaches we have used for image data to address biological problems.

As mentioned in Section 2.3, one of the largest benefits of using GFP reporter constructs is that they are capable of visualizing gene expression over a continuous time period. This allows biologists to visualize changing gene expression patterns of dynamic systems over time. One particular example of this is the short-root \textit{Arabidopsis} mutant. The short-root mutant has a well defined phenotype: the root lacks a distinct cortex/endodermis tissue, and instead, has a larger ground tissue. This merging of layers causes many disruptions in cellular processes—in particular, the auxin transport system is disrupted; which, since the auxin transport system is responsible for maintaining the gravitropism of the root, disrupts the ability of the root to detect gravity.

The short-root mutant example is an illustration of a dynamic system where an
Figure 7.1: Expression of Scarecrow (SCR) in Shortroot (SHR) inducible mutants. The induction of SHR causes the up-regulation of SCR in the endo/cortex. Over time, the induction causes a reversal and change to the wild type phenotype.

External induction of gene expression can cause a characteristic phenotypic response. SHR expression occurs through an inducible promoter, where, over time, it will cause the transcription of SCR. In collaboration with Todd Twigg, reporter constructs were used to monitor the expression of SCR and are shown in Figure 7.1. Over time, expression of SHR increases in the endocortex tissue—the tissue then begins to differentiate into separate endodermis and cortex tissues, which allows the root to recover its wild type phenotype and become gravitropic.

Another large dynamic system in *Arabidopsis* roots is the response of the root to environmental stress. As previously discussed in Section 1.1.2, plants must be capable of responding to the environment where they are growing. In many cases, environmental responses trigger many similar pathways, providing another opportunity for GFP reporter constructs to be used. In collaboration with Jose Dinneny, we
identified several genes using reporter constructs that are involved in the response of *Arabidopsis* to salt stress as shown in Figure 7.2.

In addition to environmental and inducible systems, there are also many dynamic systems that are not the result of a response to a “triggered” stimulus. One of these dynamical systems occurs in the natural periodic functions that occur along the length of the root. In collaboration with Siobhan Brady, we used GFP reporter constructs to monitor the expression of several genes with varying longitudinal expression. As shown in Figure 7.3, some genes are known to have expression patterns that vary along the longitude over time.
7.2 Conclusions

In this dissertation, I have developed methods that have addressed three core issues for dealing with biological image data. First, I have developed novel methods for image segmentation that have been used to segment and identify growing roots in the Rootarray. Additionally, informatics approaches have been developed to address both the integration of the segmentation with existing Zeiss microscopy software for full automation, as well as for the transfer and storage of image data. Second, I have presented methods for the registration and quantification of spatial expression patterns from *Arabidopsis* and *Drosophila* images. Lastly, I have validated the *Arabidopsis* data against known tissue sorted expression data, and developed new approaches in *Drosophila* for comparing and calculating the significance between spatial expression patterns.

7.3 Future Directions

Over the last twenty years, research in computer vision has progressed, resulting in new methods for registering and normalizing image data. Much of this research takes place in the biomedical research community, where registration methods are either semi-automated or validated by hand. Unfortunately, images obtained from biomolecular experiments can grow large in number (the Berkeley *Drosophila* Genome Project already contains > 70,000 images)—making it intractable for any human involvement in the registration/validation process. To overcome these issues, methods must be developed for image registration that account for the uncertainty of the image registration process—allowing for the automatic identification and tagging of failed image analysis methods. Recent work with shape models and statistical frameworks for image registration pose to be the most promising areas for dealing with these problems.
There are also many additional areas of research in terms of developing customized approaches for imaging. In particular, addressing the joint issue of morphology/genotype relationships has been largely ignored. For example, in Arabidopsis, the initiation of root hairs and subsequent change in shape of the cell walls are intimately tied to underlying changes in gene expression. Developing methods to track and monitor the change of shape in conjunction with the change in gene expression would provide large opportunities for understanding the mechanisms behind the change in morphology.
Bibliography


120


Biography

Daniel Mace was born on January 12, 1981 in New London, CT—an exceptionally cold morning. He received his High School Diploma from East Lyme High School in 1999, and his undergraduate degree from Pace University majoring in Computer Science and Mathematics, with a minor in Biology. In the fall of 2004, he matriculated into the Computational Biology and Bioinformatics program at Duke University (formerly Bioinformatics and Genome Technology). In the summer of 2005, he affiliated with the Uwe Ohler lab and began his thesis work developing image analysis methods for computational biology. In the Summer of 2006, he was invited to ISMB to give a talk on his research related to Arabidopsis image analysis. His research includes the following work:


