Regulation of Global Transcription
Dynamics During Cell Division
and Root Development

by

David A. Orlando

Department of Computational Biology & Bioinformatics
Duke University

Date: __________________________

Approved:

______________________________
Dr. Philip N. Benfey, Co-Advisor

______________________________
Dr. Alexander J. Hartemink, Co-Advisor

______________________________
Dr. Steven B. Haase

______________________________
Dr. Herbert Edelsbrunner

______________________________
Dr. Jason D. Lieb

______________________________
Dr. Sayan Mukherjee

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

Regulation of Global Transcription Dynamics During Cell Division and Root Development

by

David A. Orlando

Department of Computational Biology & Bioinformatics
Duke University

Date: __________________________

Approved:

______________________________
Dr. Philip N. Benfey, Co-Advisor

______________________________
Dr. Alexander J. Hartemink, Co-Advisor

______________________________
Dr. Steven B. Haase

______________________________
Dr. Herbert Edelsbrunner

______________________________
Dr. Jason D. Lieb

______________________________
Dr. Sayan Mukherjee

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

The successful completion of many critical biological processes depends on the proper execution of complex spatial and temporal gene expression programs. With the advent of high-throughput microarray technology, it is now possible to measure the dynamics of these expression programs on a genome-wide level. In this thesis we present work focused on utilizing this technology, in combination with novel computational techniques, to examine the role of transcriptional regulatory mechanisms in controlling the complex gene expression programs underlying two fundamental biological processes—the cell cycle and the development and differentiation of an organ.

We generate a dataset describing the genomic expression program which occurs during the cell division cycle of *Saccharomyces cerevisiae*. By concurrently measuring the dynamics in both wild-type and mutant cells that do not express either S-phase or mitotic cyclins, we quantify the relative contributions of cyclin-CDK complexes and transcriptional regulatory networks in the regulation the cell cycle expression program. We show that CDKs are not the sole regulators of periodic transcription contradicting previously accepted models; and we hypothesize an oscillating transcriptional regulatory network which could work independent of, or in tandem with, the CDK oscillator to control the cell-cell expression program.

To understand the acquisition of cellular identity, we generate a nearly complete gene expression map of the *Arabidopsis Thaliana* root at the resolution of individual cell-types and developmental stages. An analysis of this data reveals a representative set of dominant
expression patterns which are used to begin defining the spatiotemporal transcriptional programs that control development within the root.

Additionally, we develop computational tools that improve the interpretation of these data. We present CLOCCS, a model for the dynamics of population synchrony loss in time-series experiments. We demonstrate the utility of CLOCCS in integrating disparate datasets and present a CLOCCS based deconvolution of the cell-cycle expression data. A deconvolution method is also developed for the Arabidopsis dataset, increasing its resolution to cell-type/section subregion specificity. Finally, a method for identifying biological processes occurring on multiple timescales is presented and applied to both datasets.

It is through the combination of these new genome-wide expression studies and computational tools that we begin to elucidate the transcriptional regulatory mechanisms controlling fundamental biological processes.
# Contents

Abstract ........................................................................................................ iv
List of Tables ............................................................................................... x
List of Figures ............................................................................................ xiii
Acknowledgements .................................................................................... xxvii

## 1 Introduction

1.1 Regulation of expression ................................................................. 2
1.2 Transcriptional regulatory networks .............................................. 4
  1.2.1 Periodic regulation ................................................................. 5
  1.2.2 Spatiotemporal regulation ..................................................... 6
1.3 Genomic tools ................................................................................. 7
  1.3.1 Gene expression microarrays ............................................... 7
  1.3.2 ChIP-chip ............................................................................. 8
1.4 Thesis roadmap ................................................................................ 10

## 2 Saccharomyces cerevisiae and the cell cycle

2.1 Cell cycle basics .............................................................................. 12
  2.1.1 Features of cell cycle progression ......................................... 13
  2.1.2 Synchronization methods ....................................................... 14
2.2 Regulation of the cell cycle ............................................................. 16
### 3 Modeling cell cycle population dynamics

#### 3.1 Introduction

3.1.1 Previous approaches

#### 3.2 CLOCCS model

3.2.1 Mathematical formulation

#### 3.3 CLOCCS Sampling models

3.3.1 Sampling model for budding index data
3.3.2 Sampling model for DNA-content data

#### 3.4 CLOCCS model results

3.4.1 Prior distribution
3.4.2 Estimates from budding index data
3.4.3 Estimates from DNA-content data
3.4.4 Importance of each source of asynchrony

#### 3.5 Applications

3.5.1 CLOCCS-based alignment between datasets
3.5.2 Deconvolution

#### 3.6 Summary

### 4 Transcriptional regulation of the cell cycle

#### 4.1 Introduction

#### 4.2 Global control of cell-cycle transcription by coupled CDK and network oscillators

4.2.1 Biological data generated
4.2.2 Absolute change and Pearson correlation analyses
4.2.3 Dual-clustering & over-representation analysis
4.2.4 Construction of transcriptional regulatory networks
4.2.5 Boolean model construction
4.3 Challenges in data analysis ........................................ 76
  4.3.1 Alignment of wild-type and cyclin mutant datasets ....... 77
  4.3.2 Periodic gene detection ........................................ 78
4.4 Summary .............................................................. 81

5 Arabidopsis thaliana and root development .................. 84
  5.1 A. thaliana as a model for root development ................. 85
  5.2 Importance of transcriptional regulatory networks in root development .... 87
  5.3 Generating expression data ....................................... 89

6 Transcriptional regulation of root development ............ 92
  6.1 Introduction ......................................................... 92
  6.2 High-Resolution root spatiotemporal expression map ....... 93
    6.2.1 Supervised data analysis .................................... 94
    6.2.2 Unsupervised data analysis ................................ 95
    6.2.3 Root variability .............................................. 104
    6.2.4 Integrating the radial and longitudinal datasets ...... 106
    6.2.5 Transcriptional regulatory modules identification ...... 111
  6.3 Data generated ................................................... 114
  6.4 Dominant differential expression pattern ................... 115
    6.4.1 Assignment of probe sets to patterns ................... 116
  6.5 Root expression visualization .................................. 117

7 Improving the resolution of expression data ............... 119
  7.1 Introduction ....................................................... 119
  7.2 Improving the resolution S. cerevisiae data ............... 120
    7.2.1 The deconvolution model .................................... 121
    7.2.2 Summary of the mathematical implementation .......... 122
7.2.3 Deconvolving budding index data ........................................ 124
7.2.4 Results .............................................................................. 125
7.2.5 Conclusion ......................................................................... 129
7.3 Improving the resolution of the *A. thaliana* data ......................... 129
  7.3.1 Previous approaches .......................................................... 130
  7.3.2 Bi-linear Method ............................................................... 133

8 Detecting separate time scales in gene expression data ....................... 154
  8.1 Introduction ........................................................................... 154
  8.2 Methods ............................................................................... 156
  8.3 Results .................................................................................. 158
    8.3.1 Detection of separate time scales in *S. cerevisiae* ................ 158
    8.3.2 Detection of separate time scales in the *A. thaliana* root ......... 163
  8.4 Discussion ............................................................................. 168

9 Discussion ................................................................................. 170
  9.1 Contributions of this thesis ..................................................... 170
  9.2 Future directions .................................................................... 172
    9.2.1 Advances in CLOCCS ......................................................... 172
    9.2.2 Advances in *S. cerevisiae* cell cycle ................................. 174
    9.2.3 Improvements in *A. thaliana* ........................................... 177
  9.3 Conclusion ............................................................................. 178

Appendix ......................................................................................... 180
References ...................................................................................... 187
Biography ....................................................................................... 203
List of Tables

3.1 Parameter estimates for multiple experiments. Each column corresponds to a different experiment, synchronized by either centrifugal elutriation or \( \alpha \)-factor arrest, and grown at either 30\(^\circ\)C or 21\(^\circ\)C. The observed budding curves, as well as the model predictions for each experiment, are shown in Figure 3.4. Each cell contains the mean value for that row’s parameter, given 250,000 iterations of the Markov chain used to fit the column’s experiment. Below each mean is the 95% confidence range for that parameter in that experiment. .................................................. 38

3.2 Prior (column 1) and marginal posterior summaries given both the DNA-content and budding index data (column 2), given the DNA-content data only (column 3) and given the budding index data only (column 4). .................. 39

3.3 Estimates of log Bayes factors (lBFs) for various nested model comparisons given the budding index data alone. The model indexed by an entry’s column is the larger of the models and is represented in the numerator of the lBFs in that column; the model indexed by an entry’s row is the smaller of the two. The last two rows of the table provide the average and standard deviation of the RMSE of the model’s fitted values to the observed budding index data over a sample of 1000 draws from the posterior. ........ 42

3.4 Parameter estimates for \textit{S. cerevisiae} and \textit{S. pombe} experiments. The CLOCCS parameter estimates for the experiments aligned in Figure 3.7. Each cell contains the mean value for that row’s parameter, given 250,000 iterations of the Markov chain used to fit the column’s experiment. Below each mean is the 95% confidence range for that parameter in that experiment. 47

4.1 Over-represented transcription factors in Figure 4.6 by cluster/subcluster. Each row corresponds to a wild-type cluster (indexed by letters A-AD) and cyclin-mutant subcluster (indexed by numbers 1-56) from Figure 4.6 for which there are over-represented TFs. Each cell of the table—for the genes in the corresponding cluster or subcluster—contains the TFs with \( q \)-values \( \leq 0.005 \), sorted from most to least significant. ....................... 61
4.2 Boolean logic used in model shown in Figure 4.7c. An initial choice of a set of Boolean logic functions on the variables which produce the five state oscillating attractor referred to as Cycle 1. 63

4.3 The attractors of model given different logics. The percentages of all possible starting states that end in the four identified attractors for all simple functions of the activating logic rules for the repressive inputs to CLN3 and SBF (YOX1 and YHP1), and activators of SFF (SBF and HCM1), and CLN3 (SWI5 and ACE2) variables. 64

4.4 The on/off states of the transcription factors (TFs) in the each attractor. Cycle 1 attractor as used to color Figure 4.7c and shown in Figure 4.11b. The on/off states of the TFs in the Cycle 2 attractor as shown in Figure 4.11c. The on/off states of the TFs in the Cycle 3 attractor as shown in Figure 4.11d 64

4.5 CLOCCS parameter estimates. Each row corresponds to a different arrest-release experiment. The observed budding curves, as well as the model predictions for each experiment, are shown in Fig. Figure 4.12. Each cell contains the mean value for that column’s parameter, given 250,000 iterations of the Markov chain used to fit that row’s experiment. Below each mean is the 95% confidence range for that parameter in that experiment. Since the values of $\beta$ and $\delta$ were fixed to be 0.15 and 0, respectively, when fitting CLOCCS in the cyclin-mutant experiments, no confidence intervals exist for these parameters in those experiments. 80

5.1 Tissue and cell-types in the A. thaliana Root. This table defines the relationships between tissues and cell-types in the A. thaliana root. The rows are grouped by which set of initials they arise from. 86

5.2 Description of cell-types, with associated markers. This table defines the relationship between cell-type and marker line, including which longitudinal sections of each cell-type is marked by the line. Note that, differing from Table 5.1, the Lateral Root Primordia is included as a cell-type in this table, eventhough it may be a collection of multiple immature cell types. Also, because there are also no markers that differentiate between metaxylem and protoxylem or between metaphloem and protophloem, those cell-types are labeled as Xylem and Phloem respectively. 90

6.1 Percentage of genes that are co-regulated similarly in the second root replicate. Each longitudinal pattern and the percentage of genes assigned to it which show co-regulation between root replicates are shown. The ‘*’ indicates a pattern which shows potential phasing. 107
7.1 PTR values of convolved and deconvolved profiles. The first two rows of the table contain the average PTRs for the population-level and deconvolved profiles respectively. Average values are calculated across all 1208 cell-cycle regulated genes (column 1), only genes peaking in the common cell-cycle interval C (column 2), and only genes peaking in the A or Gd intervals (column 3). The third row is the ratio of the first two rows and is a measure of the fold increase in signal strength between the convolved and deconvolved profiles.

7.2 Mean expression values and scaling factors of ubiquitously, evenly expressed probesets across longitudinal and marker-lines.

7.3 The 14 cell-types in the *Arabidopsis* root and the 17 marker-lines which mark them. For markers that only mark the cell type in some of the sections, these sections are indicated by the range in parenthesis.

7.4 The cell count matrix gives the number of cells in each cell-type/section subregion. The 13 rows correspond to longitudinal sections 1 through 13. From left to right, the 14 columns correspond to the following cell-type/section subregions: quiescent center, columella, lateral root cap, hair cell, non-hair cell, cortex, endodermis, xylem pole pericycle, phloem pole pericycle, phloem, phloem companion cells, xylem, lateral root primordia, and procambium.

7.5 Root mean square percentage error rates in the reconstruction of simulated data. The first column is under a model of comparable but varying expression levels across all sections and cell-types. The second type is the error rate when that section or cell-type has its expression level raised by a factor of 10. The third and fourth columns show models in which the bilinear assumption is violated in one of the sections or one of the cell-types respectively. In all cases, 3% measurement error has been added to the expression levels.

8.1 Distribution of significant shifts in the cell-cycle dataset. The 2nd row gives the number of genes which have a significant shift (p-value ≤ 0.001) at the indicated shift (a gene can have more than one significant shift). The 3rd row gives the number of genes with their maximally significant shift at the indicated shift (a gene can only belong to one shift). Genes in the “Not Significant” category had no shifts with a significant p-value.

8.2 GO term analysis result. Every set of genes, defined by the position of their maximally significant shift, was analyzed for over-represented GO terms [1]. Each shift (column) and term (row) found to be significant are shown.
## List of Figures

1.1 Sample transcriptional regulatory network. a) A small transcriptional regulatory network with five interconnected transcription factors. Each node represents a TF and a directed edge between two nodes represents a regulatory interaction (green=activating, red=repressing). b) A four-state periodic pattern is an emergent property of the network. Active TFs at each state are shown in yellow. .......................................................... 5

1.2 Schematic overview of oligonucleotide microarrays and ChIP-chip. a) A summary of the gene expression measurement procedure using high-density oligonucleotide microarrays. b) A summary of the ChIP-chip procedure. Panels and legends adapted from Schulze and Downward [2], Buck and Lieb [3]. .......................................................... 9

2.1 The *S. cerevisiae* cell cycle and measurable cell-cycle progression features. a) Over the course of its life, the cell repeatedly traverses the cell cycle, which is divided by landmark events into the G1, S and G2/M phases. The transition from G1 to S is marked by the development of a bud (dark gray) and the start of DNA replication (denoted by the appearance of a second red bar). The S/G2 transition is marked by the completion of DNA synthesis. The cell cycle is completed when the daughter cell (dark gray) separates from the mother cell at the end of G2/M. b) A typical budding index curve. c) A typical DNA content histogram as measured by flow cytometry for an asynchronous population. ......................................................... 15

2.2 The transcription repressor Whi5 inhibits the activity of the SBF transcription factor. Phosphorylation of Whi5 by Cln3Cdc28 induces the nuclear export of Whi5 and activates SBF, which induces the transcription of the genes that encode Cln1, Cln2, Clb5 and Clb6. Cln-Cdc28 and Clb-Cdc28 phosphorylate Whi5, which might provide a positive-feedback loop. Image and legend adapted from Bloom and Cross [4]. ......................................................... 18
3.1 The lifeline representation of the cell cycle. A linearized version of Figure 2.1a, where instead of moving clockwise, a cell progress from left to right across the lifeline. The linear lifeline allows for each transcript of the cell cycle to be represented individually, enabling events that occur only once to be accounted for.

3.2 Graphical representation of the branching dynamics of the cell population. a) The position distributions of the cohorts, indexed by \{g, r\}, in the population at four points in time, each color-coded. b) The population at the last time point (green) projected onto a common lifeline.

3.3 Plot of expected flow cytometry channel for a cell given its lifeline position in units of \(\lambda\) (black curve, left vertical axis). An indicator function for the cell’s budding status is also plotted (gray dashed curve, right vertical axis).

3.4 Model fits of multiple experiments. Observed budding index curves (black lines) for cells synchronized by elutriation (a and b) or \(\alpha\)-factor (c and d) and then incubated at 30°C. Experiments synchronized by \(\alpha\)-factor and then incubated at 21°C are shown (e and f). One hundred random realizations from the Markov chain used to fit each experiment were used as parameterizations for the model, and the resulting predicted budding curves are shown (green lines). The width of the band created by the Markov chains is indicative of the effect of uncertainty in the model parameters. The model parameterizations from the Markov chains for each of the experiments are shown in Table 3.1.

3.5 a) Plot of observed budding index curve (black) and 95% pointwise interval estimates from budding index only analysis (green) and budding index/DNA-content analysis (red). b–f: DNA-content densities (gray) and their posterior mean estimates (red) at five points in time, highlighting the population’s transition from G1 (b), through S phase to G2/M (c and d) and the effect of its growing asynchrony (e and f). The corresponding time points are labeled above the budding index curve.

3.6 Plot of observed budding index curve (black) and posterior mean fitted curves under each of the competing models for the budding index data. The full model is plotted in red; the competing models are obtained by constraining the parameter(s) indicated in the figure legend to be zero. Quantitative summaries of these fits can be found in Table 3.3.
3.7 Model-based alignment of time course data from synchrony experiments. 
a) The lifeline distributions at three time points for cells synchronized with afactor incubated at either 30°C (left side) or 21°C. 
b–c) Alignment of budding curves for the two experiments (red line=30°C, blue line=21°C), based on clock time. Time points at which the two populations exhibit similar distributions are highlighted (green dots; 90 min for 30°C, 180 min for 21°C). 
d) Alignment of budding curves to the lifeline using the position of the mean of the \{0, 0\} cohort for mapping.

3.8 CLOCCS based alignment across organisms. Model estimates were generated based on the budding index curve for S. cerevisiae [5] (a, blue line) and the septation index curve for S. pombe [6] (b, red line). The light blue regions in the lifelines above the curves show the areas of the cell cycle where budding and septation occur, respectively. Using the mapping from time to \{0, 0\} position for both experiments (c) the two curves were aligned (d). Since budding and septation do not occur at the same points in the cell cycle we do not expect these two curves to overlap. Using the same mapping function (e), we align the mean-normalized log$_2$-transformed transcriptional profiles for a set of gene homologs with the S. cerevisiae gene name on the left and the expression profile shown in blue, and the S. pombe gene name on the right and the expression profile shown in red.

4.1 Reproducibility of replicate experiments. Scatter plots of mean normalized absolute expression levels over all timepoints within an experiment for each mapped probe for both a, wild-type and b, cyclin-mutant cells. Each probe’s X and Y coordinate is defined by its mean normalized absolute expression/1000 in replicate 1 and 2, respectively. A linear model was fit to each plot (gray line) and adjusted $r^2$ values are given. To more clearly reveal the density of points near the origin, each point is colored according to the number of other points contained within a square of length 500 centered on that point.

4.2 Reproducibility of replicate experiments. Scatter plots of mean normalized absolute expression levels over all timepoints within an experiment for each mapped probe for both a, wild-type and b, cyclin-mutant cells. Each probe’s X and Y coordinate is defined by its mean normalized absolute expression/1000 in replicate 1 and 2, respectively. A linear model was fit to each plot (gray line) and adjusted $r^2$ values are given. To more clearly reveal the density of points near the origin, each point is colored according to the number of other points contained within a square of length 500 centered on that point.
4.3 Comparison of periodic genes identified by three studies. A Venn diagram of the overlap between the 1271 unique genes identified as periodic in wild-type cells by this study, the 991 from Pramila et al. [5], and the 800 from Spellman et al. [7]. The set of 991 periodic genes from Pramila et al. [5] contains all those with PBM5 rankings of 1000 or less, from the three datasets available at [8]. The set of 800 periodic genes from Spellman et al. [7] was obtained directly from [9].

4.4 Transcription dynamics of established cyclin–CDK-regulated genes. Absolute transcript levels (normalized Affymetrix intensity units/1000) are shown for the genes CLN2 (a) and RNRI (b), which are regulated by SBF and MBF, respectively; the Ace2/Swi5-regulated genes SIC1 (c) and NIS1 (d); and the Clb2-cluster genes CDC20 (e) and ACE2 (f). Solid lines, wild-type cells; dashed lines, cyclin-mutant cells.

4.5 Peak to trough ratio for all genes. a) The average wild-type peak to trough ratio (PTR) (x-axis) is plotted against the average cyclin-mutant PTR (y-axis) for each gene/probe identified as periodic by all three studies (green, see Figure 4.3) [5, 7], by this study only (red), or not identified as periodic by this study (black). b) The average PTR in wild-type (Column 1) or cyclin-mutant cells (Column 2), as well as the average wild-type to cyclin-mutant PTR ratio (Column 3). Colored dots to the right indicate which genes from (a) are included in each calculation. A histogram of the average PTR for each of the WTPER genes in c, wild-type and d, cyclin-mutant cells. e) Histogram of ratios of the PTRs of each WTPER gene in wild-type vs. cyclin-mutant cells.

4.6 Genes showing altered behaviors in cyclin-mutant cells. a) Clusters of genes with similar expression patterns in wild-type cells. b) Subclusters of genes with similarly altered expression patterns in cyclin-mutant cells. Each row in a and b represents data for the same gene. Transcript levels are depicted as in Figure 4.2. Up to five over-represented transcription factors for each cluster are shown (see Table 4.1 for complete lists).

4.7 The periodic transcription program is largely intact in cyclin-mutant cells that arrest at the G1/S border. a–b) Genes maintaining periodic expression in cyclin-mutant cells (a) show similar dynamics in wild-type cells (b). Each row in a and b represents the same gene. Transcript levels are depicted as in Figure 4.2. c) Synchronously updating boolean network model. Transcription factors are arranged on the basis of the time of peak transcript levels in cyclin-mutant cells. Arrows indicate transcription factor/promoter interaction. Activating interactions, outer rings; repressive interactions, inner rings. Coloring indicates activity in one of five successive states; SBF and YHP1 are active in two states (Table 4.3).
4.8 Wild-type vs. cyclin-mutant PTRs for transcripts of all transcription factors in the Boolean network model. The average wild-type PTR (x-axis) plotted against the average cyclin-mutant PTR (y-axis) for each transcription factor (black circle, blue text) in the model depicted in Figure 4.7c. The dashed red line is drawn where the ratio between wild-type and cyclin-mutant PTR is one. 66

4.9 Histograms of the ratios of PTRs in wild-type vs. cyclin-mutant cells for the periodic targets of network transcription factors that exhibit reduced expression in cyclin-mutant cells. Histograms of the ratio between the wild-type and cyclin-mutant PTRs for all genes which were identified as periodic by this study, and are documented targets of a, Ace2, b, Swi5, and c, Yox1 according to the YEASTRACT database [10]. 67

4.10 Full transcription factor network diagrams for wild-type and cyclin-mutant cells. Network diagrams (a,c) are similar to timeline representations of the circular diagram in Figure 4.7c. Periodically expressed transcription factors are placed on the cell-cycle timeline on the basis of the time of peak transcript levels in our data sets. Arrows indicate a documented interaction between a transcription factor and promoter elements upstream of a gene encoding another transcription factor. Transcriptional activators are depicted in green, repressors in red, and the cyclin Cln3 in blue. The actual b, wild-type and d, cyclin-mutant expression of the transcription factors in the networks are shown. Transcript levels are depicted as in Figure 4.2. The scale of the cell-cycle timeline has been altered to accommodate multiple G1 events. 68

4.11 A synchronously updating Boolean model can reproduce the sequential order of transcription factor expression. a) The actual expression of the variables in Figure 4.7c (with MBP1, SWI4, and FKH2 as proxies for SBF, MBF, and SFF, respectively), as compared to the on/off (yellow/cyan) states of those variables in b, Cycle 1, c, Cycle 2, and d, Cycle 3 (Table 4.4). Transcript levels are depicted as in Figure 4.2. The scale of the cell-cycle timeline has been altered to accommodate multiple G1 events. 69
4.12 CLOCCS model fits for the four time-series. Fraction of cells with a single bud (black lines, filled circles) and two or more buds (black lines, open circles) for (a,b), wild-type cells and (c,d), cyclin-mutant cells. One hundred random realizations from the Markov chain used to fit each experiment were used as parameterizations for the model, and the resulting predicted budding curves for one bud (green lines) and two or more buds (red lines) are shown. The width of the band reflects the degree of posterior uncertainty in the budding curve and can be interpreted as forming a confidence band for the curve on the basis of the uncertainty in the parameters. Parameter estimates for each experiment are given in Table 4.5.

4.13 Expression dynamics of \textit{CHA1}(a) and \textit{SSA1}(b) during cell cycle progression. Transcript levels in: Orlando et al. [11], black; Pramila et al. [5], broken gray; Spellman et al. [12], solid gray. Time points from each experiment were mapped onto a common cell cycle timeline using the CLOCCS model. Relative expression is defined as the ratio of transcript levels at each time point to the minimum level during the interval shown.

5.1 The structure of the \textit{A. thaliana} root. A representation of the \textit{A. thaliana} root (see Section 6.5 for details on digital root representation) with each cell-type identified by its own color. The three developmental zones are demarcated on the left. The lateral root primordia is not counted in the 15 cell-types as it can be a mix of multiple cell types. The meta/proto xylem and phloem are counted as four cell-types in total.

5.2 Microarray expression profiles of 19 GFP-marked lines were analyzed. The colors associated with each marker line reflect the developmental stage and cell-types sampled (Table 5.2, CC=companion cells). Thirteen transverse sections were sampled along the root’s longitudinal axis (red lines).

6.1 Summary of significant array annotation term enrichment by cell-type. The hypergeometric distribution P-value is $\log_{10}$ transformed.

6.2 Summary of significant \textit{cis}-element enrichment by cell-type. \textit{Cis}-element enrichment was determined by ATHENA [13]. The hypergeometric distribution P-value is $\log_{10}$ transformed.

6.3 Putative novel function is assigned to xylem cells in the meristematic zone. Significantly enriched GO categories (top panel), array annotations (middle panel) and \textit{cis}-elements (bottom panel) associated with xylem cells in the meristematic zone (hypergeometric distribution P-value <1e$^{-3}$).

xviii
The majority of enriched GO terms (hierarchically clustered) are associated with individual cell-types (blue bar). a) A smaller number are present across multiple cell types (green bar). b) GO category enrichment for hair cells confirms a previous report [14]. Enriched cis-elements and an enriched TF family were also identified. c) From the top 50% of varying probe sets, 51 dominant radial patterns were identified. Pattern expression values were mean-normalized (rows) and log\(_2\) transformed to yield relative expression indices for each marker line (columns). Marker line order is the same for all figures; see Table 5.2 for marker line abbreviations. d) Pattern expression peaks were found across one to five cell-types. e–g) Patterns where expression is enriched in single and multiple cell-types support transcriptional regulation of auxin flux and synthesis. In all heat maps with probe sets, expression values were mean-normalized and log\(_2\) transformed. Expression is false-colored in representations of a root transverse section, a cut-away of a root tip, and in a lateral root primordium. e) Auxin biosynthetic genes (CYP79B2, CYP79B3, SUPERROOT1, and SUPERROOT2) are transcriptionally enriched in the QC, lateral root primordia, pericycle, and phloem-pole pericycle (P = 1.99e\(^{-11}\), pattern 5). All AGI identifiers and TAIR descriptions are found in Supplemental Table 14 of [15]. f) Auxin amido-synthases GH3.6 and GH3.17 that play a role in auxin homeostasis show enriched expression in the columella, just below the predicted auxin biosynthetic center of the QC (P = 8.82e\(^{-4}\), pattern 13). g) The expression of the auxin transporter, PIN-FORMED2, and auxin transport regulators (PINOID, WAG1) are enriched in the columella, hair cells, and cortex (P = 1.03e\(^{-4}\), pattern 31).
6.5 Forty diverse and distinct patterns were identified from sections along the longitudinal axis. **b)** Most patterns show an expression increase across two longitudinal sections. **c)** Two pattern categories were identified: a single expression peak (top) and fluctuating expression in developmental time (bottom). **d–f)** Co-regulation of probe set expression between roots. Patterns were identified from root 1. Root 2 was treated as the replicate. Probe sets are listed in the same order. Red dashed lines indicate co-regulated groups. Note that there is no columella section in root 2. **d)** Robust co-regulation of expression for a single expression peak (pattern 25). **e)** Robust co-regulation of expression for a fluctuating pattern (pattern 19). **f)** Extensive between-root variation with multiple co-regulated groups (pattern 37). **g–j)** Validation of longitudinal expression patterns with transcriptional GFP fusions. The expression profiles are able to resolve *in vivo* expression at high resolution: *AGAMOUS-LIKE21* **g)** expression validates a peak in section 1, whereas *WEREWOLF* **h)** shows a peak in section 2; **i)** At4g05170 (pattern 39) confirms the microarray data with a fluorescence peak in the maturation zone. **j)** Expression conferred by the At5g60200 promoter confirms a fluctuating expression profile (pattern 37) with two peaks of expression in the meristematic region and the distal elongation region. This image contains portions distal to the sections dissected for microarray analysis. A third peak of expression is found in this distal maturation zone. Details of microscopy and image analysis can be found in [15].

6.6 Overlap between genes identified as cell-type-enriched and genes present in the 51 dominant expression profiles. This was performed to ensure that the computational method identified genes enriched in individual cell-types as assigned by our first supervised approach.
6.7 Patterns show enrichment of expression in tissues that are ontologically (a) or spatially (b) related, or in cell-types that are spatially separated (c and d). a) Pattern 42 shows enrichment in the xylem tissue in all developmental stages. Genes assigned to these patterns show a strong enrichment for genes involved in microtubule-based processes ($P = 4.03 \times 10^{-6}$) b) Pattern 49 shows enrichment in the more mature xylem and in phloem tissue and a corresponding enrichment in ceramidase activity ($P = 6.79 \times 10^{-15}$) and microtubule binding ($P = 1.5 \times 10^{-4}$). c) Pattern 35 shows enrichment in the columella and in the mature xylem and an enrichment for genes involved in proteolysis ($P = 7.21 \times 10^{-4}$). These tissues are spatially separated by many cells. d) Pattern 33 shows high enrichment in the hair cells and in developing xylem. Enrichment analysis suggests a shared developmental pathway, presumably related to new cell wall deposition; (COPII vesicle coat ($P = 4.07 \times 10^{-5}$), protein amino acid glycosylation ($P = 6.5 \times 10^{-5}$), cell wall ($P = 1.44 \times 10^{-4}$).

6.8 Probe set expression between replicates is reproducible. The distribution of root mean squared differences between true probe set pairs (black) and each probe set and 10 randomly selected probe sets (red). Probe sets analyzed are the top 50% of expressed probe sets. For this analysis we normalized the data using mean=0 and variance=1. The two distributions are separate, demonstrating that the data between replicates is reproducible and not random.

6.9 Probe set co-expression between roots for each pattern. Most probe set members maintain greater than 90% co-expression in a second root. For each pattern, the largest coregulated group was considered.

6.10 Expression conferred by the At3g43430 promoter, a zinc finger TF, shows fluctuation along the root longitudinal axis that varies between individual roots. a) Expression values (y axis) of At3g43430 in both roots demonstrate varying fluctuation between both root samples. b–e) Four images of roots display fluctuations of expression that differ between individual roots and that validate the observations made with microarray expression profiling. (Top) Quantification of GFP expression levels along the roots longitudinal axis. Each root was divided into 50 equidistant units and these are represented on the x axis. Along the y axis are GFP intensity units normalized by the total area of each section. (Bottom) The imaged root whose GFP expression was quantified.
6.11 Combining radial and longitudinal expression data to identify gene expression in space and time using dominant expression patterns. a) Radial (x axis) and longitudinal (y axis) patterns are combined by conditional addition to yield relative enriched peaks of expression by cell-type and developmental time. The heat map predicts an enriched relative peak of expression in the cortex and more weakly in the endodermis at the distal region of the meristematic zone and elongation zone. b) A visual representation of the conditional addition in a (left). The At3g05150 transcriptional GFP fusion confirms this expression. See Section 6.5 for details of visualization. Details of microscopy and image analysis can be found in [15].

6.12 Inference of transcriptional regulatory modules. a) A MYB transcription factor, At5g60890 (ATR1) is present in a set of genes enriched for auxin biosynthesis function (Figure 6.4e) and for MYB-binding sites. All genes containing a MYB-binding site are indicated as downstream of ATR1. All genes annotated as having a function in Trp-dependent auxin biosynthesis are indicated with a blue box and yellow shading. b) A putative ARF-regulated module in the columella. At1g30330 (ARF6) and At2g28350 (ARF10) are strongly co-expressed in the columella. A set of genes also strongly co-expressed with these two members contain the binding site for ARFs and are indicated as downstream of ARF6 and ARF10. c) Putative WRKY-regulated transcriptional modules identified in different stages of root hair development from intersections of the radial and longitudinal data sets. WRKY TF binding sites (W-boxes) were enriched in root hairs (P < 1e−10). The left panel indicates a subset of this hair cell enrichment in the basal meristematic zone. The WRKY TF At1g68150 is present in this gene group and three potential targets of this WRKY TF were inferred as being downstream as they contain a W-box and are co-expressed with At1g68150. The right panel indicates an additional group of genes which show fluctuation of expression in hair cells, first in the elongation zone and further in the maturation zone. This group of genes contains the WRKY TF At1g68150 and its two potential co-expressed downstream targets.

7.1 CLOCCS and deconvolution lifeline models. A) The CLOCCS linear lifeline model. B) The deconvolution model for cells progressing through the cell cycle in a synchronized population. After recovery (Gr), cells progress through a common mitotic cell cycle (C) encompassing G1, S, and G2/M. Cells then remain attached to each other after the completion of M with mother and daughter cells potentially operating separate transcriptional programs (A). Finally, after cytokinesis newly created mother cells begin C, while newly created daughter cells go through a daughter specific growth period (Gd).
7.2 The effect synchrony loss has on observed population-level budding index measurements. Under perfect synchrony, all cells should start budding at the beginning of S and should become unbudded at the end of M (dashed green line), as opposed to the observed, convolved curve (gray). Applying the deconvolution algorithm to the observed budding index measurements recovers the cell budding index profile nearly perfectly (red line). Figure and caption adapted from [16].

7.3 Heatmap of the 1208 cell cycle transcriptionally regulated genes in the high confidence set with PTR ≥ 2.5. The deconvolution clearly provides a more precisely timed picture of transcription across the cell cycle. Figure and caption taken from [16].

7.4 Deconvolved mRNA expression profiles for 4 genes whose peak expression are in different phases of the cell cycle. Swi4 is a G1 expressed transcript. Fkh1 is an S expressed transcript. Clb2 is a G2/M transcript. Dip5 is predicted to be a daughter specific transcript (Gd). Each of the 4 gene panels shows the observed population-level measurements in replicate (top of each panel), followed by the resulting single cell high resolution profiles obtained by deconvolution. Single cell deconvolved profiles are shown for the initial cell, resulting mother, and daughter cells. Since the common cell cycle interval C is the same for the initial, mother, and daughter cells, this interval is depicted thrice in each panel. In addition, the recovery interval Gr is depicted for an initial cell, the daughter specific interval Gd is depicted for a daughter cell, and the attachment interval A is depicted along with the daughter cell although it may include both mother or daughter specific transcriptional programs. Figure and caption adapted from [16].

7.5 A graphical depiction of the $C_{mts}$ function used by the cell-type/section subregion expression prediction method in [17].

7.6 Arabidopsis root template used for expression pattern overlays.

7.7 a) Expression of AT2G18380 in all developmental stages of the phloem was predicted by our method and visualized in a representation of the Arabidopsis root. Phloem cells are shown external to the root. b) GFP expression in the longitudinal axis and c) expression in cross-section of expression driven by the AT2G18380 promoter validate the prediction.
7.8 a) Our method correctly predicts specific expression of AT4G37940 in a cell-type, procambium, that is only covered by a general tissue marker, WOL. Expression conferred by the AT4G37940 promoter fused to GFP as a reporter was visualized in the columella and in the procambium by a longitudinal section (b) and a cross section (c). The label X indicates the xylem axis. The expression also validates a maximal peak in the meristematic zone.

7.9 a) Our method correctly predicts expression of AT5G43040 in trichoblast cells in the meristematic zone, which are not currently covered by any marker-lines. Furthermore, the algorithm was able to predict differential expression within epidermal tissue with high expression in non-hair cells and atrichoblasts (immature non-hair cells in the meristematic and elongation zone) and decreased expression in hair cells or trichoblasts (immature hair cells in the meristematic and elongation zone). Expression conferred by the AT5G43040 promoter fused GFP as a reporter was visualized in the epidermis in a longitudinal section (b) and was specifically identified as high in atrichoblasts, and lower in trichoblasts (marked with an asterisk) in cross-section (c). Trichoblasts or hair cells differentiate at the junction between two underlying cortical cells.

8.1 We search for time scale separation in microarray expression data by detecting patterns which show a temporal shift between experimental replicates relative to the synchronization by a given biological process. In this example, Gene A is related to the synchronization mechanism and shows very similar expression patterns in both replicates. The expression level of Gene B on the other hand is also changing in a reproducible way, but on a different time scale, independent of the synchronization process. By calculating the similarity for each possible value of the shift s, and its statistical significance, we can determine whether a given expression pattern is likely to be evidence of a process operating on a separate biological time scale.

8.2 Distribution of maximally significant shifts. Our method identified 3974 genes which had significant shifts at or below a p-value of 0.001. Each of those genes was assigned to the shift in which it had the most significant (lowest) p-value.

8.3 Distribution of periodic genes identified in Section 4.3.2. Each of the 1274 periodic genes was assigned to its maximally significant shift. Genes which did not have a significant p-value (≤ 0.001) were assigned to the “Not Significant” category.
8.4 Distribution of maximally significant shifts in the root dataset. Our method identified 5592 genes which had significant shifts at or below a p-value of 0.01. Each of those genes was assigned to the shift in which it had the most significant (lowest) p-value.

8.5 a) Gene ontology (GO) terms associated with shifted profiles. \( \log_{10} \text{p-value} \) is indicated. All patterns except for a shift of +5 show an enriched associated biological process. b) Cell-type enrichment, and enrichment of genes associated with lateral root initiation (LRI or LRIS). Genes activated or repressed within the xylem pole pericycle cells (JO121) during lateral root initiation were also enriched.

8.6 Clustering of shift profiles identifies spatiotemporally regulated modules of genes for shifts of +2. Relative expression by marker line is visualized in the left heatmap, and relative expression by longitudinal section in the two roots is visualized in the right heatmaps. The relative expression scale is visualized on the right. If clusters with greater than ten members were identified, these are indicated on the left side of each heatmap.

8.7 For cases where multiple spatiotemporally regulatory modules were identified, statistically significant enrichment expression within cell types was also tested. The p-value scale obtained using the hypergeometric distribution is indicated on the top of each panel. a) Shift of +2. Many clusters display distinct cell type enrichment profiles. b) Shift of +3. At least two significantly cell type-enriched clusters were identified. c) Shift of -2. Among many clusters, specific cell-type enrichment was identified in only a single cluster.

9.1 Relationship between observable markers and cell-cycle position. The colored portion of each band represents the presence of that marker during the cell cycle. The gradient of color in the DNA content band (orange) indicates the ability of the marker to resolve positions in S phase as well as G2/M.

9.2 cloCCS lifeline representations. a) The current cloCCS linear lifeline representation which allows for the modeling of one-time events, but requires Gr and Gd marker behaviors (budding, gray areas) to be the same. b) An improved cloCCS lifeline which allows for more detailed modeling including distinct behaviors for recovering, mother, and daughter cells.
9.3 *Cis*-elements regulating spatiotemporal expression. A hypothetical example of four genes showing their expression patterns (red=high expression) and their promoter sequence. In this example there are four *cis*-elements regulating expression: ATTTA (red) activates expression in hair-cells, ACCTA (purple) activates expression in the elongation zone, GACGT (green) activates expression in the maturation zone, and CCTGG (blue) activates expression in the cortex. The additive combinatorial control of these elements produces the observed expression patterns.
Acknowledgements

Computational biology is an intensely collaborative field, and the work in this thesis would not have been possible without the contributions and assistance from many different wonderful people. Most importantly I would like to thank my three main advisers: Dr. Philip Benfey, Dr. Steven Haase, and Dr. Alexander Hartemink for their support, encouragement, occasional instigation, and countless insightful suggestions. I was fortunate and honored to be able to collaborate with many other great scientists during the course of the various research projects presented herein. I would like to thank Dr. Edwin Iversen for his infinite patience in teaching me statistics and for his invaluable contributions in the creation of the CLOCCS model. Dr. Allister Bernard was instrumental in the development of an efficient implementation of CLOCCS and spearheaded the development of the CLOCCS-based deconvolution method. Jurgen Sladeczek and Josh Robinson of the Hartemink lab also made important contributions in developing a JAVA based version of CLOCCS. I am indebted to Dr. Siobhan Brady for her tireless work and insights during our collaborations which produced the expression map of the *A. thaliana* root as well as the deconvolution method for that data (also in collaboration with Dustin Cartwright) and the method for identifying multiple time-scales in expression data (in collaboration with Dr. Sebastian Ahnert).

I would also like to thank Charles Lin and the rest of the Haase lab for all their contributions and for their camaraderie which made coming into the office everyday something to look forward to. Also making the time I spent at Duke much more enjoyable were my fellow students in the Computational Biology and Bioinformatics program, especially Jeff
Headd, who somehow managed to put up with me as a roommate for our entire graduate careers, and Elena Edelman who was always there to remind me of life outside research (and Joe Volpe who kept me fed when neither of them would cook). I cannot adequately express my gratitude for your friendship, encouragement, cooking, and support.

Finally, I would like to thank my parents for sharing their unlimited intellectual passion and curiosity with me. None of this work would have been possible without your unwavering support and guidance throughout the years.
Biology is replete with diverse, complex, and carefully orchestrated processes ranging from the replication of a single cell, to the development of an entire organism. The instructions for performing all the processes necessary for an organism’s morphogenesis and survival are encoded in its genome. The genome is a ‘simple’ sequence (or set of sequences) composed of the nucleotides \( A \) (adenine), \( T \) (thymine), \( C \) (cytosine), and \( G \) (guanine). The functional units of the genome are genes, most of which encode for a protein when expressed (i.e. transcribed from DNA to RNA and subsequently translated into protein). It is primarily through the coordinated activity of these proteins that an organism is able to perform the complex processes we observe. The successful completion of these processes depends on the carefully regulated production and utilization of proteins at precise times and in precise locations within the organism.

Regulating a protein’s biological activity by controlling the expression of its gene at the level of transcription is a common theme in virtually all organisms. Although protein activities are often regulated post-transcriptionally, both transcription and translation are energetically expensive. Thus, regulating transcription is an efficient mechanism for
controlling cellular processes. Advances in high-throughput microarray technologies have given researchers a powerful tool for monitoring the transcription levels of nearly every gene in an organism as it progresses through diverse biological processes. The focus of this thesis is on using whole genome microarray technology and computational techniques for understanding how an organism regulates the complex transcriptional programs essential for the successful completion of fundamental biological processes—the cell cycle and the acquisition of cellular identity.

We begin chapter with a description of the mechanisms used by eukaryotic cells to regulate expression. We then discuss the role of transcriptional regulatory networks in two very different, albeit fundamental processes: the cell division cycle and organ development, using the yeast Saccharomyces cerevisiae and the Arabidopsis thaliana root as model systems for these two processes, respectively. Finally, we describe some of the genomic tools used to measure transcriptional regulatory networks and conclude with a brief outline for the remainder of this thesis.

1.1 Regulation of expression

Although the regulation of a gene and its corresponding product(s) can take place at many levels, for the purposes of this discussion we restrict our focus to the regulation of transcription of a gene’s DNA sequence into its primary mRNA transcript. Indeed, mRNA abundance is a commonly used, simple yet successful, proxy for protein activity. This definition excludes critical post-transcriptional regulatory processes, including mRNA processing, differential degradation, selective translation, and all post-translational processes which may be necessary for production of active protein. Details of those important processes are reviewed in Alberts et al. [19].

Transcription of a gene is generally controlled by the interaction between regulatory trans-acting proteins called transcription factors (TFs) and the cis-acting regulatory re-
region (promoter) of the gene. TFs have specialized structural DNA-binding domains which allow them to selectively recognize and bind to specific short DNA motifs, known as cis-elements, which may be present in the promoters of each genes. It is the binding of TFs to the promoter which modulates the activity of the basal transcriptional machinery and RNA polymerase which is responsible for the actual transcription of DNA into primary RNA transcript. In eukaryotes, DNA is packaged by histones into a nucleoprotein complex known as chromatin. The basic unit of chromatin is the nucleosome, which comprises a short region of DNA wrapped around a histone octamer. This packaging of DNA plays an important regulatory role in transcription as it controls the accessibility of the promoter sequence to TFs and the basal transcriptional machinery, and is therefore commonly thought to repress transcription [20, 21].

Apart from this regulation by chromatin, the rate of a genes transcription depends on the positive and negative effects of the TFs binding to the cis-elements in its promoter. The activity of TFs fall into two simple classes: activators and repressors. Activators usually contain an activator domain which aids in the localization of the RNA polymerase to the proper place on the promoter to induce transcription [22]. Repressors can inhibit transcription in multiple ways (reviewed in Latchman [23]). They can inhibit the activity of activator TFs by sterically blocking the binding of activator through binding to the same or nearby cis-elements, or they can interfere with the activator’s activity directly. Additionally, repressors can interact directly with the basal transcriptional machinery to prevent transcription.

The precise transcriptional regulation of individual genes is achieved through a combination of activators and repressors. A classic example of combinatorial transcriptional regulation is the regulation of the genes required for galactose metabolism in S. cerevisiae. These genes (GAL1, GAL2, GAL7, GAL10 and MEL1) are induced by the binding of the activating transcription factor Gal4 to a specific cis-element present in the promoters of the genes. Repression of these genes is achieved by the direct binding of the repressive
TF *GAL80* to *GAL4*, preventing it from activating transcription[24]. A more complex example is the regulation of the *Drosophila melanogaster* gene *even-skipped (eve)* which is expressed in seven distinct stripes in the *D. melanogaster* embryo (reviewed in Howard and Davidson [25]). The regulation of *eve* stripe 2 is the most well characterized. This stripe of expression is induced by the binding of the activating TFs *Bicoid* and *Hunchback* and its anterior/posterior spatial domain is regulated by the binding of the repressive TFs *Giant* and *Kruppel*. These are just two examples of transcriptional regulation for individual genes, but many of the complex processes that occur in an organism require whole suites of genes to be transcribed in a precise spatial and temporal order. In the next section, we discuss the role of transcriptional regulatory networks in directing these comprehensive expression programs.

### 1.2 Transcriptional regulatory networks

Transcriptional regulatory networks are comprised of transcription factors (nodes) that regulate the expression (edges) of one another and together regulate the expression of downstream genes (Figure 1.1a). Transcriptional regulatory networks have been shown to be important in the regulation of many processes essential to an organism [26]. The careful, experimentally supported, study of transcriptional regulatory networks has thus far been limited to the study of small targeted networks. This is due to the high complexity present in, and the difficulty in carrying out experiments on, larger networks which regulate more general processes. However, in this thesis, we focus on transcriptional regulatory networks underlying two different types of very general processes: regulation of periodic expression during the cell cycle and the regulation of spatiotemporal expression during development.
Figure 1.1: Sample transcriptional regulatory network. a) A small transcriptional regulatory network with five interconnected transcription factors. Each node represents a TF and a directed edge between two nodes represents a regulatory interaction (green=activating, red=repressing). b) A four-state periodic pattern is an emergent property of the network. Active TFs at each state are shown in yellow.

1.2.1 Periodic regulation

Periodic processes, in which the final state is the same as the initial state (Figure 1.1b) are common in biological systems. Regulating one of these periodic processes requires a mechanism which intrinsically resets itself in order to ensure proper regulation during the next cycle. Transcriptional regulatory networks can function as oscillators; allowing for the regulation of periodic processes. Elowitz and Leibler [27] constructed a synthetic three-gene transcriptional regulatory network, called the repressilator, that exhibits oscillatory behavior in *E. coli*. In this very simple transcriptional regulatory network, oscillations are generated by the serial expression of transcriptional repressors, with the last gene in the series inhibiting the expression of the first. A relatively simple feedback-regulated transcriptional regulatory network forms the basis of the circadian oscillator that regulates
periodic events in a diverse collection of organisms (reviewed in Dunlap [28]). This transcriptional oscillator can function autonomously, but its function is normally reinforced by periodic environmental cues such as light/dark cycles.

Periodic events during cell division in the bacteria, *Caulobacter crescentus*, appear to be controlled by a transcriptional regulatory network oscillator. Three TFs promote the waves of transcription which regulate progression through the major events in the cell cycle including DNA replication and fission. In addition, these TFs mutually regulate their transcription to create a cell-cycle oscillator (reviewed in Holtzendorff et al. [29]).

The existence of a cell-cycle program driven by transcription in bacteria suggests that an ancestral eukaryotic cell cycle may have been regulated by a transcriptional oscillator. It has been hypothesized that a transcriptional regulatory network forms a cell-cycle oscillator in modern eukaryotes. However, the existence of a biochemical oscillator centered around cyclin/CDK activity has confounded experimentally addressing this hypothesis (discussed in more detail in Chapter 2). However, as detailed in Chapter 4, by genetically manipulating the model eukaryote *S. cerevisiae*, we can study cell-cycle oscillations in the absence of this confounding CDK oscillator and examine the validity and nature of a transcriptional regulatory network regulating the periodic expression during the eukaryotic cell cycle.

1.2.2 Spatiotemporal regulation

There are examples of both temporal and spatial control of transcription by transcriptional regulatory networks for particular phases of development in a variety of model organisms (reviewed in Levine and Davidson [30], Pourquié [31]). For example, transcriptional regulatory networks control the specification of cell fate in early sea urchin embryos as well as the dorsal-ventral patterning of *Drosophila melanogaster* (reviewed in Davidson et al. [32]). Work on these systems demonstrates that transcriptional regulatory networks play a role during development, but has not measured the complete developmental trajectory.
from beginning to end. To understand the nature of a transcriptional regulatory network underlying the complete development and differentiation trajectory of an organ, it is important to obtain measurements of each tissue at each stage of development. While this is a daunting technical challenge in most systems, by using the *A. thaliana* root as a model developmental system (see Chapter 5 for details) we are able to generate a complete map of expression through the development and differentiation of an entire organ and investigate transcriptional regulatory networks regulating that development (Chapter 6).

### 1.3 Genomic tools

The study of any transcriptional regulatory network is difficult because the complex pattern of regulation it creates is an emergent property of the network. Emergence is the way a complex pattern arises through the combination of many simple interactions (i.e. complex patterns of expression arising from simple TF–TF interactions) (Figure 1.1b). Since an emergent property cannot be understood by examining each of the component interactions in isolation, a global view must be taken in order to understand the emergent property of complex transcriptional regulation. With the advent of high-throughput genomic tools, we can now measure systems at the global level. In the subsections below, we describe two of these high-throughput technologies and illustrate how they can be used to study transcriptional regulatory networks.

#### 1.3.1 Gene expression microarrays

Gene expression microarrays are used to obtain a system wide measurement of gene expression. Since transcriptional regulatory networks act by dynamically regulating the transcription of genes, measurements of the gene expression of every gene are measurements of the entire output of the network. Further, because the TFs within the transcriptional regulatory network are themselves transcriptionally regulated, the same microarray measurements also describe the state of the network itself.
A microarray is a rectangular slide with thousands of short stretches of DNA, called probes, chemically bonded to it. Each probe is specific to a sequence from some gene. mRNA from a sample is extracted, converted to complementary DNA (cDNA), labeled, hybridized to the probes, and measured by either phospho-imaging or fluorescence scanning (for a detailed review of microarray technology see Schulze and Downward [2]). The measured hybridization intensity of labeled cDNAs to each probe is positively correlated with the amount of mRNA from the corresponding gene in the sample. The two most common microarray technologies are cDNA and oligonucleotide microarrays (Figure 1.2a). cDNA arrays use probes generated from cDNA libraries which are robotically printed onto glass slides as defined spots. Oligonucleotide arrays use short 20-25bp probes synthesized in-situ, by photolithography (Affymetrix®) or by ink-jet technology (Agilent®). While the cost of cDNA arrays are considerably lower, the technology only allows the relative difference in expression level between two samples to be measured, while oligonucleotide arrays can be used to measure absolute levels of expression within a single sample. All of the work in this thesis uses the Yeast 2.0© and ATH1© oligonucleotide arrays from Affymetrix® which cover almost every gene in the S. cerevisiae and A. thaliana genomes, respectively.

1.3.2 ChIP-chip

Chromatin immunoprecipitation (ChIP) followed by microarray hybridization (chip) assays are used to measure the system-wide binding of TFs in vivo. A sample of cells expressing a TF of interest is fixed with formaldehyde which crosslinks the TF to the DNA it is bound to in vivo (Figure 1.2b, for a more complete review see Buck and Lieb [3]). The DNA is then extracted and fragmented into small pieces. The tagged TF is subsequently immunoprecipitated with an antibody that specifically recognizes the TF, separating bound DNA from the remainder of the genome. The crosslinking is then reversed, producing a collection of DNA fragments which are enriched for binding by the TF of interest. The
relative level of enrichment of these fragments can be measured by hybridization to a whole-genome DNA tiling microarray. This ChIP-chip data provides a direct measure of the global interactions of a TF with the promoters of all other genes. This type of data can be used in conjunction with microarray expression data to generate hypothetical transcriptional regulatory networks [5, 33, 34]. A large collection of ChIP-chip data for 200+ transcription factors in *S. cerevisiae* is available, providing a valuable resource for understanding the individual interactions that lead to the emergent properties of transcriptional regulatory networks in that organism [35].
1.4 Thesis roadmap

This thesis is organized into three main parts: *S. cerevisiae*, *A. thaliana*, and computational methods applicable to both systems. The first part, comprised of Chapters 2, 3, and 4, discusses work concerning the study of the *S. cerevisiae* cell cycle. Chapter 2 provides an introduction to *S. cerevisiae* and the cell cycle. Chapter 3 describes the CLOCCS model and its application to studying the cell cycle. Chapter 4 details experimental work which provides evidence for a transcriptional regulatory network controlling periodic expression in the *S. cerevisiae* cell cycle. The second part, comprised of Chapters 5 and 6, relates to work in *A. thaliana* root development. Chapter 5 provides an introduction to *A. thaliana* and the reasons for using its root as a model for development. Chapter 6 details the generation of the complete expression map of the *A. thaliana* root and provides some initial analysis of the complex transcriptional regulation underlying its development. The third part, comprised of Chapters 7 and 8, focuses on the three computational tools which are applicable to both organisms. Chapter 7 describes two deconvolution approaches and their application to data generated in Chapters 4 and 6. Chapter 8 discusses a technique for reducing noise in these datasets by identifying biological processes occurring at multiple time-scales. Finally, Chapter 9 provides a summary of the contributions of this thesis and describes some potential next steps for the work presented herein.

It is important to note that all of the work presented herein was done in collaboration with other scientists and their contributions are recognized in the acknowledgements. Additionally, much of the work described has been published previously and portions of those publications are reproduced throughout [11, 15, 36, 37, 38, 39].
Saccharomyces cerevisiae, more commonly known as baker’s or budding yeast, is a classic model system for studying eukaryotic cell biology. It is small unicellular fungi of the phylum Ascomycetes and class Saccharomycetes. It has a compact, fully sequenced, genome with 16 chromosomes (1.2 Mb total) and approximately 6600 genes (4753 verified) [40]. It has become a classic model system for a variety of reasons, including its rapid growth (doubling time approximately 80 min at 30°C), informative cellular morphology, but most importantly, because of the awesome power of yeast genetics. With stable haploid and diploid states and viability under a large number of auxotrophic and antibiotic markers, S. cerevisiae is an extraordinarily versatile and efficient platform for DNA transformation and manipulation, far better than any other non-yeast eukaryote [41]. An endless variety of ghoulish tinkering (genetic manipulations) can be carried out in S. cerevisiae due to the fact that it exclusively uses homologous recombination for integrating transformed DNA. It is the ease with which genes can be manipulated, in addition to the high degree of conservation of metabolic and regulatory mechanisms between yeast and humans [41, 42] that
have made *S. cerevisiae* the model system of choice for studying eukaryotic cell biology.

For the purposes of this thesis, we use *S. cerevisiae* as a model system to study the transcriptional regulation during the cell division cycle. In Section 2.1 we give an overview of the basic *Saccharomyces cerevisiae* cell cycle and highlight particular features which will be relevant in later chapters. In Section 2.2 we provide an overview of the current models for transcriptional regulation during the cell cycle.

It should be noted that there are many common laboratory stains of yeast. S288C [43] is the strain sequenced for the *Saccharomyces* Genome Database, while other strains such as W303 [44], SK1 [45], and BF264-15D [46] are used extensively to study the yeast cell cycle. For the rest of this work, when we refer to *S. cerevisiae*, we are referring to the BF264-15D strain.

### 2.1 Cell cycle basics

The *S. cerevisiae* cell cycle, as with most other organisms, is traditionally divided into four phases: G1, S, G2, and M (Figure 2.1a). These phases are named for the processes that occur within them. G1 and G2 stand for growth or gap phases, S for DNA synthesis, and M for mitosis. During G1 the cell grows in size while monitoring its environment. Once a cell has grown past a critical size and determines that its environment is conducive to division, it will pass through the G1 Start restriction point. Thus the cell commits itself to progressing through the cell cycle [47, 48]. The cell then transits through S-phase when replication of its genomic DNA occurs, before proceeding with mitosis during M phase. Coincident with the transition from G1 into S is the formation of the bud. This bud continues to grow throughout S and G2 phase and will become a new cell at the completion of mitosis. Unlike many other eukaryotic cells, *S. cerevisiae* have a closed mitosis where the nuclear envelope remains intact. The nuclear envelope is pinched into two new nuclei and one is shuttled into the bud. At the completion of mitosis the cell undergoes
cytokinesis when the cytoplasm shared between the between the bud and the original cell is divided and both cells begin to operate independently. However, it is important to note that for a period of time, these two independent cells will be physically joined until a new cell wall is formed and cell separation occurs, thereby producing two separated cells. For the rest of this work, we refer to the progenitor cell from which the bud emerges as the mother cell and what had been the bud as the daughter cell. In *S. cerevisiae* cell division is asymmetric, with the daughter cell being smaller than the mother cell. As there is a minimal size which much be achieved before progression through *Start* can occur, this asymmetric division can result in unequal G1 phase lengths between mother and daughter cells [47, 48].

### 2.1.1 Features of cell cycle progression

A variety of landmarks have been defined that can be used to determine the cell-cycle state of individual cells [49, 50, 51]. Figure 2.1a depicts two of these landmarks, budding and DNA content. The first, bud emergence, is a distinct morphological landmark easily detected by simple light microscopy. It first appears near the time that a cell transitions from G1 into S phase. Cells become unbudded after the completion of mitosis (M) when the mother cell and its bud separate. Thus, the passage of a population of cells from G1 (unbudded) to S (budded), or from M (budded) into G1 (unbudded) can be measured by plotting the percentage of budded cells (referred to as the budding index) as a function of time (Figure 2.1b). Cell cycle position can also be determined by measuring the genomic DNA content of the cell [52, 53, 54]. Haploid yeast cells begin the cell cycle with one copy of genomic DNA (1C, red bar in Figure 2.1a). During S phase, DNA is replicated such that, at the completion of S phase, the cell has two copies of genomic DNA (2C). The DNA content of thousands of cells can be rapidly measured using a flow cytometer. The genomic DNA is labeled with a fluorescent dye (such as SYTOX™ green, Invitrogen) and then the flow cytometer bins each cell into one of 1024 ordered channels on basis of its
fluorescent intensity which is proportional to its DNA content [55]. An example of typical DNA content flow cytometry data is shown in Figure 2.1c.

Although some information about the cell cycle position of a cell can be learned by determining budding index and DNA content separately, the resolving power of these landmark data is limited. Budding index can only tell us what fraction of cells is in G1. DNA content can discriminate between cells in G1, S, and G2/M, but cannot distinguish between cells in sub-compartments of G1, G2, or M (e.g., early vs. late G1, or G2 vs. M). In Chapter 3 we develop a mathematical model which uses these landmark data and can thus more accurately predict the cell-cycle position of a population of synchronized cells.

2.1.2 Synchronization methods

A variety of methods have been established for synchronizing a yeast cell population at various points in the cell cycle. The two most common are \( \alpha \)-factor block-release and centrifugal elutriation. Both methods are capable of producing large populations of G1-phase synchronized cells and each are discussed below.

\( \alpha \)-factor

Synchronization by \( \alpha \)-factor block-release is achieved by adding the \( \alpha \) mating pheromone to an asynchronous culture (the block) and then subsequently removing the pheromone (the release). The \( \alpha \) mating pheromone is a 13 amino acid residue peptide which binds to the receptor Ste2 in MAT\( a \) cells and induces a cascade which results in the inactivation of the G1 cyclin/CDK kinase complexes leading to a G1-phase arrest [56]. To synchronize cells with \( \alpha \)-factor, a small quantity of the pheromone is added to a culture of MAT\( a \) cells. When the culture has a budding index > 90%, the \( \alpha \)-factor is removed and the culture is released. Populations synchronized with \( \alpha \)-factor tend to maintain synchrony for longer periods of time than populations synchronized by centrifugal elutriation. This is due to differences in initial cell size between the two methods. As discussed below, centrifugal
FIGURE 2.1: The *S. cerevisiae* cell cycle and measurable cell-cycle progression features. a) Over the course of its life, the cell repeatedly traverses the cell cycle, which is divided by landmark events into the G1, S and G2/M phases. The transition from G1 to S is marked by the development of a bud (dark gray) and the start of DNA replication (denoted by the appearance of a second red bar). The S/G2 transition is marked by the completion of DNA synthesis. The cell cycle is completed when the daughter cell (dark gray) separates from the mother cell at the end of G2/M. b) A typical budding index curve. c) A typical DNA content histogram as measured by flow cytometry for an asynchronous population.

elutriation generates a population of small cells, while cells in α-factor synchronized populations are very large (due to their continued growth between block and release). Since large cells have large daughters, synchrony loss due to the effects of asymmetric division is reduced in α-factor synchronized cells. However, because the α-factor synchrony method relies on an extra-cellular signal, it induces large cellular changes in addition to the cell-cycle arrest, including significantly altering the G1 transcriptional program [57].
Centrifugal elutriation

An alternative to the α-factor block-release methods is synchronization by centrifugal elutriation. Centrifugal elutriation is a size-based method which takes advantage of the fact that the smallest cells in a culture will be newborn daughters in G1-phase. Cells in liquid media are pumped into a rapidly spinning chamber. The centrifugal forces cause a gradient to form with cells sedimenting at the bottom (outside) of the chamber and fluid eluting out the top (inside) through the exit port. Given that small cells have a higher surface/volume ratio than larger cells their sedimentation are relatively more affected by the rate of fluid flow. By carefully adjusting the pump and centrifuge speed and measuring the budding index of the output fractions, small unbudded (G1) cells can be selectively washed out of the chamber and collected. The extreme forces involved in this size selection cause cells to have a longer period of recovery time before beginning the cell cycle than in other methods. Additionally, as mentioned before, populations synchronized by centrifugal elutriation will lose synchrony faster than other methods because the small cells are significantly affected by the asymmetric nature of cell division in S. cerevisiae. However, due to the fact that centrifugal elutriation is a size-based selection, and not an induced arrest, there is theoretically very little transcriptional alteration of G1 events.

2.2 Regulation of the cell cycle

The cell division cycle can be viewed as the ordered execution of duplication and segregation events. Years of intense research has been focused on understanding the mechanisms that control and regulate those events. Early work in frog and marine invertebrate embryos suggested that cell cycle events are triggered by the activity of a biochemical oscillator centered on cyclin/CDK complexes [58, 59, 60] (reviewed in [61, 62]). These complexes are composed of a cyclin protein that binds to a serine/threonine kinase (CDK). Cyclins are important to activate CDK and direct the kinase to specific targets. Periodic events
are triggered as cyclin/CDK activities rise and fall, and are coincident with oscillations in cyclin levels. As there is no significant transcription in early embryonic cells [61, 62], cyclin oscillations are driven primarily by the CDK-mediated activation and deactivation of cyclin protein degradation (reviewed in Murray [63]). Thus, the cyclin/CDK complex and its regulators can function as a self-limiting cell-cycle oscillator even in the absence of transcription.

In *S. cerevisiae* there is one CDK, Cdc28/Cdk1, and nine major cyclins: G1 cyclins Cln1–3 and B-type cyclins Clb1–6. All nine of these cyclins bind to and activate Cdc28. At least one of the G1 cyclins is required for progression through G1 phase. In a ∆*CLN1-3* strain, cells arrest in G1 without completing bud emergence or spindle pole body duplication [64]. Progression through S-phase requires at least one of the B-type cyclins (Clb1–6) (although in ∆*CLB5,6* strain the start of DNA replication is delayed). Finally, the progression through a successful mitosis requires the activity of at least one of Clb1–4 [4].

When the genes encoding cyclins and CDKs were identified in genetic screens as critical cell-cycle regulators in yeasts [65, 66] it seemed clear that they would also form the central cell-cycle oscillating mechanism in this non-embryonic system. This view of cell-cycle regulation was challenged by evidence for a cyclin/CDK-independent oscillator in budding yeast [67]. Subsequent studies, including the work discussed in Chapter 4, have proposed that this cyclin-independent regulation could be achieved through the oscillations of a transcriptional regulatory network [11, 34, 35]. In these proposed networks, transcription factors expressed in one cell-cycle phase bind to the promoters of genes encoding transcription factors that function in a subsequent phase. Thus, the temporal program of transcription could be controlled by sequential waves of transcription factor expression.

This hypothetical transcriptional oscillator is not inconsistent with the previous model of cyclin/CDK cell-cycle regulation. In fact, the notion that cells are driven through the cycle by successive waves of cyclin expression [68] necessitates a mechanism for establishing these waves of expression. As an example, it is known that the expression of
cyclins early-expressed in the cell cycle are controlled by the heterodimeric transcription factor complexes SBF and MBF, which are composed of Swi6 and either Swi4 or Mbp1 respectively. CLN1 and CLN2 expression is primarily regulated by SBF [69, 70], while activation of CLB5 and CLB6 transcription is regulated by MBF [71]. Preceding this in the regulatory network, the transcription factor Mcm1 which is expressed late in M/early G1 induces the expression of the early G1 cyclin, CLN3. Cln3, through association with Cdc28, then phosphorylates the transcriptional repressor of SBF, Whi5. The phosphorylation of Whi5 disassociates it from SBF allowing SBF expression to increase and subsequently the induction of later cyclins (and G1-specific transcription) [72, 73]. This small regulatory circuit is shown in Figure 2.2. This regulatory circuit is of special interest because it is very similar to the regulation of the metazoan G1/S transcription factor E2F by the retinoblastoma tumor suppressor protein, Rb, suggesting that these transcriptional circuits may be a fundamental mechanism found in all eukaryotes [74, 75].

**FIGURE 2.2:** The transcription repressor Whi5 inhibits the activity of the SBF transcription factor. Phosphorylation of Whi5 by Cln3Cdc28 induces the nuclear export of Whi5 and activates SBF, which induces the transcription of the genes that encode Cln1, Cln2, Clb5 and Clb6. Cln-Cdc28 and Clb-Cdc28 phosphorylate Whi5, which might provide a positive-feedback loop. Image and legend adapted from Bloom and Cross [4]

The regulatory circuit shown in Figure 2.2 is a mix of the two regulatory models discussed earlier: regulation by cyclin/CDK (phosphorylation of Whi5 by Cln3-Cdc28), and a transcriptional regulatory network (Mcm1 indirectly represses Whi5 which is a repressor of SBF). This interplay between the models makes the study and validation of either model difficult to perform in wild-type cells. In Chapter 4 we alleviate this problem by
studying a mutant strain in which the majority of the cyclin/CDK regulatory machinery has been disabled. Thus, any transcriptional regulation in that strain must be a product of the transcriptional regulatory network. By combining carefully designed experiments utilizing genetically engineered yeast strains with the accurate measurements of expression through the cell cycle, we can begin to understand the transcriptional regulatory mechanisms that regulate cell division in eukaryotes.
Modeling cell cycle population dynamics

3.1 Introduction

To fully understand the mechanisms that regulate cell division, it would be ideal if quantitative measurements of the dynamic processes could be made on a single cell as it progresses through the cell division cycle. In many cases, accurate measurements cannot be made on single cells because the quantitative methods lack the sensitivity to detect small numbers of biomolecules. For example, accurate quantitative measurements of genome-wide transcript levels by microarray require more mRNA than is physically available within a single cell. Thus, researchers are forced to work with populations of cells that have been synchronized to a discrete cell-cycle state.

Two distinct problems arise in these synchrony/time-series experiments. First, synchronized populations are never completely synchronous to begin with, and tend to lose synchrony over time. This lack of perfect synchrony at any given time leads to a convolution of the measurements that reflects the distribution of cells over different cell-cycle states. Second, multiple synchrony experiments are often needed to measure different aspects of a process, and it is often desirable to compare the temporal dynamics of these
aspects. However, synchrony/time–series experiments, even in the best of experimental circumstances, exhibit considerable variability which make cross-experiment time-point to time-point comparisons imprecise. Thus a mechanism is required to accurately align the data collected from each of the synchrony/time–series experiments. The model we describe in this chapter addresses both of these problems.

In this chapter we describe the CLOCCS population model (in reference to its ability to Characterize Loss Of Cell Cycle Synchrony), budding index and DNA content observation models (used to connect the population model to observed data), and show that the combinations of the two components is able to accurately describe population synchrony loss under many different conditions. We also show that the unique sources of asynchrony used by CLOCCS are a significant improvement over previously available models. We demonstrate the practical applications of the model by aligning data collected under different conditions and across organisms, and briefly discussion the use of CLOCCS for deconvolution (details of this application can be found in Chapter 7.2). We conclude with a discussion of the contributions of the model. The CLOCCS model has been published in Orlando et al. [36].

3.1.1 Previous approaches

There have been many previous efforts to model the dynamics of population synchrony loss in the literature. Most of the numerous models for cell population dynamics in synchrony/time–series experiments fall into two related classes: population balance (PB) and branching process (BP) models. PB models are usually formulated as partial-integro-differential equations and are often very difficult to work with except under special conditions [76, 77]. BP models are stochastic models for population dynamics that have been used to study both the asymptotic [78] and short term behaviors [36, 79] of populations; certain BP models have PB analogues [80]. Several models that do not explicitly account for reproduction, and hence are neither PB or BP models, have also been used to model
data from asynchrony experiments [81, 82].

The most critical distinction between models, however, is in the sources of synchrony loss the model includes. Most describe synchrony loss as the result of a single parameter, equivalent to a distribution over division times [79, 81, 83]. In contrast, the CLOCCS model we describe here is the only model to account for variability in cell-division time, initial asynchrony in the starting population, and variability due to asymmetric cell division [83], all of which we will show to be important. The CLOCCS model is based on a novel branching process construction and can be written in closed form. Its formulation is constructive, based on an accounting of unique cohorts in the population at any given time. Hence, the model can attribute one-time effects to specific subsets of the population, demonstrating flexibility not available using the PB and BP models historically applied to these populations. Further, the model’s construction allows for full Bayesian inference without the use of approximations to the likelihood. The Bayesian approach to inference has the additional advantage that it sidesteps many of the difficulties encountered by frequentist inference for BP models [84].

3.2 CLOCCS model

To describe synchrony loss during the cell cycle, we created a probabilistic model framework, CLOCCS (Characterizing Loss Of Cell Cycle Synchrony). In our framework, we visualize the movement of a synchronized population of cells using a linear graph we term a ‘lifeline’ (Figure 3.1). Although the cell cycle is often graphically depicted in a circular form (as in Figure 2.1a), a linear representation more easily allows us to consider events that occur only once in the lifetime of each cell, such as effects related to the synchrony procedure or the daughter-specific delay in early G1 (when a cell divides, we refer to the progenitor cell as the mother and what had been the bud as the daughter) [47].

The CLOCCS model specifies the distribution of cell positions over an abstract cell-
cycle lifeline as a function of time. We define \( \lambda \) to be the amount of time, in minutes, required by a typical mother cell to undergo one full cell cycle. We divide the lifeline into \( \lambda \) units, thus the average cell will move one lifeline unit per minute. We model position as having three independent sources of variability: the velocity with which the cell traverses the cell cycle, the time it spends recovering from the synchronization procedure, and the additional time spent by a daughter cell as it traverses its first cell cycle. It is well known that cells in synchrony experiments progress through the cell cycle with varying speeds. We assume that each cell moves at a constant velocity along the lifeline, and that this velocity is random, following a normal distribution. While this is technically inappropriate as velocities must be positive, in practice it is reasonable: fitted distributions give almost no mass to the negative half line. We measure velocity, \( V \), in lifeline units per minute; by definition, the mean cell velocity is 1.0. The velocity distribution’s variance, \( \sigma^2_v \), is unknown.

When released from synchrony, cells spend more time in their first G1 phase than they spend in G1 during subsequent cell cycles. The added time reflects a period of recovery from the synchronization process, whose length varies from cell to cell. We term this recovery period Gr as if it were a distinct cell-cycle phase. We model this effect as a random offset, \( P_0 \), in the starting position on the lifeline. While this offset should be strictly positive, we let \( P_0 \) be distributed \( N(\mu_0, \sigma^2_0) \) for convenience. In Chapter 9 we comment further on this choice. In \( S. \) cerevisiae, cell division is often asymmetric: the mother cell is often larger and progresses more quickly through the cell cycle than the daughter [47]. There-
fore daughter cells require additional time in G1 before they begin to divide. We term this daughter-specific period of growth Gd, and model it by introducing a fixed offset, $\delta$, to the cell’s lifeline position.

With each wave of division, the population expands in size. If cells in the culture remained synchronous, the population would branch and double in size every $\lambda$ minutes after an initial delay of $\mu_0$ minutes. Because they do not, the dynamics of this expansion is more complex: at any point in time, the population may represent a number of distinct cohorts, each defined by its lineage. Cohorts are determined by $g$, their ‘generation’—the number of daughters stages in their lineage—and $r$, their ‘reproductive instance’—the wave of division that gave rise to the cohort. Figure 3.2a depicts the branching dynamics of this process and a snapshot in time projected onto a common lifeline (Figure 3.2b). In Figure 3.2a, four distinct time periods are color coded with each cohort distribution labeled with its $\{g, r\}$ index. At time zero (black) there is a single cohort, $\{0, 0\}$, whose position distribution is located in Gr and centered at $-\mu_0$. As time passes (red), this cohort enters its second cell cycle and spawns a daughter cohort, labeled $\{1, 1\}$, which begins on its own lifeline in Gd. Later (blue), cohort $\{0, 0\}$ gives rise at its second reproductive instance to another first generation cohort, $\{1, 2\}$. At the same time, cohort $\{1, 1\}$ cells are progressing through G2/M. At the last depicted time point (green), the population is comprised of four distinct cohorts, representing three generations of cells arising at three distinct reproductive instances. Figure 3.2b is a plot of the population at this time point on the common lifeline. The CLOCCS model is a distribution over position along this common lifeline as a function of time.
Figure 3.2: Graphical representation of the branching dynamics of the cell population.  

(a) The position distributions of the cohorts, indexed by \( \{g, r\} \), in the population at four points in time, each color-coded.  

(b) The population at the last time point (green) projected onto a common lifeline.
3.2.1 Mathematical formulation

Since cells in the \( \{0, 0\} \) cohort are unaffected by the daughter specific delay, \( \delta \), their positions, \( P_t \), at time \( t \) are determined only by their starting positions \( P_0 \) and their velocity, \( V \). For these cells, \( P_t = P_0 + V t \). Hence, \( p(P_t|\Theta, R = 0, G = 0, t) \), is normal with mean \(-\mu_0 + t\) and standard deviation \( \sqrt{\sigma_0^2 + t^2 \cdot \sigma_v^2} \). In contrast, cohorts at generations greater than zero have their position distributions truncated at the beginning of \( G_d, -\delta \), on the lifeline, and include \( g \) daughter offsets and \( r \) cell-cycle offsets. For these, \( P_t = P_0 + V t - g \delta - r \lambda \), hence position, \( p(P_t|\Theta, G \geq 0, R \geq G, t) \), is normal with mean \(-\mu_0 + t - g\delta - r\lambda\) and standard deviation \( \sqrt{\sigma_0^2 + t^2 \cdot \sigma_v^2} \) truncated so that \( P_t \geq -\delta \).

Thus we write the model for position, \( P_t \), given time, \( t \), in closed form by enumerating the population’s cohorts using the latent variables \( r \) and \( g \). In particular,

\[
p(P_t|\Theta, t) = \sum_C p(P_t|\Theta, r, g, t)p(g, r|\Theta, t),
\]

where \( \Theta = (\mu_0, \sigma_0^2, \sigma_v^2, \lambda, \delta) \) and where the sum is over possible cohorts,

\( C = \{g, r\} : (g = 0 \land r = 0) \lor (0 < g \leq r \leq R) \}. \) While the number of cohorts represented in the population could theoretically be large, in practice their number is limited by the number of cell cycles that cohort \( \{0, 0\} \) is able to undergo during the experimental period. In most cases, synchrony experiments are terminated after 2 or 3 cycles, so choosing \( R = 4, 5 \) or 6 is usually sufficient. For notational clarity we use \( C \) to represent the sufficient number of cell cycles examined.

The marginal probability of drawing a representative of cohort \( \{g, r\} \) from the population at time \( t \) is \( p(g, r|\Theta, t) \). In the scenario depicted in Figure 3.2b, \( p(1, 1|\Theta, t) \) is the ratio of the mass under the cohort \( \{1, 1\} \) density to the total mass under all of the cohort densities present on the lifeline.
**Calculation of $M_\Theta(g, r, t)$**

The derivation for the mass of a cohort $\{g, r\}$ at time $t$, denoted as $M_\Theta(g, r, t)$, is one of the fundamental concepts of the CLOCCS model and as such deserves an extended description. The problem of determining the mass of any cohort, can be broken into three distinct problems: determining the mass of the $\{0, 0\}$ cohort, determining the mass of direct $\{0, 0\}$ daughters cohorts ($\{1, r\}$), and determining the mass of any other cohort ($\{g \geq 1, r\}$). The first case is trivial, the mass of the $\{0, 0\}$ never changes and can be treated as a constant.

For convenience we fix the mass of the $\{0, 0\}$ cohort as 1.0. The second case, the mass of any $\{1, r\}$ cohort, is also simple to derive. Due to the nature of the cohort generation and the linear lifeline the cohort $\{1, r\}$ is the result of the $\{0, 0\}$ cohort passing over the $r^{th}$ M/G1 transition on the lifeline. Thus, the mass of any $\{1, r\}$ cohort, is exactly equal to the percentage of the $\{0, 0\}$ cohort past the lifeline position $r\lambda$. The third case is complicated because the members of each cohort could be a collection of direct daughters from a number of different cohorts of generation $g - 1$ each passing different G1/M transitions. For example the cells of cohort $\{2, 3\}$ are a mixture between daughters of $\{1, 1\}$ from its second division and daughters of $\{1, 2\}$ from its first division. The mass of cohort $\{2, 3\}$ would therefore be the percentage of $\{1, 1\}$ past lifeline position $2\lambda$ multiplied by its mass plus the percentage of $\{1, 2\}$ past lifeline position $\lambda$ multiplied by its mass. Generalizing we write the mass of cohort $\{g > 0, r \geq g\}$ as

$$
\sum_{i=g-1}^{r-1} [F((r - i), (g - 1), i, t) \cdot M_\Theta(g - 1, i, t)]
$$

where $F(d, g, r, t)$ is the percentage of cohort $\{g, r\}$ past the $d^{th}$ M/G1 transition at time $t$. 

Incorporating the first two cases we write:

\[
M_{\Theta}(g, r, t) = \begin{cases}
1 & g = 0, \ r = 0 \\
F(r, 0, 0, t) & g = 1, \ r \geq g \\
\sum_{i=g-1}^{r-1} [F((r-i), (g-1), i, t) \cdot M_{\Theta}(g-1, i, t)] & g > 1, \ r \geq g \\
0 & \text{else,}
\end{cases}
\] (3.2)

Through work in conjunction with Allister Bernard (detailed in [16]) the recursion in equation 3.2 can be removed and \(M_{\Theta}(g, r, t)\) can be simplified and re-written as:

\[
M_{\Theta}(g, r, t) = \begin{cases}
1 & g = 0, \ r = 0 \\
\left(1 - \Phi\left(\frac{\mu_\theta - t + r \cdot \lambda + (g-1) \cdot \delta}{\sqrt{\sigma^2_\theta + t^2 \cdot \sigma^2_\delta}}\right)\right) \cdot \binom{r-1}{g-1} & g \geq 1, \ r \geq g \\
0 & \text{else,}
\end{cases}
\]

where \(\Phi()\) denotes the standard normal cumulative distribution function.

Also, let \(Q_{\Theta}(t)\) denote the mass under all cohort distributions in the population at that time,

\[
Q_{\Theta}(t) = \sum_{c} M_{\Theta}(g, r, t).
\]

Therefore, in general, \(p(g, r|\Theta, t) = \frac{M_{\Theta}(g, r, t)}{Q_{\Theta}(t)}\).

### 3.3 CLOCCS Sampling models

To utilize the CLOCCS model, it is necessary to relate distributions over the artificial cell cycle lifeline to observable cell features. In the next two sections we present two sampling models which allow CLOCCS to utilize commonly collected landmark data, budding index and DNA-content data. Details of budding index and DNA-content data can be found in section 2.1.1. While time series of budding index and DNA-content data are each sufficient to estimate the CLOCCS parameters, \(\Theta\), they provide complementary information on the cell cycle timing of distinct landmark events. Timing of these events is of independent
interest, and estimates of the same may improve the utility of the model as a tool for deconvolution of transcription data and other types of downstream analysis.

3.3.1 Sampling model for budding index data

Presence or absence of a bud is an easily measured landmark tied to a cell’s progression through the cell cycle (Figure 2.1a). Buds emerge and become detectable at the transition between G1 and S phases, at a fraction $\beta$ of the way through the normal cell cycle and split off as daughter cells at cell cycle completion (Figure 3.3, dashed line).

Assume that budding index samples are drawn at $T$ time points, $t_i$, $i = 1, \ldots, T$, and that $n_i$ cells are counted at time $t_i$. Let $b_{ji} = 1$ if the $j$th cell at time $t_i$ is budded and $b_{ji} = 0$ otherwise. The event $b_{ji} = 1$ implies that the position of the $j$th cell at time $t_i$, $P_{ji}$, falls into the lifeline interval $\left( (c+\beta)\lambda, (c+1)\lambda \right]$ for some cell cycle $c \geq 0$; the probability of this is dictated by the CLOCCS population model.

Following the development of Section 3.2, we calculate $p(b_{ji} = 1|\beta, \Theta, t_i)$ by introducing cohorts and marginalizing over them. In particular, let

$$p(b_{ji} = 1|\beta, \Theta, t_i) = \sum_{c} p(b_{ji} = 1|\beta, \Theta, g, r, t_i)p(g, r|\Theta, t_i),$$

where $p(b_{ji} = 1|\beta, \Theta, g, r, t_i)$ is the probability that a cell randomly sampled from cohort $\{g, r\}$ is budded at time $t_i$. For the progenitor cohort, $\{0, 0\}$,

$$p(b_{ji} = 1|\beta, \Theta, g, r, t_i) = \sum_{c=0}^{C} \left[ \Phi \left( \frac{\lambda \cdot (c+1) - (-\mu_0 + t_i)}{\sqrt{\sigma_0^2 + t_i^2 \cdot \sigma_v^2}} \right) - \Phi \left( \frac{\lambda \cdot (c + \beta) - (-\mu_0 + t_i)}{\sqrt{\sigma_0^2 + t_i^2 \cdot \sigma_v^2}} \right) \right],$$

where $\Phi$ is the cumulative distribution function of the standard normal distribution.
while, for subsequent cohorts, $0 < g \leq r$,

$$p(b_{ji} = 1|\beta, \Theta, g, r, t_i) = \sum_{c=0}^{C} \left[ \Phi \left( \frac{\lambda(c+1) - (-\mu_0 + t_i - r\cdot\lambda - g\cdot\delta)}{\sqrt{\sigma_0^2 + t_i^2 \cdot \sigma_0^2}} \right) - \Phi \left( \frac{\lambda(c+\beta) - (-\mu_0 + t_i - r\cdot\lambda - g\cdot\delta)}{\sqrt{\sigma_0^2 + t_i^2 \cdot \sigma_0^2}} \right) \right] \cdot \left[ 1 - \Phi \left( \frac{-\delta - (-\mu_0 + t_i - r\cdot\lambda - g\cdot\delta)}{\sqrt{\sigma_0^2 + t_i^2 \cdot \sigma_0^2}} \right) \right].$$

We model bud presence as a Bernoulli random variable with success probability $p(b_{ji} = 1|\beta, \Theta, t_i)$ and assume that samples drawn at the various time periods are independent conditional on the CLOCCS model.

### 3.3.2 Sampling model for DNA-content data

DNA-content data measured by flow cytometry provides an ordinal measurement of the DNA content of each cell in a sample: each cell appears in one of 1024 ordered channels on basis of its fluorescence, which is proportional to its DNA content (Figure 2.1c) [55]. In practice, channel number is often $\log_2$ transformed and treated as a continuous measurement.

Adapting the CLOCCS model to DNA-content data requires that we annotate the lifeline with the positions, measured as fraction of cell cycle length, at which S phase begins and ends. We denote these locations $\gamma_1$ and $\gamma_2$, respectively. As the population loses synchrony, the distribution of cells over channels will typically be bimodal, with one mode corresponding to cells in G1 (centered at $\alpha_1$), and the another corresponding to G2/M (centered at $\alpha_1 + \alpha_2$). Cells transiting S phase will fall between these points in expectation.

Further, we assume that DNA content increases linearly over the course of S phase. In particular, the expected DNA content of a cell is

$$\begin{align*}
\sum_{c=0}^{C} \begin{cases} 
\alpha_{1t} & c\lambda \leq P_t < (c + \gamma_1)\lambda & \text{G1} \\
\omega_{1t} P_t + \omega_{0t} (c) & (c + \gamma_1)\lambda \leq P_t < (c + \gamma_2)\lambda & \text{S-Phase} \\
\alpha_{2t} + \alpha_{1t} & (c + \gamma_2)\lambda \leq P_t < (c + 1)\lambda & \text{G2/M},
\end{cases}
\end{align*}$$

(3.3)
where \( \omega_{1t} = \frac{\alpha_2}{\lambda(\gamma_2 - \gamma_1)} \) and \( \omega_{0t}(c) = \frac{\alpha_1(\gamma_2 - \gamma_1) - \alpha_2(\gamma_1 + c)}{\gamma_2 - \gamma_1} \). The black line in Figure 3.3 is a plot of this curve.

**Figure 3.3:** Plot of expected flow cytometry channel for a cell given its lifeline position in units of \( \lambda \) (black curve, left vertical axis). An indicator function for the cell’s budding status is also plotted (gray dashed curve, right vertical axis).

Measurement of DNA content by flow cytometry is imprecise. Machine noise, variation in the cell’s orientation incident to the laser beam and variation in the performance of the fluorescent stain each contribute to measurement error [55]. Hence, a flow cytometry measurement made on a sample of cells drawn at a particular time point will be a sample from the convolution of a noise distribution and the CLOCCS position distribution. In particular,

\[
p(f_{ji} | \Psi, \Theta, t) = \int_{-\infty}^{\infty} p(f_{ji} | P_t, \Psi, \Theta, t)p(P_t | \Theta, t) dP_t
\]

where \( f_{ji} \) denotes the log fluorescence intensity of cell \( j \) at time \( t_i \) and where \( \Psi \) denotes the vector of parameters in the model for \( f_{ji} \) not in \( \Theta \). From above it follows that \( p(f_{ji} | P_t, \Psi, \Theta, t) \) can be modeled as a normal with mean given in Equation 3.3 and variance \( \tau_i^2 \). The log normal distribution is a common choice in this setting [85]. Additionally, the noise characteristics of the flow cytometer typically vary from one sample to the next, causing the locations of the G1 (\( \alpha_{1t} \)) and G2/M (\( \alpha_{1t} + \alpha_{2t} \)) modes, as well as the level of machine noise (\( \tau \)) to vary. Hence we allow the parameters of the DNA-content sampling distribution, \( p(f_{ji} | P_t, \Psi, \Theta, t) \), to vary across time periods. Note that Equation 3.4 can be
written as
\[ p(f_{ji}|\Psi, \Theta, t) = \sum_C I_{gr}(f_{ji})p(g, r|\Theta, t), \]

where
\[ I_{gr}(f_{ji}) = \int_{-\infty}^{\infty} p(f_{ji}|P_t, \Psi, \Theta, t)p(P_t|\Theta, r, g, t)dP_t \]
is a convolution of two normals, one of which is truncated.

Let \( l_{gr} \) denote the left limit to the support of cohort \( \{g, r\} \)'s position distribution, where \( l_{gr} = -\infty \) if \( g = r = 0 \) and \( l_{gr} = -\delta \) otherwise. Further, let \( G_{grt}(x) \) denote the normal cumulative distribution function with mean \(-\mu_0 + t - r \cdot \lambda - g \cdot \delta\) and variance \(\sigma_0^2 + t^2 \cdot \sigma_v^2\) evaluated at \(x\) and let \( S_{cgrt}(x) \) denote the normal cumulative distribution function with mean
\[
\frac{(\sigma_0^2 + t^2 \cdot \sigma_v^2)}{(\omega_{1t})^2} \left( \frac{f_{ji}}{\omega_{1t}} - \frac{\omega_{1t}(c)}{\omega_{1t}} \right) + \left( \frac{\tau^2}{\omega_{1t}^2} \right) (-\mu_0 + t - r \cdot \lambda - g \cdot \delta)
\]
and variance
\[
\frac{(\sigma_0^2 + t^2 \cdot \sigma_v^2) \frac{\tau^2}{\omega_{1t}^2}}{\sigma_0^2 + t^2 \cdot \sigma_v^2 + \frac{\tau^2}{\omega_{1t}^2}}
\]
evaluated at \(x\). It can be shown that \( I_{gr}(f_{ji}) = I^*_{gr}(f_{ji})/(1 - G_{grt}(l_{gr})) \) where
$$I_{gr}^c(f_{ji}) = \frac{1}{\tau} \phi \left( \frac{f_{ji} - \alpha_{1t}}{\tau} \right) \left[ G_{grt}(\gamma_1 \lambda) - G_{grt}((\gamma_1 + \gamma_2) \lambda) \right] + \sum_{c=0}^{C} (G_{grt}((c + \gamma_1) \lambda) - G_{grt}((c + \gamma_2) \lambda))$$

$$+ \frac{1}{\tau} \phi \left( \frac{f_{ji} - \alpha_{1t} - \alpha_{2t}}{\tau} \right) \sum_{c=0}^{C} (G_{grt}((c + 1) \lambda) - G_{grt}((c + \gamma_2) \lambda))$$

$$+ \sum_{c=0}^{C} \phi \left( \frac{f_{ji} - \alpha_{1t}}{\tau} + \frac{-\omega_1(c)}{\omega_1(t_{1t})} \frac{-(-\mu_0 + t - r - \lambda - g \cdot \delta)}{\sqrt{\sigma_0^2 + t^2 \cdot \sigma_1^2 + \frac{\sigma_2^2}{\tau}}} \right) \left( S_{cgrt}((c + \gamma_2) \lambda) - S_{cgrt}((c + \gamma_1) \lambda) \right)$$

and where \( \phi() \) is the standard normal density function. In Equation 3.5 above, the first line of the right-hand side corresponds to cells in G1, the second to cells in G2 or M, and the third to cells in S.

We assume that cell-level DNA-content measurements are conditionally independent within and between samples drawn at the various time points conditional on the CLOCCS model, \( \Psi \) and the sampling times. DNA-content and budding index measurements are made on separate samples drawn from a population’s culture, sometimes at the same points in time, sometimes not. Because they are distinct samples, we model the DNA-content and budding index data as conditionally independent given the CLOCCS parameters \( \Theta \), the budding parameter \( \beta \), the DNA-content parameters \( \Psi \) and sampling times.

### 3.4 CLOCCS model results

In what follows, we utilize the model to analyze budding index data from a collection of \textit{S. cerevisiae} synchrony/time–series experiments. We then use the model to fit DNA-content data gathered concurrently from one of those experiments. In addition, using only the budding index data, we estimate Bayes factors for the full CLOCCS model to submodels obtained by systematically removing each novel source of asynchrony, \( \delta \), \( \mu_0 \),
and \( \sigma^2_0 \) separately and in combination. Details of the strain and growth conditions used can be found in [36].

We use a random walk Metropolis [86, 87] algorithm for each model fit. In each case, the algorithm was tuned to mix well and the chain was given a lengthy burn-in period. Subsequent to this, we ran the chain for >250,000 iterations. Plots of sampled values appear stationary, and the Raftery and Lewis diagnostic [88], implemented in the R package CODA, indicates that the sample is sufficient to estimate the 0.025th quantile of any marginal posterior to within 0.01 with probability 0.95. All coefficients and associated interval estimates are based on summary statistics of marginal sample distributions. We tested our implementation of the model and the Markov chain Monte Carlo sampler by analyzing simulated data sets. Parameter estimates derived from these analyzes were consistent with their true values. What follows is a description of, and justification for, the prior choices used in our model fitting.

3.4.1 Prior distribution

Lord and Wheals [48] estimate \( S. \textit{cerevisiae} \) cell cycle length in culture at 30°C to be 78.2 min with a standard deviation of 9.1 min. To allow for differences in experimental protocol, we place a normal, mean 78.2, standard deviation 18.2 prior on cell cycle length, \( \lambda \). In \( S. \textit{cerevisiae} \), duration of S phase, \((\gamma_2 - \gamma_1)\lambda\), is about one quarter of the cell cycle; it begins a short time before buds can be visually detected and continues until mother and daughter cells separate [89]. Hence we expect \( \gamma_1 < \beta < \gamma_2 \). Based on an analysis of 30 DNA-content measurements made on an asynchronous population conducted using the same protocol as used in the synchrony experiment described in the next section, we estimate that \( \gamma_1 \) is approximately 0.1 and that \( \beta \) is approximately 0.12. With this in mind, we let \( \gamma_1 \sim \text{Beta}(2,18) \), \( \beta \sim \text{Beta}(2.4, 17.6) \) and \( \gamma_2 \sim \text{Beta}(7,13) \), constrained as above. Bar-Joseph et al. [81] estimates the standard deviation of the velocity distribution in \( S. \textit{cerevisiae} \) to be 0.09 and observed a range of values 0.07 to 0.11 across 3 experiments.
For this reason, we place an independent inverse-gamma(12, 1) prior distribution on $\sigma_v$. Aspects of experimental protocol, most notably the method used to synchronize the population, have a strong influence on the parameters of the starting position distribution and on duration of the daughter-specific offset, $\delta$. As described in Section 2.1.2 centrifugal elutriation selects for small unbudded cells while other methods, such as $\alpha$-factor arrest, do not. Because of their size, elutriated cells tend to spend more time in Gr and their daughters spend more time in Gd than their counterparts in $\alpha$-factor experiments [47]. We have chosen to specify our prior distributions on these parameters to accommodate—not condition on—this source of protocol dependent uncertainty. In particular we place an inverse-gamma distribution with shape parameter 2 and mean 78.2/3 on $\sigma_0$ and the minimally informative exponential, mean 78.2 prior distribution on $\mu_0$. The former reflects our belief that almost all cells will be in Gr at release; the latter places highest prior likelihood on a short Gr, as is expected in an $\alpha$-factor experiment, but allows for the longer Gr that is expected in elutriation experiments. Similar reasoning was behind our choice of an exponential mean 55 prior distribution on $\delta$: in $\alpha$-factor experiments, $\delta$ can be very brief, while in elutriation experiments it can exceed 40% of the length of a typical cell cycle [47, 48].

In the DNA-content distributions, flow cytometer fluorescence noise, as measured by $\tau_i$, and location of the G1 and G2/M modes, as measured by $\alpha_{1i}$ and $\alpha_{2i}$, respectively, vary randomly from assay to assay over time. We model this variability hierarchically: first placing independent normal prior distributions on $\log(\tau_i)$, $\alpha_{1i}$, and $\alpha_{2i}$, $i = 1, \ldots, T$, followed by independent conjugate normal-inverse-chi-square hyperprior distributions on the parameters of the normal distributions. The latter are parametrized as in Gelman et al. [90]. In particular,

$$\log(\tau_i) \sim \text{i.i.d. } N(\mu_\tau, \sigma_\tau^2), \quad \alpha_{1i} \sim \text{i.i.d. } N(\mu_{\alpha_1}, \sigma_{\alpha_1}^2), \quad \alpha_{2i} \sim \text{i.i.d. } N(\mu_{\alpha_2}, \sigma_{\alpha_2}^2),$$

$$\mu_\tau | \sigma_\tau^2 \sim N(\eta_{\tau}, \sigma_\tau^2/\kappa_{\tau}), \quad \sigma_{\alpha_1}^2 | \alpha_{\alpha_1} \sim \text{Inv-}\chi^2(\nu_{\alpha_1}, \gamma_{\alpha_1}^2), \quad \sigma_{\alpha_2}^2 | \alpha_{\alpha_2} \sim \text{Inv-}\chi^2(\nu_{\alpha_2}, \gamma_{\alpha_2}^2),$$

where $\text{Inv-}\chi^2(\nu, \gamma^2)$ denotes the scaled inverse $\chi^2$ distribution with $\nu$ degrees of freedom.
and scale parameter $\gamma$. Given this specification, we define

$$
\Psi = (\tau_1, \ldots, \tau_T, \alpha_{11}, \ldots, \alpha_{1T}, \alpha_{21}, \ldots, \alpha_{2T}, \mu_\tau, \sigma_\tau^2, \mu_{\alpha_1}, \sigma_{\alpha_1}^2, \mu_{\alpha_2}, \sigma_{\alpha_2}^2).
$$

We chose the hyperparameters of the above hierarchical model on basis of an exploratory analysis of the same asynchronous DNA-content data used above. We set $\eta_{\alpha_1} = 7.58$, $\eta_{\alpha_2} = 0.82$ and $\eta_\tau = -1.91$, the average of the observed estimates of $\alpha_1$, $\alpha_2$ and $\tau$, respectively. We set each of the prior sample size parameters, $\kappa_\tau$, $\kappa_{\alpha_1}$ and $\kappa_{\alpha_2}$, and each of the prior degrees of freedom parameters, $\nu_\tau$, $\nu_{\alpha_1}$ and $\nu_{\alpha_2}$, equal to 2 to keep these margins of the prior distribution relatively diffuse. Finally, we set $\gamma_\tau^2 = 0.13$, $\gamma_{\alpha_1}^2 = 0.065$ and $\gamma_{\alpha_2}^2 = 0.0089$—in each case 16 times the observed variance in the asynchronous experiment.

### 3.4.2 Estimates from budding index data

Parameters were estimated for multiple independent synchrony experiments using both elutriation and $\alpha$-factor synchrony protocols (Figure 3.4). Budding index was measured by microscopically assessing at least 200 cells for the presence of a bud. Budding index curves for each of the experiments are shown, and overlaid on those curves (green) are the predictions from models parameterized by one hundred realizations of the Markov chain. These reflect the degree of posterior uncertainty in the budding curve and can be interpreted as forming a confidence band for the curve based on the uncertainty in the parameters.

Estimates of posterior means of the parameters as well as the $2.5^{th}$ and $97.5^{th}$ percentiles for each parameter in multiple experiments are detailed in Table 3.1. The best-fit values for each of the parameters are remarkably similar for experiments performed under the same conditions. Notably, little change is observed in the variance in velocity of cells ($\sigma_0^2$), even across experiments performed under significantly different conditions. The small values observed for $\sigma_0^2$ suggest that once cells enter the cell cycle, they progress
at very similar rates, and that much of the observed loss of synchrony cannot explained by this factor alone.

![Figure 3.4: Model fits of multiple experiments. Observed budding index curves (black lines) for cells synchronized by elutriation (a and b) or α-factor (c and d) and then incubated at 30°C. Experiments synchronized by α-factor and then incubated at 21°C are shown (e and f). One hundred random realizations from the Markov chain used to fit each experiment were used as parameterizations for the model, and the resulting predicted budding curves are shown (green lines). The width of the band created by the Markov chains is indicative of the effect of uncertainty in the model parameters. The model parameterizations from the Markov chains for each of the experiments are shown in Table 3.1.]

As expected, significant differences are observed for some parameters when experimental conditions are changed. The length of the cell cycle \( \lambda \) is extended as expected when α-factor synchrony experiments are run at lower temperatures (Table 3.1 30°C vs. 21°C). We also observed that the values for \( \mu_0 \) are significantly larger in elutriated popula-
tions than in α-factor synchronized cells. Unlike cells synchronized by α-factor, elutriated populations were exposed to an osmotic shock resulting from their release into growth medium containing 1M sorbitol (see Materials and Methods in [36]), and pathways that transiently delay cell cycle progression in response to osmotic stresses are expected to increase μ₀ [91]. The increase in estimated values of δ in elutriated populations suggests that the daughter-specific delay in G1 (Gd) is significantly longer than in populations synchronized with α-factor as expected due to differences in average cell size.

<table>
<thead>
<tr>
<th></th>
<th>Elutriation</th>
<th>α-Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>30°C</td>
<td>30°C</td>
</tr>
<tr>
<td>μ₀</td>
<td>95.8 (92.7–98.6)</td>
<td>20.3 (14.2–24.5)</td>
</tr>
<tr>
<td>λ</td>
<td>77.1 (72.8–81.7)</td>
<td>62.9 (58.3–69.6)</td>
</tr>
<tr>
<td>δ</td>
<td>41.4 (36.2–46.1)</td>
<td>13.6 (15.5–21.3)</td>
</tr>
<tr>
<td>σ₀</td>
<td>15.3 (14.0–16.6)</td>
<td>12.1 (11.0–13.2)</td>
</tr>
<tr>
<td>σᵥ</td>
<td>0.06 (0.04–0.08)</td>
<td>0.08 (0.05–0.11)</td>
</tr>
<tr>
<td>β</td>
<td>0.14 (0.12–0.17)</td>
<td>0.28 (0.23–0.34)</td>
</tr>
</tbody>
</table>

**Table 3.1:** Parameter estimates for multiple experiments. Each column corresponds to a different experiment, synchronized by either centrifugal elutriation or α-factor arrest, and grown at either 30°C or 21°C. The observed budding curves, as well as the model predictions for each experiment, are shown in Figure 3.4. Each cell contains the mean value for that row’s parameter, given 250,000 iterations of the Markov chain used to fit the column’s experiment. Below each mean is the 95% confidence range for that parameter in that experiment.

### 3.4.3 Estimates from DNA-content data

We also used the CLOCCS model to fit DNA-content data gathered from the same synchrony/time–series experiment used to gather the budding index data fit in Table 3.1a. The relative DNA content of 10000 cells at each time point was measured by flow cytometry as described previously [52]. The observed fluorescence values for each measured cell in each sample were log₂ transformed prior to analysis. Here, we compare parameter
estimates given both the budding index and DNA-content data, given the DNA-content data alone and given the budding index data alone.

<table>
<thead>
<tr>
<th>Prior DNA-content &amp; Budding</th>
<th>DNA-content Only</th>
<th>Budding Only</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_0$ 78.2 (22.5,298.5)</td>
<td>-94.4 (−94.8,−94.0)</td>
<td>-94.3 (−94.7,−93.9)</td>
</tr>
<tr>
<td>$\delta$ 55.0 (15.8,202.9)</td>
<td>44.3 (43.8,44.8)</td>
<td>43.9 (43.9,44.8)</td>
</tr>
<tr>
<td>$\sigma_0$ 26.1</td>
<td>18.0 (17.8,18.2)</td>
<td>18.0 (17.8,18.2)</td>
</tr>
<tr>
<td>$\sigma_v$ (0.05,0.16)</td>
<td>0.03 (0.02,0.03)</td>
<td>0.03 (0.02,0.03)</td>
</tr>
<tr>
<td>$\lambda$ 78.2 (65.9,113.9)</td>
<td>79.5 (79.0,80.0)</td>
<td>79.7 (79.1,80.2)</td>
</tr>
<tr>
<td>$\beta$ 0.14 (0.05,0.29)</td>
<td>0.15 (0.14,0.17)</td>
<td>0.05 (0.11,0.17)</td>
</tr>
<tr>
<td>$\gamma_1$ (0.01,0.17)</td>
<td>0.05 (0.05,0.05)</td>
<td>0.05 (0.05,0.05)</td>
</tr>
<tr>
<td>$\gamma_2$ 0.36 (0.18,0.57)</td>
<td>0.35 (0.35,0.35)</td>
<td>0.35 (0.35,0.35)</td>
</tr>
<tr>
<td>$\mu_{a1}$ 7.6 (6.5,8.7)</td>
<td>8.2 (8.2,8.3)</td>
<td>8.2 (8.2,8.3)</td>
</tr>
<tr>
<td>$\sigma_{\alpha1}$ 0.07 (0.01,0.27)</td>
<td>0.04 (0.02,0.06)</td>
<td>0.04 (0.02,0.06)</td>
</tr>
<tr>
<td>$\mu_{a2}$ 0.8 (0.4,1.2)</td>
<td>1.0 (0.8,1.3)</td>
<td>1.0 (0.8,1.2)</td>
</tr>
<tr>
<td>$\sigma_{\alpha2}$ 0.01 (0.00,0.04)</td>
<td>0.33 (0.17,0.60)</td>
<td>0.33 (0.17,0.58)</td>
</tr>
<tr>
<td>$\mu_\tau$ -1.9 (-3.5,−0.4)</td>
<td>-2.1 (−2.2,−2.0)</td>
<td>-2.1 (−2.2,−2.0)</td>
</tr>
<tr>
<td>$\sigma_\tau$ 0.13 (0.02,0.54)</td>
<td>0.03 (0.02,0.05)</td>
<td>0.03 (0.02,0.05)</td>
</tr>
</tbody>
</table>

**Table 3.2:** Prior (column 1) and marginal posterior summaries given both the DNA-content and budding index data (column 2), given the DNA-content data only (column 3) and given the budding index data only (column 4).

Table 3.2 provides marginal summaries of the prior (column 1, see Section 3.4.1) and of the posterior distributions after fitting the model to both the DNA-content and budding index data (column 2), to the DNA-content data only (column 3) and to the budding index data only (column 4). Note that point and interval estimates of common parameters derived using both the budding index and DNA-content data are very close to their counterparts fit only to the DNA-content data. This is not surprising given the information rich nature of the DNA-content data: at each time period approximately 10000 cells are assayed for DNA content while only approximately 200 are assayed for presence of a bud. On average, point estimates of the common parameters differ by less than 1% and the associated posterior
interval estimates are only about 2% narrower when the budding index data is added. The parameter $\beta$ can only be estimated with budding index data, but it is estimated more accurately when DNA-content data is included owing to the fact that it is constrained by $\gamma_1$ and $\gamma_2$.

Figure 3.5a is a plot of the observed budding index curve (black) overlayed with 95% pointwise interval estimates from the analysis of only the budding index data (green) and of both the budding index and DNA-content data (red). The latter analysis estimates the recovery period (Gr) to be slightly shorter and more variable and estimates cell cycle length to be longer and less variable than estimated with the budding index data alone. This is evident in the red confidence bands positioned to the left of the green between 70 and 100 min and to the right of the green between 190 and 225 min experimental time.

Figure 3.5b–f plot observed DNA-content densities (gray) and their posterior mean estimates (red) at five experimental time points selected to highlight the population’s transition from G1 (B), through S phase to G2/M (C and D) and the effect of its growing asynchrony (E and F). The corresponding time points are identified by labels on the budding index curve Figure 3.5a. The observed DNA-content densities are discrete and unsmoothed. They are calculated by normalizing the raw DNA-content channel counts and transforming them, via the change of variables formula, to the $\log_2$ scale. The estimates are extremely good: in all cases, the G1 and G2/M modes are accurately scaled and located and capture the shape of the distributions between the modes, suggesting that the model is accurately accounting for the cells transiting S phase.

### 3.4.4 Importance of each source of asynchrony

In order to accurately capture the synchrony loss in populations the CLOCCS model introduced three novel sources of asynchrony, $\mu_0$, $\sigma_0^2$, and $\delta$ in conjunction with the previously recognized source, $\sigma_v^2$. To verify that the three novel sources of asynchrony are a significant improvement to the model, we estimate Bayes factors (BFs) [92] for a series of pairs
Figure 3.5: a) Plot of observed budding index curve (black) and 95% pointwise interval estimates from budding index only analysis (green) and budding index/DNA-content analysis (red). b–f: DNA-content densities (gray) and their posterior mean estimates (red) at five points in time, highlighting the population’s transition from G1 (b), through S phase to G2/M (c and d) and the effect of its growing asynchrony (e and f). The corresponding time points are labeled above the budding index curve.

of models nested under the fully parametrized CLOCCS model using importance sampling. These quantities allow us to measure the weight of evidence in the budding index data in favor of alternate parametrizations of the model, including variants that drop the any or all of the novel asynchrony sources. The hierarchy of models we examine is not complete but accounts for all reasonable alternatives to the full model. The simplest model, where we set $\mu_0 = 0$, $\sigma_0^2 = 0$ and $\delta = 0$, corresponds to a branching process version of the Bar-Joseph et al. [81] model. We employed a separate sampler to estimate each marginal likelihood and used 100 degrees-of-freedom multivariate $t$ densities as the importance densities, each with mean and covariance matrix matching that estimated from a Markov chain Monte Carlo analysis of the associated model. For purposes of this calculation, we used only the budding index data to inform the model and drew 10000 importance samples for each calculation. The variance of the normalized weights was less than 1.45 in all cases. Hence the effective sample size [93] for estimating the marginal likelihood
was never smaller than 4,000.

<table>
<thead>
<tr>
<th>Submodel</th>
<th>Full Model</th>
<th>$\mu_0 = 0$</th>
<th>$\delta = 0$</th>
<th>$\sigma_0^2 = 0$</th>
<th>$\mu_0 = 0$</th>
<th>$\delta = 0$</th>
<th>$\sigma_0^2 = 0$</th>
<th>$\mu_0 = 0$</th>
<th>$\delta = 0$</th>
<th>$\sigma_0^2 = 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_0 = 0$</td>
<td>368.62</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\delta = 0$</td>
<td>18.75</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\mu_0 = \delta = 0$</td>
<td>364.22</td>
<td>-4.40</td>
<td>345.47</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\mu_0 = \sigma_0^2 = 0$</td>
<td>363.78</td>
<td>-4.84</td>
<td>332.78</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\delta = \sigma_0^2 = 0$</td>
<td>31.04</td>
<td>12.28</td>
<td>0.04</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\mu_0 = \delta = \sigma_0^2 = 0$</td>
<td>359.37</td>
<td>-9.25</td>
<td>340.62</td>
<td>328.37</td>
<td>-4.85</td>
<td>-4.41</td>
<td>328.33</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SD(RMSE)</td>
<td>0.16</td>
<td>0.10</td>
<td>0.19</td>
<td>0.15</td>
<td>0.10</td>
<td>0.09</td>
<td>0.14</td>
<td>0.09</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.3: Estimates of log Bayes factors (lBFs) for various nested model comparisons given the budding index data alone. The model indexed by an entry’s column is the larger of the models and is represented in the numerator of the lBFs in that column; the model indexed by an entry’s row is the smaller of the two. The last two rows of the table provide the average and standard deviation of the RMSE of the model’s fitted values to the observed budding index data over a sample of 1000 draws from the posterior.

Table 3.3 reports estimates of log Bayes factors (lBFs) for various nested model comparisons given the budding index data. In this table, the model indexed by an entry’s column is the larger of the models and is represented in the numerator of the lBFs in that column; the model indexed by an entry’s row is the smaller of the two. As a guide to interpreting these numbers, Kass and Raftery [92] classify lBFs between 0 and 1 as ‘not worth more than a bare mention,’ those from 1 to 3 ‘positive,’ those from 3 to 5 ‘strong’ and those greater than 5 ‘very strong.’ Using this scale as a guide, the full CLOCCS model is very strongly preferred to all alternatives, including the model of Bar-Joseph et al. [81]. The worst alternative sets only $\mu_0 = 0$. When $\mu_0$ is constrained to be zero, better fits to the data are achieved by setting one or the other, or preferably both, of $\delta$ and $\sigma_0^2$ to zero.

Figure 3.6 depicts posterior mean fits to the budding index data under each of the competing models. We estimated the posterior means using the same MCMC output that was used to determine the importance distributions. Each MCMC analysis followed the same procedure used for the primary analyzes. Note that the fits achieved by all model variants that set $\mu_0 = 0$ are visually indistinguishable and markedly inferior to any variant that allows $\mu_0 > 0$. The last two rows of Table 3.3 provides estimates of root mean squared
error (RMSE) of the fits to the budding index data achieved by each model’s posterior mean curve. These estimates reinforce what is evident from the marginal likelihood and graphical analyses; namely, that models that do not allow for a nonzero location in the distribution of initial cell position are markedly inferior to those that do and that accounting for a mother/daughter offset is particularly important, at least in the case where the cell population was synchronized using centrifugal elutriation.

**Figure 3.6:** Plot of observed budding index curve (black) and posterior mean fitted curves under each of the competing models for the budding index data. The full model is plotted in red; the competing models are obtained by constraining the parameter(s) indicated in the figure legend to be zero. Quantitative summaries of these fits can be found in Table 3.3.

### 3.5 Applications

In the sections above we have shown that the CLOCCS model is able to accurately model the synchrony loss in synchrony/time–series experiments performed under a variety of conditions and that the model can be fit using two commonly collected types of data, budding index and DNA-content data. While the model is a significant improvement over
existing models, in and of itself, the ability to accurately model population dynamics is only of theoretical interest. However, we can use the model to solve the practical problems of alignment of multiple datasets and for deconvolution within a dataset.

### 3.5.1 CLOCCS-based alignment between datasets

Qualitatively, measurements of dynamic processes during the cell cycle are generally reproducible. However, even when experimental conditions are tightly controlled, populations in synchrony experiments rarely progress into and through the cell cycle with identical kinetics. Several parameters often vary from experiment to experiment including recovery time ($\mu_0$) and cell cycle length ($\lambda$), and daughter-specific delay ($\delta$). Even small variations these times can have profound effects when attempting to compare measured values from time point to time point across different experiments.

Rather than directly comparing values as a function of clock time, we can exploit CLOCCS’s ability to predict cell cycle distributions for different experiments, and use these distributions to align data based on population distribution state rather than clock time. As a proof of principle experiment, we use the model to predict lifeline distributions for two $\alpha$-factor synchrony experiments performed at different temperatures (Figure 3.7). It is immediately apparent that the cell cycle distributions at the same clock times (Figure 3.7a) are very different between the two experiments, and that aligning the budding data from these two experiments based on clock time (Figure 3.7b) does not allow for accurate biological inference.

To align these data we map values from one experiment into the other, using the mean lifeline position of the $\{0, 0\}$ cohort, normalized by cell cycle length. Where a value of 1.5 would describe a population where the mean of the $\{0, 0\}$ cohort is halfway through its second cell cycle, whereas a value of 0.75 would describe a population only 75% through its first cell cycle. The cell-cycle distributions of two populations are comparable at time points where the location of the $\{0, 0\}$ cohort, and therefore the values for the descriptor,
are the same. The ability to accurately align experiments based on the position of the \{0, 0\} cohort diminishes as the estimates for \(\delta\) begin to vary significantly between experiments; as \(\delta\) diverges, the correlation between \{0, 0\} position and overall population distribution between two experiments becomes worse, especially at later time points. Nevertheless, we constructed curves for \(\alpha\)-factor synchrony experiments performed at 30°C and 21°C, defining the \{0, 0\} position as a function of clock time. These curves illustrate how two different experiments can have the same cell-cycle distributions at very different clock times (Figure 3.7c, dashed lines). Using these curves, we realigned the budding data (Figure 3.7c) so that relevant comparisons could be made based on lifeline position, rather than clock time.

**Figure 3.7:** Model-based alignment of time course data from synchrony experiments. a) The lifeline distributions at three time points for cells synchronized with \(\alpha\) factor incubated at either 30°C (left side) or 21°C. b–c) Alignment of budding curves for the two experiments (red line=30°C, blue line=21°C), based on clock time. Time points at which the two populations exhibit similar distributions are highlighted (green dots; 90 min for 30°C, 180 min for 21°C). d) Alignment of budding curves to the lifeline using the position of the mean of the \{0, 0\} cohort for mapping.
Aligning time-series transcription data from *S. cerevisiae* & *S. pombe*

We have described the CLOCCS model in the context of learning cell cycle distributions based on the budding index or DNA-content data from *S. cerevisiae*. However, the flexibility of the model allows it to be fit using any cell-cycle event that can be measured on single cells, and where we know when and for how long in the cell cycle this event occurs.

We fit the CLOCCS model to time series data collected from the unrelated yeast species, *Schizosaccharomyces pombe*. Despite a high degree of conservation in the cell-cycle machinery, cell division in *S. pombe* is fundamentally different from *S. cerevisiae*. First, *S. pombe* does not bud, but reproduces by binary fission. Thus, divisions are not asymmetric. The cell cycle length is somewhat longer than *S. cerevisiae* under standard growth conditions, and unlike *S. cerevisiae*, *S. pombe* cells spend the majority of their time in the G2 phase of the cell cycle. Nonetheless, cell-cycle position in *S. pombe* can be monitored by measuring septation index. *S. pombe* cells begin to form a septum, which is a precursor to the cytokinetic ring, at the end of mitosis. Septation persists through the short G1, and is lost part way into S-phase when cells separate. We fit CLOCCS model parameters to septation data from the Elutriation 2 time series experiment of Rustici et al. [6].

The resulting model parameter estimates are shown in Table 3.4. The interpretation of the parameters for *S. pombe* is slightly different than for *S. cerevisiae*. The negative value for $\mu_0$ recapitulates the fact that elutriated *S. pombe* cells are in G2 and are therefore already in their first cell-cycle and thus the recovery time (Gr) to the first G1 is, in a sense, negative. The meaning of $\sigma_0$, $\sigma_0^2$, and $\lambda$ are the same across organisms. The value for $\delta$ is constrained to be zero in *S. pombe* because division is symmetric and therefore the difference between mothers and daughters is zero. The interpretation of $\beta$ is also changed. Where $\beta$ in *S. cerevisiae* was the percentage of the cell cycle where cells are unbudded, in *S. pombe* it captures the percentage of the cell cycle where cells are unseptated.

We also fit CLOCCS parameters to budding data from a published time-series transcrip-
tion experiment in *S. cerevisiae* (Pramila et al. [5]). Unfortunately, budding index data was not gathered by Pramila and colleagues for the synchrony used to generate their microarray dataset. As an approximation, we used the average of wild-type budding curves shown in Figures 2 and 5 of Pramila et al. [5]. Model fits for budding in *S. cerevisiae* experiments and separation in *S. pombe* experiments are shown in Figure 3.8a and b (respectively) and parameter estimates are in Table 3.4.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_0$</td>
<td>24.8</td>
<td>-75.9</td>
</tr>
<tr>
<td></td>
<td>(16.9,29.1)</td>
<td>(−91.6,−70.4)</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>52.3</td>
<td>151.0</td>
</tr>
<tr>
<td></td>
<td>(45.5,63.9)</td>
<td>(149.2,163.1)</td>
</tr>
<tr>
<td>$\delta$</td>
<td>21.6</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>(7.1,28.9)</td>
<td>(0.0,0.0)</td>
</tr>
<tr>
<td>$\sigma_0$</td>
<td>10.7</td>
<td>22.8</td>
</tr>
<tr>
<td></td>
<td>(9.5,11.8)</td>
<td>(20.2,26.2)</td>
</tr>
<tr>
<td>$\sigma_v$</td>
<td>0.09</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>(0.05,0.16)</td>
<td>(0.04,0.09)</td>
</tr>
<tr>
<td>$\beta$</td>
<td>0.14</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>(0.07,0.24)</td>
<td>(0.86,0.89)</td>
</tr>
</tbody>
</table>

**TABLE 3.4:** Parameter estimates for *S. cerevisiae* and *S. pombe* experiments. The CLOCCS parameter estimates for the experiments aligned in Figure 3.7. Each cell contains the mean value for that row’s parameter, given 250,000 iterations of the Markov chain used to fit the column’s experiment. Below each mean is the 95% confidence range for that parameter in that experiment.

To align *S. pombe*, we used the knowledge that *S. pombe* cells begin septating late in mitosis, corresponding to a lifeline position $\beta \lambda$ on the *S. pombe* lifeline (as opposed to *S. cerevisiae* lifeline where position $\beta \lambda$ defines the G1/S border) and that *S. pombe* cells divide approximately halfway through S-phase, corresponding to a *S. pombe* lifeline position $\lambda$ (in *S. cerevisiae*, this position would be defined as $(\beta + \frac{0.25}{2}) \cdot \lambda$, where S-phase is 25% of the cell-cycle). Using these sets of corresponding cell-cycle phase/lifeline points budding and septation curves were then mapped onto the same cell cycle lifeline as described above, with blue curves representing budding in *S. cerevisiae*, and red curves representing septation in *S. pombe* (Figure 3.8c and d). Because budding and septation occur in different phases of the cell cycle, these curves are not expected to align.

Using the model based alignment, we were able to directly compare the transcription
curves for several cell-cycle-regulated orthologs. By visually comparing transcription data from *S. pombe* and *S. cerevisiae*, published reports suggested that several orthologs were transcribed in similar cell cycle phases [6, 94, 95]. Our results demonstrate a good correlation between the expression of orthologs in G1 (*RNR1/CDC22* and *CLB6/CIG2*), S-phase (*HHF1/HHF1* and *HHT1/HHT1*), and late mitosis (*SIC1/RUM1*) (Figure 3.8e). However, our alignments suggest that genes expressed in G2 (*CDC20/SLP1* and *CDC5/PLO1*) as well as some members of the late mitotic cluster (*DBF2/SID2*) are transcribed earlier in *S. cerevisiae* than in *S. pombe*. 
**Figure 3.8**: CLOCCS based alignment across organisms. Model estimates were generated based on the budding index curve for *S. cerevisiae* [5] (a, blue line) and the septation index curve for *S. pombe* [6] (b, red line). The light blue regions in the lifelines above the curves show the areas of the cell cycle where budding and septation occur, respectively. Using the mapping from time to \{0, 0\} position for both experiments (c) the two curves were aligned (d). Since budding and septation do not occur at the same points in the cell cycle we do not expect these two curves to overlap. Using the same mapping function (e), we align the mean-normalized log$_2$-transformed transcriptional profiles for a set of gene homologs with the *S. cerevisiae* gene name on the left and the expression profile shown in blue, and the *S. pombe* gene name on the right and the expression profile shown in red.
3.5.2 Deconvolution

The very fact that synchronized populations lose synchrony over time also prevents the accurate measurement of cell cycle processes that occur in discrete cell cycle phases. Because populations are often distributed across multiple cell cycle phases at each time point, the measured value of any parameter represents a convolution of both the true value for cells at a given cell cycle position, along with values associated with cells distributed in other cell cycle phases. By learning the distribution of cells at each time point with CLOCCS, the convolution kernel can be determined. It is then possible to derive estimates for any process at distinct cell cycle positions by employing a deconvolution algorithm. A detailed description of a CLOCCS based deconvolution algorithm as well as its application in deconvolving \textit{S. cerevisiae} mRNA expression data can be found in Section 7.2.

3.6 Summary

Synchrony/time–series experiments on populations of cells are essential for understanding the dynamic processes associated with the cell cycle. In this chapter, we have described the CLOCCS model, sampling models for fitting this model to both budding index and DNA-content data, and a detailed model evaluation. We have demonstrated that accurate model fits can be obtained using budding index, DNA-content data, or both. While previous models only account for one source of asynchrony, namely variation cell cycle length ($\sigma_v^2$) [77, 79, 81, 83], the CLOCCS model adds three novel sources of asynchrony. These are variation in ($\sigma_0^2$), and duration of recovery from initial synchrony ($\mu_0$), and the variation due to asymmetric cell division ($\delta$). In Section 3.4.4 we showed that the CLOCCS model is very strongly preferred to all nested alternatives including a branching process version of the model of Bar-Joseph et al. [81].

By using the model to map time-series data onto a common cell-cycle lifeline, different data types (e.g. mRNA levels, protein levels, protein localization) from multiple
synchrony/time-series experiments can be aligned and the dynamics of multiple events can be temporally compared. Furthermore, DNA-content measurements are commonly used to measure cell-cycle position in organisms from yeast to mammals. Thus, the model permits the alignment and comparison of dynamics of cell cycle events across species, potentially providing an accurate view of evolutionary changes in cell cycle progression and regulation. Additionally, the more accurate description of population dynamics achieved by the CLOCCS model will allow more accurate deconvolution of dynamic measurements such as transcript abundance.

The model’s parameter estimates are also interpretable in terms of biological quantities associated with the cell cycle, so their estimates are of independent interest. For example, the measure of initial synchrony, $\sigma_0$, can be used to tune synchrony protocols for optimal results. When using budding index data, $\lambda$ and $\beta$ allow researchers to map temporal events to pre- or post-G1 cell cycle phases. When DNA-content data is used, this resolution is increased and events can be placed accurately into the G1, S, or G2/M phases of the cell cycle.

The CLOCCS model is unique, to our knowledge, for providing a closed form expression for the likelihood function in a complex branching process. This expression is written by enumerating and then marginalizing over the distinct cohorts present in the population at a given time. The explicit accounting of cohorts allows for extensions of the model that introduce cohort dependent effects such as one–time events and effects, such as the mother–daughter offset. The approach we describe is very general and has the potential to provide a flexible and efficient alternative in a range of problems where population balance or branching process models are used to describe the short term dynamics of a branching population. It is adaptable to new experimental measurements, and given its ability to use DNA-content data, is already applicable to virtually all biological systems where synchronized populations are studied, most notably human cell-culture systems. Further integration of the model with deconvolution and alignment algorithms will pro-
vide researchers with a powerful new tool to aid in the study of dynamic processes during the cell division cycle.
Transcriptional regulation of the cell cycle

4.1 Introduction

The biochemical oscillator controlling periodic events during the cell cycle is centered on the activity of CDKs (reviewed in [62]). The cyclin–CDK oscillator governs the major events of the cell cycle, and in embryonic systems this oscillator functions in the absence of transcription, relying only on maternal stockpiles of messenger RNAs and proteins. CDKs are also thought to act as the central oscillator in somatic cells and yeast, and directed studies suggest that they are important for controlling the temporally ordered program of transcription (reviewed in [96, 97]). However, systems-level analyses using high-throughput technologies [5, 12, 98, 99] have suggested alternative models for the regulation of periodic transcription during the yeast cell cycle [5, 33, 34]. By correlating genome-wide transcription data with global transcription factor binding data, models have been constructed in which periodic transcription is an emergent property of a transcriptional regulatory network [5, 33, 34]. In these networks, transcription factors expressed in one cell-cycle phase bind to the promoters of genes encoding transcription factors that function in a subsequent phase. Thus, the temporal program of transcription could be
controlled by sequential waves of transcription factor expression, even in the absence of extrinsic control by cyclin–CDK complexes.

4.2 Global control of cell-cycle transcription by coupled CDK and network oscillators

The validity and relevance of the hypotheses regarding intrinsically oscillatory networks of transcription factors remain uncertain, because for the limited number of periodic genes that have been dissected in detail, periodic transcription was found to be governed by CDKs (reviewed in [96]). We therefore sought to determine to what extent CDKs and transcriptional regulatory networks contribute to global regulation of the cell-cycle transcription program. To this end, we investigated the dynamics of genome-wide transcription in budding yeast cells disrupted for all S-phase and mitotic cyclins ($clb1,2,3,4,5,6$). These cyclin-mutant cells are unable to replicate DNA, separate SPBs, undergo isotropic bud growth, or complete nuclear division, indicating that they are devoid of functional Clb–CDK complexes [67, 100, 101]. So, by conventional cell-cycle measures, $clb1,2,3,4,5,6$ cells arrest at the G1/S border. It has been shown previously that $clb1,2,3,4,5,6$ cells trigger G1 events cyclically [67], including the activation of G1-specific transcription and bud emergence. Nevertheless, if Clb–CDK activities are essential for triggering the transcriptional program, then periodic expression of S-phase-specific and G2/M-specific genes should not be observed.

We examined global transcription dynamics in synchronized populations of both wild-type cells and cyclin-mutant cells. Synchronous populations of early G1 cells were collected by centrifugal elutriation. Cell aliquots were then harvested at 16 min intervals for 270 min (equivalent to about two cell cycles in the wild type and about 1.5 cell cycles in the cyclin mutant). Transcript levels were measured genome-wide for each time point with the use of Yeast 2.0© oligonucleotide arrays. Results from two independent experiments each for both wild-type and cyclin-mutant cells were highly reproducible, with adjusted
$r^2$ values of 0.995 and 0.989, respectively (Figure 4.1). All statistical analyses were performed with replicate data sets; however, to facilitate illustration, single data sets were used for all graphical representations.

**Figure 4.1:** Reproducibility of replicate experiments. Scatter plots of mean normalized absolute expression levels over all timepoints within an experiment for each mapped probe for both **a**, wild-type and **b**, cyclin-mutant cells. Each probe’s X and Y coordinate is defined by its mean normalized absolute expression/1000 in replicate 1 and 2, respectively. A linear model was fit to each plot (gray line) and adjusted $r^2$ values are given. To more clearly reveal the density of points near the origin, each point is colored according to the number of other points contained within a square of length 500 centered on that point.

To identify periodically transcribed genes, we applied a modification of a method developed previously [102] to data acquired from our wild-type cells (See Section 4.3.2 for details on periodic gene detection). We established a set of 1,271 genes that were transcribed periodically (Figure 4.2a and Supplemental Table S1 in [11]). This set of periodic genes shares 510 and 577 genes with those sets previously identified as periodic by Spellman et al. [12] and Pramila et al. [5], respectively (Figure 4.3), with 440 consensus periodic genes identified by all three studies. We then examined the transcriptional dynamics of our set of 1,271 periodic genes in the cyclin mutant (Figure 4.2b). The behavior of many genes changed significantly in the cyclin mutant, supporting previous findings. However, despite the fact that cyclin-mutant cells arrest at the G1/S border, a large fraction of periodic genes in all cell-cycle phases continued to be expressed on schedule (Figure 4.2b).
FIGURE 4.2: Reproducibility of replicate experiments. Scatter plots of mean normalized absolute expression levels over all timepoints within an experiment for each mapped probe for both a, wild-type and b, cyclin-mutant cells. Each probe’s X and Y coordinate is defined by its mean normalized absolute expression/1000 in replicate 1 and 2, respectively. A linear model was fit to each plot (gray line) and adjusted $r^2$ values are given. To more clearly reveal the density of points near the origin, each point is colored according to the number of other points contained within a square of length 500 centered on that point.

Using absolute change and Pearson correlation analyses (Section 4.2.2), we determined that 833 of the periodic genes showed changes in expression behavior in the cyclin mutant and are therefore likely to be directly or indirectly regulated by B-cyclin–CDK.

Our genome-level experiments accurately reproduced previous findings on several well-studied B-cyclin–CDK-regulated genes (Figure 4.4). We observed that a subset of late G1 transcripts (SBF-regulated genes such as $CLN2$ but not MBF-regulated genes such
Figure 4.3: Comparison of periodic genes identified by three studies. A Venn diagram of the overlap between the 1271 unique genes identified as periodic in wild-type cells by this study, the 991 from Pramila et al. [5], and the 800 from Spellman et al. [7]. The set of 991 periodic genes from Pramila et al. [5] contains all those with PBM5 rankings of 1000 or less, from the three datasets available at [8]. The set of 800 periodic genes from Spellman et al. [7] was obtained directly from [9].

...as RNRI) were not fully repressed (Figure 4.4a,b) as expected in mitotic cyclin-mutant cells [68, 103]. A subset of M/G1 transcripts (including SIC1 and NIS1) are targets of the transcription factors Swi5 and Ace2, which are normally excluded from the nucleus by CDK phosphorylation until late mitosis [104, 105, 106, 107]. SIC1 and NIS1 were expressed earlier in the cyclin mutant (Figure 4.4c,d), presumably because nuclear exclusion of Swi5 and Ace2 is lost in cyclin-mutant cells. The modest degree of shift in the timing of SIC1 and NIS1 transcription probably reflects the fact that SWI5 and ACE2 transcripts do not accumulate to maximal levels in cyclin-mutant cells as expected for Clb2-cluster genes (including CDC20) (Figure 4.4e,f) [68, 108, 109]. Although a significant fraction of periodic genes showed changes in the amplitude of expression (increased or decreased), a statistical analysis of the dynamic range of expression across all periodic genes revealed that most genes in cyclin-mutant cells show only modest changes, if any, in comparison with wild-type cells (Figure 4.5).
Figure 4.4: Transcription dynamics of established cyclin–CDK-regulated genes. Absolute transcript levels (normalized Affymetrix intensity units/1000) are shown for the genes CLN2 (a) and RNR1 (b), which are regulated by SBF and MBF, respectively; the Ace2/Swi5-regulated genes SIC1 (c) and NIS1 (d); and the Clb2-cluster genes CDC20 (e) and ACE2 (f). Solid lines, wild-type cells; dashed lines, cyclin-mutant cells.

To identify new subsets of co-regulated genes on the basis of transcriptional behaviors observed in both wild-type and cyclin-mutant cells, we employed the affinity propagation algorithm [110], first to cluster genes based on expression in wild-type cells, and then to subcluster genes on the basis of their behavior in cyclin-mutant cells (Figure 4.6). Section 4.2.3 details this dual-clustering approach. Of the 833 cyclin-regulated genes, 513 were assigned to 30 discrete clusters showing similar behaviors in wild-type cells (Figure 4.6), and these were then subclustered into 56 clusters on the basis of their transcription profiles in cyclin-mutant cells (Figure 4.6b). Using data from global tran-
Figure 4.5: Peak to trough ratio for all genes. a) The average wild-type peak to trough ratio (PTR) (x-axis) is plotted against the average cyclin-mutant PTR (y-axis) for each gene/probe identified as periodic by all three studies (green, see Figure 4.3) [5, 7], by this study only (red), or not identified as periodic by this study (black). b) The average PTR in wild-type (Column 1) or cyclin-mutant cells (Column 2), as well as the average wild-type to cyclin-mutant PTR ratio (Column 3). Colored dots to the right indicate which genes from (a) are included in each calculation. A histogram of the average PTR for each of the WTPER genes in c, wild-type and d, cyclin-mutant cells. e) Histogram of ratios of the PTRs of each WTPER gene in wild-type vs. cyclin-mutant cells.

scription factor localization studies [35], we identified subsets of transcription factors that may regulate these subclusters with the use of over-representation analyses (Figure 4.6 and Table 4.1). See Section 4.2.3 for details of over-represented TF detection. On the basis of their association with the promoters of genes in cyclin-regulated subclusters, these factors are likely to be directly or indirectly regulated by cyclins. Consistent with this hypothesis are previous demonstrations that several of these factors are CDK targets [68, 72, 103, 106, 107, 111, 112, 113, 114, 115]. These findings lay the groundwork...
for elucidating the full range of mechanisms by which cyclin–CDKs regulate transcription during the cell cycle.

**Figure 4.6**: Genes showing altered behaviors in cyclin-mutant cells. **a)** Clusters of genes with similar expression patterns in wild-type cells. **b)** Subclusters of genes with similarly altered expression patterns in cyclin-mutant cells. Each row in **a** and **b** represents data for the same gene. Transcript levels are depicted as in Figure 4.2. Up to five over-represented transcription factors for each cluster are shown (see Table 4.1 for complete lists).
<table>
<thead>
<tr>
<th>WT Cluster</th>
<th>MT Cluster</th>
<th>Over-represented TFs with Q-Value ≤ 0.005</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>-</td>
<td>RLM1</td>
</tr>
<tr>
<td>E</td>
<td>-</td>
<td>SWI6,MBP1,STB1,SWI4,TEC1</td>
</tr>
<tr>
<td>E</td>
<td>6</td>
<td>SWI6,MBP1,SWI4</td>
</tr>
<tr>
<td>E</td>
<td>8</td>
<td>STB1,TEC1,SWI4,SWI6</td>
</tr>
<tr>
<td>F</td>
<td>9</td>
<td>SWI4</td>
</tr>
<tr>
<td>F</td>
<td>11</td>
<td>MET31</td>
</tr>
<tr>
<td>I</td>
<td>18</td>
<td>SWI6,SWI16</td>
</tr>
<tr>
<td>K</td>
<td>-</td>
<td>FKH1,FKH2</td>
</tr>
<tr>
<td>K</td>
<td>22</td>
<td>FKH1</td>
</tr>
<tr>
<td>L</td>
<td>-</td>
<td>FKH1,SPT23,FKH2,YOX1</td>
</tr>
<tr>
<td>L</td>
<td>26</td>
<td>FKH1</td>
</tr>
<tr>
<td>M</td>
<td>-</td>
<td>FKH2,SMP1,NDD1,RLM1,SWI6,SWI4</td>
</tr>
<tr>
<td>M</td>
<td>28</td>
<td>FKH2,NDD1,SWI6,RLM1,SWI4</td>
</tr>
<tr>
<td>O</td>
<td>-</td>
<td>NDD1,FKH2,MC1,FKH1</td>
</tr>
<tr>
<td>O</td>
<td>30</td>
<td>NDD1,FKH2,MC1,FKH1</td>
</tr>
<tr>
<td>V</td>
<td>-</td>
<td>YOX1,SWI5</td>
</tr>
<tr>
<td>V</td>
<td>40</td>
<td>YOX1</td>
</tr>
<tr>
<td>V</td>
<td>42</td>
<td>SWI5</td>
</tr>
<tr>
<td>W</td>
<td>-</td>
<td>YOX1,MC1,HIR2,DIG1,ACE2,SWI5</td>
</tr>
<tr>
<td>W</td>
<td>44</td>
<td>YOX1,MC1,HIR2,DIG1,ACE2,SWI5</td>
</tr>
<tr>
<td>X</td>
<td>-</td>
<td>ARR1</td>
</tr>
<tr>
<td>X</td>
<td>45</td>
<td>ARR1</td>
</tr>
<tr>
<td>Y</td>
<td>-</td>
<td>SWI4</td>
</tr>
<tr>
<td>Z</td>
<td>-</td>
<td>SWI5,ACE2</td>
</tr>
<tr>
<td>Z</td>
<td>48</td>
<td>SWI5</td>
</tr>
<tr>
<td>AB</td>
<td>-</td>
<td>SWI5,RME1,RLR1,ACE2,SPT2,NDD1,MBP1,SWI4,UME6</td>
</tr>
<tr>
<td>AB</td>
<td>52</td>
<td>SWI5,RME1,RLR1,ACE2,SPT2,NDD1,MBP1,SWI4,UME6</td>
</tr>
</tbody>
</table>

Table 4.1: Over-represented transcription factors in Figure 4.6 by cluster/subcluster. Each row corresponds to a wild-type cluster (indexed by letters A-AD) and cyclin-mutant subcluster (indexed by numbers 1-56) from Figure 4.6 for which there are over-represented TFs. Each cell of the table—for the genes in the corresponding cluster or subcluster—contains the TFs with \( q \)-values ≤ 0.005, sorted from most to least significant.
Of the genes identified as periodic in wild-type cells, 882 continued to be expressed on schedule in cyclin-mutant cells despite cell-cycle ‘arrest’ at the G1/S border (Figure 4.7b). Some of these genes (450 in total) showed minor changes in transcript behavior but continued to be expressed at the proper time, as shown for ACE2 (Figure 4.4f). Therefore some genes that were cyclin-regulated are also included in the set of genes that maintain periodicity. Nevertheless, a statistical analysis of the dynamic range of expression of these genes in wild-type and cyclin-mutant cells indicates that the amplitude changes for most of these genes are quite modest (Figure 4.8 and Figure 4.9). The finding that nearly 70% of the genes identified as periodic in wild-type cells are still expressed on schedule in cyclin-mutant cells demonstrates the existence of a cyclin–CDK-independent mechanism that regulates temporal transcription dynamics during the cell cycle.

In principle, a transcriptional regulatory network defined by sequential waves of expression of transcription factors [5, 33, 34] might function independently of any extrinsic control by CDKs. To determine whether a transcriptional regulatory network could account for cyclin–CDK-independent periodic transcription, we constructed a synchronously updating boolean network model and determined that such a model can indeed explain the periodic expression patterns we observed in cyclin-mutant cells (Figure 4.7c). Transcription factors that maintained periodicity in the cyclin mutant were placed on a circularized cell-cycle timeline on the basis of their peak time of transcription in the cyclin mutant. Connections were drawn on the basis of documented physical interactions [10, 35] between a transcription factor and the promoter region of a gene encoding a transcription factor expressed subsequently. The architecture of the network in cyclin-mutant cells is virtually identical to that in wild-type cells (Figure 4.10) and is also remarkably similar to models based on wild-type expression data from previous studies [5, 33, 34]. See Section 4.2.4 for details on construction of transcription factor networks.

When the network is endowed with boolean logic functions (Table 4.2), synchronous updating of the model leads to a cycle that produces successive waves of transcription by
progressing through five distinct states before returning to the initial state (Figure 4.11a,b).

Thus, the model functions as an oscillator and produces a correctly sequenced temporal program of transcription. For details on the boolean modeling see Section 4.2.5.

<table>
<thead>
<tr>
<th>TF</th>
<th>Activation Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>MBF</td>
<td>CLN3</td>
</tr>
<tr>
<td>SBF</td>
<td>((\text{CLN3} \lor \text{MBF}) \land \lnot(\text{YOX1} \land \text{YHP1}))</td>
</tr>
<tr>
<td>YOX1</td>
<td>\text{MBF} \land \text{SBF}</td>
</tr>
<tr>
<td>HCM1</td>
<td>\text{MBF} \land \text{SBF}</td>
</tr>
<tr>
<td>YHP1</td>
<td>\text{MBF} \lor \text{SBF}</td>
</tr>
<tr>
<td>SFF</td>
<td>\text{SBF} \land \text{HCM1}</td>
</tr>
<tr>
<td>ACE2</td>
<td>\text{SFF}</td>
</tr>
<tr>
<td>SWI5</td>
<td>\text{SFF}</td>
</tr>
<tr>
<td>CLN3</td>
<td>((\text{SWI5} \land \text{ACE2}) \land \lnot(\text{YOX1} \land \text{YHP1}))</td>
</tr>
</tbody>
</table>

**Table 4.2:** Boolean logic used in model shown in Figure 4.7c. An initial choice of a set of Boolean logic functions on the variables which produce the five state oscillating attractor referred to as Cycle 1.

To examine the robustness of the network oscillator, we evaluated outcomes when initializing the network from all possible starting states. More than 80% of the 512 starting states entered the oscillatory cycle depicted in Figure 4.7c, with the remainder terminating in a steady state in which all genes were transcriptionally inactive (Table 4.3 and Table 4.4). We also examined whether the oscillations were sensitive to the choice of the boolean logic functions assigned to nodes with multiple inputs, specifically the activating inputs to Cln3 and SFF, and the repressors of SBF and Cln3. For most of the logic functions, the predominant outcome was again the oscillatory cycle depicted in Figure 4.7c, but in some cases the model entered two qualitatively similar cycles (Figure 4.11cd, Table 4.3, and Table 4.4), with the remainder again terminating in a transcriptionally inactive steady state. Several boolean logic functions were found to produce the same cycles (Table 4.3), so the model cannot precisely determine the true logic of the network connections. Nevertheless, the fact that the model can produce qualitatively similar cycles, and that these cycles can be reached from many initial states, suggests that robust oscillation is an emer-
gent property of the network architecture.

<table>
<thead>
<tr>
<th>Repressor</th>
<th>SFF</th>
<th>CLN3</th>
<th>All Off</th>
<th>Cycle 1</th>
<th>Cycle 2</th>
<th>Cycle 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>(YOX1 ∧ YHP1)</td>
<td>SBF ∧ HCM1</td>
<td>SWI5 ∧ ACE2</td>
<td>19.7%</td>
<td>80.3%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>(YOX1 ∧ YHP1)</td>
<td>SBF ∧ HCM1</td>
<td>SWI5 ∨ ACE2</td>
<td>13.9%</td>
<td>86.1%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>(YOX1 ∧ YHP1)</td>
<td>SBF ∨ HCM1</td>
<td>SWI5 ∧ ACE2</td>
<td>3.3%</td>
<td>0%</td>
<td>0%</td>
<td>96.7%</td>
</tr>
<tr>
<td>(YOX1 ∧ YHP1)</td>
<td>SBF ∨ HCM1</td>
<td>SWI5 ∨ ACE2</td>
<td>2.1%</td>
<td>0%</td>
<td>0%</td>
<td>97.9%</td>
</tr>
<tr>
<td>(YOX1 ∨ YHP1)</td>
<td>SBF ∧ HCM1</td>
<td>SWI5 ∧ ACE2</td>
<td>76.6%</td>
<td>23.4%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>(YOX1 ∨ YHP1)</td>
<td>SBF ∧ HCM1</td>
<td>SWI5 ∨ ACE2</td>
<td>79.7%</td>
<td>20.3%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>(YOX1 ∨ YHP1)</td>
<td>SBF ∨ HCM1</td>
<td>SWI5 ∧ ACE2</td>
<td>34.8%</td>
<td>0%</td>
<td>45.3%</td>
<td>19.9%</td>
</tr>
<tr>
<td>(YOX1 ∨ YHP1)</td>
<td>SBF ∨ HCM1</td>
<td>SWI5 ∨ ACE2</td>
<td>38.7%</td>
<td>0%</td>
<td>45.3%</td>
<td>16.0%</td>
</tr>
</tbody>
</table>

**Table 4.3:** The attractors of model given different logics. The percentages of all possible starting states that end in the four identified attractors for all simple functions of the activating logic rules for the repressive inputs to CLN3 and SBF (YOX1 and YHP1), and activators of SFF (SBF and HCM1), and CLN3 (SWI5 and ACE2) variables.

<table>
<thead>
<tr>
<th>Cycle 1</th>
<th>Cycle 2</th>
<th>Cycle 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1 S2 S3 S4 S5</td>
<td>S1 S2 S3 S4 S5</td>
<td>S1 S2 S3 S4 S5</td>
</tr>
<tr>
<td>MBF</td>
<td>1 0 0 0 0</td>
<td>1 0 0 0 0</td>
</tr>
<tr>
<td>SBF</td>
<td>1 1 0 0 0</td>
<td>1 1 0 0 0</td>
</tr>
<tr>
<td>YOX1</td>
<td>0 1 0 0 0</td>
<td>0 1 0 0 0</td>
</tr>
<tr>
<td>HCM1</td>
<td>0 1 0 0 0</td>
<td>0 1 0 0 0</td>
</tr>
<tr>
<td>YHP1</td>
<td>0 1 1 0 0</td>
<td>0 1 1 0 0</td>
</tr>
<tr>
<td>SFF</td>
<td>0 0 1 0 0</td>
<td>0 1 1 0 0</td>
</tr>
<tr>
<td>ACE2</td>
<td>0 0 0 1 0</td>
<td>0 0 1 1 0</td>
</tr>
<tr>
<td>SWI5</td>
<td>0 0 0 1 0</td>
<td>0 0 1 1 0</td>
</tr>
<tr>
<td>CLN3</td>
<td>0 0 0 0 1</td>
<td>0 0 0 0 1</td>
</tr>
</tbody>
</table>

**Table 4.4:** The on/off states of the transcription factors (TFs) in the each attractor. Cycle 1 attractor as used to color Figure 4.7c and shown in Figure 4.11b. The on/off states of the TFs in the Cycle 2 attractor as shown in Figure 4.11c. The on/off states of the TFs in the Cycle 3 attractor as shown in Figure 4.11d.
**Figure 4.7:** The periodic transcription program is largely intact in cyclin-mutant cells that arrest at the G1/S border.  

a–b) Genes maintaining periodic expression in cyclin-mutant cells (a) show similar dynamics in wild-type cells (b). Each row in a and b represents the same gene. Transcript levels are depicted as in Figure 4.2.  

c) Synchronously updating boolean network model. Transcription factors are arranged on the basis of the time of peak transcript levels in cyclin-mutant cells. Arrows indicate transcription factor/promoter interaction. Activating interactions, outer rings; repressive interactions, inner rings. Coloring indicates activity in one of five successive states; SBF and YHP1 are active in two states (Table 4.3).
**Figure 4.8:** Wild-type vs. cyclin-mutant PTRs for transcripts of all transcription factors in the Boolean network model. The average wild-type PTR (x-axis) plotted against the average cyclin-mutant PTR (y-axis) for each transcription factor (black circle, blue text) in the model depicted in Figure 4.7c. The dashed red line is drawn where the ratio between wild-type and cyclin-mutant PTR is one.
Figure 4.9: Histograms of the ratios of PTRs in wild-type vs. cyclin-mutant cells for the periodic targets of network transcription factors that exhibit reduced expression in cyclin-mutant cells. Histograms of the ratio between the wild-type and cyclin-mutant PTRs for all genes which were identified as periodic by this study, and are documented targets of Ace2, Swi5, and Yox1 according to the YEASTRACT database [10].
Figure 4.10: Full transcription factor network diagrams for wild-type and cyclin-mutant cells. Network diagrams (a,c) are similar to timeline representations of the circular diagram in Figure 4.7c. Periodically expressed transcription factors are placed on the cell-cycle timeline on the basis of the time of peak transcript levels in our data sets. Arrows indicate a documented interaction between a transcription factor and promoter elements upstream of a gene encoding another transcription factor. Transcriptional activators are depicted in green, repressors in red, and the cyclin Cln3 in blue. The actual b, wild-type and d, cyclin-mutant expression of the transcription factors in the networks are shown. Transcript levels are depicted as in Figure 4.2. The scale of the cell-cycle timeline has been altered to accommodate multiple G1 events.
FIGURE 4.11: A synchronously updating Boolean model can reproduce the sequential order of transcription factor expression. a) The actual expression of the variables in Figure 4.7c (with MBP1, SWI4, and FKH2 as proxies for SBF, MBF, and SFF, respectively), as compared to the on/off (yellow/cyan) states of those variables in b, Cycle 1, c, Cycle 2, and d, Cycle 3 (Table 4.4). Transcript levels are depicted as in Figure 4.2. The scale of the cell-cycle timeline has been altered to accommodate multiple G1 events.
Previous studies proposed that a cyclin/CDK-independent oscillator could trigger some periodic events, including bud emergence [67]. The robust oscillating character of our model indicates that a transcriptional regulatory network may function as this cyclin–CDK-independent oscillator. Because cyclin genes are themselves among the periodic genes targeted by this network, and because cyclin–CDKs can, in turn, influence the behavior of transcription factors in the network, precise cell-cycle control could be achieved by coupling a transcriptional regulatory network oscillator with the cyclin–CDK oscillator. The existence of coupled oscillators could explain why the cell cycle is so robust to significant perturbations in gene expression or cyclin–CDK activity. Our findings also indicate that the properly scheduled expression of genes required for cell-cycle regulated processes such as DNA synthesis and mitosis is not sufficient for triggering these events. The execution of cell-cycle events in wild-type cells is likely to require both properly timed transcription and post-transcriptional modifications mediated by CDKs.

4.2.1 Biological data generated

Yeast strains were grown in rich YEP medium (1% yeast extract, 2% peptone, 0.012% adenine, 0.006% uracil) containing 2% galactose. At 45 min before elutriation, dextrose was added to YEP 2% galactose medium to terminate CLB1 expression from the GAL1 promoter. After elutriation, wild-type and clb1,2,3,4,5,6 GAL1–CLB1 cells were grown in rich YEP 2% dextrose, 1M sorbitol at 30°C at a density of 10^7 ml^{−1}. Sorbitol was added to stabilize cells with elongated buds. Aliquots of 50 ml (cell density 10^7 ml^{−1}) were harvested every 8 min for 4 hrs. Budding index was determined microscopically by counting at least 200 cells for each time point and CLOCCS model (See Chapter 3) parameter estimates were generated for each time-series. Additional cell aliquots were harvested at 16 min intervals for 270 min (equivalent to about two cell cycles in the wild type and about 1.5 cell cycles in the cyclin mutant) for microarray quantitation. Transcript levels were measured genome-wide for each time point with the use of Yeast 2.0 oligonucleotide
arrays. CEL files from all 60 oligonucleotide arrays were normalized, and summarized with the dChip method [116] as implemented in the *affy* package (v.1.8.1) within Bioconductor using default parameters. The data is available from the NCBI Gene Expression Omnibus (Accession: GSE8799). Although the synchrony/time–series experiments used to generate this data were performed very carefully, due to the unavoidable loss of synchrony that occurs after release, the reported data is convolved, causing later time-points to be dampened. A CLOCCS model based deconvolution of this data is presented in Chapter 7.2.

### 4.2.2 Absolute change and Pearson correlation analyses

To determine if a gene changes its expression pattern in cyclin-mutant cells, we developed two scores: a Pearson correlation score (PCS) and an absolute change score (ACS). The PCS detects changes in timing of expression, but is relatively insensitive to changes in amplitude. The ACS is more sensitive to changes in amplitude of expression.

Each score is computed on the basis of two statistics, WTMT and MTMT. The exact form of these statistics for each score (WTMT\_P and MTMT\_P for PCS; WTMT\_A and MTMT\_A for ACS) is given below. The WTMT statistics capture change in gene expression from wild-type to cyclin-mutant cells, while the MTMT statistics capture variability in gene expression between cyclin-mutant replicates. As such, the MTMT statistics for all genes comprise a background distribution against which WTMT statistics can be compared. Scores for the change in expression for gene *i* relative to the corresponding background distribution are computed in each case by calculating the fraction of MTMT statistics that are more extreme (lower in the case of PCS and higher in the case of ACS) than the WTMT statistic for gene *i*. Thus for gene *i* the PCS is calculated as the fraction of all MTMT\_P statistics lower than WTMT\_P, and the ACS is calculated as the fraction of all MTMT\_A statistics greater than WTMT\_A. A gene is called different if either of the scores is less than or equal to 0.2 (when analyzing WTPER, a PCS equal to 0.2 roughly
corresponds to Pearson correlation of 0.75). This cutoff resulted in a set of 833 genes that are called different between wild-type and cyclin-mutant cells (denoted DIFF).

We also took advantage of the amplitude insensitivity of the PCS to select genes that, although they may change in amplitude, do not change in timing of expression and thus can be said to retain their periodicity. A gene is called similar if its PCS was greater than 0.1 (roughly corresponding to a Pearson correlation of 0.5 given $\text{WTPER}$). This cutoff resulted in a set of 882 genes that retain their periodicity between wild-type and cyclin-mutant cells (Figure 4.7). The calculated scores for each gene in $\text{WTPER}$ can be found in Supplemental Table S1 of [11].

**Pearson correlation statistics**

Let $w^1_i$ and $w^2_i$ represent the splined wild-type expression profiles for gene $i$ in wild-type replicates 1 and 2, respectively. Similarly, let $m^1_i$ and $m^2_i$ represent the splined cyclin-mutant expression profiles for gene $i$ in cyclin-mutant replicates 1 and 2, respectively. The WTMT Pearson correlation statistic for gene $i$ is defined as:

$$\text{WTMT}_P = \max(\rho(w^1_i, m^1_i), \rho(w^1_i, m^2_i), \rho(w^2_i, m^1_i), \rho(w^2_i, m^2_i))$$

and the MTMT Pearson correlation statistic for gene $i$ is defined as:

$$\text{MTMT}_P = \rho(m^1_i, m^2_i)$$

where $\rho(f, g)$ is the Pearson correlation between $f$ and $g$.

**Absolute change statistics**

The WTMT absolute change statistic for gene $i$ is defined as:

$$\text{WTMT}_A = \min(\tau(w^1_i, m^1_i), \tau(w^1_i, m^2_i), \tau(w^2_i, m^1_i), \tau(w^2_i, m^2_i))$$

and the MTMT absolute change statistic for gene $i$ is defined as:

$$\text{MTMT}_A = \tau(m^1_i, m^2_i)$$
where $\tau(f,g) = \frac{||f-g||_1}{2||f+g||_1}$. 

### 4.2.3 Dual-clustering & over-representation analysis

To further characterize how transcription changes in response to the loss of B-cyclins, we performed a two-step clustering of the 883 genes in DIFF. Both clustering steps use the affinity propagation algorithm of Frey and Dueck [110]. The first step clustered the 883 genes using only their wild-type expression profiles, defining the similarity between genes $i$ and $j$ as:

$$\rho(w^1_i, w^1_j) + \rho(w^2_i, w^2_j)$$

with $S_{kk}$ set to the value of $-1$ for all $k$. This first round of clustering produced 32 wild-type clusters. Each wild-type cluster was then subclustered in a second round of clustering based only on cyclin-mutant expression profiles, defining the similarity between genes $i$ and $j$ as:

$$\rho(m^1_i, m^1_j) + \rho(m^2_i, m^2_j)$$

with $S_{kk}$ again set to the value of $-1$ for all $k$. This second round of clustering produced a total of 109 subclusters within the 32 wild-type clusters. To ensure that subclusters had a sufficiently coherent expression pattern, subclusters whose average Pearson correlation of cyclin-mutant expression was less than 0.65 were removed. Fifty three subclusters—including all the subclusters of two wild-type clusters—were removed by this filter, resulting in a final set of 513 genes clustered into 30 wild-type clusters and 56 cyclin-mutant subclusters (Figure 4.6). These resulting clusters have a high average Pearson correlation (within cluster) of 0.808 and 0.790 for wild-type clusters and cyclin-mutant subclusters, respectively suggesting a high degree of co-regulation and coherent expression within each cluster.

If these coherent expression patterns are regulated by common transcription factors...
(TFs), we may expect to find certain TFs significantly associated with each cluster or subcluster. We tested this hypothesis by looking for over-represented TF binding to promoters of genes within each cluster and subcluster using the ChIP-chip data of Harbison et al. [35]. This dataset provides *p*-values for a total of 206 TFs binding to the intergenic (promoter) regions of most genes in *S. cerevisiae*. We chose all *p*-values ≤0.01 as evidence of a binding interaction between a TF and an intergenic region. We identified clusters or subclusters enriched for the given set of 206 TFs using a hypergeometric test. To increase the likelihood that the TF/cluster pairs we identified are biologically (and not just statistically) significant, we removed TFs that bound less than 3, or less than 10% of the genes in a cluster or subcluster. To correct for multiple hypothesis testing (since more than one TF could be enriched for a cluster or subcluster), we used the *q*-value method of Storey [117] to calculate *q*-values that act as FDR-corrected enrichment scores for each TF for every cluster and subcluster. A TF was designated as being over-represented in a given cluster or subcluster if its *q*-value was ≤0.005. The expression heatmaps and over-represented TFs for all clusters and subclusters can be found in Figure 4.6 and Table 4.1.

### 4.2.4 Construction of transcriptional regulatory networks

We constructed transcriptional regulatory networks using the set of periodically expressed TFs for wild-type cells and the subset of those which maintained periodic expression for cyclin-mutant cells. The set of all TF genes was defined by taking the union of all genes profiled by Harbison et al. [35] or listed as TFs in the YEASTRACT database [10]. In addition to this set, we added *MBP1, WHI5,* and *CLN3*. Mbp1 is a TF that exhibited periodic behavior, albeit below our periodicity score threshold, which along with Swi6 forms the MBF complex that is known to regulate G1 transcription [103]. Whi5 is a repressor of MBF and SBF transcription factor complexes, and Cln3 is known to directly inhibit Whi5 and thereby activate MBF and SBF complexes [72, 73]. While the genes encoding most transcription factors in the network are transcribed at normal levels in cyclin-mutant
cells, some (ACE2, SWI5, and YOX1) are expressed at reduced levels (Figure 4.4f and Figure 4.8). However, the fact that their target genes (e.g., the Ace2 targets SIC1 and NIS1; see Figure 4.4c and d, and Figure 4.9) are expressed at near-normal levels indicates that the periodic expression of these transcription factors is still biologically relevant.

A node corresponding to each TF was placed on the cell-cycle lifeline according to its time of peak expression (temporal ordering of nodes is preserved, but the scale within G1 is expanded to accommodate the depiction of multiple G1 events). Edges between nodes corresponding to TF-promoter interactions were drawn if the ChIP-chip data of [35] reported a p-value less than or equal to 0.01 or if the interaction had been documented in the literature [5, 10, 33, 34, 35, 72, 73, 108, 109, 113, 118, 119, 120, 121, 122, 123, 124, 125, 126]. Edges were classified as either activating (green) or repressing (red) on the basis of the function of the TF (those TFs whose function was ambiguous (e.g., RAP1) were pruned from the network). Self-activating edges and activating edges whose length was more than 40% of the length of the cell cycle were pruned. Nodes that had neither input nor output edges were pruned. TFs that act in known complexes (MBF, SBF, and SFF, reviewed in Wittenberg and Reed [96]) were collapsed into a single node. The cyclin-mutant transcriptional regulatory network was constructed using the same procedure, but TFs that did not maintain their timing in cyclin-mutant cells were excluded. The resulting wild-type and cyclin-mutant networks are shown in Figure 4.10a and c.

4.2.5 Boolean model construction

A synchronously updating Boolean model was constructed for all TFs that remained periodic in cyclin-mutant cells. The transcriptional regulatory network architecture was constructed as described above. TFs that did not have both input and output edges with high-confidence literature support (Supplemental Table S5 of [11]) were pruned from the network (see Figure 4.10 for full network diagrams in wild-type and cyclin-mutant cells). The remaining TFs were treated as variables in our model and the set of controlling Boolean
logic functions were determined as the strictest (i.e., AND wherever possible) combinatorial functions over incoming edges that reproduced the observed data (Figure 4.11). Initial choices of logic functions are shown in Table 4.2. To examine robustness, we determined the attractors for all possible starting states given these initial choices of logic functions. Results are shown in Row 1 of Table 4.3. We found that 80.3% of all the 512 possible starting states enter a cycle containing five states (Cycle 1). The values of all TFs in each of the five states in Cycle 1 are shown in Table 4.4.

To determine whether this cyclic attractor was dependent on our initial choice of logic functions (Table 4.2), we examined models generated with all possible simple logic functions over the activating inputs to the CLN3, SFF, and repressor variables. The repressor variable represents YOX1 and YHP1, and the activating logics determine if these repressors can inhibit their targets only when both components are active (AND logic) or when either component is active (OR logic). For each of the models, we determined the final attractor for all possible starting states (Table 4.3). This analysis revealed the presence of two additional five-state cycles, Cycle 2 and Cycle 3. Cycle 2 and Cycle 3 (Table 4.4) are qualitatively similar to Cycle 1 (Table 4.4). These cycles maintain the same temporal order of expression as Cycle 1 (Figure 4.11b), and differ only in the duration of expression of certain TFs (Figure 4.11c and d).

4.3 Challenges in data analysis

The analysis presented in the sections above relies on being able to directly compare the expression patterns of genes between wild-type and cyclin-mutant cells through CLOCCS alignment. The analysis also relies on the accurate determination of the set of periodically expressed genes. In the sections below we discuss the details of the alignment of wild-type and cyclin mutant datasets, and comment on the difficulties in determining a true set of periodic genes.
4.3.1 Alignment of wild-type and cyclin mutant datasets

To accurately compare cell-cycle gene expression profiles across our two wild-type and two cyclin-mutant datasets, we need to ensure that we are comparing relevant data together. In the cell-cycle/synchrony setting, this means comparing data points where the underlying populations are in the same cell-cycle stage. We used the CLOCCS model [36] to correct for timing differences inherent in synchrony/time-series experiments, and for the modest differences in the cell-cycle time between wild-type and cyclin-mutant cells. Only minor modifications to the model are required to predict cell-cycle position and distribution of cyclin-mutant populations (see Chapter 3 for details on the CLOCCS model and alignment).

In terms of modeling, the most important difference between wild-type and cyclin-mutant populations is that cyclin-mutant cells do not divide and hence no new daughter cell cohorts can arise. The CLOCCS model for cyclin-mutant cells thus ignores cohorts other than the initial \(\{0, 0\}\) cohort. As a result, the mother/daughter offset parameter \(\delta\) is no longer used because it relates only to the effects of asymmetric cell division. The modified CLOCCS model for cyclin-mutant cells is:

\[
\Pr(P_t|\Theta, t) = \sum_G \sum_R \Pr(P_t|\Theta, G, R, t) \Pr(G|\Theta, R, t) \Pr(R|\Theta, t)
\]

\[
= \Pr(P_t|\Theta, G = 0, R = 0, t)
\]

\[
= \phi \left( \frac{P_t - (-\mu_0 + t)}{\sqrt{\sigma_0^2 + t^2 \cdot \sigma_v^2}} \right)
\]

where \(\phi()\) is the standard normal density function, \(\Theta = (\sigma_0^2, \sigma_v^2, \mu_0, \lambda, \beta)\), and all variables are defined exactly as in the original CLOCCS model (Chapter 3 and [36]).

Unlike in a wild-type population where cells are either budded or unbudded, cyclin-mutant cells never lose their first bud, but grow an additional bud at the G1/S border of each subsequent cell cycle. Hence, before their first S phase \((P_t < \lambda\beta)\), cyclin-mutant cells
will have no buds, between their first and second S phases \((\lambda \beta \leq P_t < \lambda(1 + \beta))\), they will have one bud, between their second and third S phases \((\lambda(1 + \beta) \leq P_t < \lambda(2 + \beta))\), they will have two buds, and so on. We use this to calculate the expected percentages of cells with \(j\) buds at time \(t\):

\[
\pi_{jt} = \begin{cases} 
\Phi \left( \frac{\lambda(j+\beta)-(-\mu_0+t)}{\sigma_0^2 + t^2 \cdot \sigma_v^2} \right) - \Phi \left( \frac{\lambda(j-1+\beta)-(-\mu_0+t)}{\sigma_0^2 + t^2 \cdot \sigma_v^2} \right) & \text{if } j \geq 1 \\
\Phi \left( \frac{\lambda \beta-(-\mu_0+t)}{\sigma_0^2 + t^2 \cdot \sigma_v^2} \right) & \text{if } j = 0
\end{cases}
\]

where \(\Phi()\) is the standard normal cumulative density function. We can estimate the parameters \(\Theta\) of the modified CLOCCS model by maximizing the multinomial likelihood function:

\[
L(\Theta) \propto \prod_{t=1}^{T} \prod_{j=0}^{\infty} \pi_{jt}^{c_{jt}}
\]

where \(c_{jt} = \sum_i X_{ijt}\) is the count of the number of cells with \(j\) buds at time \(t\), and \(T\) is the number of timepoints at which budded cells are counted.

The CLOCCS model was fit on wild-type datasets using 32 timepoints of budding index data measured every 8 minutes, starting 30 minutes after release (Figure 4.12a and b). Similarly, the modified CLOCCS model was fit on cyclin-mutant datasets using 30 timepoints in one replicate (Figure 4.12c) and 29 timepoints in the other (Figure 4.12d) of budding index data measured every 8 minutes, starting 30 minutes after release. When fitting the cyclin-mutant datasets, since \(\beta\) and \(\mu_0\) are not simultaneously identifiable from the cyclin-mutant budding index data, \(\beta\) was fixed to be 0.15. CLOCCS parameter estimates are given in Table 4.5.

### 4.3.2 Periodic gene detection

In order to fully understand the functions of a transcriptional regulatory network oscillator, it is important to define the subset of genes that comprise and are controlled by the
Figure 4.12: CLOCCS model fits for the four time-series. Fraction of cells with a single bud (black lines, filled circles) and two or more buds (black lines, open circles) for (a,b), wild-type cells and (c,d), cyclin-mutant cells. One hundred random realizations from the Markov chain used to fit each experiment were used as parameterizations for the model, and the resulting predicted budding curves for one bud (green lines) and two or more buds (red lines) are shown. The width of the band reflects the degree of posterior uncertainty in the budding curve and can be interpreted as forming a confidence band for the curve on the basis of the uncertainty in the parameters. Parameter estimates for each experiment are given in Table 4.5.

oscillator. However, identifying the set of periodically expressed genes is a non-trivial endeavor and has been the subject of intense study [5, 11, 12, 102, 127, 128]. For the analysis described in Section 4.3.2, genes exhibiting periodic behavior were determined using the method of de Lichtenberg et al. [102]. Probe-level wild-type data from both replicates—with the CLOCCS recovery phase (Gr) removed—were used as input. Permutation \( p \)-values assessing ‘degree of regulation’ and ‘degree of periodicity’ were calculated for each probe in each wild-type replicate (one million permutations were used for each).
These four $p$-values (two from each replicate) were normalized and combined into a final score for each probe. We chose to identify periodic probes by defining a threshold on the basis of the final score rather than on the basis of rank. A probe with all four $p$-values of 0.2 would be given a final score of 26,020,272, and we used this as our score threshold. This threshold resulted in a total of 1274 probes being labeled as periodic. These probes were then mapped to their respective genes and since some probes map to the same gene, this resulted in a list of 1270 periodic genes. This mapping step is a critical step, and unfortunately, one that is rather ill-defined. We used the mapping file provided by the Saccharomyces Genome Database at the time of the analysis [40]. However, as the genome annotation changes, these mapping files may change as well, resulting in slightly different periodic gene lists. The gene $MCM1$ was not in our list of 1270 mapped periodic genes, but because of its documented role in regulating cell-cycle transcription [123] and its periodic expression by visual inspection, it was added to the our list, resulting in the 1271 periodic wild-type genes (WTPER) shown in Figure 4.2.

In addition to this list of 1271 periodic genes, two separate studies reporting global time-series transcription data across the mitotic cell cycle have generated unique lists of

<table>
<thead>
<tr>
<th>Experiment</th>
<th>$\mu_0$</th>
<th>$\delta$</th>
<th>$\sigma_0$</th>
<th>$\sigma_v$</th>
<th>$\lambda$</th>
<th>$\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wild-type 1</td>
<td>95.8</td>
<td>41.4</td>
<td>15.3</td>
<td>0.06</td>
<td>77.1</td>
<td>0.14</td>
</tr>
<tr>
<td></td>
<td>(93.2, 98.2)</td>
<td>(37.1, 45.4)</td>
<td>(14.2, 16.4)</td>
<td>(0.04, 0.08)</td>
<td>(73.4, 80.9)</td>
<td>(0.12, 0.17)</td>
</tr>
<tr>
<td>Wild-type 2</td>
<td>100.3</td>
<td>35.1</td>
<td>20.8</td>
<td>0.05</td>
<td>85.0</td>
<td>0.18</td>
</tr>
<tr>
<td></td>
<td>(96.8, 103.3)</td>
<td>(28.7, 40.7)</td>
<td>(19.8, 21.8)</td>
<td>(0.04, 0.07)</td>
<td>(80.6, 89.9)</td>
<td>(0.15, 0.21)</td>
</tr>
<tr>
<td>Cyclin-mutant 1</td>
<td>95.6</td>
<td>0</td>
<td>11.2</td>
<td>0.15</td>
<td>120.1</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>(94.4, 96.9)</td>
<td></td>
<td>(9.4, 13.1)</td>
<td>(0.14, 0.16)</td>
<td>(118.1, 122.1)</td>
<td></td>
</tr>
<tr>
<td>Cyclin-mutant 2</td>
<td>89.3</td>
<td>0</td>
<td>23.0</td>
<td>0.13</td>
<td>114.8</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>(87.7, 90.9)</td>
<td></td>
<td>(21.2, 24.9)</td>
<td>(0.11, 0.14)</td>
<td>(112.8, 116.7)</td>
<td></td>
</tr>
</tbody>
</table>

**TABLE 4.5:** CLOCCS parameter estimates. Each row corresponds to a different arrest-release experiment. The observed budding curves, as well as the model predictions for each experiment, are shown in Fig. Figure 4.12. Each cell contains the mean value for that column’s parameter, given 250,000 iterations of the Markov chain used to fit that row’s experiment. Below each mean is the 95% confidence range for that parameter in that experiment. Since the values of $\beta$ and $\delta$ were fixed to be 0.15 and 0, respectively, when fitting CLOCCS in the cyclin-mutant experiments, no confidence intervals exist for these parameters in those experiments.
periodically transcribed genes [5, 12]. The overlap between these three lists is surprisingly small (Figure 4.3). Of the 1831 genes collectively identified as periodic by all studies, only 24% are identified by all three, and less than 50% are identified by at least two of the studies.

The surprising lack of overlap in these studies highlights the difficulties in unambiguously identifying periodic genes, and likely reflects both the lack of a consistent definition of a cell-cycle-regulated gene, as well as variation in noise levels during data collection and analysis. By visual inspection of transcript abundance curves, it seems relatively easy (albeit very subjective) to determine if a gene is periodically transcribed. The challenge is translating human intuition into a set of objective mathematical features that can be utilized computationally to analyze large data sets.

Most algorithms designed to identify cell-cycle-regulated genes simply provide a ranking of genes from most to least ‘periodic’ however, they do not provide guidance for determining the cutoff between periodic and non-periodic genes. Regardless of the diligence used in manually choosing a cutoff, every set of periodic genes identified will contain some false positives and some periodic genes will be excluded. For example, Figure 4.13 shows two genes (CHA1 and SSA1) on either side of the cutoff used to generate WTPER. By visual examination of transcript dynamics in our dataset, it is hard to argue conclusively that one is truly periodic and one is not. However, by visually examining data from the other two studies [5, 12] as well, one could argue that the behavior of CHA1 is more consistently periodic than that of SSA1 (Figure 4.13). Thus, confidence in the periodic nature of any specific gene can be bolstered by examining data from multiple studies.

4.4 Summary

In this chapter we examined the mechanisms regulating the periodic gene expression which occurs during the S. cerevisiae cell cycle. It had been the long-standing view that
Figure 4.13: Expression dynamics of CHA1(a) and SSA1(b) during cell cycle progression. Transcript levels in: Orlando et al. [11], black; Pramila et al. [5], broken gray; Spellman et al. [12], solid gray. Time points from each experiment were mapped onto a common cell cycle timeline using the CLOCCS model. Relative expression is defined as the ratio of transcript levels at each time point to the minimum level during the interval shown.

This periodic behavior was exclusively regulated by the activities of the cyclin/CDK complex [65, 66]. However, systems-level analyses using high-throughput technologies have suggested an alternative model in which the regulation of periodic transcription is an emergent property of a transcriptional regulatory network [5, 12, 33, 34, 98, 99]. In these networks, the temporal program of transcription could be controlled by sequential waves of transcription factor expression, even in the absence of extrinsic control by cyclin–CDK complexes.

We sought to determine to what extent CDKs and transcription factor networks contribute to global regulation of the cell-cycle transcription program. To this end, we investigated the dynamics of genome-wide transcription in budding yeast cells disrupted for all S-phase and mitotic cyclins (clb1,2,3,4,5,6). If Clb–CDK activities are essential for
triggering the transcriptional program, then periodic expression of S-phase-specific and G2/M-specific genes should not be observed, and if Clb–CDK activity is unimportant then periodic expression will be unchanged.

We found that eventhough a significant fraction of periodic genes are aberrantly expressed, 70% of periodic genes continued to be expressed periodically and on schedule. That suggests that although CDKs clearly have a function in the regulation of cell-cycle transcription, they are not solely responsible for establishing the global periodic transcription program. To test whether the remaining 70% of periodic expression could be the result of a transcriptional regulatory network we constructed a synchronously updating boolean transcriptional regulatory network model. We showed that this model, under a variety of boolean logic functions, produce a highly robust 5-state cycle which largely mimics the actual observed expression dynamics of the nodes. The robust oscillatory nature of the transcriptional regulatory network model suggests that periodic expression could, at least in part, be regulated by a transcriptional oscillator.

Our proposal of an underlying transcriptional oscillator is not inconsistent with previous models of cell-cycle regulation in eukaryotes. We propose that the cyclin/CDK oscillator is coupled to the transcriptional oscillator such that timed expression of cyclins is controlled by the transcriptional oscillator and, in turn, cyclin/CDKs modulate the activity of the transcriptional oscillator by phosphorylating transcription factors. Just as the circadian oscillator can be reinforced by light/dark cycles, we proposed that coupling of the two cell cycle oscillators could explain the robustness of cell cycle progression to perturbations in either gene expression or cyclin/CDK activity. The model we propose is not the complete, or even likely correct, transcriptional regulatory network. It is simply a proof-of-concept model to demonstrate the validity of cyclin/CDK independent periodic behavior as an emergent property of a transcriptional regulatory network. The detailed nature of the complete transcription-factor oscillator and its interplay with the cyclin/CDK oscillator, awaits further investigation.
Arabidopsis thaliana and root development

Arabidopsis thaliana is a small weed in the mustard family (Brassicaceae), whose members include cabbage and radish. First studied by Laibach in 1907 [129], it has become a model system in plant biology research, especially for studying plant organogenesis. A. thaliana is well suited as a model plant because of its small genome size, rapid life cycle, prolific seed production, and ease of transformation and growth. It has a fully sequenced genome organized into five chromosomes (excluding chloroplast and mitochondrial chromosomes) containing about 27000 protein coding genes [130]. With a total size of 125 Mb, the A. thaliana genome is the smallest known among higher plants making genetic studies easy [130, 131]. For comparison, the genome sizes of other commonly studied plants, Oryza sativa (rice), Zea mays (corn), and Nicotiana tabacum (tobacco), are ~412 Mb, ~2800 Mb, and ~4500 Mb respectively [132, 133]. The Arabidopsis life cycle, from seed germination through maturation of the first new seeds, is completed within six weeks allowing for facile genetic analysis. There are also efficient whole-plant transformation methods using Agrobacterium tumefaciens for introducting DNA constructs into A. thaliana. This ease of transformation coupled with the rapid life cycle makes the effi-
cient construction of mutant lines possible [134]. The plants themselves are small (mature plants are rarely taller than 20 cm) and can be grown densely on agar plates or soil and grow well in both greenhouses and under fluorescent laboratory lighting [134]. While all of these features make *A. thaliana* a good model system for almost any question in plant biology, its simplistic root structure makes it especially tractable as a model system for studying organogenesis and development.

### 5.1 *A. thaliana* as a model for root development

The simplicity of the *Arabidopsis* root makes it a good developmental model system. The *Arabidopsis* root is a nearly radially symmetric structure organized along a longitudinal axis divided into three major developmental zones (Figure 5.1). It is comprised of ten tissues with a total of 15 cell types (Table 5.1). These cell-types arise through a series of stereotyped divisions of four sets of initials (stem cells) at the distal tip of the root. The epidermis and lateral root cap cell-types arise from one set of initials. The columella from a second set. The ground tissue, the cortex and endodermis, arise from a third set. And the stele, comprised of the pericycle, procambium, and vascular tissues (xylem and phloem) arise from a fourth set of initials. These four sets of initials surround a mitotically inactive set of cells known as the quiescent center (QC). Laser ablation studies have suggested that the QC acts to maintain the stem-cell-like undifferentiated identity of the initials by ensuring that each time an initial cell divides, only one of the progeny differentiates, namely the one not adjoining the QC [135]. This ensures that new cells are produced from only one location at the tip of the root. Coupled with the fact that plant cells do not move in relation to one another, this feature creates an organ in which cell-types are constrained within cell files along the longitudinal (vertical) axis. At each stem cell division, the newly produced cell successively displaces an older cell distal to the initials. As cells move along the longitudinal axis, they pass through three zones: the meristematic, elongation, and
maturation zones, in which the cells undergo division, elongation, and differentiation respectively. Thus, these cell files are a spatial ‘time-course’ of development with younger, less developed cells closer to the root tip and older, more differentiated cells further up the longitudinal axis.

<table>
<thead>
<tr>
<th>Tissue</th>
<th>Cell-Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lateral Root Cap</td>
<td>Lateral Root Cap</td>
</tr>
<tr>
<td>Epidermis</td>
<td>Hair Cell</td>
</tr>
<tr>
<td></td>
<td>Non-hair cell</td>
</tr>
<tr>
<td>Columella</td>
<td>Columella</td>
</tr>
<tr>
<td>Cortex</td>
<td>Cortex</td>
</tr>
<tr>
<td>Endodermis</td>
<td>Endodermis</td>
</tr>
<tr>
<td>Procambium</td>
<td>Procambium</td>
</tr>
<tr>
<td>Pericycle</td>
<td>Xylem Pole Pericycle</td>
</tr>
<tr>
<td></td>
<td>Phloem Pole Pericycle</td>
</tr>
<tr>
<td>Xylem</td>
<td>Metaxylem</td>
</tr>
<tr>
<td></td>
<td>Protoxylem</td>
</tr>
<tr>
<td>Phloem</td>
<td>Metaphloem</td>
</tr>
<tr>
<td></td>
<td>Protophloem</td>
</tr>
<tr>
<td>Quiescent Centre</td>
<td>Quiescent Centre</td>
</tr>
</tbody>
</table>

Table 5.1: Tissue and cell-types in the *A. thaliana* Root. This table defines the relationships between tissues and cell-types in the *A. thaliana* root. The rows are grouped by which set of initials they arise from.

Normally, differentiation and development occurs in a four-dimensional space, with a cells cell-type specified in three-dimensional space (x-y-z) and developing through time. The cellular organization of the *Arabidopsis* root reduces this complexity to two dimensions. The radial symmetry of the cell-types allows for the representation of cell identity in the one-dimensional radial axis, while the constrained cell files allow for the temporal development to be represented by position along the one-dimensional longitudinal axis. Having the entire organogenesis of the root represented in a two-dimensional structure is a tremendous advantage and makes the challenge of elucidating the mechanisms controlling the development of the *A. thaliana* root significantly more tractable.
Figure 5.1: The structure of the *A. thaliana* root. A representation of the *A. thaliana* root (see Section 6.5 for details on digital root representation) with each cell-type identified by its own color. The three developmental zones are demarcated on the left. The lateral root primordia is not counted in the 15 cell-types as it can be a mix of multiple cell types. The meta/proto xylem and phloem are counted as four cell-types in total.

## 5.2 Importance of transcriptional regulatory networks in root development

The development of the *A. thaliana* root is controlled by a variety of signals and mechanisms, such as hormone signaling and transcriptional regulation. Perturbations in the distribution or utilization of the hormone auxin are known to have profound developmental effects [136]. While the effects of auxin and other plant hormones are well documented, they are also promiscuous and varied. This makes it difficult to explain all of development on the basis of hormone signalling. In addition, there is no high-throughput, experimen-
tally tractable method for identifying signaling proteins.

Transcriptional regulation is known to be important in controlling the development and differentiation. In contrast to auxin’s wide ranging influences, it has been possible in certain cases to tie specific developmental steps to the actions of a particular transcription factor or factors. The Myb coiled-coil transcription factor, ALTERED PHLOEM DEVELOPMENT (APL) is known to be required for proper phloem specification [137] and may have an additional role in repression of xylem formation [138]. Members of the plant-specific GRAS family of putative transcription factors, SHORT-ROOT (SHR) and SCARECROW (SCR) are known to play key roles in the specification of the cortex, endodermis initial, and endodermis cell-types [135]. SHR has been shown to be responsible for specifying endodermis and to be required for the transcriptional activation of SCR. SCR is required for the division of the initial cell to form the cortex and endodermis. The determination of the hair and non-hair cell-types within the epidermis is regulated by the activity of WEREWOLF (WER), CAPRICE (CPC), and GLABRA2 (GL2) genes, among others. WER and CPC are MYB transcription factors, while GL2 is a homeodomain transcription factor. Cells with high levels of WER relative to CPC will become non-hair cells due to the transcriptional activation of GL2 which inhibits hair cell formation. These cells also activate the transcription of CPC which then moves laterally into the adjacent cells where it represses WER and CPC allowing hair cell formation.

There are many other examples of transcription factors being expressed in a cell-type specific manner suggesting that transcriptional regulation may be a fundamental developmental mechanism for specifying cell-type identity in the A. thaliana root. By measuring the expression level of each of these potential developmental regulators and their targets in each cell-type it may be possible to elucidate a set of transcriptional regulatory networks controlling root development.
5.3 Generating expression data

The development of the Affymetrix™ Arabidopsis ATH1 22K microarray and cell sorting technologies, have made it possible to measure the expression of approximately 22000 genes in almost every cell-type in the A. thaliana root. To profile expression in individual cell-types, the roots of a set of marker lines (transgenic plants expressing GFP in specific cell-types) are collected and protoplasted (enzymatic digestion of their cell walls) into single cells. These GFP-expressing cells are selectively collected using a fluorescence-activated cell sorter. RNA is extracted from this collection of marked cells and analyzed with the ATH1 microarray. This technique is high-throughput, allowing for the isolation of about 10 million cells in about 1.5 hrs. Through the work described in Chapter 6 and previous efforts [139, 140, 141, 142], there now exists a collection of 19 marker lines that cover 14 nonoverlapping cell-types (including the lateral root primoridia which may contain multiple immature cell-types). This collection is used to generate expression profiles for nearly all differentiated cells in the root (Chapter 6). In parallel to measuring transcription across cell differentiation it is also possible to generate a complete developmental time-course of transcription. This is achieved by microdissecting single roots along the longitudinal axis into samples for analysis by microarray. The resolution of this developmental time-course is not limited by the three developmental zones shown in Figure 5.1 since multiple consecutive sections can be take within each. The relationship between the marker lines and the cell-types and sections they mark is shown in Figure 5.2 and Table 5.2.

It is clear from Figure 5.2 that every cell-type in every section is not marked uniquely by a marker line. This means that when a sample of marked cells is analyzed, it represents a convolved mix of multiple cell-types. This is also true for the longitudinal sections, where the measurement represents a convolution of every cell-type present in that section. In Section 7.3 we discuss a new deconvolution technique for learning the true cell-type by
<table>
<thead>
<tr>
<th>Cell type</th>
<th>Markers</th>
<th>Longitudinal Section</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lateral root cap</td>
<td>LRC</td>
<td>0–5</td>
</tr>
<tr>
<td>Columella</td>
<td>PET111</td>
<td>0</td>
</tr>
<tr>
<td>Quiescent centre</td>
<td>AGL42</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>RM1000</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>SCR5</td>
<td>1</td>
</tr>
<tr>
<td>Hair cell</td>
<td>N/A</td>
<td>1–6</td>
</tr>
<tr>
<td></td>
<td>COBL9</td>
<td>7–12</td>
</tr>
<tr>
<td>Non-hair cell</td>
<td>GL2</td>
<td>1–12</td>
</tr>
<tr>
<td>Cortex</td>
<td>J0571</td>
<td>1–12</td>
</tr>
<tr>
<td></td>
<td>CORTEX</td>
<td>6–12</td>
</tr>
<tr>
<td>Endodermis</td>
<td>J0571</td>
<td>1–12</td>
</tr>
<tr>
<td></td>
<td>SCR5</td>
<td>1–12</td>
</tr>
<tr>
<td>Xylem pole pericycle</td>
<td>WOL</td>
<td>1–8</td>
</tr>
<tr>
<td></td>
<td>JO121</td>
<td>8–12</td>
</tr>
<tr>
<td></td>
<td>J2661</td>
<td>12</td>
</tr>
<tr>
<td>Phloem pole pericycle</td>
<td>WOL</td>
<td>1–8</td>
</tr>
<tr>
<td></td>
<td>S17</td>
<td>7–12</td>
</tr>
<tr>
<td></td>
<td>J2661</td>
<td>12</td>
</tr>
<tr>
<td>Phloem</td>
<td>S32</td>
<td>1–12</td>
</tr>
<tr>
<td></td>
<td>WOL</td>
<td>1–8</td>
</tr>
<tr>
<td>Phloem ccs</td>
<td>SUC2</td>
<td>9–12</td>
</tr>
<tr>
<td></td>
<td>WOL</td>
<td>1–8</td>
</tr>
<tr>
<td>Xylem</td>
<td>S4</td>
<td>1–6</td>
</tr>
<tr>
<td></td>
<td>S18</td>
<td>7–12</td>
</tr>
<tr>
<td></td>
<td>WOL</td>
<td>1–8</td>
</tr>
<tr>
<td>Lateral root primordia</td>
<td>RM1000</td>
<td>11</td>
</tr>
<tr>
<td>Procambium</td>
<td>WOL</td>
<td>1–8</td>
</tr>
</tbody>
</table>

Table 5.2: Description of cell-types, with associated markers. This table defines the relationship between cell-type and marker line, including which longitudinal sections of each cell-type is marked by the line. Note that, differing from Table 5.1, the Lateral Root Primordia is included as a cell-type in this table, even though it may be a collection of multiple immature cell types. Also, because there are also no markers that differentiate between metaxylem and protoxylem or between metaphloem and protophloem, those cell-types are labeled as Xylem and Phloem respectively.

Section specific expression profile. Even without deconvolution, the ability to gather high quality cell-type, and developmental section, gene expression data makes the A. thaliana an ideal model system for the study of transcriptional programs regulating plant organo-
Figure 5.2: Microarray expression profiles of 19 GFP-marked lines were analyzed. The colors associated with each marker line reflect the developmental stage and cell-types sampled (Table 5.2, CC=companion cells). Thirteen transverse sections were sampled along the root’s longitudinal axis (red lines).
6

Transcriptional regulation of root development

6.1 Introduction

The development of multicellular organisms is regulated by transcriptional programs that act to specify cell types and to provide positional information [30]. Elucidating these programs that underlie development is essential to understanding the acquisition of cell and tissue identity and requires detailed knowledge of the transcription at a resolution of cell-type at a specific developmental stage.

Through a collaboration with Dr. Siobhan Brady, we present microarray expression profiles of a high-resolution set of developmental time points within a single Arabidopsis Thaliana root and a comprehensive map of nearly all root cell-types; see Chapter 5 for an introduction to A. thaliana. These cell-type-specific transcriptional profiles can often be used to predict novel cellular functions. Computationally identified dominant expression patterns demonstrate transcriptional similarity between disparate cell-types. Dominant expression patterns along the root’s longitudinal axis do not strictly correlate with previously defined developmental zones, and in many cases, we observe expression fluctuation along
this axis. Both robust co-regulation of gene expression and potential phasing of gene expression were identified between individual roots. Methods that combine these profiles demonstrate transcriptionally rich and complex programs that define Arabidopsis root development in both space and time.

6.2 High-Resolution root spatiotemporal expression map

Here, we present, at high resolution, the first microarray expression map of a single organ profiling expression of nearly all Arabidopsis root cell-types. A previous analysis combining fluorescence-activated cell sorting of green fluorescent protein (GFP)-marked cell populations with microarray analysis to describe expression profiles of five tissue types and three developmental zones in the root [139]. These tissue and developmental zone profiles revealed a greater transcriptional complexity than profiles of the whole organ alone [139]. To accurately describe all transcriptional patterns that occur in the root, however, requires a higher-resolution data set profiling all cell-types and developmental stages within an organ. Using the fluorescence cell-sorting expression analysis method [139], we obtain expression profiles of eight new GFP-marked cell populations (S17, S32, COBL9, JO121, S4, SUC2, J2501, and RM1000) [140, 143, 144, 145]. We combine these with 11 previously published marker line profiles (PET111, AGL42, LRC, GL2, CORTEX, SCR, J0571, J2661, APL, WOL, and S18) [139, 140, 141, 142] to form a comprehensive data set of 19 experiments (referred to as the radial data set) profiling expression of 14 nonoverlapping cell-types in the Arabidopsis root (Table 5.2). To profile developmental stages in the root, we microdissect a single root into 13 sections, with each section encompassing three to five cells along the longitudinal axis, and we measure the expression of each by microarray (referred to as the longitudinal data set). Because temporal expression variation in shoot tissue has been demonstrated, and noise in gene expression between genetically identical organisms can determine cell fate [146, 147], we also microdissect a second root
to assess expression variation between roots in developmental time. See Section 6.3 for more details on the data.

6.2.1 Supervised data analysis

We first ask how expression between individual cell-types differs, by using a mixed-model ANOVA (analysis of variance) to determine significant differential expression. We selected ten marker lines that profile individual cell-types, two lines that explore expression variation within the pericycle, and four lines that profile varying developmental stages of xylem and phloem (Table 5.2). A gene was determined to be enriched in a cell-type if it was 1.2-fold enriched with a q-value of less than 0.001 compared to all other non-overlapping cell-types. Furthermore, a gene was also determined to be enriched in a cell-type if it is 2-fold enriched with no p-value or q-value threshold as this was determined to identify valid in vivo cell-type expression enrichment [140]. Biological functions of individual cell-types have been characterized by genetic or physiological studies, but no comprehensive analysis has described the putative functions of each cell-type as defined by its transcriptome. We therefore examined Gene Ontology (GO) term enrichment using software written by Jeremy Koch (described here [148] and available at www.arexdb.org). This software utilizes the hypergeometric distribution and corrects for multiple hypothesis testing to infer cellular function [117]. A number of genes have also been associated with biological processes through microarray expression analysis. We therefore mined the literature for such gene associations and for features that link genes to transcriptional regulatory networks through cis-element enrichment (Figure 6.1 and Figure 6.2). Two interesting trends were observed for GO term enrichment: First, the majority of enriched GO terms were associated with individual cell-types, and second, those GO terms associated with multiple cell-types represent more general biological processes (e.g., peroxisomal activity and membrane localization) (Figure 6.4a, and Supplemental Figure 2 and Table 3 of [15]) [13].
As a measure of the reliability of our method in identifying cell-type-enriched genes and of our ability to correctly annotate biological processes to a cell-type, we generated a list of genes enriched in root hair cells and compared it to a previous study that profiled root hair morphogenesis [14]. In accordance with Jones et al. [14], we found root hair cell differentiation genes, kinases, and cell wall-loosening genes enriched in hair cells (P = 2.1e−4, 5.3e−6, 3.8e−4) (Figure 6.4b). Our analysis further increased the spectrum of biological processes associated with root hair development to include calcium ion transport, vesicle docking during exocytosis, and nicotinamide adenine dinucleotide and/or nicotinamide adenine dinucleotide phosphate (NAD+/NADP+) activity (P = 2e−4, 1.5e−4, 1.1e−4, 5e−4). With this added confidence in our method, we then used enriched gene categories to infer the function of other cell-types (Figure 6.1 and Figure 6.2). Of particular note were xylem cells in the meristematic zone that were enriched for translation initiation and elongation, RNA binding and processing, and mitosis (P = 7.2e−8, 5e−4, 6.1e−13, 1.8e−5, 7.3e−10) [149] (Figure 6.3). Strong co-localization of the first two enzymes in flavonoid biosynthesis in the cortex has been demonstrated by immunolocalization [150]. We provide further evidence for cortex-enriched flavonoid biosynthesis with enriched expression of four of the five flavonoid biosynthetic enzymes [P = 6.12e−6]. Another striking example is the enrichment of genes involved in defense, heat, and oxidative stress responses in the endodermis [P = 4.3e−4, 8.97e−5, 8.23e−6].

6.2.2 Unsupervised data analysis

We were also interested in identifying complex spatial expression patterns that are specific to two or more cell-types, and therefore, we developed an unsupervised approach to identify a representative set of dominant, relative expression patterns across cell-types and along developmental time. If one does not know the a priori expression patterns present in a biological context, there are two methods that can identify them. The first enumerates all possible patterns and the second uses clustering. However, given the length and complex-
Figure 6.1: Summary of significant array annotation term enrichment by cell-type. The hypergeometric distribution P-value is $log_{10}$ transformed.

Figure 6.2: Summary of significant cis-element enrichment by cell-type. Cis-element enrichment was determined by ATHENA [13]. The hypergeometric distribution P-value is $log_{10}$ transformed.

ity of the patterns we are looking for, the number of patterns would be intractably large. Consequently, our method uses fuzzy k-means clustering to generate an initial collection of patterns with strong support (See Section 6.4 and [148] for further details). Heuristic filters were applied to reduce the collection to a smaller set of distinct representative expression patterns, and 51 distinct, dominant radial and 40 dominant longitudinal patterns resulted (Figure 6.4c, Figure 6.5a, and Supplemental Table 4 of [15]). In comparison, in a
previous analysis of five root tissues and three developmental zones, only eight dominant expression patterns were identified using Principle Component Analysis on differentially expressed genes [139]. The increased number of identified patterns clearly demonstrates the value of higher-resolution data sets and of the need for more sophisticated computational methods to elucidate underlying transcriptional programs.

We predicted that the radial patterns we identify using our unsupervised approach should contain the cell-type-enriched probe sets we previously identified in our supervised approach. To assess how well the probe sets assigned to the 51 radial patterns related to the cell-type-enriched genes, we compared the lists for each of the patterns from the two methods. For the majority of cases, there is a clear overlap between one of the cell-type-specific pattern lists and one of the lists from the 51 unsupervised patterns (Figure 6.6), suggesting that the unsupervised approach can identify these cell-type-specific patterns as
Figure 6.4: The majority of enriched GO terms (hierarchically clustered) are associated with individual cell-types (blue bar). a) A smaller number are present across multiple cell types (green bar). b) GO category enrichment for hair cells confirms a previous report [14]. Enriched cis-elements and an enriched TF family were also identified. c) From the top 50% of varying probe sets, 51 dominant radial patterns were identified. Pattern expression values were mean-normalized (rows) and \( \log_2 \) transformed to yield relative expression indices for each marker line (columns). Marker line order is the same for all figures; see Table 5.2 for marker line abbreviations. d) Pattern expression peaks were found across one to five cell-types. e–g) Patterns where expression is enriched in single and multiple cell-types support transcriptional regulation of auxin flux and synthesis. In all heat maps with probe sets, expression values were mean-normalized and \( \log_2 \) transformed. Expression is false-colored in representations of a root transverse section, a cut-away of a root tip, and in a lateral root primordium. e) Auxin biosynthetic genes (\( CYP79B2, CYP79B3, SUPERROOT1 \), and \( SUPERROOT2 \)) are transcriptionally enriched in the QC, lateral root primordia, pericycle, and phloem-pole pericycle (\( P = 1.99 \times 10^{-11} \), pattern 5). All AGI identifiers and TAIR descriptions are found in Supplemental Table 14 of [15]. f) Auxin amido-synthases \( GH3.6 \) and \( GH3.17 \) that play a role in auxin homeostasis show enriched expression in the columella, just below the predicted auxin biosynthetic center of the QC (\( P = 8.82 \times 10^{-4} \), pattern 13). g) The expression of the auxin transporter, \( PIN-FORMED2 \), and auxin transport regulators (\( PINOID, WAG1 \)) are enriched in the columella, hair cells, and cortex (\( P = 1.03 \times 10^{-4} \), pattern 31).

We analyzed these patterns to determine the breadth of expression pattern distributions. By visual inspection, 17 dominant patterns identified from the radial data set show enriched expression in a single cell-type (Figure 6.4d). In some of the 34 patterns where expression was enriched in multiple cell-types, peaks of expression occurred in ontologically or spatially related tissues; in other cases, they were found in cell-types that are spatially separated and have no known ontological relationship (Figure 6.4e–g, and Fig-
Figure 6.5: Forty diverse and distinct patterns were identified from sections along the longitudinal axis. b) Most patterns show an expression increase across two longitudinal sections. c) Two pattern categories were identified: a single expression peak (top) and fluctuating expression in developmental time (bottom). d–f) Co-regulation of probe set expression between roots. Patterns were identified from root 1. Root 2 was treated as the replicate. Probe sets are listed in the same order. Red dashed lines indicate co-regulated groups. Note that there is no columella section in root 2. d) Robust co-regulation of expression for a single expression peak (pattern 25). e) Robust co-regulation of expression for a fluctuating pattern (pattern 19). f) Extensive between-root variation with multiple co-regulated groups (pattern 37). g–j) Validation of longitudinal expression patterns with transcriptional GFP fusions. The expression profiles are able to resolve in vivo expression at high resolution: AGAMOUS-LIKE21 g) expression validates a peak in section 1, whereas WEREWOLF h) shows a peak in section 2; i) At4g05170 (pattern 39) confirms the microarray data with a fluorescence peak in the maturation zone. j) Expression conferred by the At5g60200 promoter confirms a fluctuating expression profile (pattern 37) with two peaks of expression in the meristematic region and the distal elongation region. This image contains portions distal to the sections dissected for microarray analysis. A third peak of expression is found in this distal maturation zone. Details of microscopy and image analysis can be found in [15].
Figure 6.6: Overlap between genes identified as cell-type-enriched and genes present in the 51 dominant expression profiles. This was performed to ensure that the computational method identified genes enriched in individual cell-types as assigned by our first supervised approach.

Among the 40 patterns identified from the longitudinal data set (Figure 6.5a, the majority showed maximal expression across two or more contiguous root sections (Figure 6.5b). This suggests that, in most cases, the sections we sampled are smaller than the region in which expression is being regulated, which, in turn, indicates that we are at the necessary resolution for detecting differential expression along the longitudinal axis. This also demonstrates a greater complexity of transcriptional programs that underlie previously defined root developmental zones. Development is usually described as a progressive uni-
Figure 6.7: Patterns show enrichment of expression in tissues that are ontologically (a) or spatially (b) related, or in cell-types that are spatially separated (c and d). a) Pattern 42 shows enrichment in the xylem tissue in all developmental stages. Genes assigned to these patterns show a strong enrichment for genes involved in microtubule-based processes ($P = 4.03 \times 10^{-6}$). b) Pattern 49 shows enrichment in the more mature xylem and in phloem tissue and a corresponding enrichment in ceramidase activity ($P = 6.79 \times 10^{-15}$) and microtubule binding ($P = 1.5 \times 10^{-4}$). c) Pattern 35 shows enrichment in the culmella and in the mature xylem and an enrichment for genes involved in proteolysis ($P = 7.21 \times 10^{-4}$). These tissues are spatially separated by many cells. d) Pattern 33 shows high enrichment in the hair cells and in developing xylem. Enrichment analysis suggests a shared developmental pathway, presumably related to new cell wall deposition; (COPII vesicle coat ($P = 4.07 \times 10^{-5}$), protein amino acid glycosylation ($P = 6.5 \times 10^{-5}$), cell wall ($P = 1.44 \times 10^{-4}$).

directional process, and it was therefore surprising that 17 of the 40 dominant expression patterns show expression changes that fluctuate over developmental time, presenting multiple peaks of expression (Figure 6.5c). Although oscillatory transcriptional mechanisms have been described [151, 152], until recently they had not been described during root development [153].

To determine the biological relevance of these transcriptional programs, we identified
sets of genes that correlated well with the dominant patterns and assessed these lists for GO or biological process enrichment (Supplemental Tables 5–10 of [15]). Co-expressed genes assigned to three patterns support transcriptional mechanisms in the regulation of auxin flux across single and multiple cell-types. The expression of tryptophan-dependent auxin biosynthetic genes is concordantly enriched in the QC, lateral root primordia, and pericycle [P = 1.99e−11 (Figure 6.4e)]; polar auxin transporters and genes that regulate transporter polarity show enrichment in the columella, root hair cells, and cortex (P = 1.03e−4); and genes that regulate auxin homeostasis are specifically enriched in the columella [P = 8.82e−4 (Figure 6.4f,g)]. In addition to these genes of known function, co-expressed genes of unknown function can be hypothesized to play a role in these processes. In cases where expression is enriched in cell-types that have no known ontological or spatial relationship, the biological significance of a shared transcriptional pathway that links these cell-types can be inferred. An example is found in hair cells and xylem, which undergo cell wall deposition when they undergo terminal differentiation, and in which a corresponding enrichment of genes involved in cell wall deposition was identified [P = 1.44e−4 (Figure 6.7)]. Hair cells deposit cell wall materials during hair morphogenesis when the cell rapidly expands, and xylem cells, during the deposition of secondary cell wall. Identifying the functional importance of the dominant expression patterns that span multiple cell-types should contribute important insights into the transcriptional regulatory networks that regulate Arabidopsis root development.

Analysis of the genes assigned to longitudinal patterns allowed us to infer and confirm biological processes that occur during root development. Starch-containing amyloplasts in columella cells play a role in gravity sensing and a corresponding enrichment of starch catabolism genes was identified here (P = 2.59e−5), whereas genes enriched in meristem determinacy and mitosis-specific expression co-localize with the sections containing the QC and just above, the dividing initials (P = 2.22e−4, 3.79e−5). By contrast, only one of the patterns that showed fluctuating expression in developmental time correlated with a
known biological process—cell cycle activity with expression peaks associated with the root apical meristem and the initiation of presumptive lateral root meristems (Figure 6.5e). For the remaining patterns with fluctuating expression, the majority are enriched in gene functions associated with metabolic or transport activity (See Supplemental Figure 9 and Supplemental Table 9 of [15]).
6.2.3 Root variability

The pattern-finding described above was performed on expression profiles from a single root because of variability in root growth. This variability means that individual sections from each root could not be taken from precisely the same places with the same sizes. Thus, there is no obvious choice of similarity measure to perform a time-series alignment of sections across replicates and assess for variability between root replicates.

To assess the reproducibility of expression across the two different root replicates, we calculated the root mean squared deviation (RMSD) between the expression curves from root 1 (excluding the columella section) and root 2 for all probe sets used in the pattern identification. When the distribution of these true pair RMSDs was compared against the random set generated by calculating the RMSD for each probe set in root 1 (excluding the columella section) to ten random probe sets in root 2, the distributions were clearly different indicating that the expression pattern of a probe set between roots is not random (Figure 6.8).

Next, we asked how robust co-expression is for probe sets assigned to our dominant expression patterns. Methods used to analyze other microarray time series are not suitable because the sections are not identical between roots and the spacing between developmental time points is unknown [154]. We therefore determined the extent to which probe sets assigned to dominant patterns in our first root were co-expressed in the second root (Figure 6.5d–f). The probe sets assigned to each pattern from root 1 were hierarchically clustered with the distance between two probe sets equal to $d = 1 - r$, where $r$ is the Pearson correlation of those two probe sets expression in root 2. The resulting tree was cut into groups at a height of 0.5, corresponding to a Pearson correlation value of 0.5. This value corresponds to the average inter-probe set correlation for probe sets assigned to a pattern from root 1. As a measure of co-expression, we used the size of the largest group selected by the clustering as a percentage of the size of the entire probe set. In the ma-
Figure 6.8: Probe set expression between replicates is reproducible. The distribution of root mean squared differences between true probe set pairs (black) and each probe set and 10 randomly selected probe sets (red). Probe sets analyzed are the top 50% of expressed probe sets. For this analysis we normalized the data using mean=0 and variance=1. The two distributions are separate, demonstrating that the data between replicates is reproducible and not random.

Majority of patterns, greater than 90% of the probe sets maintained coexpression (Figure 6.9 and Table 6.1), which indicated that patterns are quite robust between replicates. In most cases where low coexpression was observed (Table 6.1), subsets of probe sets showed remarkable co-regulation or phasing in several discrete regions along the root. This is reminiscent of oscillatory expression dependent on the phase of a developmental process (Figure 6.5f) [151]. In Chapter 8 we discuss a new computational method for robustly
identifying genes like these, whose expression is regulated on a different time-scale than standard differentiation and development.

Figure 6.9: Probe set co-expression between roots for each pattern. Most probe set members maintain greater than 90% co-expression in a second root. For each pattern, the largest coregulated group was considered.

To ensure that these observations were not due to artifacts introduced through quantitation by microarray, the longitudinal expression profiles were validated by analyzing expression conferred by upstream regulatory regions [140] of representative genes fused to GFP. Expression matched microarray profiles with relative peaks of high expression along the longitudinal axis in all four cases examined as quantified by image analysis (Figure 6.5g–j). A gene that displays potentially oscillatory dynamic behavior was also validated (Figure 6.10), which further demonstrated the reliability of this approach in determining accurate in vivo temporal expression programs.

6.2.4 Integrating the radial and longitudinal datasets

Together, the data sets described above represent the averaged expression within a cell-type population along portions of the longitudinal axis or among multiple cell-types at specific developmental time points. To identify genes with high, relative expression at a
<table>
<thead>
<tr>
<th>Pattern #</th>
<th>Percentage Co-regulated</th>
<th>Pattern #</th>
<th>Percentage Co-regulated</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>63%</td>
<td>21</td>
<td>64%</td>
</tr>
<tr>
<td>2</td>
<td>66%</td>
<td>22</td>
<td>75%</td>
</tr>
<tr>
<td>3</td>
<td>61%</td>
<td>23</td>
<td>99%</td>
</tr>
<tr>
<td>4</td>
<td>80%</td>
<td>24</td>
<td>85%</td>
</tr>
<tr>
<td>5</td>
<td>83%</td>
<td>25</td>
<td>99%</td>
</tr>
<tr>
<td>6</td>
<td>96%</td>
<td>26</td>
<td>67%</td>
</tr>
<tr>
<td>7</td>
<td>96%</td>
<td>27</td>
<td>93%</td>
</tr>
<tr>
<td>8</td>
<td>98%</td>
<td>28</td>
<td>50%</td>
</tr>
<tr>
<td>9</td>
<td>95%</td>
<td>29</td>
<td>54%</td>
</tr>
<tr>
<td>10</td>
<td>97%</td>
<td>30</td>
<td>97%</td>
</tr>
<tr>
<td>11</td>
<td>94%</td>
<td>31</td>
<td>90%</td>
</tr>
<tr>
<td>12</td>
<td>97%</td>
<td>32</td>
<td>100%</td>
</tr>
<tr>
<td>13</td>
<td>93%</td>
<td>33</td>
<td>97%</td>
</tr>
<tr>
<td>14</td>
<td>51%</td>
<td>34*</td>
<td>46.58%</td>
</tr>
<tr>
<td>15*</td>
<td>33.56%</td>
<td>35</td>
<td>95%</td>
</tr>
<tr>
<td>16</td>
<td>92%</td>
<td>36*</td>
<td>45%</td>
</tr>
<tr>
<td>17</td>
<td>70%</td>
<td>37*</td>
<td>37.5%</td>
</tr>
<tr>
<td>18*</td>
<td>21.21%</td>
<td>38</td>
<td>69%</td>
</tr>
<tr>
<td>19</td>
<td>94%</td>
<td>39</td>
<td>100%</td>
</tr>
<tr>
<td>20*</td>
<td>39.93%</td>
<td>40*</td>
<td>28%</td>
</tr>
</tbody>
</table>

Table 6.1: Percentage of genes that are co-regulated similarly in the second root replicate. Each longitudinal pattern and the percentage of genes assigned to it which show co-regulation between root replicates are shown. The ‘*’ indicates a pattern which shows potential phasing.

cell-type/section subregion resolution, we conditionally added the \( \log_2 \)-normalized values of the radial and longitudinal dominant expression patterns (Figure 6.11a,b). This conditional adding constrains the area in which expression can be present. This was based on the following rationale: if a radial pattern shows no or very low expression in a specific cell-type, then the genes associated with that pattern are unlikely to be expressed in a longitudinal section that contains this cell-type. Similarly, if a pattern shows high expression in a specific tissue then expression must be present in at least one longitudinal section which contains that cell-type. A conditional add ensures that both these cases will be true. A standard addition could place expression in a cell-type and longitudinal
Figure 6.10: Expression conferred by the At3g43430 promoter, a zinc finger TF, shows fluctuation along the root longitudinal axis that varies between individual roots. a) Expression values (y axis) of At3g43430 in both roots demonstrate varying fluctuation between both root samples. b–e) Four images of roots display fluctuations of expression that differ between individual roots and that validate the observations made with microarray expression profiling. (Top) Quantification of GFP expression levels along the roots longitudinal axis. Each root was divided into 50 equidistant units and these are represented on the x axis. Along the y axis are GFP intensity units normalized by the total area of each section. (Bottom) The imaged root whose GFP expression was quantified.
cross section where one data set was already specified to have low or no expression, vi-
olating the first case described. The resulting high relative peaks at the intersections of
expression were visualized in a heat map of the root (Figure 6.11b). See Section 6.5 for
further discussion on visualizing root expression. Good correlations were found between
expression from the transcriptional fusions and the predicted spatiotemporal expression
patterns (Figure 6.11b) from this method. In Section 7.3 we develop a more advanced
method of predicting cell-type/section subregion resolution expression levels that is capa-
bile of estimating absolute values, not only locations of relative enrichment. Together, the
data obtained from this high-resolution data set both highlight and revise our view of the
rich and complex network of transcriptional programs that underlie cell-type and temporal
aspects of root development.
FIGURE 6.11: Combining radial and longitudinal expression data to identify gene expression in space and time using dominant expression patterns. a) Radial (x axis) and longitudinal (y axis) patterns are combined by conditional addition to yield relative enriched peaks of expression by cell-type and developmental time. The heat map predicts an enriched relative peak of expression in the cortex and more weakly in the endodermis at the distal region of the meristematic zone and elongation zone. b) A visual representation of the conditional addition in a (left). The At3g05150 transcriptional GFP fusion confirms this expression. See Section 6.5 for details of visualization. Details of microscopy and image analysis can be found in [15].
6.2.5 Transcriptional regulatory modules identification

Expression data associated with our identified patterns has been able to resolve transcriptional programs at high spatial and temporal resolution. We should then be able to use these groups of co-expressed genes to infer transcriptional regulatory modules. Cis-element enrichment within co-expressed gene groups can provide clues as to upstream regulatory transcription factors. We inferred putative network connections by analyzing enriched cis-elements and TFs contained within our identified co-expressed gene groups on the premise that expression co-regulation was due to direct physical interaction of a TF with its target cis-element. Below are three examples of transcriptional regulatory modules that were inferred from the expression data from this simple analysis. More sophisticated algorithms should be able to extend this analysis to identify the complement of transcriptional regulatory networks important for root development.

**MYB promotion of auxin biosynthesis**

In radial transcription pattern 5, which shows coexpression of Trp-dependent auxin biosynthetic genes (Figure 6.4e), the MYB binding site was identified as enriched \((p<1e^{-4})\) using the ATHENA TF binding site enrichment tool [13]. A single MYB domain TF is present in this pattern, *ALTERED TRYPTOPHAN REGULATION1 (ATR1)*. We can therefore infer the following network module for auxin biosynthesis in cell-types previously presumed to be unrelated (Figure 6.12a). Note that among the genes containing MYB binding sites, several are key regulators of Trp-dependent auxin biosynthesis. Furthermore, a loss-of-function allele of ATR1 shows altered Trp-dependent auxin biosynthesis and indole glucosinolate biosynthesis [155].

**Putative auxin response factor-regulated gene expression in the columella**

In radial transcription pattern 13 (Figure 6.4f), columella-specific expression is found for genes involved in auxin homeostasis. Furthermore, two auxin response factors (ARFs)
**Figure 6.12:** Inference of transcriptional regulatory modules. 

**a)** A MYB transcription factor, At5g60890 (ATR1) is present in a set of genes enriched for auxin biosynthesis function (Figure 6.4e) and for MYB-binding sites. All genes containing a MYB-binding site are indicated as downstream of ATR1. All genes annotated as having a function in Trp-dependent auxin biosynthesis are indicated with a blue box and yellow shading. 

**b)** A putative ARF-regulated module in the columella. At1g30330 (ARF6) and At2g28350 (ARF10) are strongly co-expressed in the columella. A set of genes also strongly co-expressed with these two members contain the binding site for ARFs and are indicated as downstream of ARF6 and ARF10. 

**c)** Putative WRKY-regulated transcriptional modules identified in different stages of root hair development from intersections of the radial and longitudinal data sets. WRKY TF binding sites (W-boxes) were enriched in root hairs ($P < 1e^{-10}$). The left panel indicates a subset of this hair cell enrichment in the basal meristematic zone. The WRKY TF At1g68150 is present in this gene group and three potential targets of this WRKY TF were inferred as being downstream as they contain a W-box and are co-expressed with At1g68150. The right panel indicates an additional group of genes which show fluctuation of expression in hair cells, first in the elongation zone and further in the maturation zone. This group of genes contains the WRKY TF At1g68150 and its two potential co-expressed downstream targets.

are found within this set of genes. ARF transcription factors are known to regulate gene expression through binding to the *cis*-element, TGTCTC, and ARFs are known to het-
erodimerize [156]. The expression patterns of these ARFs are further regulated by micro-
RNA mediated mechanisms [157]. All transcriptional interactions between these ARFs
have not yet been identified. Transcriptional co-regulation of ARF gene expression can
provide clues as to specific spatial ARF-regulated modules. ARF10 has been demonstrated
to play a role in lateral root cap development (lateral root cap and columella) in conjunc-
tion with ARF16, and ARF6 plays a role in floral maturation and development [158, 159].
We find strong co-expression of ARF10 and ARF6 in radial pattern 13. Furthermore, we
have identified five putative targets of these ARFs which contain the TGTCTC element to
infer a potential columella-specific auxin-regulated module (Figure 6.12b). Note that one
of these genes encodes SUPERROOT2 (At4g31500), a gene responsible for auxin home-
ostasis. This module could form a feedback loop to regulate auxin levels (auxin regulates
action of ARF genes, which in turn regulates expression of target genes required for main-
taining its homeostasis).

**Putative WRKY-regulated modules in hair cells**

We identified an enrichment of W-box promoter elements (binding sites for WRKY TFs)
and the WRKY TF family in hair cells (p<1e^{-10} and p<1e^{-5}.) We analyzed our groups
of genes resulting from the intersection of the radial and longitudinal data sets to delineate
putative WRKY-regulated modules in hair cells along the roots longitudinal axis. In inter-
section pattern 150 (Figure 6.12c), the WRKY9 TF has a peak of expression in the basal meristematic zone. Three genes with W-boxes are closely correlated with expression of
this WRKY TF—At5g01320, At5g42860 and At5g46230. Further into the root elonga-
tion zone, in intersection pattern 155 (Figure 6.12C), WRKY65 shows a peak of expression,
and again a further peak of expression in the maturation zone. Two different genes that
contain W-boxes (At5g16910, At5g20050) are potential targets of this WKRY due to their
spatiotemporal co-expression.
6.3 Data generated

*A. thaliana* lines in the Columbia (Col-0) (COBL9, S17, S32, S4, SUC2) and C24 (JO121, J2501, RM1000) ecotypes were used for the microarray analysis. Plants to be used for microarray analysis were plated on nylon mesh as in [142]. All plants were grown vertically on 1X Murashige and Skoog salt mixture, 1% sucrose, and 2.3 mM 2-(N-morpholino)ethanesulfonic acid (pH 5.8) in 1% agar. Roots are cut approximately \( \frac{3}{4} \) of the way up, and then treated with protoplasting solution as described previously [139, 160]. The S17, S32 and S4 lines are described in [140]. COBL9, SUC2 and JO121 are described in [143, 144, 145], respectively. J2501 is an enhancer trap from the Jim Haselhoff collection [161] and the microarray CEL files from this line were a gift from Ken Birnbaum. The RM1000 line was a gift from Laurent Nussaume. All tissues sampled for the radial data set were five to six days old. Roots sampled for the longitudinal data set were seven days old and from the Col-0 ecotype. Transcriptional GFP fusion lines (*AT4G23790*, *AT5G14750*, *AT4G05170*, *AT5G60200*, *AT3G05150*, *AT3G29035*, *AT3G43430*, *AT3G15500*) in the Col-0 ecotype are described in [140] and were prepared for microscopy at five days of age. Seeds were sterilized as in [140].

Two to three biological replicates were performed for each marker line in the radial data set. GFP-expressing cells from roots of marker lines were sorted according to [139]. Cells marked in each of these marker lines are described in Table 5.2. The sorted cells were frozen immediately after collection. Samples for the longitudinal data set were dissected in the following steps.

1. The columnella section was taken from the most extreme point of the root tip.
2. Six sections of approximately equal size were dissected from the meristematic zone.
3. Two sections of approximately equal size were dissected from the elongation zone where cells transition from being optically dense to optically transparent as they
begin to elongate [139].

4. Four approximately equal sized sections in the maturation zone were dissected beginning at the point of visible root hair elongation.

5. The final section was taken approximately midway along the root longitudinal axis.

Sections were collected into RNA extraction buffer and immediately frozen. Total RNA was isolated from the frozen material using the Qiagen RNeasy kit (Valencia, California, United States). RNA probes were labeled using the GeneChip™ Eukaryotic Small Sample Target Labeling Assay Version II and hybridized on the Affymetrix ATH1™ GeneChip. The mixed-model software used to globally normalize all arrays and to identify differentially expressed probe sets is described in citeplevesque2006wga. The new data generated by this work is available from the NCBI Gene Expression Omnibus (Accession: GSE8934). The data analysis also included the data of Birnbaum et al. [139] (Accession: GSE5749).

6.4 Dominant differential expression pattern

The informatic task of grouping genes with similar expression patterns is commonly referred to as clustering. Clustering can be more rigorously defined as the task of separating a large set of elements (genes) into distinct subsets (groups/clusters of genes) such that all elements in a subset share a common feature (similar expression pattern). There are a wide variety of clustering algorithms which use different strategies to separate the full set of elements into subsets, with the most common being the hierarchical and K-means algorithms.

There are advantages and disadvantages to using one clustering algorithm over another. The K-means algorithm is useful for finding subsets with many members, but can force elements into subsets where they may not belong simply because all the other subsets are
worse matches. Furthermore, the K-means algorithm begins with randomly chosen cluster centers, and different runs of the algorithm can result in subsets containing different members. Finally, K-means clustering will split the full set of elements into exactly K subsets, which is user-determined and may not be optimal. In contrast, hierarchical methods are good at identifying relationships between single elements or small groups but it is often difficult to determine the proper subsets as the set of elements grows large.

We developed a computational method to find representative distinct dominant differential expression patterns from the radial or longitudinal input data. We define a ‘dominant expression pattern’ to be a pattern that has strong support (i.e. many genes exhibit the pattern) within the data. Thus our method does not try to necessarily try to find all the different expression patterns present in a dataset, only those which represent large groups of co-expressed genes. The method utilizes a variant of the K-means algorithm, Fuzzy K-means [162], to create a preliminary set of groups, from which initial patterns can be defined. Using a K-means based method ensures that the initial set of patterns will be ‘dominant’, since the K-means algorithm will preferentially create subsets with many members. We then use hierarchical clustering, and its strengths in identifying relationships between a small number of elements, to cluster and collapse initial patterns that are similar to each other, resulting in a final set of unique dominant expression patterns. Once these unique dominant expression patterns have been identified, clusters of genes associated with each of the patterns can be assembled and analyzed for biological significance. More details on the computational method can be found in [148] and [15].

6.4.1 Assignment of probe sets to patterns

To assess the potential biological roles for the identified patterns, probe sets were assigned to the patterns. A probe set was assigned to a pattern if the Pearson correlation between the probe sets expression pattern and the identified dominant expression pattern was greater than or equal to 0.85. The high value of 0.85 was chosen so that only strongly correlated
probe sets would be assigned to the patterns. One potential problem in using Pearson correlation is that in cases where probe sets have relatively low or invariant expression, spuriously high correlation values can be obtained. To avoid this problem, we only assigned probe sets which are expressed and varying in the root; the same probe sets used in pattern identification. Probe-set lists were built for each of the 51 radial and 40 longitudinal patterns. Genes were then mapped to these probe sets using the Affy_ATH1_array_elements-20060106.txt naming file and these lists were analyzed for potential biological significance (Table 6.1 and Supplemental Tables 5–10 of [15]).

6.5 Root expression visualization

Having intuitive visualization tools for gene expression data gathered across marker lines and longitudinal sections is an important tool that can be a significant aid in research. Two different representations are shown in Figure 6.11. In Figure 6.11a, expression is visualized in a rectangular heatmap, with each cell of the heatmaps representing estimated expression in the subregion marked by the intersection of the marker line (x-axis) and longitudinal section (y-axis) (see Section 6.2.4 for details of the estimation method). While this representation is straight-forward it does not use all the biological knowledge available. For an example it would allow for expression in the columella in any section, eventhough the columella is only present at the most distal end of the root.

A more biological accurate representation is used in Figure 6.11b. In this representation expression is displayed on a false-colored Arabidopsis root. This root heatmap is was generated from an annotated Arabidopsis root template with 128 distinct cell-type/section subregions and each subregion was colored according its the estimated level of expression. While the estimates used in Figure 6.11b were the same as those used in Figure 6.11a, because the root template is biologically accurate, expression in impossible cell-type/section subregions combinations is never shown. In Section 7.3 we present a
deconvolution method which also integrates this biological knowledge and estimates absolute expression levels specific to each of these 128 distinct cell-type/section subregions.
7

Improving the resolution of expression data

7.1 Introduction

The Affymetrix oligonucleotide expression datasets generated for the *S. cerevisiae* cell cycle (Section 4.2.1) and *A. thaliana* root development (Section 6.3) are some of the highest-quality expression datasets available in their respective areas. Nevertheless, the observed signal intensities in both of these datasets are damped due to convolution from population heterogeneity in their samples. In the case of the *S. cerevisiae* data, this population heterogeneity is a distribution over cell-cycle states caused by synchrony loss. In the *A. thaliana* dataset, the population heterogeneity is a function of the experimental protocol itself. Marker line samples are a heterogeneous mix of specific cell-types across many developmental stages, while the developmental section samples are a heterogeneous mix of all cell-types present at that developmental timepoint.

In this chapter we describe deconvolution algorithms, specific to each dataset, that correct for the heterogeneity in the measure samples and produce new deconvolved data at significantly higher resolution. In Section 7.2 we present a CLOCCS-based deconvolution method (developed in collaboration with Dr. Allister Bernard) that deconvolves the
*S. cerevisiae* cell cycle expression data (Section 4.2.1) and recovers single cell expression profiles. In Section 7.3 we present a deconvolution method (developed in collaboration with Dr. Siobhan Brady and Dustin Cartwright) for the *A. thaliana* expression data described in Section 6.3. This method jointly deconvolves the marker-line and developmental section data to produce cell-type/section subregion level expression profiles.

### 7.2 Improving the resolution *S. cerevisiae* data

Measurements taken of a samples from a population during a synchrony/time-series experiment are convolved due to the distribution of cells over different cell-cycle states within the same. In this section we present a deconvolution algorithm using CLOCCS population distribution estimates (Chapter 3) to correct for this convolution. Previous deconvolution approaches have been developed utilizing discrete predictions of population distributions at each time point [163, 164], but not using a continuous population model like CLOCCS. The deconvolution algorithm described provides a high-resolution single cell view of the cell cycle, with mother cell, daughter cell, and recovery-specific intervals all resolved separately. The algorithm is general and can be used to deconvolve data from nearly any time-series cell cycle measurements. Here, we apply the algorithm to budding index and wild-type mRNA data collected in the experiments described in Section 4.2.1 and published in Orlando et al. [11]. The deconvolution algorithm can simultaneously learn deconvolved profiles from multiple replicates allowing for a more robust deconvolved estimate. We show that the resultant deconvolved profiles have much higher amplitudes and a finer temporal resolution than the convolved measured data. Refinement in the amplitudes of expression make it easier to detect genes regulated at later stages in the cell cycle, as the measured observations at later timepoints are often damped and are hard to distinguish from measurement noise. Increasing the temporal resolution allows for a precise ordering of events across the cell cycle, and is very useful for most downstream
analyzes of transcription factors, transcriptional control, and transcriptional regulatory networks [165, 166]. Finally we show that the deconvolution algorithm is able to accurately identify genes specific to cell-cycle intervals, including the identification of daughter-cell specific transcripts.

### 7.2.1 The deconvolution model

Paralleling the linear lifeline representation used by CLOCCS to model cell-cycle progression, the deconvolution algorithm needs a model capable of capturing the distinct transcriptional events a cell undergoes during the cell-cycle. The deconvolution algorithm model is comprised of a recovery interval (Gr) for initial stress/recovery related processes, a standard cell cycle interval (C) encompassing G1, S, G2/M phases common to both mother and daughter cells, an attached mother and daughter cell interval (A), where mother and daughter cell have completed mitosis and may be operating separate transcriptional programs by have not separated physically, and a daughter specific interval (Gd) for processes specific to the daughter cell after cell separation. A graphical depiction of this model is shown in Figure 7.1.

The major difference between the lifeline models used by CLOCCS and deconvolution is the inclusion of the A interval in the deconvolution model. This new interval is included because the completion of mitosis and cytokinesis is not instantly followed by the complete physical separation of the two new cells. The mother and daughter cells remain physically attached for a period of time after completion of mitosis during which the mother and daughter cells could potentially be operating different transcriptional programs. The interval A is used to model this period of uncertainty by allowing for the presence of, and contribution from, potentially independent mother and daughter transcriptional programs even though the cells are still ‘undivided’ according to the CLOCCS model. This disconnect between the two models is due to CLOCCS being fit with budding index data where a cell ‘divides’ and becomes two new cells at the completion of cell separation when the
Figure 7.1: cloccs and deconvolution lifeline models. A) The cloccs linear lifeline model. B) The deconvolution model for cells progressing through the cell cycle in a synchronized population. After recovery (Gr), cells progress through a common mitotic cell cycle (C) encompassing G1, S, and G2/M. Cells then remain attached to each other after the completion of M with mother and daughter cells potentially operating separate transcriptional programs (A). Finally, after cytokinesis newly created mother cells begin C, while newly created daughter cells go through a daughter specific growth period (Gd) before bud disappears, not at the completion of cytokinesis when independent transcription may begin.

7.2.2 Summary of the mathematical implementation

In this section we summarize the major points of the deconvolution algorithm’s theoretical underpinnings to aid in the readers understanding of the approach. This is not intended to be a complete description, and the interested reader should see Chapter 5 of Bernard [16] for a more detailed and thorough discussion.

Deconvolution of a single dataset

Starting with the parametric estimates generated by CLOCs for an time-series experiment, we calculate a discretized representation of the continuous positional distribution of the population at each point in time. Because of the continuous nature of CLOCs, we can calculate fractional estimates for a discretized representation with many sub-intervals...
within the four cell cycle intervals (Gr, C, Gd, A), greatly increasing the resolution of the deconvolved profile. These fractional population estimates form the convolution kernel $H$, where each column of $H$ is a sub-interval along either Gr, C, A, or Gd, and each row is a time-point.

Now, let $g$ be the observed profile which is the result of convolving the desired deconvolved single-cell profile $f$ with the convolution kernel $H$, such that $g = H \ast f$. Because it is desirable to learn an $f$ with higher resolution (equal to the number of columns of $H$) than $g$ (equal to the number of time-points) the problem of estimating $f$ is often ill-posed. This deconvolution algorithm uses wavelet basis regularization [167, 168] to address this and provide a smooth estimate for $f$ without overfitting.

The objective function we optimize in the deconvolution algorithm is:

$$\min_f \|Hf - g\|^2_2 + \gamma(\|W_1[f_{Gr}, f_C]\|^1_1 + \|W_2[f_{A}, f_{Gd}, f_C]\|^1_1) \text{ s.t. } f \geq 0 \quad (7.1)$$

where $\|\cdot\|^1_1$ and $\|\cdot\|^2_2$ are the $l_1$ and $l_2$ norms, respectively, $\gamma$ is a regularization control parameter, and $W_1, W_2$ and $W_3$ are orthonormal wavelet basis matrices. The regularization constraints $\|W_1[f_{Gr}, f_C]\|^1_1$ and $\|W_2[f_{A}, f_{Gd}, f_C]\|^1_1$ ensure smooth transitions from Gr to C for initial cells and from A to Gd to C for daughter cells. We specifically do not force smooth transitions from C to A to allow for the possibility of unequal division of mRNA at the end of mitosis. Given these constraints, and because equation 7.1 is convex, the deconvolved profile $f$ can be easily recovered using standard methods [169] and is computationally inexpensive, only taking a few minutes on current hardware.

**Joint deconvolution of multiple datasets**

It is common practice when doing high-throughput research to generate biological replicates of any gathered data. The deconvolution algorithm is easy to extend to allow for learning of deconvolved profiles utilizing multiple replicates jointly, increasing the robustness of the estimated profile. Consider the simple two replicate case. To extend the
deconvolution method to two replicates, we construct two convolution kernels \( H_1 \) and \( H_2 \) for the two replicates using their respective CLOCCS parametric estimates as in Section 7.2.2. However we constrain the construction of \( H_1 \) and \( H_2 \) such that columns in \( H_1 \) and \( H_2 \) represents the fractional population estimates for the same sub-intervals of the cell cycle. Given those two kernels we construct the joint convolution kernel \( H_J \), such that \( H_J^T = [H_1^T \, H_2^T] \), where \( T \) is the transpose operator. Similarly, we construct a new joint population-level time-series measurement \( g_J \), such that \( g_J^T = [g_1^T \, g_2^T] \), where \( g_1 \) and \( g_2 \) are the observed population-level time-series measurements in the replicates. We now use the objective function from Equation 7.1, substituting \( H \) and \( g \) with \( H_J \) and \( g_J \) respectively. This two case example can be extended trivially to any number of replicates, allowing the deconvolution algorithm to easily utilize information from multiple replicates to provide a more robust estimate of the deconvolved profile \( f \).

7.2.3 Deconvolving budding index data

As a proof of principle, we deconvolved \textit{S. cerevisiae} budding index measurements from the time-series experiment shown in Figure 4.12a. The CLOCCS parameter estimates are in Table 3.1a. Deconvolving budding index data is an appropriate test because the true single-cell profile for budding is known. As described in Chapter 2, a bud emerges at the G1 to S transition and disappears at the completion of cell separation. The deconvolved profile of budding should therefore be a square wave with transitions at the G1/S and M/G1 cell-cycle transitions. Figure 7.2 shows that the algorithm very accurately recovers this true profile indicating that the algorithm is working properly and is capable of recovering true single-cell profiles from population level data.

Fixing the length of A interval

Before we can deconvolve the mRNA expression data described in Section 4.2.1, we need to determine the length of the attachment interval \( A \), denoted as \( \alpha \). However, because the
Figure 7.2: The effect synchrony loss has on observed population-level budding index measurements. Under perfect synchrony, all cells should start budding at the beginning of S and should become unbudded at the end of M (dashed green line), as opposed to the observed, convolved curve (gray). Applying the deconvolution algorithm to the observed budding index measurements recovers the cell budding index profile nearly perfectly (red line). Figure and caption adapted from [16].

experiments the be deconvolved do not have any morphological data which can be used to determine $\alpha$, we instead rely on the expression patterns of known daughter specific genes. The Daughter Specific Expression or DSE genes ($DSE1$, $DSE2$, $DSE3$, $DSE4$) are known to be specifically expressed only in the daughter cells during the transition from M into the next cell cycle (Gd in our model) [170]. We therefore chose the smallest value of $\alpha$ such that the entire deconvolved expression profile for all four DSE genes was in the A interval. While this constraint is conservative and may lead to over-estimated values for $\alpha$, it ensures we do not incorrectly deconvolve early G1 genes. This approach resulted in $\alpha$’s of 20 and 26 for the two wild-type replicates, respectively.

7.2.4 Results

We applied the joint learning deconvolution algorithm to the wild-type replicate mRNA expression data from Section 4.2.1. Using a PTR-based selection method a set of 1208 cell-cycle regulated genes were selected (see [16] for details). This set shared 704 (55%)
genes in common with the set of 1271 periodic genes identified in Section 4.3.2. Heatmaps of the convolved and deconvolved profiles for the set of 1208 genes are shown in Figure 7.3 and plots for four selected genes are shown in Figure 7.4. It is clear from these figures that the deconvolved profiles have significantly increased amplitudes and that the timing of expression is much more finely resolved. The average PTR of a deconvolved profile is 2.7 and 16.9 times greater than the corresponding observed PTR for genes expressed in the C, or A and Gd intervals respectively (Table 7.1).

**Figure 7.3:** Heatmap of the 1208 cell cycle transcriptionally regulated genes in the high confidence set with PTR ≥ 2.5. The deconvolution clearly provides a more precisely timed picture of transcription across the cell cycle. Figure and caption taken from [16].

To determine the increase in resolution achieved by the deconvolution algorithm we generated a set of *in silico* deconvolved profiles and a set of time-shifted versions of these profiles. We convolved and sampled this set of profiles using the same convolution kernels and sampling intervals used in deconvolving the two wild-type replicates. We add Gaussian noise (mean 0) to these population-level profiles to simulate measurement noise. These noisy profiles were then deconvolved to generate predicted deconvolved profiles and the distance between the peak of the predicted profile from the peak of the known, initial test profile was used a measure of the resolution achieved by the deconvolution algorithm.
Figure 7.4: Deconvolved mRNA expression profiles for 4 genes whose peak expression are in different phases of the cell cycle. Swi4 is a G1 expressed transcript. Fkh1 is an S expressed transcript. Clb2 is a G2/M transcript. Dip5 is predicted to be a daughter specific transcript (Gd). Each of the 4 gene panels shows the observed population-level measurements in replicate (top of each panel), followed by the resulting single cell high resolution profiles obtained by deconvolution. Single cell deconvolved profiles are shown for the initial cell, resulting mother, and daughter cells. Since the common cell cycle interval C is the same for the initial, mother, and daughter cells, this interval is depicted thrice in each panel. In addition, the recovery interval Gr is depicted for an initial cell, the daughter specific interval Gd is depicted for a daughter cell, and the attachment interval A is depicted along with the daughter cell although it may include both mother or daughter specific transcriptional programs. Figure and caption adapted from [16].

The estimated resolution was between 1.6–4.5 min in general, although the resolution decreased to 4.5–7.7 min for genes expressed near the boundaries of the A interval.

In addition to providing much finer temporal resolution, the deconvolution algorithm can also identify candidate genes which have mother or daughter specific transcriptional programs. Genes which have deconvolved profiles peaking in the C or Gd intervals belong
<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>C</th>
<th>A/Gd</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed PTR</td>
<td>2.49</td>
<td>2.32</td>
<td>2.96</td>
</tr>
<tr>
<td>Deconvolved PTR</td>
<td>61.52</td>
<td>7.10</td>
<td>219.84</td>
</tr>
<tr>
<td>Observed:Deconvolved</td>
<td>6.36</td>
<td>2.74</td>
<td>16.90</td>
</tr>
</tbody>
</table>

Table 7.1: PTR values of convolved and deconvolved profiles. The first two rows of the table contain the average PTRs for the population-level and deconvolved profiles respectively. Average values are calculated across all 1208 cell-cycle regulated genes (column 1), only genes peaking in the common cell-cycle interval C (column 2), and only genes peaking in the A or Gd intervals (column 3). The third row is the ratio of the first two rows and is a measure of the fold increase in signal strength between the convolved and deconvolved profiles.

to the mother or daughter programs respectively, while genes peaking in the A interval could belong to either. Previous work aimed at elucidating the daughter specific transcriptional program identified eight genes [170] all of which fall into intervals associated with daughter specific expression. This ability to automatically identify daughter specific genes is unique to this algorithm and could aid in future biological studies.
7.2.5 Conclusion

In this section we described a deconvolution algorithm for use in deconvolving population-level time-series measurements. Deconvolution is achieved using a convex optimization algorithm with wavelet basis regularization. The objective function optimized is convex and therefore has a unique global optimum. Using a CLOCCS based convolution kernel, the algorithm can learn deconvolved profiles with much higher resolution and amplitudes than in the observed data. In addition to providing a more finely timed resolution the deconvolved profiles also provide a spatial resolution with mother cell, daughter cell, and recovery-specific intervals all resolved separately. The algorithm is easily generalized to integrate multiple replicates. While we applied the deconvolution algorithm to budding index and mRNA expression population-level measurements from *S. cerevisiae*, the algorithm is extremely general and could be used to deconvolve many other population-level measurements. By deconvolving population-level data, researchers can begin to understand behaviors at the single cell level.

7.3 Improving the resolution of the *A. thaliana* data

Developmental transcriptional regulatory networks operate in both space and time in plants and animals. Obtaining whole genome expression data at high spatiotemporal resolution is therefore essential to understanding these networks. Experimentally generating genome-wide expression data at the resolution of each cell-type at all developmental stages in the *Arabidopsis* root is not possible due to the technical limitations regarding the amount of material required for microarray measurements. In Section 6.3 we presented datasets defining expression in cell-types (approximated by marker-lines) and developmental sections independently; however, data along these two dimensions was not integrated thoroughly. We present a novel method for reconstructing cell-type/section subregion level expression through the joint deconvolution of the marker-line and developmental data de-
scribed in Section 6.3. This method relies on an iterative algorithm for finding approximate roots to systems of bilinear equations.

### 7.3.1 Previous approaches

There have been two previous approaches to predict cell-type/section subregion expression from large microarray datasets describing expression across multiple cell-types and developmental sections. The method of Birnbaum et al. [139] used data from five marker lines and three developmental sections to predict expression in 15 distinct cell-type/section subregions. In that dataset each predicted cell-type was marked uniquely by a marker line, allowing them to use the marker line expression as a direct proxy for cell-type expression. Given this simplification they estimated the deconvolved expression of gene $g$ in tissue $t$ and developmental section $s$ (denoted $E_{ts}^{g}$) as

\[
E_{ts}^{g} = T_{tg} \cdot \frac{D_{gs}}{\sum_{s} D_{gs}}
\]

where $T_{tg}$ is the expression of gene $g$ in cell-type $t$ and $D_{gs}$ is the expression in developmental section $s$. This method relies on the assumption that each cell-type is marked uniquely and completely by a marker line.

The second method, which we developed for work published in [17], removes this assumption. This method uses the data described in Section 6.3, which profiled 19 semi-overlapping marker lines marking 14 cell-types and 13 developmental sections to predict expression in 108 cell-type/section subregions. Expression of gene $g$ within the cell-type $t$, developmental section $s$ subregion ($G_{ts}$) was estimated by scaling the mean expression from all marker lines covering that cell-type in that section by a factor equal to the expression of that gene in that section relative to its mean expression over all sections. Formally the expression of $G_{ts}$ can be written as:
\[ G_{ts} = \frac{\sum_{M} M_{gm} \cdot C_{mts}}{\sum_{M} C_{mts}} \cdot \frac{D_{g}}{\bar{D}_{g}} \]

where \( M_{gm} \) is the expression of gene \( g \) in marker line \( m \), \( C_{mts} \) is a [1,0] indicator variable defining whether marker \( m \) covers tissue \( t \) in section \( s \) (Figure 7.5), and \( \bar{D}_{g} \) is the mean expression of gene \( g \) over all sections. Expression within any subregion that is not covered by any marker line (e.g. hair cells in sections 1–6) is reported as 0.

Figure 7.5: A graphical depiction of the \( C_{mts} \) function used by the cell-type/section subregion expression prediction method in [17].

While this method is is able to integrate data from multiple overlapping marker lines, approximating cell-type/section subregion expression by this method has some significant restrictions. It assumes expression within a marker line is constant throughout all developmental stages, it is unable to approximate expression in areas not covered by any marker
line, and it cannot mathematically guarantee that the estimated data would reproduce the observed measurements. The method described in the following sections is not subject to such restrictions.
7.3.2 Bi-linear Method

Transcriptional regulation plays an important role in orchestrating a host of biological processes, particularly during development (reviewed in [30, 171]). Advances in microarray and sequencing technologies have allowed biologists to capture genome-wide gene expression data; the output of this transcriptional regulation. This expression data can then be used to identify genes whose expression is correlated with a particular biological process, and to identify transcriptional regulators that coordinate the expression of groups of genes that are important for the same biological process.

The identification of such genes and transcriptional regulators is complicated by the complex heterogeneous mixture of cell-types and developmental stages that comprise each organ of an organism. Expression patterns that are found only in a subset of cell types within an organ will be diluted and may not be detectable in the collection of expression patterns obtained from RNA isolated from samples of an entire organ. Therefore techniques have been developed to enrich samples for specific cell-types or developmental stages, especially for studies in plants [172]. In the model plant, Arabidopsis thaliana, several features of the root organ reduce its developmental complexity and facilitate analysis. Specifically, most root cell-types are found within concentric cylinders moving from the outside of the root to the inside of the root (Figure 7.6). These cell type layers display rotational symmetry thus simplifying the spatial features of development. This feature has been exploited in the development of a cell-type enrichment method. This enrichment method uses green fluorescent protein (GFP)-marked transgenic lines and fluorescently-activated cell sorting (FACS) to collect cell-type enriched samples and has allowed for the identification of cell-type-specific expression patterns [173, 174]. Using this technique, high resolution expression data have been obtained for nearly all cell-types in the Arabidopsis root (herein called the marker-line dataset, see Section 6.3).

Another feature that makes the Arabidopsis root a tractable developmental model is
that cell-types are constrained in files along the root’s longitudinal axis and most of these
cells are produced from a stem cell population found at the apex of the root. This feature
allows a cell’s developmental timeline to be represented by its position along the length of
the root. To obtain a developmental time-series expression dataset individual *Arabidopsis*
roots were sectioned into thirteen pieces, each piece representing a developmental time
point (herein called the longitudinal dataset, see Section 6.3). Each of these sections,
however, contains a mixture of cell-types, and the microarray expression values obtained
are therefore the average of the expression levels over multiple cell-types present at these
specific developmental time points.

While the 19 fluorescently marked lines in described in Section 6.3 cover expression in
nearly all cell-types, they do not comprehensively mark all developmental stages of these
cell-types. Also, the procambium cell-type was not measured, as a fluorescent marker-
line that marks that cell-type did not exist at the time. However, expression from the
longitudinal dataset, does contain averaged expression of all cell-types, and may be used
to infer the missing cell type data.

Previous studies have looked at separating expression data from the heterogeneous cell
populations that make up tumors into the contributions of their constituent cell types [175, 176].
However, in that context, the difficulty comes from the fact that the mixture of
cell-types in each sample is unknown, whereas within our experimental context, the cell-
type mixture of each sample is known. Two computational methods discussed earlier in
this thesis have been developed to combine the *Arabidopsis* longitudinal and marker-line
datasets as experimentally resolving this expression with marker lines is nearly impossible.
However, neither method takes all data into account when reconstructing expression.

In this section we formulate a model for expression levels in *Arabidopsis* roots in
which cell-type and developmental stage are independent sources of variation. The mi-
croarray data specifying overall expression levels for certain mixtures of cells lead to an
overconstrained system of bilinear equations. Moreover, due to the nature of the prob-
lem, we are exclusively interested in positive real solutions. We present a new method for finding non-negative real approximate solutions to bilinear equations, based on the techniques of expectation maximization (EM) [177, Sec. 1.3] and iterative proportional fitting (IPF) [178] from likelihood maximization in statistics. Earlier work has used expectation maximization to find non-negative matrix factorizations [179], and our method is a generalization of that work.

We applied our method to estimate cell-type/section subregion expression patterns for 20,872 Arabidopsis transcripts. These patterns have identified gene expression in cell types and developmental stages which were previously unknown. Overlays of these patterns on a schematic Arabidopsis root are publicly available in a searchable database at http://www.arexdb.org.

Expression data

Our method uses the normalized expression data collected in Section 6.3. Expression levels were measured across 13 longitudinal sections in a single root (longitudinal dataset) and across 19 different markers (marker-line dataset). For simplicity, the J2501 line was removed from further analysis as it is redundant with the WOODEN-LEG marker-line. The APL marker-line was also removed, as it contains domains of expression marked by both the S32 and the SUC2 marker-lines and adds no extra information. The remaining 17 markers covering 14 cell types are listed in the second column of Table 7.3.

Due to computational constraints, the normalization of this data was performed for the longitudinal and the marker-line datasets independently. In order to account for differences caused by these separate normalization procedures, we adjusted the marker-line data by a global factor of 0.92. This factor was calculated by comparing the expression values of ubiquitous, evenly expressed probe sets between the two datasets. We assume that by comparing these probe sets, any true expression differences due to cell-type and longitudinal section specificity should be minimal and thus any differences in expression
A set of 43 probesets were identified which were expressed ubiquitously (above a normalized value of 1.0 in all samples) and whose expression did not vary significantly among samples within a dataset (ratio of min/max expression within a dataset is at most 0.5). The scaling factor necessary to make the mean expression within the marker-line dataset equal to the mean expression level is a byproduct of the separate normalization procedures.
within the longitudinal dataset was calculated for each probe set in this set. The median value of these 43 scaling factors was 0.92, which was used as the global adjustment factor (Table 7.2).

**Model**

To model the transcript expression level of an individual cell we assume that the effects of its cell-type and its section on its expression level are independent of each other. More precisely, we assume that the transcript expression level of a cell of type $j$ in section $i$ is equal to the product $x_i \cdot y_j$, where $x_i$ depends only on the section and $y_j$ depends only on the cell-type. In other words, for each transcript, there is an idealized profile of expression over different cell-types, and an idealized profile of expression over different sections. Within a given section, our assumption is that the transcript expression level varies proportionally to its cell-type profile, and within a given cell-type, proportionally to its longitudinal profile.

Each microarray sample in the two datasets, is composed of a distinct mixture of cell-types and sections. Within each sample, the measured transcript expression level is a convex linear combination of the expression levels of its constituent cells. Under the above assumptions, these measurements constitute a system of bilinear equations,

$$
\sum_{i=1}^{13} \sum_{j=1}^{14} a_{ijk} x_i y_j = b_k \text{ for } k = 1, \ldots, 30
$$

where $x_i$ and $y_j$ are the model parameters for the 13 sections and 14 cell-types respectively, and $b_k$ is the measured expression level as $k$ ranges over the 30 measured samples (13 longitudinal sections and 17 markers).

The coefficients $a_{ijk}$ are obtained by combining the cell type by marker line data (Table 7.3) with the the cell-count matrix (Table 7.4). For $k$ at most 13, the measurement with index $k$ comes from the $k$th longitudinal section. We will use $a_{**k}$ to denote the corresponding matrix, with $a_{ijk}$ in the $i$th row and $j$th column. We set this matrix to be zero
<table>
<thead>
<tr>
<th>Probe</th>
<th>Gene(s)</th>
<th>Longitudinal</th>
<th>Marker-line</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>246630_at</td>
<td>AT1G50730</td>
<td>1.073</td>
<td>1.239</td>
<td>0.866</td>
</tr>
<tr>
<td>246980_at</td>
<td>AT5G67530</td>
<td>1.676</td>
<td>1.605</td>
<td>1.045</td>
</tr>
<tr>
<td>247163_at</td>
<td>AT5G65685</td>
<td>1.423</td>
<td>1.508</td>
<td>0.944</td>
</tr>
<tr>
<td>247286_at</td>
<td>AT5G64280</td>
<td>1.146</td>
<td>1.292</td>
<td>0.887</td>
</tr>
<tr>
<td>247334_at</td>
<td>AT5G63610</td>
<td>1.75</td>
<td>1.497</td>
<td>1.169</td>
</tr>
<tr>
<td>247363_at</td>
<td>AT5G63200</td>
<td>1.225</td>
<td>1.305</td>
<td>0.938</td>
</tr>
<tr>
<td>248266_at</td>
<td>AT5G53440</td>
<td>2.514</td>
<td>1.867</td>
<td>1.347</td>
</tr>
<tr>
<td>250524_at</td>
<td>AT5G08520</td>
<td>1.691</td>
<td>1.54</td>
<td>1.099</td>
</tr>
<tr>
<td>250791_at</td>
<td>AT5G05610</td>
<td>1.151</td>
<td>1.334</td>
<td>0.863</td>
</tr>
<tr>
<td>251104_at</td>
<td>AT5G01720;AT5G01715</td>
<td>1.214</td>
<td>1.257</td>
<td>0.966</td>
</tr>
<tr>
<td>251233_at</td>
<td>AT3G62800</td>
<td>1.171</td>
<td>1.204</td>
<td>0.973</td>
</tr>
<tr>
<td>25157_at</td>
<td>AT3G50430</td>
<td>1.398</td>
<td>1.21</td>
<td>1.155</td>
</tr>
<tr>
<td>253409_at</td>
<td>AT4G32960</td>
<td>1.406</td>
<td>1.572</td>
<td>0.894</td>
</tr>
<tr>
<td>253565_at</td>
<td>AT4G31200</td>
<td>1.461</td>
<td>1.326</td>
<td>1.101</td>
</tr>
<tr>
<td>253826_s_at</td>
<td>AT4G27960;AT5G53300</td>
<td>26.884</td>
<td>26.78</td>
<td>1.004</td>
</tr>
<tr>
<td>255253_at</td>
<td>AT4G50000</td>
<td>1.167</td>
<td>1.696</td>
<td>0.688</td>
</tr>
<tr>
<td>255704_at</td>
<td>AT4G00170</td>
<td>1.35</td>
<td>1.853</td>
<td>0.729</td>
</tr>
<tr>
<td>255725_at</td>
<td>AT1G25540</td>
<td>1.507</td>
<td>1.636</td>
<td>0.921</td>
</tr>
<tr>
<td>255838_at</td>
<td>AT2G33490</td>
<td>1.691</td>
<td>1.473</td>
<td>1.148</td>
</tr>
<tr>
<td>255946_at</td>
<td>AT1G22020</td>
<td>1.335</td>
<td>1.481</td>
<td>0.901</td>
</tr>
<tr>
<td>256236_at</td>
<td>AT3G12350</td>
<td>1.293</td>
<td>1.978</td>
<td>0.654</td>
</tr>
<tr>
<td>256907_at</td>
<td>AT3G24030</td>
<td>1.601</td>
<td>1.751</td>
<td>0.915</td>
</tr>
<tr>
<td>256961_at</td>
<td>AT3G13445</td>
<td>1.383</td>
<td>1.441</td>
<td>0.96</td>
</tr>
<tr>
<td>258269_at</td>
<td>AT3G15690</td>
<td>1.794</td>
<td>1.489</td>
<td>1.205</td>
</tr>
<tr>
<td>259243_at</td>
<td>AT3G07565</td>
<td>1.678</td>
<td>1.768</td>
<td>0.949</td>
</tr>
<tr>
<td>259280_at</td>
<td>AT3G01150</td>
<td>1.302</td>
<td>1.941</td>
<td>0.671</td>
</tr>
<tr>
<td>259313_at</td>
<td>AT3G05090</td>
<td>1.275</td>
<td>1.642</td>
<td>0.776</td>
</tr>
<tr>
<td>259341_at</td>
<td>AT3G03740</td>
<td>1.342</td>
<td>1.606</td>
<td>0.835</td>
</tr>
<tr>
<td>259800_at</td>
<td>AT1G72175</td>
<td>1.159</td>
<td>1.4</td>
<td>0.828</td>
</tr>
<tr>
<td>260133_at</td>
<td>AT1G66340</td>
<td>1.278</td>
<td>1.636</td>
<td>0.781</td>
</tr>
<tr>
<td>261348_at</td>
<td>AT1G79810</td>
<td>1.16</td>
<td>1.464</td>
<td>0.792</td>
</tr>
<tr>
<td>261515_at</td>
<td>AT1G71800</td>
<td>1.236</td>
<td>1.366</td>
<td>0.905</td>
</tr>
<tr>
<td>261634_at</td>
<td>AT1G49970</td>
<td>1.468</td>
<td>1.732</td>
<td>0.847</td>
</tr>
<tr>
<td>261666_at</td>
<td>AT1G18440</td>
<td>1.258</td>
<td>1.297</td>
<td>0.97</td>
</tr>
<tr>
<td>261744_at</td>
<td>AT1G08490</td>
<td>1.265</td>
<td>1.384</td>
<td>0.914</td>
</tr>
<tr>
<td>262089_s_at</td>
<td>AT1G56000;AT1G55980</td>
<td>1.554</td>
<td>1.194</td>
<td>1.301</td>
</tr>
<tr>
<td>262379_at</td>
<td>AT1G73020</td>
<td>1.15</td>
<td>1.398</td>
<td>0.823</td>
</tr>
<tr>
<td>262672_at</td>
<td>AT1G76050</td>
<td>1.306</td>
<td>1.356</td>
<td>0.963</td>
</tr>
<tr>
<td>262860_at</td>
<td>AT1G64810</td>
<td>1.412</td>
<td>1.39</td>
<td>1.016</td>
</tr>
<tr>
<td>263984_at</td>
<td>AT2G42670</td>
<td>1.327</td>
<td>1.314</td>
<td>1.01</td>
</tr>
<tr>
<td>264307_at</td>
<td>AT1G61900</td>
<td>1.273</td>
<td>1.718</td>
<td>0.741</td>
</tr>
<tr>
<td>265129_at</td>
<td>AT1G30970</td>
<td>1.333</td>
<td>1.411</td>
<td>0.945</td>
</tr>
<tr>
<td>267401_at</td>
<td>AT2G26210</td>
<td>1.198</td>
<td>1.476</td>
<td>0.812</td>
</tr>
</tbody>
</table>

**Table 7.2**: Mean expression values and scaling factors of ubiquitously, evenly expressed probesets across longitudinal and marker-lines.
everywhere except the $k$th row, where it is proportional to the $k$th row of the cell-count matrix, but rescaled to sum to 1. For $k$ greater than 13, the measurements come from one of the 17 marker-lines. The matrix $a_{*k}$ is likewise zero except for those cell-type/section subregions marked by that marker as indicated in Table 7.3. Note that the non-zero entries of $a_{*k}$ may span multiple columns for those markers which are listed in multiple rows of Table 7.3. The non-zero entries of $a_{*k}$ are proportional to the corresponding entries of the cell matrix, but rescaled to sum to 1.

<table>
<thead>
<tr>
<th>Cell type</th>
<th>Marker-lines</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quiescent center</td>
<td>AGL42, RM1000, SCR5</td>
</tr>
<tr>
<td>Columella</td>
<td>PET111</td>
</tr>
<tr>
<td>Lateral root cap</td>
<td>LRC</td>
</tr>
<tr>
<td>Hair cell</td>
<td>COBL9 (8-13)</td>
</tr>
<tr>
<td>Non-hair cell</td>
<td>GL2</td>
</tr>
<tr>
<td>Cortex</td>
<td>J0571, CORTEX (7-13)</td>
</tr>
<tr>
<td>Endodermis</td>
<td>J0571, SCR5</td>
</tr>
<tr>
<td>Xylem pole pericycle</td>
<td>WOL (2-9), JO121 (9-13), J2661 (13)</td>
</tr>
<tr>
<td>Phloem pole pericycle</td>
<td>WOL (2-9), S17 (8-13), J2661 (13)</td>
</tr>
<tr>
<td>Phloem</td>
<td>S32, WOL (2-9)</td>
</tr>
<tr>
<td>Phloem companion cells</td>
<td>SUC2 (10-19), WOL (2-9)</td>
</tr>
<tr>
<td>Xylem</td>
<td>S4 (2-7), S18 (8-13), WOL (2-9)</td>
</tr>
<tr>
<td>Lateral root primordia</td>
<td>RM1000</td>
</tr>
<tr>
<td>Pro cambium</td>
<td>WOL (2-9)</td>
</tr>
</tbody>
</table>

Table 7.3: The 14 cell-types in the *Arabidopsis* root and the 17 marker-lines which mark them. For markers that only mark the cell type in some of the sections, these sections are indicated by the range in parenthesis.

**Cell matrix**

As described in the previous section, the coefficients $a_{ijk}$ in our model depend on the number of cells in each cell-type/section subregion. These cell number estimates were generated by visual inspection of successive optical cross-sections of *Arabidopsis* roots along the longitudinal axis using confocal laser scanning microscopy. For the xylem, phloem and procambium cell-types, cell counts were obtained from earlier experiments [180, 181].
What follows is a detailed description of this visual and literature analysis. These results are also summarized in Table 7.4.

Longitudinal section 1 encompasses two tiers of 12 columella cells, and three tiers of lateral root cap cells (15, 18 and 18 moving up from the tip).

Longitudinal section 2 contains one tier of 12 columella cells and six tiers of lateral root cap cells (20, 20, 28, 28, 28 and 28 moving up from the tip). For all other cell-types in longitudinal section 2, three tiers of cells are present. Eight trichoblast (hair cell precursor) cells and 16 atrichoblast (non-hair cell precursor) cells are present circumferentially throughout the root, resulting in 24 and 48 cells respectively in the hair cell and non-hair cell precursor files in longitudinal section 2. Throughout the root, eight cortex and eight endodermis cells are present circumferentially. However in longitudinal section 2, the cortex/endodermis initial is undergoing asymmetric periclinal divisions to produce the cortex and endodermis cell files, so we consider there to be approximately 0.5 cells of the cortex and endodermis type, resulting in 12 cells of each type in longitudinal section 2. When the Arabidopsis root is seven days old, each longitudinal section from 3–13 contains approximately five cells of each type along the root’s longitudinal axis.

In longitudinal section 2, the tangential and periclinal divisions that give rise to phloem cell files do not occur, but do occur in longitudinal section 3 [180]. Three cells are present in the main xylem axis in the first tier of cells, four cells in the second tier, and five cells in the third tier [181]. Eight procambial cells are present in the first cell tier, 12 procambial cells in the second tier, and 18 cells in the third tier resulting in 28 procambial cells in longitudinal section 2 [181]. For all sections xylem pole pericycle cells are the two cells that flank the xylem axis on either end, and phloem pole pericycle cells are considered the intervening cells. Four pericycle cells can be identified as flanking xylem cells in all three tiers of cells present in longitudinal section 2 [181]. Seven intervening phloem pole pericycle cells can be found in tier one, and eight intervening cells can be identified in the third tier [181], resulting in 22 procambial cells in longitudinal section 2.
In a seven day old root, each of the longitudinal sections 3–13 contains approximately five tiers of cells. In longitudinal section 3, columella cells can no longer be identified, and 10 tiers of lateral root cap cells exist containing 28 cells each. In sections 4–6, a lateral root cap cell is twice the length and half the width of an epidermal cell. Eighty-four cells were identified in each tier, and two and a half tiers of cells exist each for longitudinal sections 4–6 resulting in 210 cells for each longitudinal section. All other cell-types have undergone the appropriate tangential and periclinal divisions to establish their respective cell files by longitudinal section 3. Two protophloem cells, two metaphloem cells and four accompanying companion cells are present in the phloem tissue. With the combination of protophloem and metaphloem cells, 20 phloem cells and 20 companion cells exist in each longitudinal section. Approximately 40 procambial cells exist in each longitudinal section. Secondary cell growth does not occur in the developmental stages sampled, therefore, this number remains fixed throughout all developmental stages. In longitudinal section 12, a non-emerged lateral root is hypothesized to be present based on microarray expression data in Section 6.3. This lateral root is estimated to be approximately 130 cells, or one tier of cells in longitudinal section 2.

In our modelling the distinct vasculature, protophloem and metaphloem cell-types were treated as a single cell-type, as no marker-line was specific enough to differentiate clearly between these cell-types. Also, the metaxylem and protoxylem were considered as a single cell-type by the same rationale.

**Solving Bilinear Equations**

In this section we present our method for solving the system of bilinear equations given by equation 7.2. More generally, we have a system

\[ f(x, y) := \sum_{i=1}^{n} \sum_{j=1}^{m} a_{ijk} x_i y_j = b_k \quad \text{for } k = 1, \ldots, \ell. \]  

(7.3)
In our application, we have $n = 13$, $m = 14$, and $\ell = 30$. Unlike other numerical methods for solving systems of polynomial equations, our algorithm has the advantage that it finds only non-negative, real solutions. Moreover, even in systems where there are no exact solutions, as will generally be the case in an overconstrained system of equations, our method will find approximate solutions.

Our method is based on the Expectation-Maximization (EM) [177, Sec. 1.3] and Iterative Proportional Fitting (IPF) [182] algorithms used for maximum likelihood estimation in statistics. These are iterative algorithms which reduce the modified Kullback-Leibler divergence at each step:

\[
D(f(x, y) \| b) = \sum_{k=1}^{\ell} b_k \log \left( \frac{b_k}{f_k(x, y)} \right) - b_k + f_k(x, y) .
\]  

(7.4)

The traditional Kullback-Leibler divergence consists only of the first term in the summation. The other two terms are a natural generalization, which is necessary only when the

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>24</th>
<th>51</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>12</td>
<td>152</td>
<td>24</td>
<td>48</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>22</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>0</td>
<td>28</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>280</td>
<td>40</td>
<td>80</td>
<td>40</td>
<td>40</td>
<td>20</td>
<td>45</td>
<td>20</td>
<td>20</td>
<td>25</td>
<td>0</td>
<td>80</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>210</td>
<td>40</td>
<td>80</td>
<td>40</td>
<td>40</td>
<td>20</td>
<td>45</td>
<td>20</td>
<td>20</td>
<td>25</td>
<td>0</td>
<td>80</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>210</td>
<td>40</td>
<td>80</td>
<td>40</td>
<td>40</td>
<td>20</td>
<td>45</td>
<td>20</td>
<td>20</td>
<td>25</td>
<td>0</td>
<td>80</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>40</td>
<td>80</td>
<td>40</td>
<td>40</td>
<td>20</td>
<td>45</td>
<td>20</td>
<td>20</td>
<td>25</td>
<td>0</td>
<td>80</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>40</td>
<td>80</td>
<td>40</td>
<td>40</td>
<td>20</td>
<td>45</td>
<td>20</td>
<td>20</td>
<td>25</td>
<td>0</td>
<td>80</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>40</td>
<td>80</td>
<td>40</td>
<td>40</td>
<td>20</td>
<td>45</td>
<td>20</td>
<td>20</td>
<td>25</td>
<td>0</td>
<td>80</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>40</td>
<td>80</td>
<td>40</td>
<td>40</td>
<td>20</td>
<td>45</td>
<td>20</td>
<td>20</td>
<td>25</td>
<td>130</td>
<td>80</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>40</td>
<td>80</td>
<td>40</td>
<td>40</td>
<td>20</td>
<td>45</td>
<td>20</td>
<td>20</td>
<td>25</td>
<td>0</td>
<td>80</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>40</td>
<td>80</td>
<td>40</td>
<td>40</td>
<td>20</td>
<td>45</td>
<td>20</td>
<td>20</td>
<td>25</td>
<td>0</td>
<td>80</td>
</tr>
</tbody>
</table>

**Table 7.4:** The cell count matrix gives the number of cells in each cell-type/section subregion. The 13 rows correspond to longitudinal sections 1 through 13. From left to right, the 14 columns correspond to the following cell-type/section subregions: quiescent center, columella, lateral root cap, hair cell, non-hair cell, cortex, endodermis, xylem pole pericycle, phloem pole pericycle, phloem, phloem companion cells, xylem, lateral root primordia, and procambium.
vectors \( f(x, y) \) and \( b \) do not sum to one [179].

Our algorithm begins with an arbitrarily chosen starting point \((x^{(0)}, y^{(0)})\) in \( \mathbb{R}^{m+n}_{>0} \). In each iteration \( s \), the expectation step computes the quantities:

\[
w_{ijk}^{(s)} := b_k \frac{a_{ijk}x_i^{(s)}y_j^{(s)}}{\sum_{i' = 1}^n \sum_{j' = 1}^m a_{i'j'k}x_{i'}^{(s)}y_{j'}^{(s)}}
\]  

(7.5)

for all \( i, j, \) and \( k \). This quantity \( w_{ijk}^{(s)} \) is an estimate of the contribution of the \((i, j)\) term in the \( k \)th equation in the system shown in 7.3. The maximization step is an analogue of the IPF algorithm, and itself consists of an iteration. We first compute the analogues of the sufficient statistics:

\[
X_i^{(s)} = \sum_{j=1}^m \sum_{k=1}^\ell w_{ijk}^{(s)}
\]

\[
Y_j^{(s)} = \sum_{i=1}^n \sum_{k=1}^\ell w_{ijk}^{(s)}.
\]

Then we perform an iteration beginning with \( x_i^{(s,0)} = x_i^{(s)} \) and \( y_j^{(s,0)} = y_j^{(s)} \) and the update rules

\[
x_i^{(s,t+1)} := \frac{x_i^{(s,t)} X_i^{(s)}}{\sum_{j=1}^m \sum_{k=1}^\ell a_{ijk}x_i^{(s,t)}y_j^{(s,t)}}
\]

\[
y_j^{(s,t+1)} := \frac{y_j^{(s,t)} Y_j^{(s)}}{\sum_{i=1}^n \sum_{k=1}^\ell a_{ijk}x_i^{(s,t+1)}y_j^{(s,t)}}
\]

until the parameters converge. We then re-normalize and use the values from the last index \( t \) for the next step of the EM algorithm:

\[
x_i^{(s+1)} := \frac{x_i^{(s,t)}}{\sum_{i'=1}^n x_{i'}^{(s,t)}}
\]

\[
y_j^{(s+1)} := y_j^{(s,t)} \sum_{i'=1}^m x_{i'}^{(s,t)}.
\]
At each step of each of these algorithms, the Kullback-Leibler divergence defined in equation 7.4 decreases. In the statistics literature, the convergence of the EM and IPF algorithms is known under the additional assumptions that

$$\sum_{i=1}^{n} x_i = \sum_{j=1}^{m} y_j = \sum_{k=1}^{\ell} b_k = 1$$

However, relaxing these conditions does not change the convergence proof.

We repeatedly ran our EM algorithm beginning with 20 different randomly chosen starting points. For each transcript in the data, all 20 runs of the algorithm converged to the same solution, strongly suggesting that we have found a global minimum to the modified Kullback-Leibler divergence.

**Computational validation methodology** In order to validate our method, we simulated expression profiles according to various models and tested our method’s ability to reconstruct the underlying parameters. First, we simulated data according to the same independence model defined in Section 7.3.2. The underlying cell-type/section subregion expression levels were sampled from a log-normal distribution with standard deviation 0.5. The simulated measurements $b_k$ were computed from these subregion levels according to our model of the Arabidopsis root in equation 7.2. Finally, multiplicative error was added, distributed according to a log-normal distribution with standard deviation 0.03 to simulate measurement noise. This procedure created expression data with varying but comparable expression levels, which we will call the “uniform” dataset. However, since we are particularly interested in genes for which the expression levels are not uniform, we also produced simulations with the expression level for a given section or cell-type raised by a factor of 10, which we will call the “elevated” dataset. In this dataset, we only measured the error for the same section or cell-type which was elevated. These simulations measure our ability to detect a dominant expression pattern.
In addition, we designed simulations that test the robustness of the algorithm to failures of the bilinear model for root expression levels. For each section and cell-type, we simulated data in which the expression levels for cells in that section or cell-type did not follow the bilinear model, and call these the “section” and “cell type” datasets respectively. Instead, the expression levels in the given section or cell-type were chosen independently according to a log-normal distribution with standard deviation $0.5\sqrt{2}$. The factor of $\sqrt{2}$ was introduced because the product of two log-normally distributed numbers with standard deviation $0.5$ is distributed log-normally with standard deviation $0.5\sqrt{2}$.

The predictions were compared to the true expression levels across the cell-type/section subregions within each section and each cell-type. For each section and each cell-type, the expression levels in its cell-type/section subregions were averaged, ignoring those combinations which are not physically present in the root, (i.e. those whose entry in Table 7.4 is 0). The difference between the predicted and true average expressions was computed as a proportion of the true average expression. We then computed the root mean square of the proportional error over 500 simulations.

**Visualization of predicted expression patterns** Predicted expression values were colored according to an *Arabidopsis* root template (Figure 7.6). The green channel of each cell was set according to a linear mapping between the expression range shown in the template $[1, 10]$ or $[1, 5]$ to the range $[0, 255]$. Expression values above or below that range are given values of 255 or 0 respectively. The mapping is also shown to the right of the false color image in the form of a gradient key. Phloem cells by longitudinal section are visualized separately on the right hand side of the root as they are physically occluded by other cells in the left hand side representation.

**in vivo validation methodology** To validate predicted expression values, we used transgenic *Arabidopsis thaliana* lines containing transcriptional GFP fusions in the Columbia...
ecotype [140]. For each gene being validated, six plants from at least two insertion lines previously described as expressing GFP were characterized. All plants were grown vertically on 1X Murashige and Skoog salt mixture, 1% sucrose and 2.3 mM 2-(N-morpholino) ethanesulfonic acid (pH 5.7) in 1% agar. Seedlings were prepared for microscopy at 5 days of age. Confocal images were obtained using a 25x water-immersion lens on a Zeiss LSM-510 confocal laser-scanning microscope using the 488-nm laser for excitation. Roots were stained with 10 µg/mL propidium iodide for 0.5 to 2 minutes and mounted in water. GFP was rendered in green and propidium iodide in red. Images were saved in TIFF format. Images were manually stitched together in Adobe Photoshop CS2 using the Photomerge command. The black background surrounding the root was modified to ensure uniformity across figures. No other image enhancement was performed.

**Results**

**Computation validation** The root mean square percentage errors in the reconstruction of each parameter are shown in Table 7.5. In the first two columns, where the data were generated according to the bilinear model, the error rate is generally no greater than the simulated measurement error. In most cases, elevated expression led to a lower error rate. In particular, reconstruction of expression in procambium was much more accurate in the elevated dataset.

The last two columns show that the algorithm is robust to violations of the bilinear model. Also, the predicted expression level in each cell-type is generally not greatly affected by the failure of the model in other cell-types, and similarly with sections.

**In vivo validation** To determine whether our algorithm is able to accurately resolve cell-type/section subregion-level transcript expression values, it would be ideal to compare the predictions to measured microarray expression values of the same cell-type/section sub-region. However, due to technical constraints, it is not possible to measure mRNA ex-
<table>
<thead>
<tr>
<th>Variable</th>
<th>Error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>uniform</td>
</tr>
<tr>
<td>Section 1</td>
<td>2.7</td>
</tr>
<tr>
<td>Section 2</td>
<td>3.4</td>
</tr>
<tr>
<td>Section 3</td>
<td>3.3</td>
</tr>
<tr>
<td>Section 4</td>
<td>3.2</td>
</tr>
<tr>
<td>Section 5</td>
<td>3.1</td>
</tr>
<tr>
<td>Section 6</td>
<td>3.3</td>
</tr>
<tr>
<td>Section 7</td>
<td>3.1</td>
</tr>
<tr>
<td>Section 8</td>
<td>3.0</td>
</tr>
<tr>
<td>Section 9</td>
<td>3.0</td>
</tr>
<tr>
<td>Section 10</td>
<td>2.7</td>
</tr>
<tr>
<td>Section 11</td>
<td>2.9</td>
</tr>
<tr>
<td>Section 12</td>
<td>3.3</td>
</tr>
<tr>
<td>Section 13</td>
<td>2.4</td>
</tr>
<tr>
<td>Quiescent center</td>
<td>3.0</td>
</tr>
<tr>
<td>Columella</td>
<td>3.1</td>
</tr>
<tr>
<td>Lateral root cap</td>
<td>2.6</td>
</tr>
<tr>
<td>Hair cell</td>
<td>3.4</td>
</tr>
<tr>
<td>Non-hair cell</td>
<td>3.0</td>
</tr>
<tr>
<td>Cortex</td>
<td>2.9</td>
</tr>
<tr>
<td>Endodermis</td>
<td>2.8</td>
</tr>
<tr>
<td>Xylem pole pericycle</td>
<td>3.3</td>
</tr>
<tr>
<td>Phloem pole pericycle</td>
<td>3.0</td>
</tr>
<tr>
<td>Phloem</td>
<td>3.0</td>
</tr>
<tr>
<td>Phloem ccs</td>
<td>3.3</td>
</tr>
<tr>
<td>Xylem</td>
<td>2.2</td>
</tr>
<tr>
<td>Lateral root primordia</td>
<td>3.5</td>
</tr>
<tr>
<td>Procambium</td>
<td>8.3</td>
</tr>
</tbody>
</table>

Table 7.5: Root mean square percentage error rates in the reconstruction of simulated data. The first column is under a model of comparable but varying expression levels across all sections and cell-types. The second type is the error rate when that section or cell-type has its expression level raised by a factor of 10. The third and fourth columns show models in which the bilinear assumption is violated in one of the sections or one of the cell-types respectively. In all cases, 3% measurement error has been added to the expression levels.

expression to such a degree of specificity and thus we cannot validate the estimates directly. Instead, we validated the method by visually comparing the predicted pattern of expression to patterns obtained from transcriptional GFP fusions using laser scanning confocal microscopy, as described in Lee et al. [140].

For each gene validated, a false-colored root image was generated by coloring each
cell-type/section subregion of an annotated A. thaliana root template (Figure 7.6) according to the expression level in that subregion as predicted by our method. This false-colored image was then visually compared against the actual pattern of fluorescence observed in plants expressing a transcriptional GFP fusion specific for the promoter of that gene. These transcriptional GFP fusions contain up to 3 kb of regulatory sequence upstream of the translational start site of the respective gene. In many cases, this sequence is sufficient to recapitulate endogenous mRNA expression patterns as defined by cell-type resolution microarray data [140]. This comparative method of validation allows us to assess the accuracy of cell-type/section subregion expression predictions in an efficient and technically feasible way.

As a benchmark validation test, a set of three transcriptional fusions which were used to obtain some of the marker-line dataset were examined: S18 (AT5G12870), S4 (AT3G25710), and S32 (AT2G18380). These fusions were originally selected for use in profiling because they exhibited enriched cell-type expression as observed by laser scanning confocal microscopy and subsequently confirmed in the microarray expression data. The expression predictions from our method accurately recapitulated the observed pattern of all three benchmark genes (Figure 7.7 and data not shown).

To assess the novel predictive ability of our method to reconstruct in vivo expression patterns given missing data, we selected transcriptional fusions for genes for which our method predicts expression in cell-types or in cell-type/section subregions that were not marked by fluorescent marker-lines in the original dataset. At least two lines per transcriptional fusion were monitored. With respect to an unmarked cell-type, our method predicted that AT4G37940 was highly expressed in the columella and developing procambium. Imaging of a transcriptional fusion of this gene confirmed this expression (Figure 7.8).

To determine if our method could correctly differentiate expression in a specific developmental stage of a cell-type, we selected AT5G43040 for further analysis.
Figure 7.7: a) Expression of AT2G18380 in all developmental stages of the phloem was predicted by our method and visualized in a representation of the Arabidopsis root. Phloem cells are shown external to the root. b) GFP expression in the longitudinal axis and c) expression in cross-section of expression driven by the AT2G18380 promoter validate the prediction.
FIGURE 7.8: a) Our method correctly predicts specific expression of AT4G37940 in a cell-type, procambium, that is only covered by a general tissue marker, WOL. Expression conferred by the AT4G37940 promoter fused to GFP as a reporter was visualized in the columella and in the procambium by a longitudinal section (b) and a cross section (c). The label X indicates the xylem axis. The expression also validates a maximal peak in the meristematic zone.

of marker-lines used to generate the original dataset included a marker for all developmental stages of non-hair cells, composed of their precursors (atrichoblasts) and fully developed non-hair cells. However, the marker-line used for hair cells only marks mature hair cells, and not their precursors (trichoblasts). Our method predicts AT5G43040 ex-
pression throughout the epidermis—in mature hair cell, trichoblast, mature non-hair cell and atrichoblast cell files—with higher expression predicted in non-hair cells than in hair cells. This differential expression was validated using the $AT5G43040$ transcriptional fusion (Figure 7.9) demonstrating that our method is not only able to identify expression in a developmental stage of a cell-type not marked by the marker-line data, but also to accurately differentiate relative levels of a transcript. However, it should be noted that expression in the transcriptional fusion did not fully corroborate the expression predicted by our algorithm—specifically, expression was found in the lateral root cap which was not predicted by our algorithm.

Examination of the raw microarray expression data revealed that expression was not elevated in the lateral root cap in the input microarray data. Most likely, the presence of GFP is not indicative of erroneous reconstruction of $AT5G43040$ expression in this case. Instead, the transcriptional fusion does not contain sufficient regulatory elements to direct the appropriate expression as described in [140], perhaps within downstream sequences. For this reason, a comparison of the ratio between raw marker line and section expression data can be obtained as a link for each gene so that the user can simultaneously assess raw expression data with the reconstructed expression patterns.

**Discussion**

We have shown that spatiotemporal patterns of gene expression in the *Arabidopsis* root can be reconstructed using information from the marker-line and longitudinal datasets. Current experimental techniques are limited in their ability to rapidly and accurately microdissect organs into all component cell types at all developmental stages. Our computational technique helps to overcome these limitations. We fully integrate the marker-line and longitudinal data sets into a comprehensive expression pattern, across both space and time. In particular, this method has enabled the identification of *Arabidopsis* root procambium and trichoblast-specific genes, which have been previously experimentally intractable cell
FIGURE 7.9: a) Our method correctly predicts expression of *AT5G43040* in trichoblast cells in the meristematic zone, which are not currently covered by any marker-lines. Furthermore, the algorithm was able to predict differential expression within epidermal tissue with high expression in non-hair cells and atrichoblasts (immature non-hair cells in the meristematic and elongation zone) and decreased expression in hair cells or trichoblasts (immature hair cells in the meristematic and elongation zone). Expression conferred by the *AT5G43040* promoter fused GFP as a reporter was visualized in the epidermis in a longitudinal section (b) and was specifically identified as high in atrichoblasts, and lower in trichoblasts (marked with an asterisk) in cross-section (c). Trichoblasts or hair cells differentiate at the junction between two underlying cortical cells.
Our high-resolution expression patterns will allow us to better understand the regulatory logic that controls developmental processes of the *Arabidopsis* root. These transcriptional regulatory networks are key to understanding developmental processes and environmental responses. With only a portion of these genes and fewer cell-types, high-resolution spatiotemporal data has been used to identify transcriptional regulatory modules (Section 6.2.5). Our more accurate and complete dataset will allow a more comprehensive discovery of regulatory networks across additional cell types.

Moreover, we expect that our algorithm and the model which underlies it are applicable to time course experiments on other heterogeneous cell mixtures. Measurements in multicellular organisms are taken from complex cell mixtures of organs, tissues, heterogeneous cell lines, or cancerous samples. When precise histological characterization of these samples can estimate underlying cell type composition, our method can be used to reconstruct the underlying cell-type-specific gene expression patterns or any other type of quantitative data, such as high-throughput protein abundance measurements. Theoretically, this algorithm can be applied to identify missing data in any experimental system that captures data in two or more dimensions which are assumed to be independent of one another.
8

Detecting separate time scales in gene expression data

8.1 Introduction

Biological processes in the living organism occur on a vast range of time scales, $10^{-9}$ s (nanoseconds) to $10^8$ s (decades), many of them taking place simultaneously. The advent of high-throughput technologies has given biologists the ability to measure system wide gene expression, potentially capturing many of these processes at once. As a result, one of the biggest challenges of biological data analysis is the separation of these processes and their time scales. In many cases it is not even known how many processes underlie the measured signal, and what their respective timescales are. These questions are particularly relevant in the field of developmental biology. Developmental studies focus on systems such as mammalian embryos [151] and root development in plants [15], where processes such as growth, segmentation and differentiation can all take place simultaneously, but on different time scales, making the analysis of expression data difficult.

Here we introduce a way of detecting the presence of different timescales in time series gene expression data. Our approach is based on two assumptions, which hold for many, if
not all time series data sets: First, that we have two replicate measurements, and second, that there is at least one time-dependent process for which the timescale is known. This known process allows us to synchronize the replicates, and is most often the timescale on which the data was measured.

In astronomy, the separate timescales on which planets and stars move across the sky are distinguished using a blink comparator. Two photographs of the night sky are taken on different days and the stars aligned so that they occupy the same position on both photographs. The comparator then alternates between the two images. Any object that is not a star, such as a planet, will jump back and forth, because it moves on a different timescale relative to the stellar background (which only moves due to the Earth’s motion). The two astronomical photographs correspond to the biological replicates, the stellar background to the known biological process, and the object which changes position represents another biological process on a different timescale.

Similarly, we propose to detect biological processes on timescales other than the known one (used to synchronize the replicates) by searching for temporal expression patterns which are similar in the two replicates, but occur at different times. In other words, these patterns are shifted (Figure 8.1). To this end, we measure the correlation of expression patterns adjusted for varying shifts. Assessing the statistical significance of this correlation is not straightforward, as the width of the comparison region varies, depending on magnitude of the shift. For each gene we can then determine the shift yielding the most significant correlation, and rank all genes by this significance to find the most prominent shifted patterns.

We apply our approach to detect time scales in two datasets. The first is a *S. cerevisiae* cell-cycle dataset [11] (see Section 4.2.1), and acts as a benchmark. We demonstrate that our method is successful at detecting processes operating on two different time scales, namely real (chronological) time and cell-cycle time in yeast. The second dataset measures gene expression through root developmental time in *A. thaliana* [15] (see Sec-
tion 6.3). Using our approach we discover many classes of statistically significant shifted patterns for the root dataset. These patterns can be further divided into tightly co-expressed spatiotemporal transcriptional modules, some of which are related to processes that occur during branching of the root, termed lateral root initiation. These patterns and modules suggest a rich and complex set of genes that act at multiple time scales to perform a range of biological functions.

![Figure 8.1](image)

**Figure 8.1:** We search for time scale separation in microarray expression data by detecting patterns which show a temporal shift between experimental replicates relative to the synchronization by a given biological process. In this example, Gene A is related to the synchronization mechanism and shows very similar expression patterns in both replicates. The expression level of Gene B on the other hand is also changing in a reproducible way, but on a different time scale, independent of the synchronization process. By calculating the similarity for each possible value of the shift \( s \), and its statistical significance, we can determine whether a given expression pattern is likely to be evidence of a process operating on a separate biological time scale.

### 8.2 Methods

Consider two datasets with \( M \) probesets or genes, and \( N \) datapoints, which we write as \( M \times N \) matrices \( d_{ij}^1 \) and \( d_{ij}^2 \). We convert these series to rank permutations for each row (probeset/gene), resulting in two new matrices \( \pi_{ij}^1 \) and \( \pi_{ij}^2 \). The conversion to permuta-
tions makes it straightforward to measure the significance of the similarity between the replicates in a computationally efficient way.

We then calculate a measure of similarity of two rank profiles for a given shift $s$:

$$
\gamma_i = \sum_{j=1}^{N-s} |\pi_{1ij} - \pi_{2i(j+s)}| \quad \text{for } s \geq 0
$$

$$
\gamma_i = \sum_{j=1}^{N+s} |\pi_{1ij} - \pi_{2i(j-s)}| \quad \text{for } s \leq 0
$$

For uniformly random permutations (which will occur for any i.i.d. randomly distributed data) we expect $\gamma_i$ to follow a Gaussian distribution with mean $\mu$ and standard deviation $\sigma$, given by:

$$
\mu = \frac{N(N - |s|)}{3}
$$

and

$$
\sigma = \frac{N\sqrt{N - |s|}}{3\sqrt{2}}
$$

The p-value (confidence value) for a given $\gamma_i$ is given by the probability that such a value, or an even rarer value, occurs by chance. Hence the p-value, for $\gamma_i < \mu$ is given by:

$$
p(\gamma_i) = \frac{1}{\sigma\sqrt{\pi}} \int_{-\infty}^{\gamma_i} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right) dx
$$

We are only interested in similarity, i.e. in cases where $\gamma_i < \mu$. Hence we can calculate a p-value for every gene $i$ and every shift $s$ under this condition. We can then record, for each gene, the shift at which the maximally significant (smallest) p-value occurs, and the p-value itself. Note that the most significant shift between two replicates does not, in general, correspond to the shift with the lowest value of $\gamma_i$ itself.
8.3 Results

8.3.1 Detection of separate time scales in *S. cerevisiae*

To validate our method we chose to analyze a dataset for which there was a known separation between the timescale of the experiment and the timescale of a biological process of interest. The dataset we chose was a recent synchrony/time–series microarray dataset from the yeast *Saccharomyces cerevisiae* measuring gene expression through the cell cycle (GEO Accession: GSE8799)\(^1\) (Section 4.2.1)

In the synchrony/time–series protocol used by the study, a population of cells are synchronized to early G1 phase. The cells are subsequently released to progress through the cell cycle, during which a time-series of microarray measurements are made. Thus, the measured timescale in the dataset is chronological time (minutes since release from the synchronization). However, as the kinetics of release from synchronization always varies from experiment to experiment \(^3\) (Chapter 3), replicate time-series experiments will not progress through the cell-cycle in exactly the same way. This introduces a separation of timescales between the measured scale, chronological time, and that of a biological process of interest, the cell-cycle.

The dataset itself consists of two replicate synchrony/time–series experiments, each with 15 Affymetrix Yeast 2.0™ microarray measurements taken at 16 min intervals after the first sample. In this dataset the start of sampling in each replicate began at slightly different times (30 min and 38 min). As our method requires directly comparable data, simple linear splines were fit to each replicate and sampled at 8 min intervals starting at 38 minutes after release, with a total of 28 samples per replicate. Any probes which were not mapped to a *S. cerevisiae* gene were excluded, leaving 5742 genes for our analysis.

We analyzed each of the 5742 genes for shifts ranging from $-48$ to $+48$ min at 8 min resolution. We found 3077 genes had a non-zero shift with a p-value at or below 0.001, and 900 genes had a significant shift at zero (Figure 8.2 and Table 8.1).
FIGURE 8.2: Distribution of maximally significant shifts. Our method identified 3974 genes which had significant shifts at or below a p-value of 0.001. Each of those genes was assigned to the shift in which it had the most significant (lowest) p-value.

<table>
<thead>
<tr>
<th>Shift</th>
<th>Not Sig.</th>
<th>-48</th>
<th>-40</th>
<th>-32</th>
<th>-24</th>
<th>-16</th>
<th>-8</th>
<th>0</th>
<th>8</th>
<th>16</th>
<th>24</th>
<th>32</th>
<th>40</th>
<th>48</th>
</tr>
</thead>
<tbody>
<tr>
<td>≤ 0.001</td>
<td>1768</td>
<td>38</td>
<td>91</td>
<td>182</td>
<td>364</td>
<td>689</td>
<td>1689</td>
<td>2970</td>
<td>3351</td>
<td>2705</td>
<td>1489</td>
<td>441</td>
<td>193</td>
<td>59</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Shift</th>
<th>Most Sig.</th>
<th>-48</th>
<th>-40</th>
<th>-32</th>
<th>-24</th>
<th>-16</th>
<th>-8</th>
<th>0</th>
<th>8</th>
<th>16</th>
<th>24</th>
<th>32</th>
<th>40</th>
<th>48</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1768</td>
<td>21</td>
<td>19</td>
<td>14</td>
<td>56</td>
<td>41</td>
<td>214</td>
<td>897</td>
<td>1992</td>
<td>402</td>
<td>237</td>
<td>32</td>
<td>33</td>
<td>16</td>
</tr>
</tbody>
</table>

TABLE 8.1: Distribution of significant shifts in the cell-cycle dataset. The 2nd row gives the number of genes which have a significant shift (p-value ≤ 0.001) at the indicated shift (a gene can have more than one significant shift). The 3rd row gives the number of genes with their maximally significant shift at the indicated shift (a gene can only belong to one shift). Genes in the “Not Significant” category had no shifts with a significant p-value.

We expected to find a large group of shifted genes related to the cell-cycle given the known timescale separation in this data. We therefore tested if any of the gene sets, defined by maximal shift, were related to the cell-cycle. Of the 1274 cell-cycle regulated probes identified in Section 4.3.2 over 60% had a significant shift at +8 minutes (referred to as Shift+8 genes), with a majority of the remaining periodic genes not having a significant p-value at any shift (Figure 8.3). We also performed a GO term enrichment analysis on the each of the gene sets to determine if we could annotate any the underlying processes [1].

159
The GO analysis results (Table 8.2) showed a huge enrichment for terms related to the cell-cycle (mitotic cell cycle: $1.2e^{-16}$, cell cycle: $5.7e^{-16}$, cell division: $4e^{-13}$) in the Shift+8 gene set, with no other gene sets showing any enrichment for cell-cycle related terms. Only two other gene sets, Shift+0 and Shift+16 have any enriched terms. The Shift+0 set is enriched for terms having to do with general biological processes such as growth and RNA processing (ribosome biogenesis: $1.9e^{-32}$, ncRNA processing: $3.7e^{-25}$) which are generally not associated with the cell-cycle. This enrichment in the Shift+0 group is not surprising, as upon release from synchrony cells would be expected activate programs associated with growth and cellular reorganization. Thus these processes would be operating on the chronological timescale of the synchronization. The GO term analysis revealed that the Shift+16 set was slightly enriched for terms related to catabolic process. The biological relevance of this enrichment remains to be tested, but suggests an additional, as yet, uncharacterized timescale.

The identification of biologically coherent sets of shifted genes strongly suggests that our method is able to successfully detect the presence of processes occurring on multiple timescales unrelated. Furthermore, by analyzing the identified genes associated with those shifts, we were able to correctly identify the other major process, associated with Shift+8, as the cell-cycle.
FIGURE 8.3: Distribution of periodic genes identified in Section 4.3.2. Each of the 1274 periodic genes was assigned to its maximally significant shift. Genes which did not have a significant p-value (≤ 0.001) were assigned to the “Not Significant” category.
<table>
<thead>
<tr>
<th>+0 min</th>
<th>+8 min</th>
<th>+16 min</th>
</tr>
</thead>
<tbody>
<tr>
<td>ribosome biogenesis</td>
<td>M phase of mitotic cell cycle</td>
<td>proteolysis involved in cellular protein catabolic process</td>
</tr>
<tr>
<td>ribonucleoprotein complex biogenesis and assembly</td>
<td>mitosis</td>
<td>ubiquitin-dependent protein catabolic process</td>
</tr>
<tr>
<td>ncRNA processing</td>
<td>mitotic cell cycle</td>
<td>modification-dependent protein catabolic process</td>
</tr>
<tr>
<td>rRNA processing</td>
<td>cell cycle</td>
<td></td>
</tr>
<tr>
<td>rRNA metabolic process</td>
<td>cell cycle phase</td>
<td></td>
</tr>
<tr>
<td>ncRNA metabolic process</td>
<td>M phase</td>
<td></td>
</tr>
<tr>
<td>maturation of 5.8S rRNA</td>
<td>cell cycle process</td>
<td></td>
</tr>
<tr>
<td>maturation of 5.8S rRNA from tricistronic rRNA transcript...</td>
<td>cell division</td>
<td></td>
</tr>
<tr>
<td>RNA processing</td>
<td>DNA replication</td>
<td></td>
</tr>
<tr>
<td>ribosomal large subunit biogenesis</td>
<td>chromosome segregation</td>
<td></td>
</tr>
<tr>
<td>organelle organization</td>
<td>sister chromatid segregation</td>
<td></td>
</tr>
<tr>
<td>maturation of SSU-rRNA from tricistronic rRNA transcript...</td>
<td>response to DNA damage stimulus</td>
<td></td>
</tr>
<tr>
<td>endonucleolytic cleavage in ITS1 to separate SSU-rRNA ...</td>
<td>mitotic sister chromatid segregation</td>
<td></td>
</tr>
<tr>
<td>cleavages during rRNA processing</td>
<td>glycoconjugate metabolic process</td>
<td></td>
</tr>
<tr>
<td>endonucleolytic cleavages during rRNA processing</td>
<td>protein amino acid glycosylation</td>
<td></td>
</tr>
<tr>
<td>endonucleolytic cleavage of tricistronic rRNA transcript...</td>
<td>biopolymer glycosylation</td>
<td></td>
</tr>
<tr>
<td>RNA metabolic process</td>
<td>carbohydrate metabolic process</td>
<td></td>
</tr>
<tr>
<td>maturation of LSU-rRNA from tricistronic rRNA transcript...</td>
<td>glycoprotein biosynthetic process</td>
<td></td>
</tr>
<tr>
<td>maturation of LSU-rRNA</td>
<td>cellular response to DNA damage stimulus</td>
<td></td>
</tr>
<tr>
<td>rRNA 3'-end processing</td>
<td>DNA repair</td>
<td></td>
</tr>
<tr>
<td>ncRNA 5'-end processing</td>
<td>microtubule-based process</td>
<td></td>
</tr>
<tr>
<td>RNA 5'-end processing</td>
<td>DNA metabolic process</td>
<td></td>
</tr>
<tr>
<td>cellular component organization</td>
<td>DNA-dependent replication</td>
<td></td>
</tr>
<tr>
<td>endonucleolytic cleavage to generate mature 5'-end ...</td>
<td>response to stimulus</td>
<td></td>
</tr>
<tr>
<td>ribosomal large subunit assembly and maintenance</td>
<td>response to stress</td>
<td></td>
</tr>
<tr>
<td>ribosome assembly</td>
<td>DNA packaging</td>
<td></td>
</tr>
<tr>
<td>endonucleolytic cleavage in 5'-ETS of tricistronic ...</td>
<td>microtubule cytoskeleton organization</td>
<td></td>
</tr>
<tr>
<td>ribosomal subunit assembly</td>
<td></td>
<td></td>
</tr>
<tr>
<td>nucleobase, nucleoside, nucleotide and nucleic acid ...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 8.2:** GO term analysis result. Every set of genes, defined by the position of their maximally significant shift, was analyzed for over-represented GO terms [1]. Each shift (column) and term (row) found to be significant are shown.
8.3.2 Detection of separate time scales in the *A. thaliana* root

The *Arabidopsis* root is an excellent model system for studying development due to its cellular anatomy. In the root, the majority of new cells are born at the root apex from a stem cell population termed the quiescent center (QC). Cell types are constrained within files, and with each new cell division, an older cell is successively displaced distal to the QC. A cell's developmental time line can therefore be tracked along the roots longitudinal axis. In Section 6.3, two developmental microarray timecourses were generated by taking 12 or 13 successive transverse sections along an *Arabidopsis* root. As per the requirements of our method, we use those two time-courses as the replicates (removing the 1st section of the 13 section time-course) and use developmental time as the known timescale.

We analyzed each of the 22746 genes in the root dataset for shifts of $-6 : +6$ sections. We used a maximal shift threshold of $\pm 6$, as larger shifts would result in an overlap window covering less than half the developmental time points. The distribution of maximally significant shifts for the 5992 genes with shifts at or below a p-value of 0.01 is shown in Figure 8.4. As sections sampled in both roots were not exact replicates, we find that the mean value of maximal shifts over all genes is 0.63 representing the approximate temporal difference in sections between both root replicates. Therefore, genes with profiles shifted greater than $+2$ sections and less than $-2$ were considered for further analysis.

While genes that display shifted expression profiles in vertebrates are known to regulate the process of somitogenesis [151], and as shown above, to regulate the cell cycle in yeast, the biological function of genes exhibiting shifted profiles in the *Arabidopsis* root was unclear. To infer their function, we tested for statistically significant enrichment of Gene Ontology (GO) terms, for enrichment of genes involved in secondary cell wall biosynthesis, in the M-phase or S-phase of mitosis, in activation or repression during lateral root initiation, in root hair morphogenesis, and for enrichment in individual root cell types using the ChipEnrich program [15, 148]. This analysis demonstrated that all shift
FIGURE 8.4: Distribution of maximally significant shifts in the root dataset. Our method identified 5592 genes which had significant shifts at or below a p-value of 0.01. Each of those genes was assigned to the shift in which it had the most significant (lowest) p-value.

gene sets, except for the gene set associated with a shift of +5, show enrichment of terms associated with at least one biological process (Figure 8.5a), and some shift gene sets showed very strong enrichment in an individual cell type (Figure 8.5b). Also, gene sets with shifts of +2 and +3 sections show very similar term enrichment suggesting either a conservation of biological function between these two groups of genes, or technical noise. It may be the case that the majority of these genes have a true shift of +2.5 sections which could not be detected based on our method given the resolution in sampling sections.

Strong enrichment of expression in a single cell type suggests that a group of genes associated with a particular shift in developmental time may in fact contain multiple spatiotemporally regulated transcriptional modules. Therefore, for each shift gene set we hierarchically clustered individual profiles according to their cell-type expression [15], and cut these trees at a Pearson correlation threshold of 0.75. These clusters displayed distinct cell type and developmental stage enrichment (Figure 6, Figure 8.7 and data not shown).
**Figure 8.5:** a) Gene ontology (GO) terms associated with shifted profiles. $\log_{10} p$-value is indicated. All patterns except for a shift of +5 show an enriched associated biological process. b) Cell-type enrichment, and enrichment of genes associated with lateral root initiation (LRI or LRIS). Genes activated or repressed within the xylem pole pericycle cells (JO121) during lateral root initiation were also enriched.

Resulting clusters with greater than 10 members were queried for GO term and biological process enrichment as described above.

Many different enriched processes were identified suggesting distinct biological functions of these gene groups. One cluster of phloem-enriched, chloroplast genome-encoded genes is associated with the translation of energy capturing proteins. Two additional clusters show enrichment of genes known to be activated or repressed during lateral root initiation. Lateral root initiation occurs within in pericycle cells located at root xylem poles, and is dependent upon xylem pole architecture, suggesting cell-cell communication between xylem and pericycle cells [183]. However, specific causal factors within the xylem have not yet been identified.

These two clusters may contain such causal xylem factors or additional genes that play an important role in regulating lateral root initiation. The first cluster shows enriched expression in phloem cells, phloem companion cells and phloem pole pericycle cells ($p=1.59e^{-4}$, $p=5.6e^{-5}$, $p=5.54e^{-5}$). These cell types are located in the same tissue as xylem cells. This gene set is shifted +2 sections within the roots maturation zone (where lateral roots are initiated), and is enriched for genes repressed during lateral root initiation.
in xylem pole pericycle cells (p=1.92e⁻⁹) (Figure 8.7). This suggests that during lateral root initiation, genes that are actively repressed in xylem pole pericycle cells must also be repressed within phloem tissue and within phloem pole pericycle cells. The second cluster, also showing a shift of +2 shows enriched expression in xylem cells in the meristematic zone (p=3.2e⁻⁶³), and an extremely strong enrichment for genes activated during lateral root initiation (p=3.1e⁻³⁹). While lateral roots are not initiated in the root meristematic zone, an oscillator positioned in the root meristematic zone has been predicted [153]. These genes show shifted behavior that would correlate well with the existence of an oscillator, and suggest that this oscillator may act within xylem cells.

**Figure 8.6:** Clustering of shift profiles identifies spatiotemporally regulated modules of genes for shifts of +2. Relative expression by marker line is visualized in the left heatmap, and relative expression by longitudinal section in the two roots is visualized in the right heatmaps. The relative expression scale is visualized on the right. If clusters with greater than ten members were identified, these are indicated on the left side of each heatmap.
FIGURE 8.7: For cases where multiple spatiotemporally regulatory modules were identified, statistically significant enrichment expression within cell types was also tested. The p-value scale obtained using the hypergeometric distribution is indicated on the top of each panel. 

a) Shift of +2. Many clusters display distinct cell type enrichment profiles.

b) Shift of +3. At least two significantly cell type-enriched clusters were identified.

c) Shift of -2. Among many clusters, specific cell-type enrichment was identified in only a single cluster.
8.4 Discussion

Given the number of genes in high-throughput datasets, the computational efficiency of any data analysis method is critically important. By converting the data to rank permutations, our method is able to use a continuous Gaussian distribution for $p(\gamma_i)$ as a close approximation to the real (discrete) probability density function of $\gamma_i$ values. Using a Monte Carlo simulation over uniformly random permutations we confirmed that this continuous distribution is an accurate approximation. Since the Gaussian distribution extends below $\gamma_i = 0$, the p-value given by the Gaussian distribution is in fact an upper bound on the true p-value for small (i.e. the most significant) $\gamma_i$, which means that the true p-value lies even lower. Therefore, our method provides an efficient, accurate, and conservative method for determining the significance of shifts in high-throughput datasets.

In our analysis of the yeast cell-cycle dataset, it is not a coincidence that the cell cycle process was identified in the Shift+8 group and that the original study adjusted the sampling times by eight min in the second biological replicate. In the original study, the authors used a model designed to use auxiliary budding index data to specifically analyze kinetics of populations releasing from synchrony/release experiments [36] (Chapter 3). They used this information to determine which of the samples to put onto microarrays. Our method, blind to this prior knowledge, successfully identified this difference as the +8 minute shift. To ensure this agreement was not due to the 8 min interval data splining used, we repeated the analysis on data splined at one minute intervals. This still identified shifts of +8 and +9 minutes as being highly enriched for known periodic genes (data not shown), thus indicating the method is not sensitive to sampling intervals and is successfully detecting the known biological shift.

Numerous biological processes have been identified in plants that occur in multiple time scales ranging from seconds (calcium signaling) to hours (circadian rhythms). In the root however, few processes acting in multiple time scales have been characterized. Two
examples are gene expression programs that fluctuate in developmental time [15], and a putative oscillatory pattern regulating lateral root initiation [153]. Our rigorous method is able to utilize the gene expression dataset measuring expression through root developmental time in Arabidopsis thaliana to identify numerous spatiotemporal transcriptional modules acting in separate time scales. Each of these modules demonstrate a strong conservation of biological association occurring during root development. Interestingly, the strongest observed associations are linked to the process of lateral root initiation. These gene sets suggest a complex regulatory mechanism for this process, occurring in different developmental stages of cells and in different cell types. One in particular supports the existence of an oscillator within the basal root meristem that regulates initiation and suggests that this oscillator is localized within xylem cells.

For all identified, uncharacterized modules in both yeast and Arabidopsis, further studies are needed to determine the relevant time scale of the observed shifts, and the nature of these shifts. Do these shifts act as part of signaling pathways that on the scale of seconds, as part of metabolic rhythms [184], or are they shifting with respect to the circadian clock? Furthermore, are these groups of co-expressed genes oscillatory in nature, or is their observed shift may part of a moving wave of expression that is not oscillatory? How are these temporal shifts generated? Finally, the functional roles of these genes acting at separate scales need to be experimentally elucidated. Regardless, the numerous spatiotemporal modules identified by our method suggest that multiple biological processes acting at distinct time scales are present within the Arabidopsis root and within yeast. Using similar large scale expression datasets, the identification of biological processes acting at multiple time scales in many organisms is now possible.
9

Discussion

9.1 Contributions of this thesis

The focus of this thesis is on integrating high-throughput genomic techniques with novel computational tools to understand the transcriptional regulation of the cell division cycle in *S. cerevisiae* and the development of the *A. thaliana* root. To that end, the main contribution of this work is the generation of two new sets of microarray expression data which capture the transcriptional dynamics involved in both of those fundamental processes and a set of computational tools which are applicable to these datasets. The work described in Chapter 4 generated the first oligonucleotide microarray expression dataset of the *S. cerevisiae* cell cycle [11] (GEO: GSE8799). An analysis of the transcriptional dynamics in the wild-type and *clb1,2,3,4,5,6* strains profiled, revealed the existence of a periodic transcription factor network which could function to regulate the majority of the cell-cycle entrained transcription independent of cyclin/CDK regulation. Complimenting this cell-cycle dataset, work done in collaboration with Dr. Siobhan Brady and detailed in Chapter 6, generated the first high-resolution microarray-based expression map of a single organ at cell-type resolution [15] (GEO: GSE8934). Computational analysis of this de-
velopmental dataset suggested the existence of a complex set of spatiotemporal regulatory patterns which act coordinately to regulate proper *Arabidopsis* root development. Both of these new cell-cycle and developmental datasets provide a rich resource for future studies examining the transcriptional dynamics underlying two fundamental processes.

In addition to the generation of these new datasets, this work also developed a collection of computational tools which were used to increase the informative power of the datasets. Chapter 3 describes the CLOCCS model of population synchrony loss. This model is unique, providing a closed form expression for the likelihood function of a complex branching process, and takes into account asynchrony induced by the synchronization procedure and by asymmetric cell division. This model, extended to utilize both budding index [36] and DNA content data [185], allows for the alignment and deconvolution of experimental data. The alignment functionality of CLOCCS allows researchers to perform experiments under different conditions and measure different types of data but still reliably compare these data in a biologically interpretable manner. This greatly increases the power of the data and allows the researcher to draw more significant conclusions. CLOCCS also allows for the removal of measurement noise caused by the loss of synchrony in the measured population over time. Section 7.2 presents work, done in collaboration with Dr. Allister Bernard, which uses the CLOCCS model to deconvolve the *S. cerevisiae* cell cycle dataset. The deconvolved profiles have much higher resolution and amplitudes than the observed data and provide a spatial resolution with mother cell, daughter cell, and recovery-specific intervals all resolved separately. Although CLOCCS cannot be used to deconvolve the developmental dataset, Section 7.3 presents work, done in collaboration with Dustin Cartwright, which deconvolves the *Arabidopsis* root dataset to provide an easily visualized expression map with expression resolution at the level of specific cell-type/section subregion. Finally, Chapter 8 details work, done in collaboration with Dr. Sebastian Ahnert, which identifies expression patterns that may be occurring on multiple time-scales. Being able to isolate the various time-scales and corresponding bio-

171
logical processes within a dataset may allow for a more precise analysis of each process, revealing novel insights which may have been obscured previously.

9.2 Future directions

Although the datasets and tools described throughout this thesis are instructive in their current forms, each of them has open questions and opportunities for improvement. In this section we briefly describe a few of these possible future directions.

9.2.1 Advances in CLOCCS

Distributional assumptions for $\mu_0$ & $\delta$

The accuracy of the CLOCCS model depends on the accuracy of the distributional assumptions of its parameters. In the formulation defined in Chapter 3, CLOCCS assumes normal distributions for the velocity, $N(1, \sigma_0^2)$, and initial position, $N(\mu_0, \sigma_0)$, distributions. It also models the cell-cycle length difference between mothers and daughters ($\delta$) as a constant. Since the velocity parameter represents the cumulative effect of many stochastic variables, a normal assumption is appropriate. The tendency of CLOCCS budding index predictions to increase with a shallower slope than the observed data (Figure 3.4) suggests that a left skewed distribution with finite support may be more realistic a choice for the initial position distribution. The assumption that the cell cycle delay due to asymmetric cell division ($\delta$) is constant does not reflect all the biological knowledge currently available. There is evidence which suggests that the magnitude of this delay may diminish as the experiment progresses for cells synchronized by centrifugal elutriation. This modelling inconsistency can be addressed with a suitably parametrized cohort-specific delay term, although the duration of a typical time-course experiment may limit the model’s power to detect this effect. Finally, while our model estimates of DNA content data (Figure 3.5) do not contradict a linear accumulation of DNA during S phase, others have suggested alternative parametrizations [79, 186]. Modifying CLOCCS to use a flexibly parametrized S phase
function will allow inference on its functional form and, by doing so, address a question of fundamental interest to the greater biological community.

**New markers**

The inherent separation of the underlying population model (Section 3.2) from the observation model (Section 3.3) which connects the population model to the observed data within CLOCCS makes it flexible and easily extended. In this thesis we developed CLOCCS observation models for both budding index (Section 3.3.1) and DNA content data (Section 3.3.2). These observation models can map temporal events to pre- and post-G1, or on to the G1, S, G2/M phases of the cell cycle respectively (Figure 9.1, green and orange). However, by creating new observation models for other measurable events such as spindle formation and elongation (short vs long spindle) or myosin-ring formation, we can increase this cell-cycle compartment resolution dramatically (Figure 9.1). Work is ongoing to develop new observation models and to create a user-friendly framework to aid in the creation of future observation models.

![Figure 9.1: Relationship between observable markers and cell-cycle position. The colored portion of each band represents the presence of that marker during the cell cycle. The gradient of color in the DNA content band (orange) indicates the ability of the marker to resolve positions in S phase as well as G2/M.](image)

173
**Improved lifeline**

One of the key innovations in the CLOCCS model is the use of the linear lifeline to model cell-cycle progression. The linear lifeline construction allows for the introduction of one-time effects, such as the recovery period following release from synchrony or the delay in cell-cycle progression of new daughter cells. When paired with certain observation models, however, there is a potential weakness. Due to the fact that the observation model maps between lifeline position and the expected value of the observable marker, and because the linear lifeline positions of cells in late Gr and Gd are the same, the current lifeline requires that the observable behavior be the same in both recovering cells and new daughters (Figure 9.2a). This may not always be the case, and creating a more flexible version of the lifeline representation (Figure 9.2b) would alleviate this problem by modeling each of the recovering, mother, and daughter lifelines separately. However, as we increase the flexibility of the lifeline we introduce new parameters, hence requiring more data to obtain accurate fits. There is ongoing work to transition the CLOCCS model to this more flexible life model but still allow models to be fit with less informative data.

### 9.2.2 Advances in *S. cerevisiae* cell cycle

In Chapter 4 we provided the first experimental evidence for maintenance of periodic cell-cycle expression without regulation by B-cyclin/CDKs [11]. Integrating our expression data with available asynchronous ChIP-chip data [35] and other high-quality data available from the literature (Supplemental Table S5 of Orlando et al. [11]), we were able to construct a synchronously updating Boolean transcriptional regulatory network model which oscillated with the proper cell-cycle period (Figure 4.7c). These results suggest that the regulation of periodic expression during the *S. cerevisiae* cell cycle could be, at least partially, the result of a transcriptional regulatory network. One way to validate this predicted network is to perturb the timing of the transcription factors which comprise the network.
Figure 9.2: CLOCCS lifeline representations. a) The current CLOCCS linear lifeline representation which allows for the modeling of one-time events, but requires Gr and Gd marker behaviors (budding, gray areas) to be the same. b) An improved CLOCCS lifeline which allows for more detailed modeling including distinct behaviors for recovering, mother, and daughter cells.

By genetically manipulating the TF’s through knockouts, promoter swaps, and constitutive or conditional expression we can begin to test the validity of predicted networks. Complementing this network validation research, many other aspects of the network are being explored and some of those areas are described below.

**Improved edge information**

The regulatory edges in the transcriptional regulatory network shown in Figure 4.7c were determined, in large part, by an analysis of asynchronous ChIP-chip data [35]. These edges could be significantly refined by performing synchrony/release time-series ChIP-chip studies of the transcription factors. These data would allow us to assign new temporal information to the edges. This temporal information would also allow us to make inference about the repressive/activating nature of the TF’s and provide insight into the combinatorial “logic” functions which regulate each TF. Finally, by performing these studies in
synchronous populations we expect to identify short-lived TF/promoter interactions that would otherwise be missed in studies based on asynchronous populations.

**Transcriptional dynamics in α-factor**

One potential experimental issue which was not addressed by the work presented in Chapter 4 was the presence of functional G1 cyclins in the clb1,2,3,4,5,6 strain. While we are confident that the G1 cyclins are not regulating expression past the G1/S border, given the data presented we cannot rule out their involvement in controlling the overall periodic behavior. To address this issue, we are currently performing synchrony/release time-series experiments with cln1,2,3 cells arrested in α-factor. Yeast cells arrested in α-factor are known to have negligible B-cyclin activity [187]. Thus, these cells are effectively devoid of Cln and Clb/Cdc28 kinase activity. Genome-wide expression measurements in these cln1,2,3 cells during an α-factor arrest could help refine the hypothesis of a cyclin/CDK independent cell-cycle oscillator.

**Coupling to other regulatory mechanisms**

One final area under exploration is how the transcriptional regulatory network oscillator collaborates with CDKs and checkpoint mechanisms to control the ordered events of the cell cycle. Checkpoint mechanisms play a role in maintaining the order of cell cycle events when progression is disrupted. Checkpoints modulate the CDK oscillations and mitotic entry in response to cellular stresses (reviewed in Elledge [188]). Evidence suggests that the transcriptional oscillator is regulated by checkpoint pathways[67]. However, it remains unclear whether checkpoint mechanisms inhibit transcriptional regulatory network oscillations or uncouple them from cell cycle events. Checkpoint pathways have been shown to regulate gene expression[189], and future studies will examine the role of checkpoint mechanisms in controlling the transcriptional oscillator.
9.2.3 Improvements in *A. thaliana*

In Chapter 6 we discussed work examining the regulation of spatiotemporal expression patterns during development in the *A. thaliana* root. After generating a microarray expression map of nearly all the cell-types through developmental time we utilized a fuzzy K-means based clustering algorithm to generate a representative set of “dominant expression patterns” [15]. While this method performed well and produced useful and validated sets of cell-type and developmental time expression patterns, there is room for improvement in the pattern-finding approach. The most obvious improvement would be to search for patterns in the deconvolved, cell-type/section subregion specific expression data (Section 7.3). This would avoid the problem of having to intersect the representative patterns and would better represent the true spatiotemporal patterns of expression. The computational method itself could also be improved. The method currently utilizes a number of heuristic filters and hard cutoffs. More advanced analysis methods of the expression data, which can reduce the reliance on these filters and cutoffs, could produce new insights that were missed in the initial analysis.

Another avenue for improvement is in the prediction of the regulatory interactions in the *Arabidopsis* root. While the global TF binding data necessary for generating the global transcription-factor/gene networks like those in *S. cerevisiae*, is not currently available in *A. thaliana*, progress can still be made in predicting the regulatory elements of individual genes. One potential approach is to combine the spatiotemporal expression data with sequence data to try and determine the *cis*-regulatory code regulating expression. In this approach, we model the expression of a gene strictly as a function of the *cis*-elements in its promoter. We assume that different combinations of *cis*-elements will produce specific expression patterns. The task is to determine what the *cis*-elements are and which expression patterns they regulate. A simple example of this approach is shown in Figure 9.3. In this example, there are four *cis*-elements: ATTTA, ACCTA, GACGT, and CCTGG that act
in combination to regulate expression in hair and cortex cells in the elongation and maturation zones. The goal would be to identify these four elements and relate them to their regulatory activities. Having a compendium of these regulatory interactions would be an invaluable aid for understanding how spatiotemporal expression is regulated at a sequence level. Moreover it would give insight into the combinatorial nature of transcriptional regulation underlying *A. thaliana* root development.

### 9.3 Conclusion

This thesis has expanded our knowledge regarding the transcriptional regulation of two fundamental biological processes: development and the cell division cycle. We have gen-
erated expression datasets describing the complete transcriptional dynamics which occur
during these processes, providing evidence that these dynamics may be regulated by tran-
scriptional regulatory networks. We hope that this work can contribute towards future
efforts that will provide a more complete understanding of how these fundamental pro-
cesses of development and cell division are regulated, and to what extent transcriptional
regulatory networks are responsible for that regulation. In conclusion; While the amount
we do understand is far outweighed by what we do not; to quote Galileo Galilei, “All truths
are easy to understand once they are discovered; the point is to discover them”.
Appendix

Many of these definitions were taken from [19, 190].

• **Arabidopsis thaliana** Small flowering weed related to mustard. Model organism for flowering plants and the primary model for studies of plant molecular genetics and root development.

• **ATH1 22k™** An Affymetrix microarray used to study *A. thaliana*.

• **branching process** Markov process that models a population in which each individual in generation $n$ produces some number of individuals in generation $n + 1$, according to a fixed probability distribution that does not vary from individual to individual.

• **budding index** The percentage of *S. cerevisiae* cells in a population which have a bud.

• **CDK** Protein kinase that has to be complexed with a cyclin protein in order to act. Different CDK-cyclin complexes trigger different steps in the cell-division cycle by phosphorylating specific target proteins.

• **cell cycle (cell-division cycle)** Reproductive cycle of a cell: the orderly sequence of events by which a cell duplicates its chromosomes and, usually, the other cell contents, and divides into two.

• **cell-type** One of fifteen distinct classes of cells in the *A. thaliana* root.
• **cell-type/section subregion** A particular cell-type at a particular developmental section.

• **centrifugal elutriation** A size-based synchronization method which takes advantage of the fact that the smallest cells in a culture will be newborn daughters in G1-phase.

• **cis-element** A DNA sequence that is specifically recognized by a transcription factor on the same side of the DNA as the gene being regulated.

• **CLOCCS** Characterize **Loss Of Cell Cycle Synchrony**. A branching process model which describes the dynamics of synchrony loss within a population during a synchrony/time–series experiment (See below for definitions of CLOCCS parameters).

• **CLOCCS lifeline** The linear representation of repeated cell cycles used by the CLOCCS model.

• **convolution** A convolution is an integral that expresses the amount of overlap of one function $g$ as it is shifted over another function $f$. It therefore "blends" one function with another.

• **cyclin** Protein that periodically rises and falls in concentration in step with the eukaryotic cell cycle. Cyclins activate crucial protein kinases (called CDK) and thereby contribute to the regulation of progression from one stage of the cell cycle to the next.

• **cyclin/CDK complex** Protein complex formed periodically during the eukaryotic cell cycle as the level of a particular cyclin increases. A cyclin-dependent kinase (CDK) then becomes partially activated.
• **cyclin mutant** Refers to a mutant *S. cerevisiae* strain deleted for all B-cyclins, *clb1,2,3,4,5,6*.

• **deconvolution** The inversion of a convolution equation. The common goal is to remove the effect of the convolving function $g$ from the function of interest $f$.

• **developmental section** One of the thirteen transverse sections taken along the *A. thaliana* root (Section 6.3).

• **developmental stage** One of three distinct morphologically distinct longitudinal zones in the *A. thaliana* root: meristmatic, elongation, and elongation.

• **DNA microarray** A large array of short DNA molecules (each of known sequence) bound to a glass microscope slide or other suitable support. Used to monitor expression of thousands of genes simultaneously: mRNA isolated from test cells is converted to cDNA, which in turn is hybridized to the microarray.

• **dominant expression profile** One of a set of computationally derived patterns of expression either across cell-types or developmental sections which is shared by many genes (dominant) and is distinct from all others in the set.

• **elutriation** See centrifugal elutriation.

• **emergent property** A characteristic of a system that derives from the interaction of its parts and is not observable or inherent in the parts considered separately.

• **expression level** Measure of the abundance of a particular transcript in a sample as measured by a microarray.

• **expression map** A complete description of expression for each gene across all regions covered by the map. In the case of the *A. thaliana* root the regions are all cell-types and longitudinal sections.
• **expression profile**  The expression level for a single gene or probe across many experimental conditions.

• **flow cytometry**  Technique for counting, examining, and sorting microscopic particles suspended in a stream of fluid. It allows simultaneous multiparametric analysis of the physical and/or chemical characteristics of single cells flowing through an optical and/or electronic detection apparatus.

• **fluorescence-activated cell sorter (FACS)**  Machine that sorts cells according to their fluorescence.

• **gene**  Region of DNA that is transcribed as a single unit and carries information for a discrete hereditary characteristic, usually corresponding to (1) a single protein or (2) a single RNA.

• **gene expression**  Production of an observable molecular product (RNA or protein) by a gene.

• **GO (gene ontology) term**  A biological concept in one of three domains: cellular component, molecular function or biological process. GO terms consist of a GO ID, in the form GO:nnnnnnn, and a term string. The majority of GO terms have a textual definition, and many have synonyms, database cross-references and/or a comment [191].

• **hybridization**  In molecular biology, the processes whereby two complementary nucleic acid strands form a base-paried duplex. Forms the basis of a powerful technique for detecting specific nucleotide sequences.

• **longitudinal dataset**  The expression of each gene in *A. thaliana* across the 13 longitudinal sections (see Section 6.3).

• **longitudinal section**  See developmental section.
• **marker line** A genetically modified *A. thaliana* plant, in which a particular gene is *marked* with GFP such that GFP is expressed in the same anatomical domain as that gene.

• **maximally significant shift** Used in Chapter 8. The shift between two replicate expression profiles at which the alignment between the two profiles is the most improved as compared to the expected alignment of two profiles shifted similarly.

• **MBF** Transcription factor protein complex composed of Mbp1 and Swi6.

• **microarray** See DNA microarray.

• **model system** A well studied system which serves as a “model” for deriving fundamental biological principles.

• **mRNA** RNA molecule that specifies the amino acid sequence of a protein. Produced in eukaryote by processing of an RNA molecule made by RNA polymerase as a complementary copy of DNA. It is translated into protein in a process catalyzed by ribosomes.

• **organogenesis** The formation and development of the organs of an organism from embryonic cells.

• **peak-to-trough ratio (PTR)** A ratio between the peak (maximum) and the trough (minimum) of an expression profile.

• **periodic gene** A gene whose expression level varies according to cell cycle position.

• **Probe** Defined fragment of RNA or DNA (or set of identical fragments), radioactively or chemically labeled, used to quantify abundance of specific nucleic acid sequences by hybridization.
• **radial dataset**  The expression of each gene in *A. thaliana* across the 19 marker lines (see Section 6.3).

• **rank permutation**  Converting a vector of real numbers to rank. Where the lowest value is replaced with 1, the next lowest 2, ..., etc.

• **root image**  A false colored template image of the *A. thaliana* root used to visualize expression of a gene.

• **Saccharomyces cerevisiae**  Widely used simple model system in the study of eukaryotic cell biology.

• **Schizosaccharomyces pombe**  Rod-shape yeast which reproduces by binary fission.

• **SBF**  Transcription factor protein complex composed of Swi4 and Swi6.

• **SFF**  Transcription factor protein complex composed of Fkh1, Fkh2, and Ndd1.

• **spatiotemporal expression**  Expression of a gene in both space and time, or across cell-types and longitudinal sections.

• **system**  A collection of interdependent components forming a unified whole. Most frequently used in this work to describe the entire collection of processes/activities required to complete a biological process.

• **TF**  See transcription factor

• **tissue**  A is a cellular organizational level intermediate between cell-type and a complete organism.

• **transcript**  RNA product of DNA transcription.

• **transcription factor**  General name for any protein that binds to a specific DNA sequence to influence the transcription of a gene.
• **transcriptional regulatory network**  See transcriptional regulatory network.

• **transcriptional regulatory module**  Small transcriptional regulatory network and its targets. Often consisting of only one or two transcription factors and their targets.

• **transcriptional regulatory network**  A collection of inter-regulated transcription factors which coordinately act to regulate a biological process.

• **WTPER**  The set of periodic genes identified in wild type *S. cerevisiae* cells in [11].

• **Yeast 2.0™**  An Affymetrix microarray used to study *S. cerevisiae* and *S. pombe*.

**CLOCCS glossary**

• $\mu_0$  The length of the recovery interval in min.

• $\sigma_0^2$  The variance of the initial position distribution.

• $\delta$  The length of the daughter specific growth period in min.

• $\lambda$  The length of the cell cycle in min.

• $\beta$  The percentage through a normal cell cycle when a bud emerges.

• $\gamma_1$  The percentage through a normal cell cycle when a DNA replication begins.

• $\gamma_2$  The length of S-phase, as a percentage of the normal cell cycle.

• $\sigma_v^2$  The variance of cell cycle length distribution.

• **cohort**  $\{g,r\}$  A subpopulation of generation $g$ and reproductive instance $r$, within which there is no asynchrony due to reproduction.

• **CLOCCS alignment**  An alignment between multiple synchrony/time-series experiments based on CLOCCS model fits to those experiments. The experiments are compared on the basis of their cell cycle position as opposed to chronological time.
Bibliography


[40] SGD project. “Saccharomyces Genome Database”
http://www.yeastgenome.org(cache/genomeSnapshot.html


[132] NCBI: Plant Genomes Central


198


[191] Gene Ontology Consortium Wiki
Biography

David A. Orlando was born on June 3rd 1981, in Falls Church, Virginia. He received dual Bachelor of Arts degrees in Computer Science and Biology from the University of Rochester in May 2003. In September of the same year, he enrolled as a member of the inaugural class of Duke University’s Computational Biology and Bioinformatics doctoral program (formerly Bioinformatics & Genome Technology). Under the supervision of his three mentors, Dr. Philip Benfey (Biology), Dr. Steven Haase (Biology), and Dr. Alexander Hartemink (Computer Science) he has led two major research efforts in the area of Systems Biology. His work was focused on using high-throughput genomic technologies to study the role of transcriptional regulatory networks in controlling the gene expression programs which underlie the cell cycle of *S. cerevisiae* and the development of the *A. thaliana* root. He also developed the CLOCCS model, describing the dynamic loss of synchrony in dividing populations. His research has resulted in the scientific publications listed below:


