ABSTRACT


Current advancements in biological experimentation allow for the collection of whole genome activity data in an organism over time. The resulting datasets contain information that can help to better understand the molecular machinery of a living cell and provide insights into possible ways of controlling cell and organism behavior. However, state of the art computational approaches for processing gene activity or transcriptome datasets is still evolving in its ability to fully harvest the information embedded in the data. Time course transcriptome datasets have high dimensionality in terms of the number of genes in the genome, substantial sample variance because of the measurement noise and biological variation, and a small number of samples because of high experimental costs, which limits most of the temporal behavior studies with tests for significant difference between time points. This leads to a need for more advanced computational techniques for temporal transcriptome data processing that could unravel the mechanisms governing molecular behavior on gene level.

The research presented in this dissertation focuses on computational approaches for transcriptome data processing, which are used to better understand the molecular mechanisms underlying iron deficiency response in model plant Arabidopsis thaliana. The work outlined here shows the steps taken for refining our knowledge about these mechanisms, from the identification of differentially expressed genes to a dynamic model of gene regulation. First, collaborators and I developed the Clustering and Differential Alignment Algorithm (CDAA) to identify regulators for known iron deficiency mediating genes based on the existing whole genome time course iron deficiency dataset with 2,754 differentially expressed genes. The algorithm takes advantage of fundamental electrical engineering concepts such as digital signal processing, unsupervised learning, and pattern recognition to group genes in activity stages and perform the inference. We then formulated a computational framework for transitioning from a collection of regulatory connections to a system of ordinary differential equations that were capable of describing gene interaction dynamics. We designed mRNA decay and knock-out mutant experiments to estimate the resulting model parameters that were not identifiable based on time course samples alone. We collected the corresponding experimental data, applied a Bayesian approach for aggregating datasets for model training, and estimated uncertainty ranges for model predictions while accounting for uneven sample variance (i.e. heteroscedasticity). Finally, we tested model predictions on the results of double mutant experiments that were not used for model training.

The presented methodologies unraveled information about various aspects of iron deficiency
response in *A. thaliana* on transcriptome level. CDAA narrowed down the search space for genes affecting known iron deficiency mediators, namely PYE, BTS, MYB10, MYB72, bHLH39, bHLH101, and bHLH115, to 7 potential candidates that were not previously linked to iron deficiency response. The following experimental validation showed that 4 of these candidates, namely ETF9, COL4, ASIL2, and MYB55, significantly affect activity levels of the known iron deficiency mediating genes. We formulated the dynamic model based on these connections and the ones from bHLH34, bHLH104, and ILR3 that were found in the existing literature to regulate the known iron deficiency mediators. The developed dynamic model for gene regulations revealed mRNA decay rate dependence of the known iron deficiency mediators on iron condition. This dependence was not statistically significant for their regulators. Model simulations predicted gene activity behavior patterns for currently unknown gene regulators of iron deficiency response. Model validation through the double mutant experiments unveiled potential synergistic effects between bHLH34 and bHLH104 when regulating BTS and between bHLH104 and ILR3 when regulating bHLH115. Overall, the presented computational methodologies were able to extend the knowledge about gene regulatory mechanisms involved in iron deficiency response in *A. thaliana* and have a potential to provide additional insights when applied to studies with different stress conditions or even in other species.
DEDICATION

I dedicate my dissertation to my parents, Sergei Koriachko and Tatiana Koriachko. Your support, help, and guidance contributed enormously to make this happen.
BIOGRAPHY

Alexandr (Sergeyevich) Koryachko was born in a small town Emva of Komi Republic on the north-west of USSR to Sergei Koriachko and Tatiana Koriachko. The family moved to Moscow when Alexandr was 11. At the age of 14, Alexandr was accepted to a school with a concentration on sciences and foreign languages. Upon successful completion of the high school, Alexandr was accepted to the Moscow Institute of Electronic Technology and chose a computer science major. Top 10% standing in the university rankings allowed Alexandr to pick a concentration of his interest which at that time was information security. Alexandr received a Bachelors of science in computer science degree with honors in 2007.

Upon completion of his B.S., Alexandr entered into the Master of Science in Computer Science program at the same university and started working at the Research Institute of Computing and Control Systems where he was developing electronic devices for various applications. Alexandr defended masters thesis on a speed measurement device for fast moving objects in 2009. While pursuing Masters degree, Alexandr was involved in teaching metrology for 5 years and electric circuits for 2 years. Alexandr developed an interactive teaching aid for the electric circuits class. The teaching aid was awarded the 2nd place at an international conference in Moscow in 2009.

In 2011 Alexandr decided to pursue an academic career and applied for a graduate program at a number of universities. Alexandr was accepted by the Department of Electrical and Computer Engineering at North Carolina State University and started his Ph.D. program in 2012. In the same year, he joined EnBiSys lab under Dr. Cranos Williams and got involved into the research associated with large scale genomics data analysis. Alexandr focused on computational frameworks allowing to extract regulatory information from transcriptome datasets and model the dynamics of gene activity in a plant. During his Ph.D. program, Alexandr presented at several conferences, was the leading author of two journal publications, gave a guest lecture on Systems Biology Modeling.
ACKNOWLEDGEMENTS

I would like to give thanks and express my deepest appreciation to my advisor, Dr. Cranos Williams, who believed in me from the very beginning, or at least showed as this was the case, spent long hours giving academic and life advises for survival at the graduate school, resolved my research, program, financial, and other issues, provided useful feedback and insights into problems I was stuck on. Your patience, dedication, wisdom, constructive criticism, unconditional optimism, and ability to give a different perspective on my ideas were a significant aid in gaining an adequate training and carrying through the 6 years of the graduate school. I would also like to thank Dr. Terri Long for biological insights and consultations, Dr. Ilse Ipsen for enforcing programming culture and LaTeX usage on me, Dr. Edgar Lobaton, Dr. Hamid Krim, Dr. Joel Ducoste, and Dr. James Tuck for their assistance throughout my research.

Thank you to Anna Matthiadis for educating me on the experimental part of the project, getting into the weeds of computational techniques to resolve occasional interdisciplinary misunderstandings, and improving my writing style. I want to thank both Anna Matthiadis and Durreshahwar Muhammad for running endless experiments to test the generated hypothesis and produce enough data for model training. I thank everyone at the EnBiSys lab for sharing struggles, successes, desk, and printer with me, for giving useful feedback on my lab presentations, for sharing research insights.

Special thanks to my spouse Yelena Sharonina for her patience with the “scientific one” in the family, vast emotional support, encouragement, and understanding. I thank my close friends Chuyue Wang, Leon Dmello, Brittany Oxendine, Sara Elouri, and other members of “The Cool Gang” for pulling me occasionally out of the PhD related stress for a dinner, conversation, trip, or birthday celebration.

I want to acknowledge the North Carolina State University for the high level education delivered to me by the professional and inspiring teachers and especially the Department of Electrical and Computer Engineering that gave me the opportunity to use the gained knowledge to advance in my research. I also acknowledge National Science Foundation for funding the research project I worked on.
# TABLE OF CONTENTS

LIST OF TABLES .................................................. viii

LIST OF FIGURES .................................................. ix

Chapter 1  Introduction ............................................. 1
  1.1 Overview .................................................. 1
  1.2 Motivation ................................................ 2
  1.3 Contributions ............................................. 4
    1.3.1 Intellectual merit .................................. 4
    1.3.2 Broader impacts .................................... 5
  1.4 Organization ............................................. 6

Chapter 2  Computational techniques for transcriptome data analysis in plants ............. 8
  2.1 Classification of inference algorithms ................................ 10
  2.2 Computational approaches ..................................... 16
    2.2.1 Relevant gene identification ............................ 18
    2.2.2 Gene function elucidation .............................. 19
    2.2.3 Gene relationship inference ............................ 19
    2.2.4 Summary ............................................. 21
  2.3 Conclusions ............................................. 22

Chapter 3  Clustering and differential alignment algorithm ................................. 24
  3.1 Background ............................................. 25
  3.2 CDAA description .......................................... 28
    3.2.1 Stage separation ..................................... 28
    3.2.2 Gene to stage assignment .............................. 30
    3.2.3 Interaction inference .................................. 31
  3.3 Results and discussion ...................................... 33
    3.3.1 Application of the CDAA ............................... 33
    3.3.2 Validation of predicted relationships .................. 37
    3.3.3 Identification of direct connections using enhanced yeast one-hybrid ........... 45
    3.3.4 Conclusion ........................................... 47
  3.4 Materials and methods ....................................... 48
    3.4.1 Plant growth and materials ............................ 48
    3.4.2 qRT-PCR ............................................. 49
    3.4.3 Plasmid construction .................................. 49

Chapter 4  Scalable approach to data driven transcriptome dynamics modeling ............ 50
  4.1 Introduction ............................................. 51
  4.2 Model formulation .......................................... 52
    4.2.1 Basic model .......................................... 52
    4.2.2 Transcription factor effect ............................ 53
    4.2.3 Condition induced effects ............................ 56
Chapter 5  Dynamic modeling of the iron deficiency modulated transcriptome response in A. thaliana roots

5.1  Background ............................................................................................................. 63
5.2  Results ..................................................................................................................... 65
  5.2.1  Model structure ................................................................................................. 65
  5.2.2  Model fitting ........................................................................................................ 69
  5.2.3  Assessment of the unknown effects function .................................................... 70
  5.2.4  Model predictions ............................................................................................... 75
5.3  Methods .................................................................................................................... 78
  5.3.1  Establishing regulatory connections ................................................................... 78
  5.3.2  Mathematical model formulation ....................................................................... 78
  5.3.3  Model parameter estimation .............................................................................. 97
  5.3.4  Model prediction uncertainty quantification ...................................................... 101
  5.3.5  Biological experimentation .............................................................................. 101

Chapter 6  Conclusions and Future Work ........................................................................... 103
6.1  Conclusions .............................................................................................................. 103
  6.1.1  Computational tools for transcriptome analysis ............................................... 104
  6.1.2  Causality inference for iron deficiency related genes ....................................... 104
  6.1.3  Computational framework for building a dynamic model ................................ 104
  6.1.4  Dynamic modeling of iron deficiency response ................................................. 105
  6.2  Future work ............................................................................................................. 106
    6.2.1  Time course experiment design optimization ............................................... 106
    6.2.2  Cell specific inference of genetic regulations ................................................. 106
    6.2.3  Combinatorial gene regulatory effect inference ............................................. 107
    6.2.4  Other model extension strategies .................................................................... 107

BIBLIOGRAPHY ............................................................................................................. 109

APPENDIX ...................................................................................................................... 128
Appendix A  Supplemental plots ........................................................................................ 129
  A.1  Dissimilarity tables for the candidate regulators of the known iron response mediators ............................................................................................................. 129
  A.2  Posterior probability distributions for the model parameters ................................ 132
  A.3  Model predictions for single regulator mutant dynamics .................................... 145
  A.4  Uncertainty regions for the currently unknown regulators’ expression patterns time course trajectories .................................................................................................................. 147
  A.5  Credible region of the modeled bHLH115 trajectory overlaid with credible regions of its known targets’ unknown effect functions ......................................................... 149
  A.6  Initial and bias compensated time course expression patterns ............................... 153
LIST OF TABLES

Table 3.1 Regulatory interactions inference algorithms. ................................. 27
Table 3.2 PYE dissimilarity tables summary. The table lists regulators that appeared in dissimilarity tables for each thresholded version of expression patterns. Regulators that appeared for at least 2 patterns (ASIL2, ETF9, WRKY57, MYB55, and COL4) were identified as potential regulators of PYE. ......................... 38

Table 5.1 List of known iron related genetic interactions. Interaction types: ac – gene activation, in – gene inhibition, pd – protein-DNA interaction ....................... 66
Table 5.2 Ranges of parameter sweep for the parameter identifiability analysis. $\mu$ and $\sigma$ indicates the mean value and standard deviation of the corresponding parameters obtained through experiments. ................................. 86
Table 5.3 Upper and lower parameter values bounds for the PRCC method. Parameters and minimum and maximum values for the corresponding uniform priors. 89
Table 5.4 Inputs to the eFAST algorithm ............................ 89
Table 5.5 List of model parameters for gene $x_n$ expression. Units are presented in expression units [$e x$], hours [$h r$], or unitless [1]. ................................. 99
LIST OF FIGURES

Figure 2.1 Conceptual view of the information flow in a computational approach. Biological data is used to identify genes of interest (Type 1 algorithm), infer connections between these genes (Type 2 algorithm), and predict types of these connections (Type 3 algorithm). .............................................. 10

Figure 2.2 Boolean network representation in graphical and functional forms. Combinations of transcription factors $g_1, g_2, g_3$ influence expression of each other and target genes. The state of $g_6$, for example, is influenced by a combination of $g_2$ and $g_3$ or by $g_1$ alone. ...................................................... 14

Figure 2.3 Type 5 algorithms output in terms of the system of ODEs and predicted gene expression dynamics ($g_i(t)$) based on experimental values ($g_i(t_j)$). In this example, the expression pattern of each gene is influenced by the expression of at least one other gene, with some genes ($g_4$) influenced by their own expression (feedback loop). .............................................. 16

Figure 2.4 Validation techniques for algorithm Types with key references. Examples shown are those typically seen in current computational research approaches, specifically for research projects in A. thaliana. ............................................... 17

Figure 2.5 Transitions between algorithms of different types. Typical experimental transitions between algorithms are indicated with blue arrows and perspective future transitions, less common but possible with more reliable supporting algorithms, are indicated with white arrows. ................................. 23

Figure 3.1 Gene to Stage Assignment. Genes active before the Initiation-Response (I-R) boundary are assigned to the INITIATION STAGE. Genes that start their activity after the I-R boundary are assigned to the RESPONSE STAGE. Primary response genes are active right after the I-R boundary and Secondary response genes are active later. ......................................................... 30

Figure 3.2 Number of genes in each gene set (cardinality). Gene set $G_n$, $n = 1, \ldots, 6$, is comprised of genes whose maximum change occurs over the interval $(t_n, t_{n+1})$. 34

Figure 3.3 4 points based clustering. Clustering based on centered expression values $g_i^1(t_k)$, $i = 1, \ldots, 2754$, $k = 1, \ldots, 4$. $n_{genes}$ – number of genes in each cluster and $n_{TF}$ – number of Transcription Factors in each cluster. .......................... 34

Figure 3.4 5 points based subclustering. Clustering based on centered expression values $g_j^R(t_k)$, $j = 1, \ldots, 1752$, $k = 1, \ldots, 5$. $n_{genes}$ – number of genes in each cluster and $n_{TF}$ – number of Transcription Factors in each cluster. ................. 35

Figure 3.5 Average number of changes above the threshold per gene. Changes in expression $(s_n(g_i, k), k = 1, \ldots, 4)$ for Initiation stage genes were thresholded with a range of cutoff values. The graph shows the average number of changes that exceed the threshold per gene out of 4 possible changes. ............... 36
Figure 3.6  **Dissimilarity tables for PYE at different thresholds.** Dissimilarity scores between PYE and its putative regulators at a range of delays based on differential expression patterns with (A) No threshold, (B) Threshold of 0.2, or (C) Threshold of 0.4. ‘-i’ signifies that smaller dissimilarity scores were obtained for inverted regulator expression \( (d(g_R, g_T, m \Delta T) < d(g_R, g_T, m \Delta T) \forall m) \); ‘-a’ signifies that smaller dissimilarity scores were obtained for non-inverted regulator expression \( (d(g_R, g_T, m \Delta T) < \hat{d}(g_R, g_T, m \Delta T) \forall m) \).

Figure 3.7  **Regulatory relationships predicted by the CDAA.** Predicted regulations between 7 early stage transcription factors and 7 known iron homeostasis transcription factors. Edges indicating positive regulations are green and edges indicating negative regulations are red.

Figure 3.8  **Location of T-DNA insertions and qRT-PCR primers in regulator genes.** Regulator genes shown with exons in blue, untranslated regions (UTR) in gray, and promoters and introns as lines. Insertion locations are indicated with triangles and lines underneath genes indicate region spanned by qRT-PCR primers.

Figure 3.9  **Expression of regulators in mutant backgrounds.** Root tissue was collected from seedlings grown 4 days on iron sufficient media and transferred to iron deficient media for 3 days. Expression values are normalized to \( \beta \)-tubulin and to WT (Col-0) expression for each gene. Error bars indicate ±SEM (n=4). Expression of (A) OBP4, (B) WRKY57, (C) ETF9, (D) COL4, (E) ASIL2, and (F) MYB55 in respective mutant regulator backgrounds. Asterisk indicates significant difference from WT (Student’s t-test, \( p < 0.05 \)).

Figure 3.10  **Expression validation in alternate allele asil2-2.** Root tissue was collected from seedlings grown 4 days on iron sufficient media and transferred to iron deficient media for 3 days. Expression values are normalized to \( \beta \)-tubulin and to WT (Ler) expression for each gene. Error bars indicate ±SEM (n=4). Expression of (A) ASIL2 regulator, (B) ASIL2 targets, and (C) negative control gene IAA27 in asil2-2 mutant background. Asterisk indicates significant difference from WT (Student’s t-test, \( p < 0.05 \)).

Figure 3.11  **No accumulation of full-length transcript in exonic insertions.** Root tissue was collected from seedlings grown 4 days on iron sufficient media and transferred to iron deficient media for 3 days. PCR was performed on cDNA using primers for full length product (TOPO F&R) for (A) ETF9, (B) ASIL2, and (C) MYB55, each shown with \( \beta \)-tubulin (bTUB) transcript as a control and run until saturation (35 cycles).

Figure 3.12  **Expression validation of predicted targets in mutant regulator backgrounds.** Root tissue was collected from seedlings grown 4 days on iron sufficient media and transferred to iron deficient media for 3 days. Expression values are normalized to \( \beta \)-tubulin and to WT (Col-0) expression for each gene. Error bars indicate ±SEM (n=4). Mutant backgrounds are (A) obp4-1, (B) wrky57-3, (C) etf9-1, (D) col4-1, (E) asil2-1, and (F) myb55-1. Asterisk indicates significant difference from WT (Student’s t-test, \( p < 0.05 \)).
Figure 3.13  **Predicted and tested relationships between regulators and known iron homeostasis transcription factors.** .......................... 45

Figure 3.14  **Expression of negative control genes in mutant backgrounds.** Root tissue was collected from seedlings grown 4 days on iron sufficient media and transferred to iron deficient media for 3 days. Expression values are normalized to $\beta$-tubulin and to WT (Col-0) expression for each gene. Error bars indicate ±SEM (n=4). Expression of (A) ERF3, (B) IAA27, (C) IAA27, (D) UPB1, (E) IAA27, and (F) UPB1 negative control genes in mutant regulator backgrounds. All values are not significantly different from WT (Student’s t-test, $p < 0.05$). .......................... 46

Figure 3.15  **Experimentally validated regulatory relationships.** Validated regulations between 4 early stage transcription factors and 7 known iron homeostasis transcription factors. Edges indicating positive regulations are green and edges indicating negative regulations are red. Edge indicating a direct connection validated by Y1H is darker. .......................... 48

Figure 4.1  **Model building progression.** A – initial model including transcription ($a$) and degradation ($b$) rates. B – extension of the initial model by incorporating known regulators’ effects. C – model extension by incorporating a condition related effect $f_u$. .......................... 54

Figure 4.2  **Regulator influence function approximations.** Experiments for estimating parameters associated with the influence function $f_r(x_r)$ when using (A) linear (lin), (B) exponential (exp), and (C) Hill-function (hill) approximations. .......................... 57

Figure 4.3  **Unknown influencing factors effects approximation with a sigmoid function.** $u_T$ – scale coefficient, $r$ – rate coefficient, and $\tau$ – delay coefficient. .......................... 58

Figure 4.4  **Condition level effect dependence modeling.** $S_m$ – measured condition level of the initial experiment, $S_p$ – proposed condition level for gene expression measurements based on the results of physiological parameter measurements during phenotyping experiments. .......................... 59

Figure 4.5  **Profile likelihood trajectories for different parameter identifiability cases.** Solid line represents the values of error function used for model fitting. Dashed line represents the significant level. .......................... 60

Figure 4.6  **Model prediction example.** A hypothetical example of constructing a predictive model from time course samples in Wild-Type (WT) and in a Condition Induced (CI) state. .......................... 62

Figure 5.1  **Modeled regulatory connections.** Connections represent significant regulatory effects based on previous publications and the results of regulator knockout experiments. Green lines stand for activation effect, red lines indicate inhibition. .......................... 68

Figure 5.2  **Decay rate data analysis.** Estimates of decay rate values and the corresponding standard errors. Genes with significantly different estimates under iron sufficiency and iron deficiency are represented with 2 corresponding uncertainty ranges for the decay rate value. .......................... 69
Figure 5.3 **Decay rate change analysis.** Linear regression for accessing the relationships between decay rate change and the fold change of gene expression under iron deficiency conditions. 70

Figure 5.4 **Parameter uncertainty range adjustments through model fitting.** A – reduction in uncertainty for a parameter with no prior knowledge \((a)\). B – uncertainty adjustment for a parameter for which prior knowledge came from additional experimental measurements \((b)\). 71

Figure 5.5 **Time course modeling for regulators.** Model fit for gene expression dynamics of regulators. Best fit under iron deficiency (solid blue line) and under iron sufficiency (solid red line) are shown. Shaded region indicate 95% credible region. Error bars indicate experimental measurements. 72

Figure 5.6 **Time course modeling for targets.** Model fit for gene expression dynamics of targets. Best fit under iron deficiency (solid blue line) and under iron sufficiency (solid red line) are shown. Shaded region indicate 95% credible region. Error bars indicate experimental measurements. 73

Figure 5.7 **Single mutant predictions.** Model prediction and experimental results for different single mutant experiments for *MYB10* under iron deficiency. Lines show the best fit trajectories produced by the model. Shaded regions are the 95% credible regions of model prediction. Error bars indicate experimental measurements from corresponding experiment. A single measurement at 72 hours of exposure to iron deficiency was collected for each mutant experiment. 74

Figure 5.8 **Unknown effects patterns for MYB10 and MYB72.** 95% credible regions for expression patterns of potential regulators of *MYB10* and *MYB72* overlaid with a scaled versions of the known influencing genes. A, B – comparison with the credible region for bHLH115; C, D – comparison with FIT expression measurements. 76

Figure 5.9 **Knock-out mutant results for the regulators of bHLH115.** *bHLH115* expression in wild type, single and double knock-out mutant backgrounds. * – significant difference with wild type (WT). 77

Figure 5.10 **Double mutant experiment predictions.** Model prediction and experimental results for double mutant experiments for A – *MYB10*, B – *MYB72*, C – *bHLH39*, D – *PYE*, E – *bHLH101*, and F – *BTS* under iron deficiency. For the mutant experiments gene expression was measured only at 72 hrs. Dashed lines show the best fit trajectories produced by the model under wild type and single mutant conditions. Shaded regions are the 95% credible regions of model prediction under double mutant scenario. Error bars indicate experimental measurements from corresponding experiments. 79

Figure 5.11 **Decision tree for regulatory connection testing.** Guidelines for accepting/rejecting a regulator-target connection based on individual test p-value \((p_l)\), multiple comparisons compensated p-value \((p_C)\), and the results of similar tests found in the literature (PMT). 80

Figure 5.12 **Heteroscedasticity mitigation.** log2 transformation of gene expression values allowed to even-up sample variation for the statistical analysis. 80
Figure 5.13 **qRT-PCR plate effect demonstration.** Graphs show expression values for samples processed in 2 different qRT-PCR runs and values for the same samples processed in the one separate qRT-PCR run for ASIL2. 82

Figure 5.14 **Plate effect compensation procedure validation for bHLH104.** Collected time course samples capture the true (fully redone) trajectories much better after applying the developed qRT-PCR plate bias compensation algorithm. 84

Figure 5.15 **Profile likelihood.** Profile likelihood analysis curves for COL4. Solid line shows the error function, dashed line is the significance threshold. Parameters $b$ and $r$ are non-identifiable since the sum of squared errors does not change significantly with the change in parameter values. 87

Figure 5.16 **OAT sensitivity analysis for the regulators.** A parameter related slope corresponds to the influence on model behavior. 88

Figure 5.17 **PRCC analysis for the regulators.** *— significant difference between model parameter and dummy parameter influence on the model output. 90

Figure 5.18 **eFAST analysis for the regulators.** Bar height corresponds to the model sensitivity to a parameter at a given time point. Intersection of a parameter’s confidence region with the one for the dummy variable indicates model insensitivity to that parameter. 91

Figure 5.19 **Housekeeping gene outlier samples removal for iron deficient (Fe–) conditions.** Housekeeping gene log transformed Ct-values before and after the outlier removal. 93

Figure 5.20 **Housekeeping gene outlier samples removal for iron sufficient (Fe+) conditions.** Housekeeping gene log transformed Ct-values before and after the outlier removal. 94

Figure 5.21 **Statistical analysis results for COL4.** $y$ axis represents housekeeping gene decay compensated log transformed $\Delta Ct$ values at iron deficient and iron sufficient conditions. Different samples colors stand for technical replicates. 96

Figure 5.22 **Statistical analysis results for the target BTS.** $y$ axis represents housekeeping gene decay compensated log transformed $\Delta Ct$ values at iron deficient and iron sufficient conditions. Different samples colors stand for technical replicates. 96

Figure 5.23 **Sample distribution histograms for regulator coefficients $c_k$ for MYB10.** 98

Figure 5.24 **qRT-PCR measurements error variance.** A – Measurement variance dependence on gene expression value reveal heteroscedasticity in the data, B – first order approximation of variance dependence on mean expression value after the corresponding $log_2$ transformation. 100
CHAPTER 1

INTRODUCTION

1.1 Overview

The constantly growing world population raises food security concerns. Plants are the principal source of nutrients in most people's diet, so advances in agriculture play a vital role in counteracting food security threats. Plant science came to the point when traditional methods of crop breeding and the use of fertilizers and pesticides struggle to provide additional increases in yield. At the same time, current advances in genomics open avenues for precise control of plant growth and development on a molecular level. Such technologies are capable of pushing the current agricultural limits. State of the art genomics experiments produce massive datasets for in-depth exploration of molecular activity, yet the existing plant science data analysis tools can hardly harvest the embedded information in full capacity. A need for more advanced techniques in genome activity analysis promotes the need for collaborations between biologists and engineers in an effort to gain a deeper understanding of molecular mechanisms governing plant behavior.

Plant research has a vast interest in studying abiotic stress influences because the associated knowledge leads to strategies for plant growth optimization despite extreme conditions in certain areas of the planet. Whole genome time course gene activity measurements had been collected for iron [Din08; Lon10; Buc09], salt [Din08; Kil07], sulfur [IP11], nitrate [Kro10], phosphate [Lin11], pH [IP11] stress conditions in the model plant Arabidopsis thaliana with an intent of gaining knowledge that can be applied to crops. Gene activity or gene expression measurements show how genes...
get turned on and off over time in response to a given condition. Aside from time course measurements, researchers have other experimental techniques including loss of gene function, gene overexpression, and kinetic parameter measurements at their disposal to explore different aspects of plant response to stress conditions. However, the data from such experiments are typically analyzed in isolation, which hinders the ability to capture a biological process as a whole.

Among other abiotic stress conditions, iron deficiency gets a great deal of attention due to significant impact on yield. A greater understanding of iron metabolism and accumulation can aid in fighting anemia through the potential biofortification of dietary plants [Mur12; Rob99]. Significant efforts have been made towards unraveling key genetic regulators involved in iron levels sensing and iron deficiency mitigation [CG04; Bau04; Jak04; Bau07; Yua08; Iva12; Pal13b; Sel15]. Time course whole genome gene expression measurements had been collected after plant transition from iron sufficient to iron deficient media [Din08; Buc09]. These data were analyzed using a differential expression technique [Chu02], which identified genes with significant changes in activity between control and condition induced samples [Lon10]. Differential expression, however, can only reduce the search space for important genes from the whole genome to a couple thousands of the most probable options. Key regulators are then usually identified through gene deletion experiments where a mutant plant with a selected gene disabled has to be engineered. Gene importance is then established based on the magnitude of alterations in the corresponding plant's physiology parameters (phenotype). Such brute force approach can take months or even years to identify key regulators due to time and effort demands per each gene candidate.

More advanced approaches, compared to differential expression, have been developed in an effort to narrow down the search space for key genetic regulators to feasible numbers [Sim09; WH14; Li08; Hac09; Usa09; PW11]. These approaches extract additional information from genomics datasets to gain insight into molecular machinery underlying the process of interest. The capabilities of such algorithms range from causal relationships inference to continuous time dynamic modeling [Hec09; Mid12; PK13]. However, little to no guidance is often given in terms of experimental data sufficiency for a given computational technique. Moreover, such techniques are typically developed for a specific experimental setup and may not generalize well to other cases of interest.

1.2 Motivation

Current advances in experimental techniques enabled exploration of iron deficiency related response on a genetic level [Din08; Buc09]. However, state of the art techniques for whole genome activity analysis, namely differential expression tools [Chu02; TS06; Ste10; Rob10], were only able to reduce the search space for key iron gene regulators from about 30,000 to a little less than 3,000 candidates [Lon10]. Time-consuming and labor-intensive experimental screening of that subset revealed several genes that significantly affect plant physiology under iron starvation [Lon10]. How-
ever, many of the early response regulators that activate those genes remain unknown. A framework for understanding the dynamics of iron deficiency mitigation at gene level does not currently exist.

A number of computational tools have been developed to discover gene regulators and infer potential causal relationships between genes in different organisms based on temporal gene expression data [Bar11a; Irr10; Win12; Shi07; Kwo03]. However, the existing techniques for time course data analysis either work only with small subsets of genes [Bar11a], or with a large number of time course samples (i.e. tens or hundreds) [Irr10; Shi07], or with time courses that have even sampling time intervals [Win12; Kwo03]. The available whole genome time course iron deficiency datasets [Din08; Buc09] have less than 8 temporal samples that were collected with the most sample density at earlier time points due to an assumption of increased expression activity right after plant’s exposure to the stress. Such an experimental setup hinders the use of the existing computational tools and raises a need for a methodology capable of identifying causal gene regulatory relationships from whole genome time course data with a small number of unevenly sampled time points.

Inferred causal relationships between genes are typically organized in a gene regulatory network graph which presents a qualitative assessment of regulatory interactions in response to a stress condition and provides structural information for further quantitative mathematical description of gene activity kinetics. A quantitative model of gene regulation has a potential for capturing molecular behavior trends and is capable of replacing time, cost, and labor-intensive experimentation with computer simulations. Moreover, a model based on a system of ordinary differential equations provides avenues for efficient behavior control and phenotype tuning strategies, which can be imposed through the corresponding genetic manipulations.

Systems of ordinary differential equations have shown their merit when applied to well studied molecular pathways [CR12; Pok12; SC10]. The theoretical basis for dynamic modeling of genetic regulations had been discussed in a set of reviews [Hec09; Mid12; PK13]. However, a typical limitation of 2 to 8 time points per gene expression dataset [Ros12] (4 and 7 time points for the existing iron deficiency datasets [Din08; Buc09]) combined with high sample variation prevent a wide application of the dynamic modeling techniques [Bre08]. Recent approaches have supplemented time course data with the results of additional experimentation like mRNA decay rate measurements [Zak03; Bar06] to resolve data limitation issues. However, neither a general scheme for incorporating additional experimental data was proposed, nor the recommendations for the types of additional experiments that can increase informational content were given. Moreover, no criteria were given for checking whether an ensemble of experimental data has a sufficient amount of information to support a selected model structure.

This research focuses on computational methodologies for gene expression data analysis to gain a deeper understanding of iron deficiency mitigation mechanisms in plants at the molecular level. As a part of an interdisciplinary research team, I developed and applied computational tools for analyzing existing and newly obtained data from iron deprivation experiments in the model
plant *Arabidopsis thaliana*. My research efforts resulted in the development of a continuous model of gene activity progression over time in response to iron deficiency in *A. thaliana* with an aim to extend the relevant findings and model building methodology to other plants and, possibly, to other species including humans. Collaborators and I started by classifying the existing computational approaches to gene expression data analysis based on inference abilities and the corresponding data demands. Next, we developed an algorithm that allowed us to obtain a testable set of gene regulatory connections based on the existing whole genome time course data for iron deficiency in *A. thaliana*. More than half of the inferred connections were experimentally validated. After that, we formulated data sufficiency criteria for creating an ODE representation of gene interactions and developed a scalable approach to dynamic model building. This approach allowed for aggregation of various types of experimental data for model training. Finally, we collected experimental data needed to meet the data sufficiency criteria and trained the dynamic model of iron deficiency induced gene interactions in *A. thaliana*. The model was able to predict expression trends of missing iron deficiency regulators and the experimental results in 4 of 6 double mutants. This work adds to the body of literature on computational approaches for gaining biological insights from transcriptome datasets and extends the knowledge of molecular mechanisms involved in the iron deprivation response in plants.

### 1.3 Contributions

#### 1.3.1 Intellectual merit

This research presents a computational framework for gene expression analysis that includes:

1. *Functional categorization of existing computational algorithms [Kor15b].* We separated the available computational tools for transcriptome data processing into 5 categories of increasing complexity and corresponding data requirements to help researchers pick an appropriate computational tool for answering questions of interest. This categorization shows which hypotheses can be tested depending on the collection of experimental data available and suggests computational techniques that are capable of performing the tests. These techniques are organized in a sequential order to illustrate how the knowledge gained from a certain level of complexity can facilitate the next level inference.

2. *Gene regulatory network inference algorithm for whole genome time course data with uneven sampling time intervals [Kor15a].* We developed a computational algorithm capable of narrowing down the number of possible regulator-target connections based on whole genome time course with varying time intervals between adjacent time samples. We applied the algorithm to the iron deficiency
dataset [Din08] with the previously identified set of around 3,000 differentially expressed genes [Lon10]. The algorithm predicted a set of 7 early response regulators for known iron deficiency mitigation genes. More than half of the predicted connections were experimentally validated. The validated relationships served as the structural basis for a dynamic model of gene regulatory activity in response to iron deficiency.

3. **Scalable approach to building a dynamic model of gene interactions [Kor18].**
   We proposed a computational framework for aggregating various types of transcriptome (gene expression) data to capture the dynamics of genetic regulations in a set of Ordinary Differential Equations (ODEs). The framework guides through ODE structures of increasing complexity and allows to judge whether a collection of available experimental data supports a given model. The guidelines illustrate ways to increase model description and prediction power by extending the model equation structure with an addition of new experimental data. We gave recommendations on the types of new experiments that can enrich the informational content of the data in an efficient way. These recommendations are mostly applicable to plant oriented studies, but similar solutions may be applicable for other species. The framework also provides tools for model training based on the collection of experimental datasets and for estimating the uncertainty associated with model predictions.

4. **Dynamic model of gene regulation under iron deficiency in A. thaliana.**
   We formulated an ODE based model for describing regulatory dynamics of iron deficiency mitigation within a group of iron responsive genes. Using parameter identifiability analysis we showed that a time course gene expression data alone was not sufficient to uniquely estimate all the parameters. We designed experiments capable of resolving parameter identifiability, collected the corresponding samples, and incorporated the information from these experiments into the model training routine. The trained model captured regulatory interactions between previously identified iron deficiency related genes over time and predicted the results of new experiments. The model also allowed us to identify the dependence of gene mRNA decay rate on the iron concentration for the known iron mediators, estimate temporal gene expression trajectories for currently unknown gene regulators involved in iron mitigation, and hypothesize synergistic effects between regulatory genes.

1.3.2 **Broader impacts**

Application of the developed computational framework to the iron deficiency network in *A. thaliana* provides a better understanding of complex mechanisms driving the behavior of living organisms at the molecular level through dynamic modeling of genetic interactions. Mathematical models of biological systems would allow for narrowing down a set of plausible hypotheses about the molecular
behavior on a cell/organ level with computer simulations to avoid excessive experimentation, which is expensive, labor intense, and time consuming. The results of computer simulations can then be verified experimentally. Insights from model simulations can open avenues for efficient genetic manipulations to enhance the desired properties of a given organism.

Computational approaches presented in this research can be used to speed up the process of identifying useful genetic engineering strategies that can address current agricultural challenges, increase crop yield and mineral content, improve stress and pathogen resistance, and provide better utilization of poor soils that are currently not used for farming. This effort is directed towards addressing food availability challenges for the increasing world population and treating micronutrient deficiencies in human beings.

1.4 Organization

Chapter 2 provides an overview of computational algorithms used for gene expression data analysis in plants systems. The presented algorithms are separated in 5 complexity levels starting from differential expression analysis and finishing with the systems of ordinary differential equations. We illustrate relationships between a complexity level, inference capabilities, and the amount of required information to support that inference. A possible path through the levels of computational complexity is proposed. Examples of successful application of the outlined computational techniques are given.

Chapter 3 presents the developed Clustering and Differential Alignment Algorithm (CDAA) for gene regulatory connections inference from unevenly sampled whole genome time course data. We state the assumptions upon which the algorithm is based and guide through the steps of data processing and causality inference. Application of the CDAA to the whole genome Microarrays time course collected from *A. thaliana* root under iron deficiency [Din08; Lon10] is presented. We demonstrate the inference abilities of the CDAA by identifying novel iron deficiency mediation related genes. Experimental validation of the inferred regulatory connections is given.

Chapter 4 presents the computational framework for building a scalable dynamic model of gene interactions. We highlight the common features of dynamic model structures that are currently utilized for transcriptome modeling and formulate a basic model with minimal data requirements. The ways of model descriptive and predictive power improvement through the increase in model structure complexity are then illustrated and data sufficiency criteria for a model structure are stated. We suggest the types of experiments that can supplement a time course dataset for increasing model complexity. Model training tools capable of aggregating different types of experimental data are proposed.

Chapter 5 presents a dynamic model of iron deficiency induced gene regulation between known and newly identified genetic mediators. We formulate a set of ordinary differential equations that
model gene expression dynamics of 7 regulators and 7 targets based on regulatory connections identified through the CDAA and found in the literature. Parameter identifiability issues associated with the dynamic model fit to the time course data alone are resolved by collecting and analyzing samples from mRNA decay rate measurements and knock-out mutant experiments. Modeled gene expression trajectories and the corresponding credible regions are compared to the training data. Model predictions on the effects of double knock-out mutations are validated with the corresponding experimental results.

Chapter 6 summarizes the progression of research from chapters 2, 3, 4, and 5 and explains further steps towards continued understanding of molecular dynamics associated with transcriptional response to iron deficiency in *A. thaliana*. 
CHAPTER

2

COMPUTATIONAL TECHNIQUES FOR TRANSCRIPTOME DATA ANALYSIS IN PLANTS

Environmental conditions in agricultural settings are highly variable, leading to suboptimal crop yields and survival rates [AU12]. The frequency and intensity of environmental extremes, particular drought, heat, and pests, are expected to increase in the near future [MB10; ST12; AU12]. A large number of stress response studies are focused on elucidating transcriptional cascades regulating responses to individual and combined stresses. Transcription factors that play important roles in modulating such cascades are candidates for genetic engineering approaches and are worthy of intensive study. Gene expression analysis is a widely proposed means of bringing a greater understanding to all abiotic stress responses for several reasons. A large number of genes have altered expression in response to stress and these alterations play an important role in adaptation [Kil07; LM08]. Expression data is also relatively cheap. Because of this, high-throughput gene expression datasets have been generated and are publicly available for a multitude of stresses, both biotic and abiotic, with examples in Arabidopsis thaliana including but not limited to pathogen infection [Dit06; O'C12; Win12], cold [Lee05], pH [IP11], salt [Din08], light [GP11], and nutrient [Din08; Buc09; Lon10; Kro10; Lin11] stress. Though these studies are comparable in theory, a few large studies have attempted to mitigate the effects of variations in experimental setup by collecting expression data under different stresses imposed with otherwise identical growth conditions [Kil07] or under combinations of stresses [Riz04; Ras13; PS13; Sew14]. Analyses of these concurrent and combinatorial experiments in particular have revealed distinct patterns of different stress responses along with some common features, inferring that both general and specific stress response pathways exist. For example, analysis of the AtGenExpress database of concurrent stress application indicates that some abiotic stresses result in sustained gene expression alterations and others in transient alterations [Kil07]. A set of early- and commonly-induced genes, representing the so called Plant Core Environmental Stress Response (PCESR), includes transcription factors, indicating that a general stress response may be transcription factor mediated and likely occurs early in stress response cascades [Kil07; Hah13]. Combinatorial studies indicate that genes responding to combined stresses are often distinct from those responding to individual stresses, highlighting a need for both more studies of this type as well as computational methods to attempt to predict these emergent behaviors [Riz04; Ras13; PS13; Sew14]. Despite these extensive analyses, limited direct predictions concerning stress pathways have been made and validated. The majority of detailed characterizations of transcription factors, including direct promoter binding and influence on target gene expression, are the result of traditional studies. These studies are time and cost intensive. Furthermore, since many key regulators have been found through phenotypic mutant screens, subtle yet important phenotypes and genes can easily be missed. Redundancy is expected in critical regulatory mechanisms [Rie02], and predictions concerning which regulators or mutants to combine in a genetic engineering strategy would be extremely valuable. A recent increase in algorithm development and utilization will help to increase the predictive power in available datasets so that regulators and combinatorial regulatory mechanisms beyond the “low hanging fruit” can be identified. In the following sections, we describe
and organize sets of algorithms and implementations thereof in experimental approaches, aiming to bring attention to the benefit of these approaches and facilitate future increases in frequency and strength of computational biology studies.

## 2.1 Classification of inference algorithms

Many computational algorithms have been developed for analyzing gene expression data. We focus here on algorithms capable of identifying stress related genes, grouping genes by function, inferring connections between genes, estimating gene interaction direction and type, and predicting gene expression states and values in interconnected regulatory networks. These algorithms differ in complexity and implied assumptions, but can be classified based on functionality. We categorize these algorithms in 5 distinct groups based on the type of insight they provide to a biological process of interest. Depending on research objectives, these algorithms can either be used separately or as a part of a systematic computational approach where inferences from algorithms of one type can be used as input for algorithms of another type. For example, a computational approach designed to predict gene interactions and their type based on time course microarray data can be comprised of 3 algorithms of different types that sequentially process input data to obtain a desired output (Figure 2.1).

![Figure 2.1 Conceptual view of the information flow in a computational approach.](image)

Figure 2.1 Conceptual view of the information flow in a computational approach. Biological data is used to identify genes of interest (Type 1 algorithm), infer connections between these genes (Type 2 algorithm), and predict types of these connections (Type 3 algorithm).

The algorithms described can be applied to transcriptomic data obtained at \( M \) time points or treatments \( (t_j, j = 1, \ldots, M) \) for a set of \( N \) genes \( (g_i, i = 1, \ldots, N) \). Examples of such datasets include the global abiotic stress expression database AtGenExpress [Kil07]. This database includes datasets for multiple abiotic stress treatments that are obtained for \( N \approx 24,000 \) genes at \( M = 7 \) time...
points using Affymetrix ATH1 GeneChip microarray analysis. Hence, the activity of each gene can be represented by a set of numbers \( g_i(t_1), g_i(t_2), \ldots, g_i(t_M) \), forming a pattern that is used by algorithms to make inferences.

**Type 1** algorithms attempt to capture genes that are relevant to a particular condition. Techniques for determining differentially expressed genes are an example of algorithms falling into this category [CC03]. Differential expression techniques work by assuming that significant change in transcript levels of a given gene under stress condition relative to its activity under normal conditions indicates that the gene plays a role in the stress response. This assumption disregards posttranscriptional modifications as alternate means of gene product regulation. Since transcript measurement precision can vary from one experimental approach to another, statistical tests are often applied to determine the significance of the change in transcript levels. Student’s t-test for 2 treatments or ANOVA for a set of treatments are commonly applied to deduce statistical significance. Other differential expression inference algorithms were developed for large scale experimental techniques such as microarrays, for which the correlation between within-array replicates can be taken into consideration [Smy05], or RNA-Seq, for which count based statistics are more appropriate [AH10; Rob10].

**Type 2** algorithms aim to identify relationships between genes. These algorithms work by assuming that genes with “similar” expression patterns are co-regulated or are part of the same regulatory pathway [Hec09]. Techniques like co-expression analysis [Aok07; ZH05; Wol05; Usa09; Lee10] fall into this category.

Common metrics that have been used to assess similarities between genes based on their expression patterns include Pearson correlation coefficient [ZH05; Gup06; Ehl08; Pol14], Spearman correlation coefficient [Bal05; Nie11; Cui10], partial correlation coefficient [KW00; Wil04; SS05], Euclidean distance [D’h05; Ma14], and mutual information [AES10; Mar06; Ste02]. These metrics typically represent a quantified measure that establishes a pair-wise comparison between the expression levels of two genes, \( g_1 \) and \( g_2 \), across time points or experimental treatments. Kumari et al. [Kum12] presented a study that evaluated the utility of Spearman rank correlation, Weighted Rank Correlation, Kendall, Hoeffding’s D measure, Theil-Sen, Rank Theil-Sen, Distance Covariance, and Pearson correlation coefficient on transcriptional data for determining gene association. The authors found that Spearman, Hoeffding, and Kendall correlation coefficients were more effective in identifying related pathway genes than others. In contrast, Ma et al. [MW12] claim that based on manual inspection of the expression patterns of several pairs of TF-target genes, the Gini correlation coefficient can compensate for the shortcomings of the Pearson, Spearman, Kendall, and Tukey’s biweight correlations in detecting transient regulatory relationships between transcription factors and their targets. Metrics such as area between expression curves [Rue08], Z-score [Tri13], and others appear in the literature but have not been extensively evaluated.

Relationships between individual genes or across established groups of genes can be gener-
ated based on these similarity metrics. A typical procedure for estimating relationships between individual genes is to set a threshold value and assign connections between genes whose pairwise similarity value is higher than a selected threshold [Rue08; Mar06]. The statistical significance of the similarity can also be taken into consideration when establishing a connection [AES10]. Groups of similarly behaving genes are in most cases identified using clustering algorithms. Clustering algorithms apply similarity metrics to isolate groups of co-expressed genes. K-means clustering [Mar07], the Markov Cluster algorithm [VD00; MW08], biclustering [Zha05], self-organizing maps [Tam99], hierarchical clustering [Eis98], and affinity propagation [FD07] are examples of clustering algorithms applied to transcriptomic data. Martin et al. [Mar07] applied k-means clustering, hierarchical clustering, and self-organizing maps to time series transcriptomic data from mice. The results suggested that k-means was able to convey comparable grouping to hierarchical clustering, and self-organizing maps (more than 80% agreement) while maintaining less of a computational load than other approaches. Frey and Dueck [FD07] showed that the affinity propagation algorithm yields more compact clusters compared to k-means in terms of the sum of intercluster distances which might imply tighter relationships between genes in the same cluster.

Clustering has also been used to reduce the complexity of building transcriptional networks by reducing high dimensional networks with many genes to lower dimensional networks of clusters of genes or “metagenes”, which represent groups of genes with similar expression activity. The expression pattern of a metagene may be defined as the cluster average or the expression pattern of the gene with the highest sum of similarities with its cluster members. Some algorithms have extracted metagene expression patterns first by applying principal component analysis (PCA) or singular value decomposition (SVD) to the overall expression dataset. The clusters are then assembled based on similarities between gene and metagene expression patterns [Li13; Liu12].

**Type 3** algorithms aim to infer causal relationships between genes. Causal inference procedures are often based on the assumption that a change in one gene (\( g_1 \)) will result in a subsequent change in another gene (\( g_2 \)) at some later time if \( g_1 \) activates or inhibits \( g_2 \) [Che99; Kwo03; Sch04; Zha06; MC08]. Thus, the approach is similar to co-expression analysis in that it aims to find genes with similar temporal expression patterns. The key difference distinguishing this approach from those in **Type 2** is the assumption that these similarities will occur at a delay, allowing for inference on the direction of regulation (which gene comes first in a regulatory cascade) in addition to a relationship connection. The equation for Pearson correlation coefficient, for example, can be modified to assess this temporal characteristic by incorporating a time delay. Equation (2.1) reflects similarity at the delay of one time unit. The algorithms capture the regulation delay for a pair of genes by selecting the time unit duration that maximizes the correlation coefficient [Sch04].

\[
\rho_{g_1 \rightarrow g_2} = \frac{\sum_{j=1}^{M-1} (g_1(t_j) - \bar{g}_1)(g_2(t_{j+1}) - \bar{g}_2)}{\sqrt{\sum_{j=1}^{M-1} (g_1(t_j) - \bar{g}_1)^2} \sqrt{\sum_{j=1}^{M-1} (g_2(t_{j+1}) - \bar{g}_2)^2}},
\] (2.1)
where \( \bar{g}_1 = \frac{1}{M-1} \sum_{j=1}^{M-1} g_1(t_j), \bar{g}_2 = \frac{1}{M-1} \sum_{j=1}^{M-1} g_2(t_{j+1}) \).

Two sets of similarity values, each corresponding to a range of delays for a certain direction of shift, are calculated to assess the strength and directionality of connection in each pair of genes. Small similarity values, corresponding to a low probability of regulation, can be removed, leaving the remaining high confidence connections to characterize genes that have potential causal relationships. Approaches that use modifications of the metric in (2.1) have been effective for single datasets with 50 and 27 time points and sampling intervals of 20 minutes [Sch04] and for a collection of 18 datasets with 7 time points in each and sampling intervals ranging from 0.5 to 12 hours [Red07]. Other sample times may be relevant depending on the features that exist in the data.

Another class of algorithms that infer regulatory interactions between genes is Bayesian networks [Hec08; Fri00]. Bayesian networks are capable of inferring regulatory connections from time course and non-time course data. These algorithms attempt to find causal connections based on Bayes’ rule by explicitly choosing a network structure that best describes experimental data. The algorithm considers a network of gene regulations as a set of dependencies where the probability of expression of a target is conditioned on the expression of its regulator. These regulations are described as conditional probabilities. Algorithms then try to find a network structure that best describes the data based on a scoring function. Identification of the network structure is a computationally intensive problem. Complexity grows exponentially with an increasing number of nodes [Sim09]. For example, around \( 10^{18} \) different topologies arise for a network of only 10 genes [Cho07]. Thus, most of the approaches using Bayesian networks concentrate on a small subset of genes (typically when some portion of a gene regulatory network is already known) or employ sub-optimal but less computationally intense solutions to handle larger networks [Sim09].

Dynamic Bayesian Networks (DBN) [MM99; Doj06] incorporate ordering information in time course data to allow for feedback loops (not allowed in standard Bayesian networks). These feedback loops are allowed by treating expression of the same gene at different time points as different nodes. Nodes corresponding to the same gene are combined after the structure inference procedure. This algorithm leads to an increase in complexity since the number of nodes involved in structure inference routine is a product of the number of genes and the number of time points.

**Type 4** algorithms aim to infer combinations of regulator expression states that are necessary to result in a particular state of target. These algorithms can be conceptualized as a search for a functional relationship between a target and its regulator(s) \( g_t = f(g_1, g_2, \ldots, g_N) \). In this case, a qualitative measure of gene behavior can be used, with gene expression values represented as either high or low, active or inactive, or “ON” or “OFF” to simplify the problem. An “ON” state of only a couple regulators may suffice to upregulate the expression of the target. This qualitative assumption allows the use of Boolean networks [Lia98] in **Type 4** inference problems. Expression
values in Boolean network inference approaches are discretized mostly in two states, representing an activity level at each time point [Alb04; Dim11; LS04]. Regulatory connection inference algorithms try to find a binary function that computes the next state of a gene based on a combination of other genes’ states using simple Boolean operations, e.g. AND (\&) if more than one regulator should have a certain state to influence a common target, OR (\|) if any of the regulator states suffice for the same purpose, and NOT (\neg) in the case of repression (Figure 2.2). The goal of this approach is to find the simplest function for each gene, which is the function that depends on the fewest regulator genes possible.

\[
\begin{align*}
g_3 &= \neg g_2; 
g_4 &= g_1; 
g_5 &= \neg g_1 \& g_2; 
g_6 &= g_1 | (g_2 \& \neg g_3); 
g_7 &= g_3
\end{align*}
\]

Figure 2.2 **Boolean network representation in graphical and functional forms.** Combinations of transcription factors \(g_1, g_2,\) and \(g_3\) influence expression of each other and target genes. The state of \(g_6,\) for example, is influenced by a combination of \(g_2\) and \(g_3\) or by \(g_1\) alone.

A direct approach to find the simplest Boolean function that satisfies a given data set is to compare all possible functions capable of generating the observed expression pattern. The number of Boolean functions that can represent the expression activity of a gene regulated by as many as \(n\) transcription factors is \(2^{2n}\) [Aku99], making the problem computationally infeasible for a large (more than 10) number of genes. Some algorithms use prior knowledge to confine the number of genes to analyze. Others rely on network structures inferred by other types of algorithms to confine the number, type and directionality of possible regulatory relationships between individual genes or groups of genes. Another factor constraining the use of Boolean networks in whole genome dataset analysis is the small number of samples (time points) associated with most datasets. These small sample sizes typically do not provide the diversity needed to uniquely define relationships across a large number of individual genes. For example, for 5 time points, which is the median number in typical gene expression datasets [Ros12], the number of genes with distinct Boolean expression patterns is limited to only \(2^5 = 32.\) Any attempted analysis of more than 32 genes with such a dataset would result in at least 2 genes with identical behavior which would limit resolution to groups of
Type 5 algorithms aim to describe dynamic behavior in a transcriptional network. The resulting network representation allows for the reconstruction of continuous changes in transcripts over time (Figure 2.3). Ordinary differential equations are commonly used to capture the dynamics associated with gene expression changes [Ber13]. These equations allow for the estimation of gene expression values at any given time point either between samples (interpolation) or beyond the last collected sample (extrapolation) [Kro10]. When a gene regulatory network is represented in terms of linear differential equations, the instantaneous change in expression of a gene is related to the sum of weighted expression values of influencing genes:

\[
\frac{d g_i}{d t} = \sum_{k=1}^{N} a_{ik} g_k,
\]

(2.2)

where \(a_{ik}\) represent influence coefficients. Coefficients for linear differential equations are often inferred using the Least Absolute Shrinkage and Selection Operator (LASSO) algorithm [Tib96], a modification of the linear regression approach. When LASSO is used for ODE inference purposes, the changes in expression, i.e. differences between expression values at consecutive time points, are approximated by a linear combination of other genes' expression values. Expression patterns for target genes are replaced with patterns of changes in expression [Gut05; Gus04; Yeu02] to infer influence coefficients. Given that biological processes are assumed to be inherently nonlinear, linear Ordinary Differential Equation (ODE) inference algorithms for transcriptional networks rely on the assumption that the system operates close to a stability point [Yeu02]. The system may not stay close to a stability point in the case of stress induced responses, where a plant may transition from one stable steady state to another. Nonlinear ODEs, though potentially more biologically relevant because they do not rely on the steady state assumption, typically require the estimation of more coefficients associated with nonlinear terms [Gus09]. Coefficient estimation routines for inference algorithms search the parameter space to find coefficients that yield solutions closest to measured expression values [Pal13a; Kab10].

All of the described algorithms require implementation and validation in biological systems in order to assess their utility. A number of validation techniques exist, depending on the type of algorithm [Ash00; Mae05; Alo03; Ses02; Ros03; Ülk08; Hil04; Coe14; Hig99; Dav03; Pal06; O’C05; BP09; Bai09; DD07; Ber09; Weh11]. These validation techniques are visualized with key references in Figure 2.4. Validation for algorithms of Types 1 and 2, which predict associations between a gene and a process or a gene and a group of genes, are limited to analysis of Gene Ontology (GO) enrichment or phenotypes in mutants of transcriptional regulators. These phenotypes range widely depending on the stress response in question, and could involve extensive experimentation to search for a phenotype of interest. A wider range of techniques exist for algorithms of Types 3-5,
Figure 2.3 Type 5 algorithms output in terms of the system of ODEs and predicted gene expression dynamics \( (g_i(t)) \) based on experimental values \( (g_i(t_j)) \). In this example, the expression pattern of each gene is influenced by the expression of at least one other gene, with some genes \( (g_4) \) influenced by their own expression (feedback loop).

algorithms that predict relationships between transcription factors and target genes. These relationships can be tested indirectly through expression profiling, computationally through promoter analysis, or directly through binding interactions. Given that no “gold standard” validation technique exists [Ols14], convincing support often involves the combination of multiple validation techniques, such as expression analysis and binding activity for a regulator and target of interest. Similarly, complex predictions such as those derived from Type 4 and Type 5 algorithms require a combination of static and dynamic validation techniques – including expression profiling at multiple time points, preferably along with determination of binding activity.

### 2.2 Computational approaches

Computational approaches are used widely to gain insight into processes underlying plant response to stress conditions. These approaches have a similar structure in terms of the types of algorithms they use and differ in the combination of and order in which these algorithms are applied. In the following examples, we describe how algorithms of different types have been combined in particular computational approaches to answer research specific questions.
Figure 2.4 Validation techniques for algorithm Types with key references. Examples shown are those typically seen in current computational research approaches, specifically for research projects in *A. thaliana.*
2.2.1 Relevant gene identification

A large number of current computational approaches are focused on identifying genes that play a key role in a process of interest. The importance of these genes is then typically tested through mutant phenotypic analysis. Ma et al. [Ma14] analyzed a set of *A. thaliana* abiotic stress response transcriptome datasets with 6 time points to identify stress related genes. The computational approach started by partitioning each stress dataset into “informative” and “noninformative” genes using differential network analysis (Type 1 algorithm). The authors stated that differential network analysis that involves machine learning and training based on *a priori* information is more sensitive than differential expression analysis, which is statistics oriented. The Gini correlation coefficient was then calculated for pairs of “informative” genes to establish significant connections (Type 2 algorithm). Stress related genes were identified from the resulting network based on the combination of 33 topology scores obtained from the network of significant connections (Type 1 algorithm). The authors validated their algorithm by performing a phenotypic screen for 89 candidates identified as salt stress related. Mutants of 2 previously unreported salt stress-related genes showed phenotypes.

Dinneny et al. [Din08] conducted DNA Microarray experiments on *A. thaliana* root response to iron deficient media with 7 time points spanning 72 hours to identify common stress response behaviour patterns. The authors applied differential expression analysis [Chu02] to identify genes having at least a 1.5-fold change in expression with a false discovery rate value less than $10^{-4}$ at a sampling time point compared to no treatment (Type 1 algorithm). The analysis showed that the strongest transcriptional response occurred after 24 hours of treatment. Dinneny et al. [Din08] then applied the affinity propagation clustering algorithm [FD07] to form groups of similarly expressed genes and thus identify general patterns of gene expression (Type 2 algorithm). Long et al. [Lon10] used the results of this analysis and screened through mutants of 38 identified genes coding for coexpressed transcription factors. The screens led to identification of important iron homeostasis regulators POPEYE (PYE) and BRUTUS (BTS).

Lin et al. [Lin11] investigated the effect of phosphate starvation on *A. thaliana* root gene signaling using a DNA microarray time course with 3 time points to infer functional modules in early transcriptional responses. The authors used differential expression analysis with the requirement of a 2-fold change in expression with a p-value cutoff of 0.05 to identify stress related genes (Type 1 algorithm). Additional information from 2,671 experimental datasets, 300 of which are root specific, was used to select 187 root specific genes (Type 1 algorithm). The authors used the Multi-Array Correlation Computation Utility (MACCU) toolbox based on thresholding pairwise Pearson correlation coefficients to obtain 3 functional modules of stress specific genes (Type 2 algorithm). To validate the results, Lin et al. [Lin11] conducted mutant screens on 31 members of a cluster where most of the genes are known to participate in root development. Only 5 tested lines did not show a statistically significant root hair length phenotype.
2.2.2 Gene function elucidation

Another group of computational approaches aim to associate genes with a specific function during a process of interest. The guilt-by-association heuristic [Wol05] is often used to assign a function to an unknown gene based on known functions of co-regulated genes (Gene Ontology enrichment). Polanski et al. [Pol14] analyzed six *A. thaliana* stress response transcriptome datasets to identify gene modules showing evidence for co-regulation. The computational approach revealed 78 modules of co-regulated genes, 71 of which were overrepresented in Gene Ontology categories and 51 of which were enriched in transcription factor binding motifs (compared to 24 and 6 of 78 randomly assigned modules, respectively). The approach used information about which genes were differentially expressed in each stress response as an input (previously determined in other publications using Type 1 algorithms). For each gene differentially expressed under at least 2 conditions, the algorithm assembled a set of correlated genes for each condition (Type 2 algorithm). A co-regulation relationship in a pair of genes was established if these genes had shared a significant number of correlated genes across stress conditions (Type 2 algorithm). The authors used Gene Ontology enrichment, promoter analysis, and yeast one-hybrid protein-DNA interactions to validate the resulting modules of co-regulated genes.

Ma and Bohnert [MB07] integrated time course and cell specific transcriptomics data with gene promoter structures to identify stress related cis-elements in *A. thaliana*. The computational approach used in this work detected known stress related cis-elements and identified secondary motifs. The authors combined abiotic and biotic stress, hormone and chemical treatment time courses and different light condition samples to create one combined expression pattern of 145 values per gene. Differentially expressed genes were identified by combining the results from fuzzy k-means clustering [GE02] applied to all gene probes and the ‘limma’ statistical program [Smy04] which identified genes differentially expressed in at least one condition (Type 1 algorithm). Fuzzy k-means clustering was again applied to the resulting set to identify stress related clusters of genes (Type 2 algorithm). The authors assigned functions to clusters based on GO enrichment. Binding motif analysis using Plant Cis-acting Regulatory DNA Elements (PLACE) database [Hig99] revealed motifs significantly overexpressed in the function related clusters. Further analysis of 22 major clusters resulted in the identification of new DNA regulatory motifs [Ma12].

2.2.3 Gene relationship inference

Computational approaches that aim to unravel influential relationships between regulators and their targets are less common but are increasing in frequency. Windram et al. [Win12] applied a computational approach to identify transcription factor families operating at different stages of *A. thaliana* pathogen defense response. The authors analyzed transcriptional profiles at 24 time points with 4 replicates per time point. The computational approach predicted gene regulatory
interactions, confirmed experimentally or by binding motif enrichment. The analysis started with assessment of differentially expressed genes based on a combination of MAANOVA (MicroArray ANAlysis Of VAriance) [Wu03], approximate F tests, GP2S (Gaussian process 2 sample) test [Ste10], and Hotelling statistic ($T^2$) [TS06] (Type 1 algorithm). Next, a SplineCluster [Hea05] algorithm separated differentially expressed genes into clusters associated with different stages of stress response (Type 2 algorithm). The clusters were validated by GO enrichment analysis. Nonparametric modification of Bayesian network inference algorithm [Kle08] was applied to cluster representatives to infer regulatory connections between clusters (Type 3 algorithm). The authors validated the regulatory effect of one of the clusters through experiments with a knockout mutant line for the transcription factor TGA3. Experimental data showed altered gene expression in predicted TGA3 target clusters in the tga3-2 mutant, whereas targets regulated by non TGA3 clusters were less affected. The effect of another transcription factor, ANAC055, was validated by binding motif enrichment in target clusters.

Redestig et al. [Red07] analyzed a set of 18 DNA microarray time series corresponding to nine different abiotic stresses with seven time points obtained from root and shoot of A. thaliana seedlings with the aim of associating stress responsive transcription factors with their targets. The authors concluded that their computational approach delivered a usable number of high-confidence target genes (12 - 59% of identified true targets) for stress related transcription factors. The computational approach identified stress related transcription factors by selecting ones with maximum overall response and maximum change in response satisfying a specific threshold criteria (Type 1 algorithm). Covariance values between a transcription factor and other genes over a set of delays were calculated for a set of conditions (Type 3 algorithm). High scores corresponded to a high probability of regulation.

Krouk et al. [Kro10] conducted DNA Microarray experiments on A. thaliana nitrate response with six time points spanning 20 minutes to capture a gene regulatory network underlying plant adaptation to nitrate provision. The inferred temporal model of the reaction process built for 20 cluster representatives resulted in 70% correct predictions of expression value direction change after the last time point in the time course. The computational approach started with ANOVA to identify nitrogen regulated genes (Type 1 algorithm). Next, MeV software [How10] was used to separate the nitrogen regulated genes into 20 clusters, eight of which appeared to have over-represented biological functions (Type 2 algorithm). The application of LASSO based algorithm to cluster representatives provided coefficients for a system of linear ODEs describing the dynamics of each cluster (Type 5 algorithm). Predictions on the direction of change obtained from ODEs were tested by comparing them with expression values from a time point that was not used for inference purposes.
2.2.4 Summary

As can be surmised from the examples given, algorithms from Type 1, Type 2, and Type 3 are more common in current experimental approaches applied to plants. The problem of dimensionality prevents the extensive use of Type 4 and Type 5 algorithms for individual genes based on whole genome datasets due to data requirements for such type of inference [Hec09]. Thus, the dimension of the problem is typically reduced by limiting a set of genes to ones known to interact or participate in the same biological process. Recent non-stress related approaches in A. thaliana have employed such techniques. Espinosa et al. [ES04] used experimentally obtained knowledge about relationships of 15 genes in A. thaliana flower development process to predict development scenarios using Boolean networks approach (Type 4 algorithm). Sankar et al. [San11] built a model to predict states of the components from auxin and brassinosteroid signaling networks in A. thaliana by applying Boolean logic approach (Type 4 algorithm) and then transformed the resulting discrete network representation to a set of ordinary differential equations (Type 5 algorithm) to obtain quantitative predictions. Cruz-Ramirez et al. [CR12] investigated the dynamics of asymmetric cell division within the A. thaliana root by analyzing a system of nonlinear differential equations for 7 interacting complexes (Type 5 output). The analysis predicted a bistable behavior of the process. Finally, Pokhilko et al. [Pok12] refined the interaction model describing circadian rhythms in A. thaliana by modeling the process with a system of nonlinear ODEs (Type 5 output).

Similarities in regulatory processes on a genomic level allow for the application of computational approaches that were developed for non-plant species. Some computational approaches are available in software packages. An extensive use of these packages shows that even if a technique was developed and tested for one species, it can be applied to a similar dataset from another species. Examples of these approaches are briefly described here. Vermeirssen et al. [Ver14] combined the Learning Module Networks algorithm [Jos09] developed for yeast, Context Likelihood of Relatedness algorithm [Fai07] tested on E. coli, and Double Two-way t-tests algorithm tested on E. coli to identify oxidative stress regulatory transcription factors in A. thaliana (Type 3 output). The Algorithm for the Reconstruction of Gene Regulatory Networks (ARACNE) [Mar06] was developed to infer transcriptional regulations in human B cells, but then used for other applications including the inference of transcriptional interactions underlying root development and physiological processes in A. thaliana [Mon14] (Type 3 output). Other software packages that showed the ability to recover gene regulatory networks from transcriptomic data include CLR [Fai07], MRNET [Mey07], C3NET [AES10], and ARTIVA [Leb10]. The Dialogue on Reverse Engineering Assessment and Methods (DREAM) project attempted to compare such GRN inference methods applied to *Escherichia coli*, *Staphylococcus aureus*, *Saccharomyces cerevisiae* and *in silico* microarray data [Mar12a]. The authors discovered that these methods have complementary advantages and limitations under different contexts. In the case of multicellular organisms, the performance of techniques has so far been measured based
on goals achieved for a specific application. Such performance is difficult to compare between methods since goals and applications are often diverse.

2.3 Conclusions

We presented a classification of computational algorithms based on the type of information they aim to infer. This structure was used to describe approaches in the literature that have been used to gain insight into biological processes of interest based on transcriptomics data. Examples of existing computational approaches applied to plant stress transcriptional datasets demonstrated a pattern of transition between algorithms of different types (displayed graphically in Figure 2.5). This progression demonstrates that the quality of predictions made by an algorithm in the scope of a computational approach often depends on the quality of predictions made by a preceding algorithm as well as on the quality of the original biological data. Based on available algorithms and example implementations, we can state that even though both stress related gene identification and grouping algorithms (Type 1 and Type 2) are still evolving, confidence in Type 2 algorithm predictions is sufficient to allow for a transition to causality inference (Type 3). Type 3 algorithms have the potential to supply Type 4 and Type 5 algorithms with information about the structure of gene regulatory networks. This information will reduce the number of possible functional relationships to consider for these types of algorithms dramatically and thus allow for the increase in scope and predictive power. Therefore, the perspective transitions shown in Figure 2.5 will likely appear more often in future computational approaches as reliability of Type 3 algorithms predictions increase.
Figure 2.5 Transitions between algorithms of different types. Typical experimental transitions between algorithms are indicated with blue arrows and perspective future transitions, less common but possible with more reliable supporting algorithms, are indicated with white arrows.
CHAPTER

3

CLUSTERING AND DIFFERENTIAL ALIGNMENT ALGORITHM

3.1 Background

Transcriptome studies are commonly used to assess differential gene activity. Differentially expressed genes identified as having DNA binding activity, termed Transcription Factors (TFs), are of interest due to their ability to control the activation and repression of gene expression, directly influencing the accumulation of RNA and proteins that control growth and stress responses. Given the importance of transcription factors in plant stress responses, development, and cell differentiation [Sin98], the identification of key plant transcriptional regulators and their targets continues to be an area of intense research. Though many high throughput time course transcriptomic datasets are available, the prediction of regulator-target relationships between individual genes from these datasets remains an on-going area of research.

Much of what has been inferred from time course transcriptomic analysis regarding transcription factor involvement in stress responses comes from visual assessment of gene expression behavior followed by mutant screens [Din08; Buc09; Lon10; Lin11; Pal13b]. These techniques are limited at inferring regulatory relationships between genes. Moreover, mutant screens in the absence of specific predictions can be time consuming and genes without mutant phenotypes are often disregarded. This lack of mutant phenotypes is because the combinatorial and often redundant function of a gene in a pathway results in the absence of a dramatic phenotype, making experimental identification and verification difficult. Computational inference approaches can increase our understanding of transcription factor involvement in stress response by creating testable hypotheses concerning regulatory relationships, revealing networks of interactions that could be easily missed when using mutant screens. Many regulatory network inference algorithms that use gene expression data start with a refined set of genes to generate predictions. These algorithms, therefore, can require extensive prior knowledge and are most appropriate for inferring structure [Bar11a; Zha06; MS13] and/or mathematical relationships [Gus09; Kab10; ZK06] based on a subset of genes consisting of known major players in the response. There remains a need for further development of computational algorithms that are able to predict gene regulatory relationships based on a full transcriptomic dataset with little prior knowledge.

We sought to develop such a computational approach to identify key regulator-target relationships involved in the iron deficiency stress response in Arabidopsis thaliana. Iron deficiency is a useful stress to help develop and test such an algorithm because: (1) iron homeostasis is tightly regulated by transcription factors [KN12]; (2) a previously published iron deficiency time course microarray data in A. thaliana roots was available [Din08; Lon10]; and (3) several transcription factors involved in iron deficiency homeostasis have been characterized and understanding the regulation of these transcription factors would be valuable to assist in the development of future applications in agriculture.

Previous iron deficiency studies have led to the identification of several key iron homeostasis
transcription factors including \textit{bHLH39} [Yua08], \textit{bHLH101} [Siv12], \textit{bHLH115} [Lon10], \textit{PYE} [Lon10], \textit{MYB10} [Pal13b], \textit{MYB11} [Pal13b], and \textit{BTS} [Lon10; Sel14; Zha15]. These genes have altered expression after 12 hours of exposure to iron deficient conditions [Lon10]. Little is known about transcription factors that are active before 12 hours or about how early regulators target or influence the expression of known iron homeostasis transcription factors. We focused on formulating and implementing a computational approach that can be applied to the iron deprivation dataset in Dinneny et al. [Din08] as well as other typical transcriptome time course datasets (microarray or RNA-Seq) to identify unknown regulator-target relationships under a series of challenges (e.g. missing prior information) that are common to other stress analyses. Given that more than 80% of biological time course stress datasets in \textit{A. thaliana} include less than 8 (typically unevenly spaced) time points [Ros12] and 3 or less replicates [Din08; Kil07; IP11], we focused on addressing the identification of relationships in low resolution, unevenly sampled, and noisy time course data. We focused on formulating an algorithm that can work on as few as 4 time points. Effectiveness of the algorithm would in all likelihood increase with additional time resolution, particularly depending on the timing of the biological process of interest and sampling point selection with respect to this process. We also wanted to create an algorithm whose output is in the form of regulator-target connections between individual genes. An algorithm of this type would identify players involved in a transcriptional response cascade. With these players known and validated, further computational tools can be used to create more complex and predictive gene regulatory networks that capture the response of corresponding biological processing over time and that can be used to make predictions on various experimental scenarios [Hec09]. A critical aspect of this is that the algorithm should result in a manageable set of putative candidates that can be experimentally validated. We emphasize here that in the case of the iron deficiency response, very few genetic players have been identified. This lack of knowledge prevents the accurate development of a dynamic gene regulatory network of the iron deficiency response. It is the case for this and many other stresses that identification of these initial set of players and relationships is a fundamental step toward the dynamic modeling and further analysis of these responses.

Although several gene regulatory connection inference algorithms exist in the literature [Win12; Nie11; Kwo03; Bar11a], the characteristics of the iron deprivation dataset and insufficient prior knowledge about interactions between iron response regulators present unique challenges that must be addressed. Gene regulatory network inference algorithms presented in the literature are shown in Table 3.1; none of which fully address the challenges associated with iron response analysis. Some algorithms require expression data from a limited set of genes [Bar11a; Zha06; MS13] where others use expression data from evenly spaced time course experiments [Kwo03; Bic05; Sch04]. Some algorithms do not resolve regulator-target interactions between individual genes and focus more on broad relationships between clusters of genes [Win12]. In particular, a recent time-course based computational approach presented in Windram et al. [Win12] looked at formulating regula-
Table 3.1 Regulatory interactions inference algorithms.

<table>
<thead>
<tr>
<th>Paper</th>
<th>Algorithm Capabilities</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Whole genome analysis</td>
</tr>
<tr>
<td>Windram et al. [Win12]</td>
<td>✓</td>
</tr>
<tr>
<td>Nie et al. [Nie11]</td>
<td>✓</td>
</tr>
<tr>
<td>Kwon et al. [Kwo03]</td>
<td>✓</td>
</tr>
<tr>
<td>Bickel et al. [Bic05]</td>
<td>✓</td>
</tr>
<tr>
<td>Schmitt et al. [Sch04]</td>
<td>✓</td>
</tr>
<tr>
<td>Barker et al. [Bar11a]</td>
<td></td>
</tr>
<tr>
<td>Zhao et al. [Zha06]</td>
<td></td>
</tr>
<tr>
<td>Misra et al. [MS13]</td>
<td></td>
</tr>
</tbody>
</table>

Network connections between plant transcription factor families in response to pathogen infection. By analyzing 24 equally spaced transcriptome samples under stressed and unstressed conditions, the authors were able to infer connections between clusters of genes that responded at different time stages. Using this approach to extract specific regulators that influence known iron homeostasis transcription factors would be challenging since inter-cluster connections do not imply relationships between all genes from the connected clusters [SS05]. Other algorithms that extract causal influences between pairs of genes, such as the Event Method algorithm in Kwon et al. [Kwo03], can be modified to analyze general datasets with uneven time course measurements. However, these algorithms can result in an extensive number of pairwise predictions. The application of a modified Event Method algorithm to the iron deprivation dataset yielded results that were unable to resolve the roles (regulator/target) for a significant number of individual gene pairs. Moreover, most connections that we found and experimentally validated were not resolved by the modified algorithm, as detailed in the Results section. Other algorithms require multiple transcriptome datasets [Red07; Shi07] or predict connections between genes based on correlation [Ram06a; ORS07], which without modification ignore temporal evidence provided by the type of dataset [Kwo03] and are likely to result in the prediction of coexpressed genes rather than regulator-target relationships.

We developed the Cluster and Differential Alignment Algorithm (CDAA) to address the unique challenges associated with better understanding regulator-target interactions in the iron deprivation stress response. Key aspects of the algorithm include co-expression analysis [Wol05] to associate each gene with a stage in the response process, relevance network inference techniques [Sim09] to identify causal relationships between genes, and thresholding [Ros12] to mitigate the effects of noise in the data. The algorithm groups genes showing transcriptional activity at different time intervals into stages and looks for similarities in expression behavior of genes in adjacent stages.
considering a delay in order to make regulatory predictions. We applied the CDAA to iron deficiency microarray time course data from Dinneny et al. [Din08] to identify putative regulators involved in the control of known iron homeostasis transcription factors. Our results revealed distinct stages of the transcriptional response during 72 hours of exposure to iron deficient conditions. We identified transcription factors that are active within the first 12 hours of iron deficiency and experimentally validated their influence on 7 known iron transcription factors using quantitative real-time PCR (qRT-PCR). A majority (53%) of such influential predictions were validated, and one relationship was shown to be a direct binding interaction through yeast one-hybrid (Y1H) analysis. The CDAA was able to make testable and valid predictions that extend our understanding of the iron deficiency transcriptional cascade and can be used on comparable datasets to obtain a better understanding of regulatory responses in a variety of conditions.

3.2 CDAA description

We developed the Cluster and Differential Alignment Algorithm (CDAA) to make testable predictions about regulatory influences based on time course transcriptome data. The CDAA contains three consecutive steps: Stage Separation, Gene to Stage Assignment, and Interaction Inference. These steps delimit temporal stages of cascaded stress response, distribute differentially expressed genes across these stages based on expression activity, and identify potential regulations between genes in adjacent stages. The CDAA uses time course transcriptome data as an input and assumes that differential expression analysis has already been implemented based on specifics associated with the experimental approach (i.e. microarray [Smy05] or RNA-Seq [AH10; Rob10]). It is important to note that as the CDAA operates solely on gene expression data, any posttranscriptional regulation will not be captured by its predictions. The algorithm starts by calculating normalized expression values to enforce compatibility across datasets obtained using different approaches:

\[ g_i(t_k) = \frac{g_i^{raw}(t_k) - \bar{g}_i^{raw}}{\sigma_{g_i^{raw}}}, \quad i = 1, \ldots, P, \quad k = 1, \ldots, N, \]  

(3.1)

where \( g_i^{raw}(t_k) \) is the raw expression value of differentially expressed gene \( i \) at the \( k \)-th time point, \( \bar{g}_i^{raw} \) and \( \sigma_{g_i^{raw}} \) are the mean and standard deviation of the raw expression values, \( P \) is the number of differentially expressed genes, and \( N \) is the number of sampling time points.

3.2.1 Stage separation

The first step of the CDAA separates a time course of all differentially expressed genes into distinct stages based on their transcriptional activity. This provides a mechanism to computationally assess the dynamic landscape of a transcriptional cascade. Stage separation is based on the assumption
that transcriptional cascades are characterized by waves of activity, with early transcription factor activity (Initiation) triggering expression activity in subsequent stages (Response). Time intervals where groups of genes exhibit high expression activity can be identified and separated. The Stage Separation step of the CDAA assigns borders between dynamic stages by identifying the time interval where the majority of differentially expressed genes have their largest change in expression. This is based on the assumption that waves of expression activity increase in magnitude as they propagate until peak activity is reached.

The CDAA first normalizes changes in expression with respect to time using the difference in sample times to account for unevenly spaced time course data, typical in available time course datasets [Din08; Kil07; IP11]. This allows the CDAA to compare small expression changes over small time intervals and large expression changes over large time intervals without bias. The normalized change in expression of gene $g_i$ over the time interval $(t_k, t_{k+1})$ is defined as:

$$s(g_i, k) = \frac{g_i(t_{k+1}) - g_i(t_k)}{t_{k+1} - t_k}, \quad k = 1, \ldots, N - 1.$$  \hfill (3.2)

Each gene $g_i$ is then assigned to a set $G_n$, $1 \leq n \leq N - 1$, if its maximum change in expression appears at the time interval $(t_n, t_{n+1})$ ($s(g_i, n) = \max_{k=1,...,N-1} s(g_i, k)$). The set $G_b$, $1 \leq b \leq N - 1$, with maximum cardinality (number of elements) represents the time interval where the majority of genes have their highest activity, leading to assignment of the time boundary $t_b$ at the time point preceding this interval. We refer to this boundary as the Initiation-Response (I-R) boundary. All time intervals to the left of the I-R boundary are denoted as the Initiation stage and all time intervals to the right of the I-R boundary are denoted as the Response stage. The Response stage can then be subdivided into Primary and Secondary response to account for genes that start exhibiting a change in expression directly after the I-R boundary or after some delay (Figure 3.1).

The approach above provides a systematic way of preliminarily partitioning genes based on the hypothesis that the activity of a few genes (Initiation stage) triggers later activity of a large set of genes (Response stage). The presence, characteristics, and duration of these stages will differ from process to process and dataset to dataset. The sampling scheme of the dataset will heavily influence the presence/duration of the Initiation stage, existence of a Secondary or even Tertiary response in the Response stage, and/or multiple I-R boundaries. The primary goal of this initial partitioning is to capture at least two distinct stages that would allow for later extraction of regulator (stage 1 gene) / target (stage 2 gene) interactions. A sampling scheme that results in less than two stages, which is highly unlikely to occur, would result in a dataset where regulator-target interactions would be difficult to predict.
Figure 3.1 Gene to Stage Assignment. Genes active before the Initiation-Response (I-R) boundary are assigned to the INITIATION STAGE. Genes that start their activity after the I-R boundary are assigned to the RESPONSE STAGE. Primary response genes are active right after the I-R boundary and Secondary response genes are active later.

3.2.2 Gene to stage assignment

The second step of the CDAA further characterizes activity in the Initiation, Primary Response, and Secondary Response stages and assigns genes to these stages based on their expression patterns. Genes are assigned to a specific stage based on the time intervals where expression activity for that gene is first seen. For example, genes active during the Initiation stage are classified as Initiation genes, regardless of their expression activity during subsequent stages. This assignment is determined using a stage specific clustering scheme. This scheme clusters expression values across the different stages, starting first with time points corresponding to the Initiation stage then iteratively adding time points from subsequent stages. This approach enables stage specific clustering, allowing for the effective partitioning of activity at each stage while eliminating the effect that dominant expression activity over small intervals can have on whole time course clustering. Clustering using Initiation stage time points starts by centering the expression values for all genes:

\[
g_i^I(t_k) = g_i(t_k) - \bar{g}_i^I, \quad i = 1, \ldots, P, \quad k = 1, \ldots, b,
\]

where \( \bar{g}_i^I \) is the mean gene expression value for gene \( g_i \) over time points \( t_k \leq t_b \). The number of clusters chosen is not fixed and varies depending on the dataset. The ultimate goal of clustering is to
partition genes into clusters that show activity during the Initiation stage and a cluster of genes that show little to no activity during the Initiation stage. This can be achieved heuristically or via cluster number defining techniques [Rue08; Mar07]. Genes belonging to clusters with activity during the Initiation stage are assigned to the Initiation stage and genes with no apparent activity are assigned to the Response stage. Response stage genes are centered again, this time using expression values corresponding to Initiation and Primary Response stages:

\[
g_{ij}^{R}(t_k) = g_{ij}(t_k) - \bar{g}_{ij}^{R}, \quad j = 1, \ldots, P - P_I, \quad k = 1, \ldots, b + c.
\]

Here \( P_I \) represents the number of genes assigned to the Initiation stage, \( \bar{g}_{ij}^{R} \) is the mean expression value for \( g_{ij} \), \( j = 1, \ldots, P - P_I \), over time points \( t_k \leq t_{b+c} \), and \( c \) stands for the number of intervals in the Primary Response stage. Clustering is applied for a second time to isolate a group of genes with no activity after extending the time range. Genes belonging to active clusters are classified as Primary Response genes and genes belonging to the inactive cluster are Secondary Response genes. This incremental approach to clustering allows for the identification of waves of activity – the first wave containing clusters of genes active during the Initiation stage, the second wave containing clusters of genes whose activity starts at the Primary Response stage, and the final wave containing clusters of genes active only at the Secondary Response stage. This process can be adjusted based on the number of stages identified in the dataset.

### 3.2.3 Interaction inference

The final step of the CDAA predicts putative regulatory relationships between genes in adjacent stages. This step is based on the assumption that the expression activity of regulator genes in one stage will be reflected in the expression activity of corresponding target genes in a subsequent stage with some delay in regulation [Sch04; Kwo03; MC08]. Regulators are selected from genes classified in one particular stage and targets are selected from genes classified in the subsequent adjacent stage (i.e. Initiation and Primary Response). The Interaction Inference procedure uses changes in expression over time rather than expression values to assess trend similarities between putative regulators and targets. Changes in expression values are first normalized with respect to maximum change:

\[
s_n(g_i, k) = \frac{s(g_i, k)}{\max_{1 \leq n \leq N-1} |s(g_i, n)|}, \quad k = 1, \ldots, N - 1.
\]

Here, \( s_n(g_i, k) \) is a signal that ranges from –1 to 1 over all \( k \), where a value of –1 (or 1) corresponds to the largest negative (or positive) change.

The signal \( s_n(g_i, k) \) is discrete (one value represents an entire time interval), which limits the assessment of delayed similarities between a target gene, \( g_{T} \), and some putative regulator, \( g_{R} \). This problem is exacerbated when samples are sparse and non-uniform. The CDAA assigns values at
intermediate time points by assuming that the change in expression is constant between sample time points. This assumption results in a zeroth-order approximation of $s_n(g_i, k)$:

$$s_n^0(g_i, t) = s_n(g_i, k), \quad t_k < t \leq t_{k+1}, \quad k = 1, \ldots, N - 1.$$  \hspace{1cm} (3.6)

Next, a dissimilarity score between the approximated expression change signal of a candidate regulator, $s_n^0(g_R, t)$, and a delayed (shifted) version of the approximated expression change signal of a candidate target, $s_n^0(g_T, t + \Delta t)$, is calculated using a modification of pattern alignment technique [MC07; Bar11a]. A smaller dissimilarity score corresponds to a higher chance that the behavior in the regulator influences the expression activity of the target. Dissimilarity scores are calculated for a candidate pair, $(g_R, g_T)$, for a set of delays:

$$d(g_R, g_T, m\Delta T) = \frac{1}{M} \sum_{i \in \mathcal{T}} |s_n^0(g_R, t_i) - s_n^0(g_T, t_i + m\Delta T)|, \quad m = 0, 1, \ldots, M - 1,$$  \hspace{1cm} (3.7)

where $\mathcal{T}$ is the set of time points in the regulator’s stage, $\Delta T$ is the largest common divisor of the time intervals in the time course data, and $M$ represents the maximum number of $\Delta T$ that can fit in each time interval corresponding to regulator’s and target’s stages. The resulting dissimilarity score quantifies likelihood of a positive influence between a regulator and its target assuming similar, yet delayed, expression behavior. Dissimilarity scores for the inverted regulator expression $\hat{d}(g_R, g_T, m\Delta T)$ are also calculated to detect possible negative influences. The smaller of $d(g_R, g_T, m\Delta T)$ and $\hat{d}(g_R, g_T, m\Delta T)$ is taken for each time delay $m\Delta T$, and the predicted influence type (positive or negative regulation) is recorded. Dissimilarity scores for a potential target are organized into a dissimilarity table where rows correspond to potential regulators and columns to delays. Rows at which the minimal dissimilarity score is achieved at a delay of 0 hrs are discarded to avoid assigning a regulatory connection between genes that are co-expressed.

Noise in expression data can often disrupt the accuracy of alignment algorithms [Ros12]. The algorithm addresses the possibility that some small changes in gene activity may be due to experimental error or noise by applying thresholding to normalized gene expression changes, $s_n^0(g_i, m\Delta T)$, to convert changes into events of upregulation (1), downregulation (–1), or no regulation (0) [Kwo03]:

$$s_{n,thr}^0(g_i, t_j) = \begin{cases} 
1, & \text{if } s_n^0(g_i, m\Delta T) > thr, \\
0, & \text{if } |s_n^0(g_i, m\Delta T)| < thr, \\
-1, & \text{if } s_n^0(g_i, m\Delta T) < -thr.
\end{cases} \hspace{1cm} (3.8)$$

Dissimilarity tables for multiple thresholded versions of the signal $s_{n,thr}^0(g_i, m\Delta T)$ along with the unthresholded version, $s_n^0(g_i, m\Delta T)$, are generated. Different threshold values assume different levels of noise and will result in different dissimilarity tables for the same potential target. A maximum
dissimilarity cutoff is used to identify candidate regulators that are more likely to influence a potential target at each threshold. Consensus over multiple thresholds results in CDAA regulatory predictions that can be experimentally validated.

3.3 Results and discussion

3.3.1 Application of the CDAA

We applied the CDAA to the iron deficiency dataset from Dinneny et al. [Din08] with $P = 2754$ differentially expressed genes sampled at $N = 7$ time points in *Arabidopsis thaliana* roots 0, 3, 6, 12, 24, 48, and 72 hours after exposure to iron deficient conditions. Differentially expressed genes were defined in Long et al. [Lon10] as genes that were at least 1.5-fold differentially regulated with a false discovery rate (Q-value) less than $10^{-4}$. We maintained this designation for application of the CDAA.

We calculated changes in expression for each differentially expressed gene using Eq. (3.2) and assembled the sets $G_n$, $n = 1, \ldots, 6$, with genes whose maximum change occurs over the interval $(t_n, t_{n+1})$. The number of genes in each set $G_n$ (cardinality) is shown in Figure 3.2. The set $G_4$, corresponding to the interval between 12 and 24 hrs, contains the maximum number of genes. We assigned the I-R boundary to the time point preceding this interval, $t_b = 12$ hrs ($b = 4$), and defined the stages as Initiation: $0 \leq t \leq 12$ hrs and Response: $12 < t \leq 72$ hrs. We assigned Primary Response (defined as the interval of high activity following the I-R boundary) to $12 < t \leq 24$ hrs. The transcriptional iron deficiency response as described by the CDAA, therefore, has at least 3 waves of activity, with the first wave ending at 12 hours.

After the Stage Separation step, we clustered all differentially expressed genes based on expression patterns during the Initiation stage for Gene to Stage Assignment (Figure 3.3). We chose k-means clustering for this procedure since it is not as computationally intensive as hierarchical clustering or self-organizing maps but is shown to produce similar results when applied to transcriptome datasets [Mar07]. Clustering revealed four behavioral patterns: decrease in expression (Figure 3.3A), increase in expression (Figure 3.3B), oscillatory behavior (Figure 3.3C), and no change in expression (Figure 3.3D). We assigned genes from Clusters 1 through 3 to the Initiation stage, and used the inactive Cluster 4 for the second round of clustering. Cluster 4 contained 1752 genes (63% of all genes in the dataset) and 97 known transcription factors (72% of all known transcription factors in the dataset). Hence, the majority of activity associated with iron deficiency occurs after 12 hours of exposure. Clusters 1 through 3 contain 36 transcription factors, none of which have so far been implicated in the iron deficiency response, meaning that these regulators may trigger the plant’s overall response to the stress.

We added the Primary Response stage time point $t_5 = 24$ hrs to the expression patterns to classify the remaining genes. The results of clustering applied to Cluster 4 after adding the 24 hrs time point
**Figure 3.2 Number of genes in each gene set (cardinality).** Gene set $G_n$, $n = 1, \ldots, 6$, is comprised of genes whose maximum change occurs over the interval $(t_n, t_{n+1})$.

**Figure 3.3 4 points based clustering.** Clustering based on centered expression values $g^k_i(t_k)$, $i = 1, \ldots, 2754$, $k = 1, \ldots, 4$. $n_{genes}$ - number of genes in each cluster and $n_{TF}$ - number of Transcription Factors in each cluster.
are shown in Figure 3.4. Genes from Clusters 4.2 and 4.4 show a rise in expression after 12 hours, genes from Cluster 4.1 show a decrease in expression, and genes from Cluster 4.3 are inactive during the whole interval from 0 to 24 hours. Thus, we assigned genes from Clusters 4.1, 4.2, and 4.4 to the Primary Response stage.

![Graphs A, B, C, and D showing expression levels for Clusters 4.1, 4.2, 4.3, and 4.4 respectively.](image)

**Figure 3.4 5 points based subclustering.** Clustering based on centered expression values $g_j^R(t_k), j = 1, \ldots, 1752, k = 1, \ldots, 5, n_{\text{genes}}$ – number of genes in each cluster and $n_{\text{TF}}$ – number of Transcription Factors in each cluster.

We selected 7 transcription factors with published roles in the iron deficiency response and used the Interaction Inference stage of the CDAA to predict relationships involving these genes as a means of focusing validation to a feasible set. We were limited to genes that were present in the Affymetrix chip used for this particular expression analysis. Therefore, transcriptome data for the master iron deficiency regulator *FIT1* and its heterodimer partner *bHLH38* were not available for analysis by the CDAA. The known iron homeostasis transcription factors chosen for this study were *bHLH39* [Yua08], *bHLH101* [Siv12], *bHLH115* [Lon10], *PYE* [Lon10], *MYB10* [Pal13b], *MYB72* [Pal13b], and *BTS* [Lon10; Sel14].
All 7 known iron related transcription factors were assigned to the Primary Response stage by the CDAA (bHLH39, bHLH101, and bHLH115 appeared in Cluster 4.2 and the remaining transcription factors appeared in Cluster 4.4). We hypothesized that regulators from the Initiation stage (Clusters 1-3) may be responsible for influencing the known iron homeostasis transcription factors. Since the Initiation stage regulators and known iron transcription factors appeared in adjacent stages, we applied the CDAA to test this hypothesis.

We calculated normalized changes in expression for each transcription factor from the Initiation stage (regulator) and each known iron transcription factor (target). The largest common divisor for time intervals in Initiation and Response stages is 3 hours, so this value served as the delay step size (i.e. $\Delta T = 3$ hours). Since each stage is 12 hours long, a maximum of $4\Delta T$ can fit in each stage (i.e. $M = 4$).

We selected thresholds for the noise reduction portion of the Interaction Inference step to account for different levels of signal fluctuations. We first applied a set of thresholds to Initiation Stage gene expression changes, $s_n(g_i, k)$, $k = 1, \ldots, 4$, to obtain the average number of changes per gene above the threshold (Figure 3.5). Based on these results, we set thresholds equal to 0.2 and 0.4 so that 25% and 50% of possible changes per gene, respectively, were attributed to noise. Using these thresholds, we produced two more versions ($s_{n,0.2}^0(g_i, m\Delta T)$ and $s_{n,0.4}^0(g_i, m\Delta T)$) of the normalized change in expression signal $s_n^0(g_i, m\Delta T)$ for each regulator from the Initiation stage and each known iron transcription factor.

![Figure 3.5 Average number of changes above the threshold per gene](image)

**Figure 3.5 Average number of changes above the threshold per gene.** Changes in expression ($s_n(g_i, k)$, $k = 1, \ldots, 4$) for Initiation stage genes were thresholded with a range of cutoff values. The graph shows the average number of changes that exceed the threshold per gene out of 4 possible changes.
We calculated dissimilarity values between all regulators from the Initiation stage and one of the targets, PYE, at multiple time delays for each threshold and organized them into dissimilarity tables (Figure 3.6). We chose a cutoff of 0.4 to remove potential regulators with high dissimilarity over all delays. This cutoff produces a testable number of predicted regulators (3-5) per target. A deviation by 0.1 from this value adds or eliminates 1 to 3 candidate regulators. The 4 regulators that appeared in 2 out of 3 dissimilarity tables were assigned as potential regulators of PYE (Table 3.2).

Using the same procedure, we determined potential regulators for the remaining targets. All 7 targets were predicted to be regulated by a set of 7 regulators (Appendix A.1). These predictions resulted in a small network of interactions containing 14 nodes and 32 edges (Figure 3.7). The majority of the edges (26 out of 32) were predicted to be positive regulations. 6 of the 7 regulators are named genes, though only 3 have been characterized (WRKY57 [Jia12; Jia14], ASIL2 [Wil11; Bar12], and LRL3 [Kar09]) and none are currently linked to iron homeostasis. The remaining regulator (At2g36720) was named Early Transcription Factor 9 (ETF9).

### 3.3.2 Validation of predicted relationships

We used quantitative real time PCR (qRT-PCR) to validate predicted regulator-target relationships by measuring transcript of targets in a background with significantly altered expression of the predicted regulator. Ideally, multiple mutant alleles could be tested for each regulator, but due to limited availability of lines with significantly altered expression, only one mutant allele per regulator was tested (with the exception of ASIL2 for which 2 lines were tested and LRL3 for which...
Table 3.2 *PYE* dissimilarity tables summary. The table lists regulators that appeared in dissimilarity tables for each thresholded version of expression patterns. Regulators that appeared for at least 2 patterns (ASIL2, ETF9, WRKY57, MYB55, and COL4) were identified as potential regulators of *PYE*.

<table>
<thead>
<tr>
<th>Regulator</th>
<th>No Thr.</th>
<th>Thr. = 0.2</th>
<th>Thr. = 0.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASIL2</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>ETF9</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>WRKY57</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>MYB55</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>GNU1</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TG</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LRL3</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WRKY26</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RD26</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>COL4</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>TGA2</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OBP4</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 3.7 Regulatory relationships predicted by the CDAA. Predicted regulations between 7 early stage transcription factors and 7 known iron homeostasis transcription factors. Edges indicating positive regulations are green and edges indicating negative regulations are red.
no suitable line was identified during validation). We sequenced insertion locations; 4 are exonic (etf9-1, asil2-1, myb55-1, and asil2-2), 2 are intronic (wrky57-3 and col4-1), and 1 is in a promoter (obp4-1) (Figure 3.8). Insertions in the introns and promoter led to reduced regulator expression and no full product was made in mutants with exon insertions (Figure 3.9, Figure 3.10, Figure 3.11). We measured transcript levels of predicted targets for each regulator in the mutant backgrounds as compared to wild-type (either Col-0 or Ler) in 7 day old seedlings, 3 days after shift from iron sufficient to deficient media (Figure 3.10, Figure 3.12). We considered target expression significantly affected if it differed from wild-type values with a p-value of 0.05 or less. We considered significantly altered target expression in either direction as support for an influential relationship and considered altered expression in the correct direction (i.e. lower target expression in the mutant of a predicted positive influencer) as support for a specific type of influential relationship. In the case of ASIL2, for which 2 mutant lines were available, we considered significantly affected expression in either mutant line as support.
Figure 3.8 Location of T-DNA insertions and qRT-PCR primers in regulator genes. Regulator genes shown with exons in blue, untranslated regions (UTR) in gray, and promoters and introns as lines. Insertion locations are indicated with triangles and lines underneath genes indicate region spanned by qRT-PCR primers.
Based on qRT-PCR results, we were able to validate 17 out of 32 influential relationships (53%) (Figure 3.13). Interestingly, though a majority of influential relationships were validated, some of the type specific (activation or inhibition) predictions were incorrect. This may be because the CDAA was based on the assumption that change in expression of a potential regulator leads to change in the expression of its target gene in isolation (a rise in a target can only be the result of a rise in its regulator if these genes have a positive influential relationship). This assumption does not take into account combinatorial effects of multiple transcription factors acting on the expression of one gene, and the resulting algorithm predictions of positive or negative influence are unable to assess to what extent each regulator controls the expression of each target in combination. Also, it may be possible that similar yet delayed expression patterns could instead be indicative of a regulator acting to continuously dampen expression of targets that are activated by another regulator. This effect is a likely explanation for the results seen for ETF9 and ASIL2, which both were predicted as positive regulators of their targets by the CDAA due to trend similarities. Experimentally measured expression of regulators and targets using qRT-PCR indicate, however, that target expression is increased in both mutant backgrounds. This may indicate that under iron deprivation, increased expression of the regulators works to constrain the expression of the targets. Thus, the assumption on the type of influence between a potential regulator and its target, widely used in gene regulatory network inference algorithms, appears to be limited in the case of our application.

Transcription factors that were not predicted to be regulated by the 7 regulators (dissimilarity value higher than 0.4 at at least 2 out of 3 thresholds) were chosen as negative control genes. Expression of each negative control gene was not significantly different in the mutant backgrounds, indicating that the expression alterations seen are specific to the predictions of the algorithm and not indicative of widespread expression alterations in the mutants (Figure 3.10, Figure 3.14).

While other algorithms have been developed to infer regulatory relationships based upon transcriptomic data, they are typically driven by substantial prior knowledge of regulator-target relationships or are of limited utility for minimal, unevenly spaced datasets. For example, the Event Method [Kwo03], similar to the CDAA, aims to infer causal relationships between genes by aligning their differential expression patterns with an assumption of a possible delay in regulation, but required modification to work with an unevenly sampled time course dataset. After implementing a linear interpolation step as a modification to the Event Method algorithm and limiting a set of genes to transcription factors, we obtained predictions for the same known iron response genes used with the CDAA. The predictions resulted in a network containing 44 nodes and 144 edges. Only 2 regulatory connections that were identified by the CDAA and experimentally validated were found in the Event Method prediction set. Thus, the CDAA is an improvement on currently available regulatory inference algorithms.
Figure 3.9 Expression of regulators in mutant backgrounds. Root tissue was collected from seedlings grown 4 days on iron sufficient media and transferred to iron deficient media for 3 days. Expression values are normalized to β-tubulin and to WT (Col-0) expression for each gene. Error bars indicate ±SEM (n=4). Expression of (A) OBP4, (B) WRKY57, (C) ETF9, (D) COL4, (E) ASIL2, and (F) MYB55 in respective mutant regulator backgrounds. Asterisk indicates significant difference from WT (Student’s t-test, p < 0.05).
Figure 3.10 Expression validation in alternate allele asil2-2. Root tissue was collected from seedlings grown 4 days on iron sufficient media and transferred to iron deficient media for 3 days. Expression values are normalized to $\beta$-tubulin and to WT (Ler) expression for each gene. Error bars indicate $\pm$SEM (n=4). Expression of (A) ASIL2 regulator, (B) ASIL2 targets, and (C) negative control gene IAA27 in asil2-2 mutant background. Asterisk indicates significant difference from WT (Student’s t-test, $p < 0.05$).

Figure 3.11 No accumulation of full-length transcript in exonic insertions. Root tissue was collected from seedlings grown 4 days on iron sufficient media and transferred to iron deficient media for 3 days. PCR was performed on cDNA using primers for full length product (TOPO F&R) for (A) ETF9, (B) ASIL2, and (C) MYB55, each shown with $\beta$-tubulin (bTUB) transcript as a control and run until saturation (35 cycles).
Figure 3.12 Expression validation of predicted targets in mutant regulator backgrounds. Root tissue was collected from seedlings grown 4 days on iron sufficient media and transferred to iron deficient media for 3 days. Expression values are normalized to β-tubulin and to WT (Col-0) expression for each gene. Error bars indicate ±SEM (n=4). Mutant backgrounds are (A) obp4-1, (B) wrky57-3, (C) etf9-1, (D) col4-1, (E) asil2-1, and (F) myb55-1. Asterisk indicates significant difference from WT (Student’s t-test, p < 0.05).
3.3.3 Identification of direct connections using enhanced yeast one-hybrid

Though the CDAA can predict influential relationships between transcription factors and their targets, it cannot differentiate between direct (binding) or indirect connections. We utilized yeast one-hybrid (Y1H) analysis to identify direct regulatory connections involving one of the target genes, PYE, and to see if any of these connections correspond to CDAA predictions. We cloned the promoter region of PYE into Y1H reporter constructs and screened it against an expanded collection of A. thaliana root specific transcription factors [Bra11; Gau11]. We identified 20 transcription factors that bind to the PYE promoter. Two of these transcription factors are differentially expressed under iron deficiency and were thus a part of CDAA analysis. It is likely that other interactions could have been missed in the Y1H analysis because this assay is conducted in vitro and independent of iron availability. Some direct interactions may require other regulatory machinery found only in plants or only under iron deficiency.

One of the two iron-responsive transcription factors that bound the PYE promoter is ASIL2, which was predicted and validated to affect the expression pattern of PYE (Figure 3.12, Figure 3.10). Interestingly, ASIL1, the close homolog to ASIL2, also binds the PYE promoter. The other iron-responsive transcription factor that targets PYE, HB-12, was not predicted to regulate PYE expression via the CDAA because the minimum dissimilarity in the alignment of HB-I2 and PYE expression occurred at a delay of 0 hrs, where the CDAA is unable to distinguish between genes affecting each other and genes that are co-expressed.

The close homolog to ASIL2, ASIL1, is known to bind to the GT-box-like-element (GTGATT) [Gao09].
Figure 3.14 Expression of negative control genes in mutant backgrounds. Root tissue was collected from seedlings grown 4 days on iron sufficient media and transferred to iron deficient media for 3 days. Expression values are normalized to $\beta$-tubulin and to WT (Col-0) expression for each gene. Error bars indicate ±SEM (n=4). Expression of (A) ERF3, (B) IAA27, (C) IAA27, (D) UPB1, (E) IAA27, and (F) UPB1 negative control genes in mutant regulator backgrounds. All values are not significantly different from WT (Student's t-test, $p < 0.05$).
This element is found in the *PYE* promoter region. Given that *PYE* was validated as a direct connection, it is possible that ASIL1 and ASIL2 share this binding element. It could also be possible that ASIL2 binds to other unidentified promoter elements. The CDAA as an expression analysis tool will therefore be particularly effective in tandem with promoter analysis and high throughput transcription factor binding data including Y1H and chromatin immunoprecipitation sequencing (ChIP-Seq). These additional experiments could improve the specificity of further predictions by revealing characteristics that are common specifically to direct connections. It is striking that even though binding predictions were not the immediate goal of the CDAA, one such connection was detected.

The 7 regulators predicted to influence known iron regulators come from distinct transcription factor families and are all previously unlinked to iron homeostasis. Several of the validated transcription factors (Figure 3.13) have known or predicted roles in stress and development. COL4 (At5g24930) has a predicted B-box zinc finger domain and CCT motif [Fin08]. Although COL4 is uncharacterized, it is closely related to COL3, involved in light signaling and root growth [Dat06]. ASIL2 (At3g14180) has been shown to play a role in regulating embryo maturation together with its close homolog ASIL1 [Wil11]. Both ASIL1 and ASIL2 are members of the trihelix transcription factor family. ASIL1 recognizes and binds to a specific element in promoter sequences, and over 1000 genes are misregulated in the *asil1-1* mutant background [Gao09]. Early chlorophyll accumulation during embryo development is seen in both *asil1* and *asil2* mutants, and more strongly in an *asil1asil2* double mutant [Wil11]. Given the requirements of iron for chlorophyll biosynthesis, as well as links between seed iron content and embryogenesis [Sta08; Ros11; Gri13; Sel14], it is possible that ASIL2’s role in embryo development is related to its role in iron homeostasis.

We did not observe visual phenotypic differences from wild type for any of the mutants when grown under iron deficient conditions (data not shown). This result is not necessarily unexpected, especially given the modest alterations in target expression seen. The algorithm assigns multiple regulators to each iron homeostasis gene of interest, indicating that combinatorial effects may be in effect. Therefore, it will likely be necessary to examine higher order mutants to observe more dramatic phenotypes.

### 3.3.4 Conclusion

The CDAA was able to make specific predictions about regulatory relationships between genes. A set of 931 potential regulatory relationships between 133 differentially expressed transcription factors and the 7 chosen targets was reduced by the CDAA to a very testable subset of 32 connections. The majority of the relationship predictions (53%) were experimentally validated by significantly altered target expression in a background with altered regulator expression. The regulators identified were previously unlinked both to a role in iron deficiency and to the predicted targets. One of the
connections predicted by this algorithm was a direct connection, validated by Y1H analysis. Together, these results yield a small network of interactions which has expanded our understanding of the iron deficiency response in *A. thaliana* to novel genes and connections (Figure 3.15). Thus, the developed CDAA is capable of making predictions with biological significance and can be used to reveal gene regulatory connections in distinct fields of study.

![Figure 3.15 Experimentally validated regulatory relationships.](#)

**Figure 3.15 Experimentally validated regulatory relationships.** Validated regulations between 4 early stage transcription factors and 7 known iron homeostasis transcription factors. Edges indicating positive regulations are green and edges indicating negative regulations are red. Edge indicating a direct connection validated by Y1H is darker.

### 3.4 Materials and methods

#### 3.4.1 Plant growth and materials

The *Arabidopsis thaliana* ecotypes Columbia (Col-0) and Landsberg erecta (Ler) were used as wild type, depending on mutant background. T-DNA insertion lines for *obp4-1* (SALK_118463), *wrky57-3* (GK-078H12), *etf9-1* (SALK_025328), *col4-1* (SALK_092012C), *asil2-1* (SAIL_258_F06), *myb55-1* (GK-460G09), and *asil2-2* (ET8777) were confirmed using primers listed in the supplementary table at [https://doi.org/10.1371/journal.pone.0136591.s009](https://doi.org/10.1371/journal.pone.0136591.s009). *A. thaliana* seedlings were grown on iron sufficient media (+Fe) containing Murashige and Skoog basal salt solution supplemented with 0.05% (w/v) MES, 1% (w/v) sucrose, and 0.1 mM Fe-EDTA in the place of iron sulfate. Iron deficient media (–Fe) is prepared as described above except 0.3 mM of the iron chelator ferrozine is substituted for iron sulfate. Prior to plating, seeds were sterilized in 70% EtOH for 5 minutes, 30% bleach and 0.02% Triton X-100 for 15 minutes, and then rinsed 3 times in dH₂O. Seeds were stratified in dH₂O for 2-3 days at 4°C. For expression analysis, seeds were sown directly on 100 µm Nitex Nylon mesh (Genesee Scientific) on square plates filled with iron sufficient media for 4 days,
and transferred to iron deficient media for 3 days.

3.4.2 qRT-PCR

Total RNA was isolated from pooled roots of *A. thaliana* seedlings using the RNeasy Plant Mini Kit (Qiagen). cDNA was synthesized using the SuperScript® III cDNA synthesis kit (Life Technologies) with oligo(dT) primers. qRT-PCR was performed using iTaq™ Universal SYBR® Green Supermix (Bio-Rad) and the StepOnePlus™ Real-Time PCR System (Applied Biosystems). Relative expression was calculated using the $2^{-\Delta\Delta C_T}$ method, normalized to $\beta$-tubulin and wild type. Statistical analysis was performed using Student’s t-test ($p < 0.05$) (n=4).

3.4.3 Plasmid construction

The *PYE* promoter construct was created as described in Long et al.[Lon10]. Briefly, 1120 bp upstream of the *PYE* start codon was amplified using primers listed in the supplementary table at https://doi.org/10.1371/journal.pone.0136591.s009 and cloned into the pDONR™ P4-P1R (Invitrogen) vector. This fragment was recombined into HIS3 and LacZ promoter:reporter vectors for enhanced yeast one-hybrid (Y1H) screening, as described in Gaudinier et al. [Gau11].
CHAPTER

4

SCALABLE APPROACH TO DATA DRIVEN TRANSCRIPTOME DYNAMICS MODELING

4.1 Introduction

Living organisms develop and respond to stimuli through a set of regulations on a molecular level. The regulatory rules are hard-written in a genome and implemented through the action of transcription factors which modulate gene activity according to a given condition. Various types of experiments are performed to gain insight into that machinery with the ultimate goal of using this knowledge to modify organisms in novel and strategic ways. Transcriptome abundance measurements are a widely utilized technique to estimate the change of gene activity over time or under a condition of interest. Methods of various complexity have been used to analyze the transcriptome (gene expression) data [PK13; WH14; Kor15b]. Out of those methods the systems of Ordinary Differential Equations (ODEs) present the most descriptive way of representing transcriptome dynamics over time within a cell. ODEs based models allow to approximate experimental data by fitting a flexible continuous curve to a given set of measurements. Hence, a multitude of experimental data presented across publications can be aggregated into a set of ODEs which approximate gene regulatory mechanisms. However, the majority of dynamic modeling approaches restrict the model training data to time course gene expression measurements, thus, limiting their perspective on the process of interest.

Ordinary Differential Equations (ODEs) gain a growing interest as a tool for modeling gene expression dynamics. The theoretical basis has been laid out in reviews [PK13; Pol09; Bre08], yet a typical limitation of 2 to 8 time samples per time course [Ros12] is still a barrier for a wide use of ODEs in practical applications [Bre08]. This limitation also leads to formulation of phenomenological models like linear models [Car09; Yeu02; Gut05; Gus04; Kro10], Standardized Qualitative Dynamical Systems (SQUAD) models [MX06; DC07; Phi09; SC10; San11], nonlinear basis functions models [Gus09], or time-variant models [Kim08; Kab10] with a small number of biologically irrelevant parameters rather than using mathematical constructs based on molecular kinetics like in S-Systems [Pal13a; Cho13; LW08; TW05; Man15; Kik03] or Hill-function kinetics based models [Kru10; Has14; Hen15]. Moreover, a wide range of proposed ODE structures for modeling transcriptome activity makes the choice of an appropriate mathematical representation challenging due to a lack of specific requirements for experimental data in the corresponding papers.

A number of studies have successfully applied ODEs to model transcriptome dynamics given a sufficient amount of information in terms of gene regulatory network graph and/or the results of various types of experiments which complemented the time course data [Gut05; Car09; SC10]. Despite the findings facilitated by such modeling and the potential of building on previous results by collecting additional data, the cases of gradual model evolution are rather an exception than a rule. One such exception is the circadian clock effect in plants, which has been the subject of a number of ODE models [Loc05b; Loc05a], continuously improved over time by the addition of new feedback loops [Loc06], post-transcriptional and post-translational regulation [Pok10], and
mutant expression data [Pok12; Pok13]. In each case the addition of new data allowed for greater
descriptive and predictive power [BD13]. However, each iteration required a reformulation of the
previous model structure to incorporate new experimental results, making the process of model
improvement long and not intuitive.

We propose a methodology for dynamic model building which allows for a gradual increase in
model complexity when new experimental data become available by expanding the corresponding
elements in the existing ODE structure rather than reformulating the entire model. In Section 4.2 we
summarize the commonly used ODE structures into levels of mathematical complexity where each
new level extends the previous one based on additional data and propose the types of experiments
allowing for an efficient transition between the levels. In Section 4.3 we propose an algorithm for
aggregating the available experimental datasets and criteria for data sufficiency for a given level of
model complexity. Thus, the resulting model will represent the outcomes of all relevant experiments
in a set of uniquely identifiable parameters, provide insights into transcriptome properties if the
parameters are biologically relevant, and allow for gene expression predictions in a wide range of
conditions combinations.

4.2 Model formulation

The scalable approach to building a dynamic model presented here provides a guide through levels
of biological complexity included in a set of ODEs while finding a balance between current biological
knowledge and descriptive capabilities of the inferred model. This approach is based on current
understanding of transcriptome dynamics and allows for an increase in model complexity when
new experimental data becomes available. We will start the description with a basic model and
build it up by incorporating additional data which can be obtained experimentally.

4.2.1 Basic model

Gene expression can be thought of as a balance between the rate of gene transcription and the
rate of the corresponding mRNA degradation. Conceptually, ODEs describe the factors that influ-
ence the increase in mRNA production and the factors that influence the decrease or reduction of
mRNA [Mid12]:

\[
\frac{dx}{dt} = \text{mRNA production} - \text{mRNA degradation},
\]  

(4.1)

where \( x \) represents gene expression of a given gene over time.

Both the production and degradation components of (4.1) can be represented by mathematical
constructs of varying complexity [Car09; Pal13a; Kru10; MX06; Kim08; Gus09]. The availability of data
dictates the level of detail that can be captured by a given model. Under a simple assumption of a con-
stant rate of expression \((\text{mRNA production} = a)\) and linear degradation \((\text{mRNA degradation} = b \times x)\), we can model gene expression dynamics with the following ODE:

\[
\frac{dx}{dt} = a - bx,
\]  

(4.2)

where \(x\) represents gene expression, \(a\) represents the transcription rate \((a > 0)\), and \(b\) represents the mRNA decay rate \((b > 0)\). With a steady state assumption (i.e. \(dx/dt = 0\)) have:

\[
\frac{dx}{dt} = a - bx_{ss} = 0.
\]

Thus, only one gene expression measurement \(x_{ss}\) would suffice to initiate the model building process and to estimate the ratio of the rates at a steady state:

\[
\frac{a}{b} = x_{ss}.
\]

Figure 4.1A shows examples of gene expression trajectories that can be obtained based on this assumption. The plot illustrates how 2 different sets of transcription and degradation rates reproduce the same measured steady state expression and, thus, cause parameter identifiability issues. Further resolution of parameters associated with changes in expression under this or any more complex scenarios requires additional experimentation. Time course data, the most common source of information in modeling approaches, can be used for this purpose if it captures a sufficient amount of gene expression dynamics. This scenario is rarely the case due to the typical sparseness of biological data. Zak et al. [Zak03] proposed solving this problem by measuring the decay rate separately. Barenco et al. [Bar06] obtained direct mRNA decay rate measurements to constrain the tumor suppressor transcription factor p53 model while fitting it to the time course data. Decay rate values may also be available in the literature [Nar07; SL14]. However, the reported values should be used with caution since decay rates are known to be condition specific [Gar07]. Moreover, most experimental protocols are invasive and might heavily affect cellular physiology [Mun11].

### 4.2.2 Transcription factor effect

Gene regulation in a cell is modulated by the activity of transcription factors that modulate the rate of gene transcription. Assuming that the transcription factors affect the target gene independently, this modulation can be reflected in (4.2) as follows:

\[
\frac{dx}{dt} = af_1(x_1)f_2(x_2)\cdots f_R(x_R) - bx,
\]  

(4.3)
Figure 4.1 Model building progression. A – initial model including transcription \((a)\) and degradation \((b)\) rates. B – extension of the initial model by incorporating known regulators’ effects. C – model extension by incorporating a condition related effect \(f_u\).
where $a$ is a scaling coefficient, $x_r \ (r = 1, 2, \ldots, R)$ is the expression of one of $R$ transcription factor genes regulating $x$, $f_r(x_r)$ is the regulator influence function which is equal to 1 when no regulation occurs, greater than 1 for activators, and between 0 and 1 for inhibitors. The simplest case of such function is a linear approximation $f_r(x_r) = 1 + c_r x_r$, where $c_r$ is a constant impact factor.

Sparse and noisy time course samples complicate the differentiation between regulators’ expression patterns and, thus, affect the influence function parameter inference. Additional sampling time points or replicates do not guarantee sufficient resolution improvements. Thus, time course data should be supplemented with additional information to estimate the influence coefficients. Experiments where target expression is measured while regulator expression is manipulated can reveal this information.

Regulator knock-out mutant experiments [Zha15; Kor15a] can uniquely define a linear approximation of the influence function $f_r(x_r) = 1 + c_r x_r$. If the regulating transcription factor $x_r$ is an activator with a measured wild-type expression $x_{rWT}$, target gene expression measurements in wild-type ($x_{WT}$) and mutant ($x_{MA}$) conditions allow to approximate the regulator-target dependence with a line (Figure 4.2A):

$$x = x_{MA} + \frac{x_{WT} - x_{MA}}{x_{rWT} x_{MA}} x_r,$$

which allows for the constant impact factor $c_r$ approximation by associating $x_{MA}$ with the scaling coefficient $a$ and rewriting the dependence in a form of the linear influence function:

$$f_{\text{lin}}(r; x_r) = 1 + \frac{x_{WT} - x_{MA}}{x_{rWT} x_{MA}} x_r. \tag{4.4}$$

Similar transformations for the case of inhibitory effects give the following equation for estimating linear regulatory effects:

$$c_r = \frac{x_{WT} - 1}{x_{rWT}} \tag{4.5}$$

Figure 4.1B shows an example of how target’s gene expression $x$ can change depending on whether regulator $x_1$ is its activator, repressor, or has no effect at all.

However, a linear construct is expected to approximate the influence in a range that does not extend far beyond the regulator’s wild type gene expression value. Otherwise, unrealistically high target expression is expected in case of activators and negative expression in case of inhibitors. One way to deal with this issue is to specify a maximum expression value ($x_{max}$) that a target can take when influenced by an activator and fit an exponential relationship to predict the regulation effects (Figure 4.2B):

$$x = x^{max} + (x_{MA} - x^{max}) e^{-p^A_r x_r} = x^{max} . f_{\text{exp}}(x_r), \tag{4.4}$$
where $p_r^A$ can be derived by substituting wild type conditions for $x$ and $x_r$. Overexpression experiments [Pre12; Lia17] present a suitable way for approximating the $x_{\text{max}}$ ($x_{\text{min}}$ for an inhibitor) if the regulator’s overexpression effect is at least several fold larger than the wild-type expression value.

Hill-function approximation [Ges12; Kru10] is another, more biologically relevant, way of representing regulator influence (Figure 4.2C):

$$x = x^M + (x_{\text{max}} - x^M) \frac{x_r^l}{x_r^l + K^l} = x^M \cdot f_{\text{hill}}.$$  

Here the dissociation constant $K$ can be estimated using knock-out mutant ($x^M$) and overexpression ($x_{\text{max}}$) experiment values. The regulator’s protein affinity $I$ can be obtained through additional experiments, for example, through fluorescence correlation spectroscopy in plants [Cla16]. Hence, each additional parameter in the regulator influence function requires an experiment to estimate it.

### 4.2.3 Condition induced effects

A host of transcriptome research studies are interested in mechanisms governing organism’s response to a certain condition like biotic or abiotic stress in plants [Lon10; Kil07; IP11; Lin11] or pathogen infection in single cells or animals [Gut05; Wu14]. Equation 4.3 would be sufficient in describing gene expression dynamics over time under such condition if the full set of regulators is known, which is almost never the case at the current stage. Thus, a model has to account for factors that remain unknown by aggregating them in a corresponding function:

$$\frac{dx}{dt} = a f_u(t) \prod_{r=1}^{R} f_r(x_r) - bx, \quad (4.6)$$

where $R$ is the number of known regulators, and $f_u(t)$ is a function representing the currently unknown influencing factors which change their activity under a condition of interest. An example of such influencing factor could be a change in a currently unknown condition induced transcription factor that binds to the target gene’s promoter. $f_u(t)$ takes positive values and turns into 1 in wild-type conditions. The shape of $f_u(t)$ can be obtained using Gaussian process approximation [Ji16; Gao08; Ras06]. Another approach to approximating the unknown effect would be to model it as a continuous shift to a new condition induced equilibrium:

$$u(t) = u_T \frac{1}{\left( \frac{\tau}{t} \right) + 1},$$
Figure 4.2 Regulator influence function approximations. Experiments for estimating parameters associated with the influence function $f_r(x_r)$ when using (A) linear (lin), (B) exponential (exp), and (C) Hill-function (hill) approximations.
where $u_T$ represents an impact coefficient ($u_T > -1$), $r$ quantifies how fast the transition between wild-type and condition induced steady states occurs ($r > 0$), and $\tau$ accounts for the transition delay (Figure 4.3). Because $u(t)$ turns to 0 when no condition is applied, an adjustment like $f_u(t) = 1 + u(t)$ is needed to represent the unknown regulatory effect function. Figure 4.1C illustrates the effects of $f_u(t)$ on simulated gene expression trajectory in (4.6). Parameters shaping $u(t)$ can be estimated by fitting the model to time course data under wild-type and the condition of interest.

Figure 4.3 Unknown influencing factors effects approximation with a sigmoid function. $u_T$ – scale coefficient, $r$ – rate coefficient, and $\tau$ – delay coefficient.

Additional experiments can help shaping the response to different levels of the applied condition if the condition levels are quantifiable and the sigmoid function is used. In this case the parameters shaping $u(t)$ might be affected by the condition level $S$:

$$u(t, S) = \frac{u_T(S)}{\left(\frac{\tau(S)}{t}\right)^{r(S)} + 1}.$$  \hspace{1cm} (4.7)

We will concentrate on shaping the magnitude parameter $u_T(S)$ while the condition dependence of other two parameters can be established in a similar fashion. Wild-type and condition induced gene expression values allow for a linear approximation through a range of condition levels, which might, in some cases, be significantly far from reality. A more reliable approximation can be obtained by sampling gene expression at intermediate condition levels. However, transcriptome measurements are resource consuming, so the condition levels should be chosen in an efficient manner to produce maximum information with minimum experimentation. If the organism of interest exhibits a quantifiable change in size, shape, or other easily accessible physiological parameters under the condition of interest, a faster and less expensive procedure of phenotyping can be used under a set of intermediate condition levels (e.g. micronutrient content level or pathogen load) to judge whether linear approximation captures the condition effect. Phenotyping results can give clues on which condition levels to choose for transcriptome measurement experiments. Figure 4.4
shows a hypothetical example of selecting the most informative sampling point and the magnitude response function based on the results of phenotyping experiments.

Figure 4.4 Condition level effect dependence modeling. $S_m$ — measured condition level of the initial experiment, $S_p$ — proposed condition level for gene expression measurements based on the results of physiological parameter measurements during phenotyping experiments.

4.3 Model fitting

Guidelines for the additional experimentation proposed in the previous section illustrate ways for increasing model complexity and, thus, its descriptive and predictive power. However, determining whether the collected data is sufficient for a given model complexity and combining various types of datasets to train the model are not trivial tasks considering that each type of experimentation has an associated measurement noise. We propose parameter identifiability analysis as a criterion for data sufficiency when scaling a model up and a Bayesian inference approach for model parameter optimization.

4.3.1 Model scalability assessment

An increase in model complexity for extending its predictive and descriptive power is usually associated with an increase in the number of model parameters. These additional parameters allow capturing the biological process of interest with more details. Hence, an ability to uniquely identify all parameters in the updated model is a key to a better understanding of the associated characteristic of the modeled biological process. However, additional parameters increase the risk of overfitting the model to the training data. In this case, wide ranges of parameter values can produce an equally good model fit, which hinders biological significance of the estimated parameter values. Overfitted models do not generalize as well as less complex structures, which makes the process of scaling the model up detrimental in terms of prediction power improvement and should, thus, be avoided. We
propose the parameter identifiability analysis for assessing biological relevance of model parameters in a given equation structure.

We require all parameters in the fitted model to be uniquely identifiable after the model equation structure is scaled up. A presence of non-identifiable parameters tends to cause unreliable model predictions [Rau11]. A parameter is identifiable if the confidence interval of its estimate is finite [Rau09]. A lack of experimental samples and an inherent measurement noise can cause a situation when almost any value of some parameter gives an equally good model fit to the data leading to an infinite confidence interval and non-identifiability. For example, any value of decay rate $b$ can be compensated with a corresponding value of the transcription rate $a$ in (4.2) if $x_{ss}$ is the only sample aside of the wild type measurement. Several methods such as Differential Algebra Identifiability of Systems (DAISY) [Sac03], Exact Arithmetic Rank (EAR) [Kar12], and Profile Likelihood (PL) [Rau09] have been used to detect parameter non-identifiability. Among these methods, PL is the only one that uses experimental data for detecting non-identifiability. We propose using the results of PL analysis for model discrimination when increasing model complexity based on additional data.

Profile Likelihood (PL) algorithm is capable of spotting structural non-identifiability, which can not be resolved with additional data, and the practical non-identifiability, which is the result of insufficient or noisy data (Figure 4.5). The algorithm fits the model to the data for a range of fixed values of the parameter of interest ($\theta_i$). The resulting collection of calculated cost function values $\mathcal{P}L(\theta_i)$ is then tested for a significant change to verify that deviations in $\theta_i$ affect the model fit. If $\mathcal{P}L(\theta_i)$ does not cross the significance threshold (dashed line on Figure 4.5) for the range of its values, the confidence interval for the parameter is assumed to be infinite and the parameter is called non-identifiable.

![Figure 4.5](image-url)

**Figure 4.5 Profile likelihood trajectories for different parameter identifiability cases.** Solid line represents the values of error function used for model fitting. Dashed line represents the significant level.
4.3.2 Parameter estimation

Bayesian inference methods allow to aggregate different sources of data by shaping prior distributions of the corresponding parameters before fitting a model to the corresponding time course. Each experiment that we proposed allows for obtaining mean and standard deviation estimates for a specific parameter. An assumption on experimental error distribution (e.g. Gaussian or Poisson) would allow to construct the corresponding prior distribution. The model fitting algorithm will sample parameter values from the corresponding prior distributions while minimizing the sum of squared differences between the time course data and gene expression pattern produced by the model. Parameter values from the regions that are far from the experimental measurements are highly unlikely to be sampled which would ensure that the model describes both the time course data and the results of the additional experiments.

Due to a nonlinear nature of the differential equations governing gene expression dynamics we suggest using the latest generation of Bayessian inference based parameter estimation algorithms, namely the Differential Evolution Adaptive Metropolis (DREAM) software package [Vru09]. It has been demonstrated that DREAM outperforms similar software in nonlinear, multimodal, and high dimensional problems [LV12]. This package has been used in solving inference problems in different fields including chemistry [Owe12], ecology [Bar11b], biology [Coe11; Zao14], and epidemiology [Mar12b]. In addition to the general MCMC scheme, DREAM software allows to account for unequal measurement variance (heteroskedasticity) which is expected in transcriptome measurement time course experiments due to an inherent exponential transformation when calculating expression values for qRT-PCR [SL08] and Microarray data [Chu02] and significant changes in expression values over the time courses. The outcome of the DREAM algorithm is a set of posterior parameter distributions exhibiting a consensus between the experimental data and the model fit in terms of inferred parameter values and the associated variances that consolidate sample variation, measurement noise, and the discrepancy between the biological process and its model representation. These posterior distributions can then be used to calculate estimates and credible intervals for parameter values and model predictions [Smi13].

4.4 Conclusions

In this chapter, we presented a methodology for sequential increase in gene expression dynamic model complexity by aggregating different types of experimental data. The methodology provides a flexible framework for accumulating the existing knowledge of a biological process of interest at the transcriptome level and proposes efficient ways for expanding this knowledge through additional experimentation. This paper aims to facilitate modeling efforts in the studies where time course experiments have been implemented and the key regulatory connections have been identified. For
example, iron deficiency response in *A. thaliana* [Kor15a] can be modeled after running a sufficient number of additional experiments to make all parameters identifiable. The proposed methodology can also be readily applied to osmotic, cold, and drought stress datasets [Kil07] in the same plant after inferring gene regulatory network graph experimentally or computationally. The stress datasets contain time course measurements under wild-type and stress induced conditions each of which can be modeled with (4.6). Equation 4.7 with the corresponding additional experimentation would allow to further increase the prediction space of the resulting model (Figure 4.6).

**Figure 4.6 Model prediction example.** A hypothetical example of constructing a predictive model from time course samples in Wild-Type (WT) and in a Condition Induced (CI) state.
5.1 Background

Plants require iron for growth and development. Severe or extended iron deficiency leads to chlorosis, reduced yield, and plant death [FF81; Cla88; SS80]. The onset of iron deficiency triggers a regulatory response involving the cascaded activation or repression of many genes, leading to the production of key proteins that modify physiological processes that make iron more available [Din08; Buc09; Lon10; Kor15b]. In Arabidopsis thaliana and other dicots, this physiological response, defined as the Strategy I response, is characterized by an increase in rhizosphere acidification, iron reductase activity, and iron transport across epidermal cell membranes [Mar86]. All of the genes involved in the Strategy I response are known to be transcriptionally induced under iron deficiency, primarily by bHLH transcription factors including FIT [CG04; Bau04; Jak04; Bau07; Yua08; Iva12]. Additional transcription factors, including POPEYE (PYE), MYB10, and MYB72, are known to modulate the mobilization and sequestration of iron and other physiological responses throughout the root.
and plant [Ram06b; Lon10; Pal13b; Sel15; Zha15; Li16; Lia17]. Mutants of many of these key iron regulators display physiological phenotypes when exposed to extreme iron deficient conditions ranging from intolerance to iron deficiency in fit and pye mutants to tolerance in bts mutants [CG04; Bau04; Jak04; Lon10; Sel15]. Intriguingly, knockdown mutants of bts are more tolerant to iron deficiency yet full knockout mutants are lethal [Lon10; Sel15]. Other data have shown that gradual changes in iron content result in more subtle changes in phenotype [Gru13; Leš17]. Some mutants exhibit noticeable phenotypic alterations only when present in combination, for example in the case of double mutants myb10myb72 and bhlh100bhlh101 [Pal13b; Siv12]. These and other results suggest that genes involved in the iron deficiency response are under precise control. By aggregating existing experimental data in a descriptive and predictive transcriptome response representation, we can further address agricultural concerns for improving plant adaptation to poor soils or for increasing iron content within a plant.

Recent studies have identified a collection of transcription factors, namely COL4, ETF9, ASIL2, MYB55, bHLH34, bHLH104, and bHLH105 (ILR3), that modulate the expression of a set of known iron response genes, namely bHLH39, bHLH101, bHLH115, PYE, MYB10, MYB72, and BTS [Kor15a; Zha15; Li16]. However, it remains unclear how these modulating transcription factors, separate or in combination, enhance or suppress the molecular phenotypes associated with the well studied iron response genes. Understanding this control architecture requires a quantitative description that captures causal influences across these transcription factors over time. Mathematical modeling can provide this quantitative description of relationships and can be used as a tool to make in-silico predictions concerning the iron deprivation response under novel genetic perturbations. Modeling can also reveal activity patterns of regulatory genes that have not yet been identified, which would facilitate identification of new iron deficiency response modulating genes.

Ordinary Differential Equations (ODEs) provide a suitable framework for developing a dynamic gene regulatory network model of the iron deprivation response in A. thaliana. A collection of research and review papers describe various approaches for modeling gene expression using this framework [Gut05; Car09; Pal13a; Wit09; SC10; PK13; Pol09; Bre08]. ODEs have been applied to model various processes in A. thaliana including Zinc uptake regulation [CCK12], flower organ formation [Mou10] and specification [SC10], and Auxin signalling [Ver11]. ODEs have distinct advantages over static representations such as Bayesian networks [Fri00; D’h00; MM99; Doj06] and other dynamic representations such as Boolean networks [SA13; CL12; SVC08; Cha06; Li06; GG10; Wan12]. These advantages include, but are not limited to, increased diversity of possible mathematical constructs, flexibility in equation complexity, and availability of diverse approaches for parameter estimation, uncertainty quantification, dynamics simulation, and model analysis.

We propose a dynamic model describing regulatory interactions of genes involved in the iron deficiency response. Specifically, we focused on the interactions between several proteins that are known to influence phenotypic changes under iron deficient conditions (bHLH39, bHLH101,
bHLH115, PYE, MYB10, MYB72, and BTS) and transcription factors (COL4, ETF9, ASIL2, and MYB55) [Kor15a] and bHLH34, bHLH104, and bHLH105 (ILR3) [Zha15; Li16]) that have been shown to modulate the response of the corresponding genes. We collected time course gene expression data under iron sufficiency and iron deficiency, mRNA decay rate measurements, and the results of single knockout mutant experiments. Information from these experiments were combined using a Bayesian parameter estimation approach to model optimization. Parameter identifiability analysis allowed us to balance model complexity with the available experimental data. We applied uncertainty quantification analysis to estimate credible regions associated with simulated model dynamics. We detected sample heteroscedasticity in experimental data and compensated for its effects on credible regions estimates. The trained model was able to describe the expression profiles of wild type and single mutant experiments within the range of calculated uncertainty and predict the expression profiles of affected genes in double mutant backgrounds. This model can serve as a first contribution towards a comprehensive mathematical representation of the iron deficiency response. Such a representation would ultimately be capable of replacing resource consuming experimentation with in-silico simulations for testing hypotheses on aspects of iron induced plant behavior. Overall, this approach provides a useful example of how diverse datasets can be used to establish a foundation for mathematical descriptions of stress responses in plant systems.

5.2 Results

5.2.1 Model structure

We started by identifying regulators that affect the genes known to influence phenotypic changes under iron deficient conditions. The available knowledge about iron induced gene interactions [Lia17; Li16; Zha15; Kor15a; Pal13b; Wan13; Yua08; CG04], summarized in Table 5.1, allowed us to identify a set of common regulators for those genes. We ran additional experiments to test each possible regulator-target connection between the identified regulators and the genes of interest. Statistical analysis of the experimental data is presented in subsection 5.3.1, and the inference results are summarized in the gene regulatory network on Figure 5.1. We used the scalable dynamic modeling framework from Chapter 4 to construct a system of Ordinary Differential Equations (ODEs) for capturing expression dynamics of the genes in the gene regulatory network under iron deficiency and iron sufficiency. For each gene $x_n$ ($n = 1, 2, \ldots, 14$) the corresponding ODE describes a change in expression $d x_n$ over a time interval $d t$, which is influenced by a host of factors. These factors include mRNA transcription rate $a_n$, decay rate $b_n$, the effect of known regulators $f_n(x_i)$, and the effect of currently unknown factors induced by iron deficiency $f_{un}(t)$. The corresponding terms describe the balance between factors that increase and decrease gene expression in regulators $x_{n_{reg}}^r$ ($n = 1, 2, \ldots, 7$).
and targets $x_n^{tg}$ ($n = 8, 9, \ldots, 14$) as follows:

\[
\frac{dx_n^{reg}}{dt} = a_n f_{u_n} - b_n x_n^{reg}, 
\]

(5.1)

\[
\frac{dx_n^{tg}}{dt} = a_n f_{u_n} \prod_{i \in \text{reg}_n} f_n(x_i^{reg}) - b_n x_n^{tg}, 
\]

(5.2)

where reg$_n$ is the set of regulators affecting the expression of target $x_n^{tg}$, $f_n(x_i^{reg})$ is the effect of a regulator on its target, and $f_{u_n}(t)$ is a cumulative effect of unknown factors. We chose a sigmoid function to represent the unknown factors effect $f_{u_n}(t)$ and the linear approximation $f_n(x_i)$ to model the effects of a regulator on its target as presented in (5.3) and (5.4) and explained in subsection 5.3.2.

\[
f_{u_n}(t) = 1 + \frac{u_{T_n}}{\left(\frac{T_n}{t}\right)^{\tau_n} + 1},
\]

(5.3)

\[
f_n(x_i) = 1 + c_n x_i.
\]

(5.4)

**Table 5.1 List of known iron related genetic interactions.** Interaction types: ac – gene activation, in – gene inhibition, pd – protein-DNA interaction

<table>
<thead>
<tr>
<th>Protein/Gene A</th>
<th>Int.</th>
<th>Protein/Gene B</th>
<th>Validation Method</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASIL2</td>
<td>in bHLH101</td>
<td>mutant qRT-PCR</td>
<td>[Kor15a]</td>
<td></td>
</tr>
<tr>
<td>ASIL2</td>
<td>in bHLH115</td>
<td>mutant qRT-PCR</td>
<td>[Kor15a]</td>
<td></td>
</tr>
<tr>
<td>ASIL2</td>
<td>in bHLH39</td>
<td>mutant qRT-PCR</td>
<td>[Kor15a]</td>
<td></td>
</tr>
<tr>
<td>ASIL2</td>
<td>in BTS</td>
<td>mutant qRT-PCR</td>
<td>[Kor15a]</td>
<td></td>
</tr>
<tr>
<td>ASIL2</td>
<td>in MYB10</td>
<td>mutant qRT-PCR</td>
<td>[Kor15a]</td>
<td></td>
</tr>
<tr>
<td>ASIL2</td>
<td>in MYB72</td>
<td>mutant qRT-PCR</td>
<td>[Kor15a]</td>
<td></td>
</tr>
<tr>
<td>ASIL2</td>
<td>in PYE</td>
<td>mutant qRT-PCR</td>
<td>[Kor15a]</td>
<td></td>
</tr>
<tr>
<td>ASIL2</td>
<td>pd PYE</td>
<td>Y1H</td>
<td>[Kor15a]</td>
<td></td>
</tr>
<tr>
<td>bHLH104</td>
<td>ac bHLH101</td>
<td>mutant and overexpression qRT-PCR</td>
<td>[Zha15]</td>
<td></td>
</tr>
<tr>
<td>bHLH104</td>
<td>ac bHLH101</td>
<td>mutant and overexpression qRT-PCR</td>
<td>[Li16]</td>
<td></td>
</tr>
<tr>
<td>bHLH104</td>
<td>pd bHLH101</td>
<td>ChIP-qPCR</td>
<td>[Zha15]</td>
<td></td>
</tr>
<tr>
<td>bHLH104</td>
<td>pd bHLH101</td>
<td>Y1H and transient expression assay</td>
<td>[Li16]</td>
<td></td>
</tr>
<tr>
<td>bHLH104</td>
<td>ac bHLH39</td>
<td>mutant and overexpression qRT-PCR</td>
<td>[Zha15]</td>
<td></td>
</tr>
<tr>
<td>bHLH104</td>
<td>ac bHLH39</td>
<td>mutant and overexpression qRT-PCR</td>
<td>[Li16]</td>
<td></td>
</tr>
<tr>
<td>bHLH104</td>
<td>pd bHLH39</td>
<td>ChIP-qPCR</td>
<td>[Zha15]</td>
<td></td>
</tr>
<tr>
<td>bHLH104</td>
<td>ac BTS</td>
<td>mutant and overexpression qRT-PCR</td>
<td>[Zha15]</td>
<td></td>
</tr>
<tr>
<td>bHLH104</td>
<td>ac MYB10</td>
<td>mutant and overexpression qRT-PCR</td>
<td>[Zha15]</td>
<td></td>
</tr>
<tr>
<td>bHLH104</td>
<td>ac MYB10</td>
<td>mutant and overexpression qRT-PCR</td>
<td>[Li16]</td>
<td></td>
</tr>
<tr>
<td>bHLH104</td>
<td>ac MYB72</td>
<td>mutant and overexpression qRT-PCR</td>
<td>[Zha15]</td>
<td></td>
</tr>
</tbody>
</table>

Continued on the next page
Table 5.1 – continued from the previous page

<table>
<thead>
<tr>
<th>Protein/Gene A</th>
<th>Int.</th>
<th>Protein/Gene B</th>
<th>Validation Method</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>bHLH104</td>
<td>ac</td>
<td>MYB72</td>
<td>mutant and overexpression qRT-PCR</td>
<td>[Li16]</td>
</tr>
<tr>
<td>bHLH104</td>
<td>ac</td>
<td>PYE</td>
<td>mutant and overexpression qRT-PCR</td>
<td>[Zha15]</td>
</tr>
<tr>
<td>bHLH104</td>
<td>pd</td>
<td>PYE</td>
<td>ChIP-qPCR</td>
<td>[Zha15]</td>
</tr>
<tr>
<td>bHLH115</td>
<td>ac</td>
<td>bHLH101</td>
<td>mutant and overexpression qRT-PCR</td>
<td>[Lia17]</td>
</tr>
<tr>
<td>bHLH115</td>
<td>pd</td>
<td>bHLH101</td>
<td>transient expression assay</td>
<td>[Lia17]</td>
</tr>
<tr>
<td>bHLH115</td>
<td>ac</td>
<td>bHLH39</td>
<td>mutant and overexpression qRT-PCR</td>
<td>[Lia17]</td>
</tr>
<tr>
<td>bHLH115</td>
<td>pd</td>
<td>bHLH39</td>
<td>transient expression assay</td>
<td>[Lia17]</td>
</tr>
<tr>
<td>bHLH115</td>
<td>ac</td>
<td>BTS</td>
<td>mutant and overexpression qRT-PCR</td>
<td>[Lia17]</td>
</tr>
<tr>
<td>bHLH115</td>
<td>ac</td>
<td>MYB10</td>
<td>mutant and overexpression qRT-PCR</td>
<td>[Lia17]</td>
</tr>
<tr>
<td>bHLH115</td>
<td>ac</td>
<td>MYB72</td>
<td>mutant and overexpression qRT-PCR</td>
<td>[Lia17]</td>
</tr>
<tr>
<td>bHLH115</td>
<td>ac</td>
<td>MYB72</td>
<td>mutant and overexpression qRT-PCR</td>
<td>[Lia17]</td>
</tr>
<tr>
<td>bHLH115</td>
<td>pd</td>
<td>MYB72</td>
<td>transient expression assay</td>
<td>[Lia17]</td>
</tr>
<tr>
<td>bHLH115</td>
<td>ac</td>
<td>MYB72</td>
<td>mutant and overexpression qRT-PCR</td>
<td>[Li16]</td>
</tr>
<tr>
<td>bHLH34</td>
<td>ac</td>
<td>bHLH101</td>
<td>mutant and overexpression qRT-PCR</td>
<td>[Li16]</td>
</tr>
<tr>
<td>bHLH34</td>
<td>pd</td>
<td>bHLH101</td>
<td>Y1H and transient expression assay</td>
<td>[Li16]</td>
</tr>
<tr>
<td>bHLH34</td>
<td>ac</td>
<td>bHLH39</td>
<td>mutant and overexpression qRT-PCR</td>
<td>[Li16]</td>
</tr>
<tr>
<td>bHLH34</td>
<td>ac</td>
<td>MYB10</td>
<td>mutant and overexpression qRT-PCR</td>
<td>[Li16]</td>
</tr>
<tr>
<td>bHLH34</td>
<td>ac</td>
<td>MYB72</td>
<td>mutant and overexpression qRT-PCR</td>
<td>[Li16]</td>
</tr>
<tr>
<td>COL4</td>
<td>ac</td>
<td>bHLH115</td>
<td>mutant qRT-PCR</td>
<td>[Kor15a]</td>
</tr>
<tr>
<td>COL4</td>
<td>ac</td>
<td>bHLH39</td>
<td>mutant qRT-PCR</td>
<td>[Kor15a]</td>
</tr>
<tr>
<td>COL4</td>
<td>ac</td>
<td>MYB10</td>
<td>mutant qRT-PCR</td>
<td>[Kor15a]</td>
</tr>
<tr>
<td>COL4</td>
<td>ac</td>
<td>MYB72</td>
<td>mutant qRT-PCR</td>
<td>[Kor15a]</td>
</tr>
<tr>
<td>ETF9</td>
<td>in</td>
<td>bHLH115</td>
<td>mutant qRT-PCR</td>
<td>[Kor15a]</td>
</tr>
<tr>
<td>ETF9</td>
<td>in</td>
<td>BTS</td>
<td>mutant qRT-PCR</td>
<td>[Kor15a]</td>
</tr>
<tr>
<td>ETF9</td>
<td>in</td>
<td>MYB10</td>
<td>mutant qRT-PCR</td>
<td>[Kor15a]</td>
</tr>
<tr>
<td>ETF9</td>
<td>in</td>
<td>MYB72</td>
<td>mutant qRT-PCR</td>
<td>[Kor15a]</td>
</tr>
<tr>
<td>FIT</td>
<td>ac</td>
<td>MYB10</td>
<td>mutant microarray</td>
<td>[CG04]</td>
</tr>
<tr>
<td>FIT</td>
<td>ac</td>
<td>MYB72</td>
<td>mutant microarray</td>
<td>[CG04]</td>
</tr>
<tr>
<td>ILR3</td>
<td>ac</td>
<td>bHLH101</td>
<td>mutant qRT-PCR</td>
<td>[Zha15]</td>
</tr>
<tr>
<td>ILR3</td>
<td>pd</td>
<td>bHLH101</td>
<td>ChIP-qPCR</td>
<td>[Zha15]</td>
</tr>
<tr>
<td>ILR3</td>
<td>pp</td>
<td>bHLH115</td>
<td>Y2H and BiFC</td>
<td>[Lia17]</td>
</tr>
<tr>
<td>ILR3</td>
<td>ac</td>
<td>bHLH39</td>
<td>mutant qRT-PCR</td>
<td>[Zha15]</td>
</tr>
<tr>
<td>ILR3</td>
<td>pd</td>
<td>PYE</td>
<td>ChIP-qPCR</td>
<td>[Zha15]</td>
</tr>
<tr>
<td>MYB55</td>
<td>in</td>
<td>bHLH101</td>
<td>mutant qRT-PCR</td>
<td>[Kor15a]</td>
</tr>
<tr>
<td>MYB55</td>
<td>in</td>
<td>BTS</td>
<td>mutant qRT-PCR</td>
<td>[Kor15a]</td>
</tr>
</tbody>
</table>

For each gene in the regulatory network (Figure 5.1) we obtained time course expression data at 0, 12, 24, 36, 48, 60, and 72 hours following transfer to iron deficient and iron sufficient media using quantitative Real-Time PCR (qRT-PCR). We used a parameter identifiability analysis to as-
Figure 5.1 Modeled regulatory connections. Connections represent significant regulatory effects based on previous publications and the results of regulator knockout experiments. Green lines stand for activation effect, red lines indicate inhibition.

We estimated mRNA decay rates by quantifying the change in gene expression over time under iron sufficient and iron deficient conditions after inhibiting transcription (see subsection 5.3.2.5). We obtained decay rate estimates with the associated uncertainty and incorporated this information as prior knowledge to stabilize the parameter $b_n$ when fitting the model to the time course data. Statistical analysis showed that decay rates were not significantly different between iron deficient and iron sufficient conditions for the regulators but were significantly different for all targets except for bHLH115 (Figure 5.2). The mechanisms associated with that difference were unknown, but had to be accounted for in the model. We fixed the regulators’ decay rate values under both iron conditions and proposed a function of decay rate change under iron deficiency for each of the 7 targets. The data showed that an increase in the decay rate was associated with an increase in mRNA abundance due to a strong correlation between log fold change in expression and a change in the measured decay rate (Figure 5.3). We also assumed that the change in decay rate only occurs if the
corresponding change in gene expression is present under iron deficiency. Thus, the only parameter that we fit in the statistical model was the slope $m$ which appeared to be significantly different from 0 ($p < 0.001$) with $R^2 = 0.8$. Hence, we modeled the change in decay rate for each target under iron deficient conditions as follows:

$$b_n^{Fe}(t) = b_n^{Fe+} + m \log_{10} \left( \frac{x_n^{Fe}(t)}{x_n^{Fe}(0)} \right),$$

(5.5)

where $b_n^{Fe+}$ is the decay rate under iron sufficiency.

Figure 5.2 Decay rate data analysis. Estimates of decay rate values and the corresponding standard errors. Genes with significantly different estimates under iron sufficiency and iron deficiency are represented with 2 corresponding uncertainty ranges for the decay rate value.

5.2.2 Model fitting

We used a-priori probability distributions for model parameters as inputs to the model fitting software. These distributions were computed based on additional experimentation or inferred based on domain experts’ knowledge. We fit the model to experimental data using the Differential Evolution Adaptive Metropolis (DREAM) software package [Vru09]. DREAM is a Markov Chain Monte Carlo (MCMC) approach that produces parameter estimates that provide the best fit of the model to a collection of experimental data. DREAM can also generate a posterior distribution for each parameter and account for the uneven sample variance (i.e. heteroscedasticity) that we detected in experi-
Figure 5.3 Decay rate change analysis. Linear regression for accessing the relationships between decay rate change and the fold change of gene expression under iron deficiency conditions.

mental data (see subsection 5.3.3). The obtained posterior distributions (Appendix A.2) showed the ranges of parameter uncertainty that remained after fitting the model to the time course data as illustrated in Figure 5.4 for parameters \(a\) and \(b\) as examples. We observed a significant decrease in uncertainty for the parameters that were not estimated through additional experimentation and adjustments in uncertainty regions for the remaining parameters. Posterior parameter uncertainty allowed us to assess confidence in model predictions. We obtained 95% credible interval regions for gene expression trajectories by propagating the parameter uncertainty through model simulations (see subsection 5.3.4). Figure 5.5 and Figure 5.6 illustrate expression patterns for regulators and targets, respectively, based on the model fit to experimental data and the calculated 95% credible interval regions. The graphs show that the experimental measurements fall within the range of the model prediction uncertainty, meaning that the model was able to successfully describe the time course data. Similarly, the model was able to capture the effects of single regulator knockout mutations on targets as shown in Figure 5.7 for MYB10 and in Appendix A.3 for all the targets. The simulated trajectories illustrate gene expression progression over time in a single regulator mutant background under iron deficiency.

5.2.3 Assessment of the unknown effects function

Parameter posterior distributions provided avenues for characterizing the behavior of currently unknown factors that influence expression dynamics of genes in our network. We used the posterior
Figure 5.4 Parameter uncertainty range adjustments through model fitting. A – reduction in uncertainty for a parameter with no prior knowledge (a). B – uncertainty adjustment for a parameter for which prior knowledge came from additional experimental measurements (b).
Figure 5.5 Time course modeling for regulators. Model fit for gene expression dynamics of regulators. Best fit under iron deficiency (solid blue line) and under iron sufficiency (solid red line) are shown. Shaded region indicate 95% credible region. Error bars indicate experimental measurements.
Figure 5.6 Time course modeling for targets. Model fit for gene expression dynamics of targets. Best fit under iron deficiency (solid blue line) and under iron sufficiency (solid red line) are shown. Shaded region indicate 95% credible region. Error bars indicate experimental measurements.
Figure 5.7 Single mutant predictions. Model prediction and experimental results for different single mutant experiments for MYB10 under iron deficiency. Lines show the best fit trajectories produced by the model. Shaded regions are the 95% credible regions of model prediction. Error bars indicate experimental measurements from corresponding experiment. A single measurement at 72 hours of exposure to iron deficiency was collected for each mutant experiment.
distributions to gain insight into the patterns of unknown effects represented by $f_{u_n}(t) = 1 + u_n(t)$ and to test whether these influences were significant. We assumed that the aggregated dynamics of unknown regulators were concentrated in $u_n(t)$ similar to $x_i$ driving the effect of a known regulator in $f_n(x_i) = 1 + c_n x_i$. Hence, we calculated 95% credible regions for $u_n(t)$ where the gene expression trajectories of currently unknown regulators were assumed to reside (Appendix A.4). The graphs show that the effect of unknown factors is significantly different from 0 for almost all genes which suggests the existence of currently unknown transcription factors that modulate the activity of the regulators and targets under iron deficiency. We used regulatory relationships hypothesized in the literature, but not included in the model, to access whether credible regions for $u_n(t)$ of the influenced genes capture the expression patterns of the corresponding regulators.

Although bHLH115 is presented in our model as a target, it has also been shown to control the expression of other targets [Lia17]. Therefore, we obtained the credible region of $x_{bHLH115}^{tg}(t)$ and overlaid it with the credible regions for the unknown effect function $u_n(t)$ of the influenced genes (Figure 5.8A,B for MYB10 and MYB72 and Appendix A.5 for all influenced genes). While the figures do show consistency between bHLH115 pattern and the unknown factors predictions, it is clear that a part of the unknown factor uncertainty is attributed to something other than bHLH115. MYB10 and MYB72 are also known to be regulated by the transcription factor FIT [CG04] which was not included in the original analysis but serves as an excellent candidate for validation. We expected that FIT regulatory effect should be captured by the corresponding unknown effects functions $u_{MYB10}(t)$ and $u_{MYB72}(t)$ which we previously compared with the credible region of bHLH115. We collected qRT-PCR time course measurements for FIT to test this hypothesis. Figure 5.8C,D shows that a scaled version of the FIT expression pattern was mostly covered by the predicted credible regions of $u_{MYB10}(t)$ and $u_{MYB72}(t)$, yet a rise in $u_{MYB72}(t)$ between 12 and 24 hours may indicate another unknown regulator with an elevated activity at that time interval. These results show that the unknown effect function credible regions for other genes can narrow down the search space for a possible regulator if the corresponding expression pattern samples are available.

### 5.2.4 Model predictions

We tested its ability to predict the combinatorial influence of multiple regulators on specific targets. We sampled the expression of target genes in available double mutants at 72 hours after exposure to iron deficiency. We measured gene expression of MYB10, MYB72, bHLH39, and bHLH101 in the bhlh104ilr3 background and PYE and BTS in the bhlh34bhlh104 background. We then modeled the results of the corresponding experiments by setting the expression of the knocked out regulators to 0 and used the parameter posteriors to generate mean trajectories and credible intervals for the targets. The model predictions for double mutant experiments along with the experimental measurements results are presented in Figure 5.10. In four out of six cases, mutant experiment results were within
Figure 5.8 Unknown effects patterns for MYB10 and MYB72. 95% credible regions for expression patterns of potential regulators of MYB10 and MYB72 overlaid with a scaled versions of the known influencing genes. A, B – comparison with the credible region for bHLH115; C, D – comparison with FIT expression measurements.
the credible interval obtained from the model (Figure 5.10A-D). For \textit{bHLH101} (Figure 5.10E), the experimental result for the double knockdown was slightly outside the credible interval while the trend was similar, which might suggest an underestimation of the credible interval. This underestimation could have been the result of modeling with the assumption of null alleles and no associated regulator expression variance while using mutants that did not have a complete loss of function. For \textit{BTS} (Figure 5.10F), the double mutant \textit{bhlh34bhlh104} model prediction indicated a more significant reduction in expression level compared to the wild type than was obtained experimentally. Experimental results also showed that the reduction in expression in the double mutants was similar to that of single mutants. This implies that the effect was due to gene interaction between the regulators because removing one of the regulators is not different from removing both of them. Thus, \textit{bHLH34} and \textit{bHLH104} are likely involved in the same pathway regulating \textit{BTS}. The model did not capture these mechanisms due to the assumption of independent regulatory action. We made this assumption because gene interaction effects including synergism and epistasis [Phi08; PV15; DS01] are less likely to occur [Cos11] but require a significant amount of additional experimental data to estimate the corresponding additional model parameters [Van14].

We also observed a possible synergistic effect while inspecting the results of other double mutant experiments. Single regulator knockout mutants \textit{bhlh104}, \textit{ilr3}, and \textit{bhlh34} did not cause a significant change in \textit{bHLH115} expression, so we did not model the corresponding influences. However, double mutants \textit{bhlh34bhlh104} and \textit{bhlh104ilr} caused a significant increase in \textit{bHLH115} expression (Figure 5.9). Similar effects were previously reported for transcription factors GCN5 and CLV1 acting on ethylene response pathway mediators ERF1 and EBF2 [PV15] although the mechanism behind this regulation remains unknown. Assuming high levels of gene duplication [PP09; MI99], we hypothesize that \textit{bHLH104} provides functional redundancy for ILR3 and \textit{bHLH34} when regulating \textit{bHLH115}, however, further work is needed to unveil the exact mode of gene interactions.

\textbf{Figure 5.9} Knock-out mutant results for the regulators of \textit{bHLH115}. \textit{bHLH115} expression in wild type, single and double knock-out mutant backgrounds. * – significant difference with wild type (WT).
5.3 Methods

5.3.1 Establishing regulatory connections

We performed a set of single regulator knock-down mutant experiments to establish regulatory connections between genes in Figure 5.1. We reused some samples from our previous work [Kor15a] while adding new replicates and connections to test. Each target’s expression was measured in each regulator’s mutant background after exposing mutant and wild type plants to 72 hours of iron deficiency. A significant difference in expression in the mutant background compared to the wild type suggested a regulatory relationship between two genes. The overall number of significance tests to perform for a network with 7 regulators and 7 targets was 49. Such number of simultaneously tested hypothesis increases chances of false positive connections and requires a compensation procedure to be run. Hence we separated experiments into 3 groups which did not share any samples in common and ran Holm-Bonferoni compensation for multiple comparisons [Hol79] in each group after obtaining the initial p-values. Moreover, we did not reject a connection if it passed the $\alpha = 0.05$ threshold for our initial significance test and was also found significant in [Zha15; Li16], because some of the connections were previously tested by the corresponding labs. The decision making diagram is shown on Figure 5.11.

We also looked at whether the assumption on uniform variance was satisfied before running the tests for statistical significance. The plot on Figure 5.12a shows that sample standard deviation grows with an increase in expression value, meaning the data is heteroscedastic [Whi80]. We log-transformed the expression data to deal with heteroscedasticity (Figure 5.12b).

Additionally, samples for some regulators were assembled from 2 measurements of 2 samples with a 2 year difference. For other regulators, 3 samples were collected. With all that in mind, we constructed a statistical model that tests for the effect significance of a regulator knock-out ($M_i, i = 1, 2, \ldots, 7$) on gene expression $x_n$ ($n = 8, 9, \ldots, 14$) and accounts for the random processing time effect ($T_j, j = 1, 2$) using log-transformed gene expression values as follows:

$$\log_2 x_n = \mu + M_i + \theta T_j + \epsilon,$$

We summarized the inference results in the Gene Regulatory Network graph shown on Figure 5.1.

5.3.2 Mathematical model formulation

We characterized gene expression dynamics in the Gene Regulatory Network graph with a set of Ordinary Differential Equations (ODEs) to capture the change in expression $dx$ over time $dt$ using (5.1) and (5.2) according to the developed computational framework in Chapter 4. The next steps were to collect the experimental data to uniquely identify model parameters and decide on
Figure 5.10 Double mutant experiment predictions. Model prediction and experimental results for double mutant experiments for A – MYB10, B – MYB72, C – bHLH39, D – PYE, E – bHLH101, and F – BTS under iron deficiency. For the mutant experiments gene expression was measured only at 72 hrs. Dashed lines show the best fit trajectories produced by the model under wild type and single mutant conditions. Shaded regions are the 95% credible regions of model prediction under double mutant scenario. Error bars indicate experimental measurements from corresponding experiments.
Figure 5.11 Decision tree for regulatory connection testing. Guidelines for accepting/rejecting a regulator-target connection based on individual test p-value ($p_I$), multiple comparisons compensated p-value ($p_C$), and the results of similar tests found in the literature (PMT).

(a) Sample standard deviation dependence on the expression values reveals heteroscedasticity
(b) Standard deviation vs. its log-transformed expression

Figure 5.12 Heteroscedasticity mitigation. log$_2$ transformation of gene expression values allowed to even-up sample variation for the statistical analysis
functional representations of \( f_{u_i}(t) \) and \( f_{n_i}(x) \). The major type of data for dynamic model training is the time course gene expression data. Time course datasets have been generated for Arabidopsis under various conditions including nitrate [Kro10], iron [Din08; Buc09], and phosphate [Lin11] starvation; salt [Din08; Kil07], pH [IP11], sulfur [IP11], osmotic [Kil07], heat [Kil07], UV-B light [Kil07], and drought [Kil07] stress. Specifically, Dinneny et al. [Din08] collected a data set of genome wide expression activity under iron deficiency in a 72 hrs span which showed that significant changes in the expression of known iron deficiency response genes occur after 24 hrs [Kor15b]. Based on this information, we designed the specifications for time course experiments with 7 evenly spaced samples over a 72 hour time span under both iron sufficient and iron deficient conditions. We captured the expression of 7 known iron deficiency response genes and 7 transcription factors from the GRN on Figure 5.1 using quantitative Real-Time PCR (qRT-PCR) experiment.

5.3.2.1 Time course data collection and preprocessing

Collection of time course measurements was conducted in phases, leading to the assessment of samples in multiple individual qRT-PCR experiments. This resulted in plate associated effects which introduced inconsistencies when merging the data. Measurements at the time points 0, 12, 24, and 36 hrs were processed separately from the ones at the time points 48, 60, and 72 hrs for each gene. Samples from different plates for the same gene appeared to have a scaling bias as was confirmed by redoing all time points on the same qRT-PCR plate (Figure 5.13 for ASIL2). This effect could have resulted from bias in Ct-values, typically compensated for by assessing fold changes among samples on the same plate. In our case samples were disconnected (plates did not share common time points). We developed the following computational procedure to (a) compensate for the variation in each plate within a time course for each gene and (b) keep the scaling bias the same for all regulators and targets. The procedure requires 2 additional samples to be run on the same plate for each gene (one from the first set of time points and one from the second set).

qRT-PCR plate effect compensation

The qRT-PCR experiment consists of the cycles of doubling the specific mRNA in the sample until it reaches a predefined threshold [SL08]. The number of doubling cycles needed to reach this threshold is called a Ct-value. Experimental sample volume is controlled by measuring the expression of two genes: the gene of interest and housekeeping gene that is assumed to have a constant expression value. Samples for the qRT-PCR measurements are distributed across a plate and processed in parallel. The qRT-PCR plate effect can be thought of as a plate \((p)\) specific bias \((\delta_{C}^p)\) added to a true Ct-value at time point \(t\) for both control/housekeeping gene measurements \((C_t C_p)\) and
Figure 5.13 qRT-PCR plate effect demonstration. Graphs show expression values for samples processed in 2 different qRT-PCR runs and values for the same samples processed in the one separate qRT-PCR run for ASIL2.

measurements of the gene of interest ($Ct^p_{t_i}$):

\[ Ct^p_{t_i} = Ct_{t_i} + \delta^p_C \]  
\[ Ct^p_{I_t} = Ct_{I_t} + \delta^p_I \]  

Thus, $\Delta Ct^p_{t_i}$ would have a plate dependent offset:

\[ \Delta Ct^p_{t_i} = Ct_{t_i} - Ct_{I_t} + \delta^p_C - \delta^p_I \]  

This offset will result in a scaling of the resulting expression value

\[ E^p_{t_i} = 2^{\Delta Ct^p_{t_i}} = 2^{Ct_{t_i} - Ct_{I_t} + \delta^p_C - \delta^p_I} \]  

which is compensated when looking for a fold change within the same plate. However, the scaling values will be different if using samples from different plates, thus, the correct fold change cannot be determined.

We propose a solution to the plate scaling issue with minimal extra experimentation by processing at least 2 time points that belong to different plates on the same additional plate. For the iron deficiency dataset we used 3 qRT-PCR plates ($p = 1, 2, 3$: 2 initial plates for different parts of the time
course and 1 for the scaling purposes) with the following \( C_t \)-values:

\[
\begin{align*}
\Delta C_t^1 &= C_t C_t - C_t I_t + \delta_C^1 - \delta_I^1 & t &= 0, 12, 24, 36 \text{ hrs} \\
\Delta C_t^2 &= C_t C_t - C_t I_t + \delta_C^2 - \delta_I^2 & t &= 48, 60, 72 \text{ hrs} \\
\Delta C_t^3 &= C_t C_t - C_t I_t + \delta_C^3 - \delta_I^3 & t &= 0, 48 \text{ hrs}
\end{align*}
\]

Then we calculated the compensation terms:

\[
\begin{align*}
\delta_{31} &= \Delta C_t^3 - \Delta C_t^1 = \delta_C^3 - \delta_I^3 - (\delta_C^1 - \delta_I^1) \\
\delta_{32} &= \Delta C_t^3 - \Delta C_t^2 = \delta_C^3 - \delta_I^3 - (\delta_C^2 - \delta_I^2)
\end{align*}
\]

These terms were added to \( \Delta C_t \)-values from the first 2 plates to set the same bias value across all 3 plates:

\[
\begin{align*}
\hat{\Delta} C_t^1 &= \Delta C_t^1 + \delta_{31} = C_t C_t - C_t I_t + \delta_C^3 - \delta_I^3 \\
\hat{\Delta} C_t^2 &= \Delta C_t^2 + \delta_{32} = C_t C_t - C_t I_t + \delta_C^3 - \delta_I^3
\end{align*}
\]

Hence, we changed the bias value for each gene to the one equal to the bias from plate 3, meaning that expression values for all genes in the GRN became of the same magnitude scale. We ran all time points samples for bHLH104 on the same plate to evaluate the performance of the developed compensation procedure. Comparison of redone and bias compensated expression patterns are shown in Figure 5.14. Initial and bias compensated expression patterns for the rest of genes can be found in Appendix A.6. This method could be useful for qRT-PCR measurements where samples are spread across multiple plates, a common case particularly in experiments with a high volume of samples or genes to test. We applied the qRT-PCR plate bias compensation technique to the time course samples before using the collected data for model training.

### 5.3.2.2 Unknown regulatory effects function

Various factors may modulate the stress response of a given gene by enhancing or dampening its expression. These factors include regulatory effects of known components and effects not represented in any of the known regulators’ expression patterns. A comprehensive list of regulators is not available in most cases, so mathematical constructs are needed to describe unknown effects when building a dynamic model [Ji16; Gao08; Ras06]. We aggregated the unknown factors that contribute to the change in gene expression in the *A. thaliana* root under iron deficiency into a stress signal \( f_u(t) \):

\[
\frac{dx}{dt} = a f_u(t) - bx, \tag{5.11}
\]
Figure 5.14 Plate effect compensation procedure validation for bHLH104. Collected time course samples capture the true (fully redone) trajectories much better after applying the developed qRT-PCR plate bias compensation algorithm.
where \( f_u(t) \) is a function representing the transient part of currently unknown regulatory factors affected by iron deficiency. \( f_u(t) \) takes positive values and turns into 1 if no stress is applied. We modeled the unknown factor effect as a monotonic shift to a new iron deficiency induced equilibrium state. We represented this continuous shift by a sigmoid function \( u(t) \), which for a given gene was defined by a scale coefficient \( u_T \), rate coefficient \( r \), and delay coefficient \( \tau \):

\[
u(t) = \frac{u_T}{(\frac{\tau}{t}) + 1}, \quad u_T > -1, \quad r > 0, \quad 0 < \tau < 72.
\]

(5.12)

Conceptually, \( u_T \) is the magnitude of the unknown factors signal (sign defines activation/inhibition effect), \( r \) represents how fast the transition between iron sufficient and iron deficient states occur for a given gene, and \( \tau \) is the time when a half saturation of \( u(t) \) occurs (Figure 4.3). We used the collected time course gene expression data under iron sufficient and iron deficient conditions to estimate coefficients shaping \( f_u(t) \) for each gene.

We chose Equation 5.11 as a basis for describing regulators’ dynamics because the regulators of those genes and other behavior influencing factors are currently unknown. However, despite the simplicity of the selected equation, it contains 6 parameters per regulator that have to be estimated. We used parameter identifiability analysis to check whether the time course data alone was sufficient to support the selected equation structure.

### 5.3.2.3 Parameter identifiability

A mechanism for determining whether a given model structure is supported by the available data is needed to make a good trade off between equation structure complexity and descriptive capabilities of the model. We require all parameters to be uniquely identifiable in the model to ensure that enough data is available to uniquely identify all parameter values. A parameter is considered identifiable if it can be estimated with a finite confidence interval [Van13; Rau14]. If a selected parameter is non-identifiable, other parameters can adjust their values to produce a good model fit no matter which value is chosen for the selected parameter. When non-identifiable parameters are present in a model, an equally good model fit can be achieved for drastically different values of the same parameter which affects the biological relevance of the corresponding estimate. A presence of non-identifiable parameters tend to cause non reliable model predictions [Rau11].

Several methods such as Differential Algebra Identifiability of Systems (DAISY) [Sac03], Exact Arithmetic Rank (EAR) [Kar12], and Profile Likelihood (PL) [Rau09] have been introduced in the literature to detect non-identifiability. We applied the PL method to detect the non-identifiable parameters in the proposed model because this approach is the only one that accounts for the quality and quantity of experimental data. First, we found optimum model parameter values with MATLAB multistart fmincon routine. Next, we specified the range of reasonable search values for
each parameter (Table 5.2). For the decay rate and initial concentration we fixed the range within the six standard deviations from experimental mean value. Other ranges were selected heuristically. Then, we were re-optimizing the model fit for each parameter value from that range to detect whether the cost function grew significantly as we went further from the optimum value of this parameter. A lack of significant cost function growth indicated the cases where other parameters were able to compensate for the change in the parameter of interest making a unique estimation of that parameter value impossible in those cases. The resulting cost function plots are illustrated in Figure 5.15 for COL4 and in Appendix A.7 for all the regulators in the GRN. Finally, we checked whether the cost function crossed the threshold on both sides from the optimal parameter value to determine whether a parameter was identifiable.

Table 5.2 Ranges of parameter sweep for the parameter identifiability analysis. μ and σ indicates the mean value and standard deviation of the corresponding parameters obtained through experiments.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>min</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>$u_T$</td>
<td>-1</td>
<td>4</td>
</tr>
<tr>
<td>$b$</td>
<td>$\mu_b - 6\sigma$</td>
<td>$\mu_b + 6\sigma$</td>
</tr>
<tr>
<td>$x_0$</td>
<td>$\mu_{x_0} - 6\sigma$</td>
<td>$\mu_{x_0} - 6\sigma$</td>
</tr>
<tr>
<td>$r$</td>
<td>1</td>
<td>50</td>
</tr>
<tr>
<td>$\tau$</td>
<td>1</td>
<td>72</td>
</tr>
</tbody>
</table>

Parameter identifiability results showed that 2 parameters were not identifiable for most of the regulators. Those parameters were the mRNA decay rate $b$ and the rate coefficient $r$ of the unknown effects function $f_u(t)$. We ran parameter sensitivity analysis to determine whether the model is sensitive to both of these parameters or any of them can be fixed while training the model.

5.3.2.4 Parameter sensitivity analysis

We assessed the influence of the parameters on the model response by performing Sensitivity Analysis (SA). Different SA techniques may yield different results based on the nature of the mathematical model [Mar08]. To guard from this possible issue we applied three commonly used methods to sensitivity analysis and compared the results. We conducted one at a time (OAT) SA [Sal08], partial rank correlation coefficient (PRCC) method [Sal00], and extended Fourier amplitude sensitivity test (eFAST) [Sal04]. Each of these SA methods estimated the influence of each parameter on the model response. The estimates were consistent across the methods.
Figure 5.15 Profile likelihood. Profile likelihood analysis curves for COL4. Solid line shows the error function, dashed line is the significance threshold. Parameters $b$ and $r$ are non-identifiable since the sum of squared errors does not change significantly with the change in parameter values.

One At a Time (OAT) sensitivity

To perform OAT SA, we first found an optimum value for the parameters (using MATLAB multi-start FMINCON method). Then we perturbed each parameter by $\pm 50\%$ from its optimum value while fixing all other parameters to the estimated optimal values. We plotted change in model response against relative parameter perturbation from the nominal value. We performed OAT SA for each time point separately. Gene expression showed a large change in response to the perturbation of the parameters $a$, $b$, $u$, and $\tau$. On the other hand, the expression did not change significantly for the parameter $r$ and for the initial gene expression $x_0$ except for the $t = 0$ time point. Figure 5.16 shows the OAT sensitivity analysis for the regulators. The OAT analysis suggested that $a$ and $b$ are more influential parameters compared to others, and $x_0$ is the only influential parameter at $t = 0$.

Partial Rank Correlation Coefficient (PRCC) method

For PRCC, we used 10,000 samples generated using Latin hypercube sampling (LHS) from uniform prior distributions to analyze the sensitivity of the six parameters $a$, $b$, $u_T$, $x_0$, $r$, and $\tau$. We compared sensitivity measures of all these parameters to a dummy parameter $d$ which had no relation to the model (purely non-influential). Unlike OAT method all the parameters changed simultaneously in PRCC. Thus the dummy parameter served as a control. If other parameters were not changing, sensitivity index of dummy parameter would be zero (change in the dummy parameter causes no
Figure 5.16 OAT sensitivity analysis for the regulators. A parameter related slope corresponds to the influence on model behavior.
change in model response). Upper and lower bounds for the priors are given in Table 5.3.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a$</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>$u_T$</td>
<td>-200</td>
<td>200</td>
</tr>
<tr>
<td>$b$</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>$x_0$</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>$r$</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>$\tau$</td>
<td>0</td>
<td>72</td>
</tr>
</tbody>
</table>

We calculated PRCC at the seven time points. The output response was only dependent to $x_0$ at initial time point $t = 0$ hrs. The PRCC value for the rate parameter $r$ was very small and almost same as the dummy parameter $d$ at all the time points. PRCC results indicated that the model response is insensitive to change in $r$. The PRCC results for the regulators are shown in Figure 5.17.

**Extended Fourier Amplitude Sensitivity Test (eFAST)**

To further investigate the sensitivity of the parameter $r$ we used the eFAST algorithm presented in [Mar08]. This method calculates a fast order sensitivity index $S_i$ and total sensitivity index $S_{Ti}$ for each parameter. The eFAST implementation [Mar08] required the input parameters given in Table 5.4.

<table>
<thead>
<tr>
<th>Input</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of search curves</td>
<td>5</td>
</tr>
<tr>
<td>No. of input factors</td>
<td>7</td>
</tr>
<tr>
<td>Samples per search curve</td>
<td>1000</td>
</tr>
</tbody>
</table>

We found that sensitivity indices of $r$ were not significantly different than that of the dummy parameter $d$ at most of the time points. The eFAST analysis for regulator parameters is shown in Figure 5.18.

The results of all 3 parameter sensitivity methods above agreed on the conclusion that the model is sensitive to the parameter $b$, but not sensitive to the parameter $r$. We conducted decay rate experiments to estimate genes' decay rate values under iron sufficient and iron deficient conditions.
Figure 5.17 PRCC analysis for the regulators. *— significant difference between model parameter and dummy parameter influence on the model output.
Figure 5.18 eFAST analysis for the regulators. Bar height corresponds to the model sensitivity to a parameter at a given time point. Intersection of a parameter’s confidence region with the one for the dummy variable indicates model insensitivity to that parameter.
while fixing the parameter $r$ to a nominal value.

5.3.2.5 Decay rate data collection and processing

We estimated mRNA decay rates for the genes in the modeled gene regulatory network by applying the transcription inhibitor, cordycepin, that prevents formation of new mRNAs (see subsection 5.3.5) and sampled gene expression at the time points following the application. Three biological samples were collected at 0, 15, 30, 45, 60, 90, 120 minutes after applying the transcriptional inhibitor to the plant exposed to 48 hours of either iron deficient or iron sufficient conditions. mRNA relative abundance was determined by qRT-PCR using at least two technical replicates. Decay rate data is comprised of control gene and gene of interest Ct-values sampled at the following conditions:

$$i = 1, \ldots, 7 \quad \text{time points (0; } \frac{1}{4}; \frac{1}{2}; 1; 1.5; 2 \text{ hrs);}$$

$$n = 1, 2 \quad \text{iron conditions (–Fe, +Fe);}$$

$$m = 1, \ldots, M \quad \text{genes;}$$

$$j(n, m) = 1, 2, 3 \quad \text{replicates per gene per iron condition;}$$

$$k(j) = 1, \ldots, K \quad \text{technical replicates per biological replicate.}$$

The main assumption behind using a control (housekeeping) gene in qRT-PCR measurements is that this gene is known to have a constant expression, which is not the case when expression is inhibited. However, the decay rate of the housekeeping gene can be estimated from its Ct-values and compensated when inferring decay rates for the genes of interest. Since the mRNA of the housekeeping gene is decaying after the transcription is terminated, we can describe the corresponding gene expression samples as follows:

$$x_{i,n,m,j,k}^h = x_0 e^{-b_h t_i}$$

where $x_0$ is the concentration of the housekeeping gene before inhibiting the gene expression and $b_h$ is the decay rate of the housekeeping gene. During a qRT-PCR run, mRNA content of a sample ($x_{i,n,m,j,k}$) is doubled Ct-value times during polymerase chain reaction until it reaches a
predefined constant threshold value $T$:

$$x_{inmjk}^h \cdot 2^{C_{inmjk}^h} = T_h$$

$$x_0^h e^{-b_n t_i} \cdot 2^{C_{inmjk}^h} = T_h$$

$$2^{C_{inmjk}^h} = \frac{T_h}{x_0^h} e^{b_h t_i}$$

$$\ln \left( 2^{C_{inmjk}^h} \right) = \ln \left( \frac{T_h}{x_0} \right) + b_n t_i$$

Hence, to estimate the decay rate of the housekeeping gene we can use the following statistical model:

$$y_{inmjk}^h = \alpha + \gamma f e_n + b_n t_i + \phi t_i f e_n + R_{j(nm)} + \epsilon_{inmjk},$$

where $y_{inmjk}^h = \ln \left( 2^{C_{inmjk}^h} \right)$, $\gamma$ represents constant iron effects, $\phi$ stands for the interaction effects between time and iron, and $R_{j(nm)}$ is a blocking factor that compensates for the run dependent bias. Before applying the statistical model we checked the resulting data for outliers at each time point. Values that exceeded 3 standard deviations from the mean at each time point for $+Fe$ and $-Fe$ conditions were removed from the dataset (Figure 5.19 and Figure 5.20). Statistical analysis showed that the decay rate for the housekeeping gene ($b_n = 0.185$) was significantly different from zero ($p_{b_n} < 0.0001$) and did not depend on iron condition ($p_\phi = 0.57$).

![Figure 5.19 Housekeeping gene outlier samples removal for iron deficient (Fe–) conditions.](image)
Figure 5.20 Housekeeping gene outlier samples removal for iron sufficient (Fe+) conditions. Housekeeping gene log transformed Ct-values before and after the outlier removal.

$\Delta Ct$-values for each gene of interest were processed in a similar way to $Ct$-values for the housekeeping gene. However, since fewer data was available for the genes of interest than for the housekeeping gene, 2 standard deviations threshold value was used for outlier removal. The statistical model for gene $m$ is based on qRT-PCR equations for the gene of interest and the housekeeping gene:

$$x_{inj}^m \cdot 2^{C_t_{inj}^m} = T_m$$
$$x_{inm}^h \cdot 2^{C_t_{inm}^h} = T_h$$

Dividing both sides of the equation for gene $m$ by both sides of the equation for the housekeeping gene we get:

$$2^{C_t_{inj}^m - C_t_{inm}^h} = \frac{T_m}{T_h} \cdot \frac{x_{inm}^h}{x_{inj}^m}$$

Substituting the decay assumptions have:

$$2^{\Delta C_t_{inj}^m} = \frac{T_m}{T_h} \cdot \frac{x_0^h e^{-b_h t_i}}{x_0^m e^{-b_m t_i}}$$

$$\ln\left(2^{\Delta C_t_{inj}^m}\right) = \ln\left(\frac{T_m}{T_h} \cdot \frac{x_0^h}{x_0^m}\right) - b_h t_i + b_m t_i$$
Since we have already inferred \( b_h \) we can rewrite the last equation as follows:

\[
\ln \left( 2^{\Delta C_t^{m_{inj}}} \right) + b_h t_i = \ln \left( \frac{T_m}{T_h} \cdot \frac{x_0^h}{x_0^m} \right) + b_m t_i
\]

So the statistical model for each gene is:

\[
y_{injk}^m = \alpha + \gamma f e_n + b t_i + \phi t_j f e_n + R_{k(j)} + \epsilon_{injk},
\]

where \( y_{injk}^m = \ln \left( 2^{\Delta C_t^{m_{inj}}} \right) + b_h t_i \) and other parameters are similar to those for the housekeeping gene.

Figure 5.21 and Figure 5.22 show the results of linear regression for the regulator COL4 and target BTS decay rates based on the stated model. Different samples colors stand for different technical replicates \((k)\) of the same biological replicates \((j)\) and emphasize the need in the blocking factor \( R_{k(j)} \) because different runs seem to have constant shifts with respect to each other as was also illustrated in subsection 5.3.2.1. We used the statistical model to test whether the iron effect is significant \((p_\gamma)\), whether the decay rate under iron deficiency is significantly different from 0 \((p_b)\), and whether the difference between decay rates under iron deficiency and iron sufficiency is significant \((p_\phi)\).

As can be observed from the graphs, the decay rate for the regulator COL4 does not depend on iron condition \((p_\gamma > 0.05)\) while for the target gene BTS the difference is significant \((p_\gamma < 0.05)\). We quantified the effect of iron condition dependence for the regulators in the Equation 5.5 to account for it in the model. Graphs for other genes in the GRN can be found in Appendix A.8. Small negative decay rates imply that the corresponding transcripts decay very slow, and the corresponding decay rate can not be detected with the given time span and the existing measurement variation. The resulting decay rate estimates and the corresponding standard errors are shown in Figure 5.2.

### 5.3.2.6 Regulator effect function

Decay rate estimates allowed us to make all the parameters for the regulator genes identifiable, so we could proceed with representing target genes’ dynamics. The gene regulatory network graph which we used as a basis for our ODE model (Figure 5.1) represents influential interactions between regulator and target genes. However, the presented set of regulators per each target is likely to be incomplete. We modeled the effects of known and unknown regulators on gene expression with Equation 5.13.

\[
\frac{dx}{dt} = af_u(t)f_1(x_1)f_2(x_2)\cdots f_K(x_K) - bx,
\]

where \( x_k \) is the expression of one of \( K \) known regulators for a given gene \((k = 1,2,\ldots,K)\), and \( f_k(x_k) \) is the regulator’s influence function which should be equal to 1 when \( x_k = 0 \), greater than
Figure 5.21 Statistical analysis results for COL4. $y$ axis represents housekeeping gene decay compensated log transformed $\Delta Ct$ values at iron deficient and iron sufficient conditions. Different samples colors stand for technical replicates.

Figure 5.22 Statistical analysis results for the target BTS. $y$ axis represents housekeeping gene decay compensated log transformed $\Delta Ct$ values at iron deficient and iron sufficient conditions. Different samples colors stand for technical replicates.
1 for activators and between 0 and 1 for inhibitors. The structure of (5.13) implies the simplifying assumption that the regulators influence the targets in an independent manner. We implemented a linear approximation of the regulator effect under the independence assumption:

\[ f_k(x_k) = 1 + c_k x_k, \tag{5.14} \]

where \( c_k \) is a constant impact factor.

Similarity in regulator gene expression patterns limited our ability to differentiate between regulator’s influences using only the available time course data. Levels of noise associated with time course samples put unfeasible demands on the number of additional time points that had to be collected to uniquely identify the impact factors \( c_k \) in (5.14). We opted for a more direct estimation of these parameters based on the results of single regulator knockout experiments which quantify the change in target expression when regulator expression is suppressed. We collected 3 replicates of each target gene expression in each regulator knockout mutant background after 72 hours of plant exposure to iron deficiency to estimate both mean effect and variation. Assuming that by that time genes had finished their transition to the new quasi steady state (i.e. \( dx/dt \approx 0 \) at \( t \geq 72 \) hrs), we estimated \( c_k \) using Equation 4.5.

We calculated mean and standard deviation of \( c_k \) for each regulator-target pair based on mean values and sample variances of \( x^M \), \( x^W \), and \( x^W_k \) and assuming a normal distribution for these terms. We generated 100,000 samples from each of the corresponding distributions and substituted them into (4.5) to estimate the distribution shape for \( c_k \) and calculate the corresponding standard deviation. Examples of such sample distributions for the target \( i = MYB10 \) and each regulator \( (n) \) are shown on Figure 5.23. Histograms estimating \( c_k \) (\( k = n_i \)) distributions for each pair of each regulator \( i \) and a target \( n \) are presented in Appendix A.9. However, only the distributions for the regulator-target pairs from Figure 5.1 were used for model training.

### 5.3.3 Model parameter estimation

We chose a Markov Chain Monte Carlo (MCMC) based parameter estimation algorithm for model fitting because it was capable of incorporating the results of additional experiments by summarizing them in terms of prior parameter distributions. This algorithm also returns posterior parameter distributions which facilitates the quantification of uncertainty associated with simulated gene expression patterns. In this work, we used the Differential Evolution Adaptive Metropolis (DREAM) software package [Vru09], which is the latest generation of MCMC based algorithms. It has been demonstrated that DREAM works well for nonlinear, multimodal, and high-dimensional problems and often outperforms other MCMC algorithms in these areas [LV12]. In addition to the general MCMC scheme, DREAM software allows to account for unequal measurement variance (heteroskedasticity) which we expected to observe from qRT-PCR experiments results due to an inherent
Figure 5.23 Sample distribution histograms for regulator coefficients $c_k$ for MYB10.
exponential transformation when calculating expression values [SL08] and significant changes in these values.

We defined the cost function for model fit as:

\[ \mathcal{L}(\theta|Y) = \frac{m}{2} \log(2\pi) + \sum_{i=1}^{m} \log(\sigma_i) + \frac{1}{2} \sum_{i=1}^{m} \left( \frac{x_i - y_i}{\sigma_i} \right)^2, \quad (5.15) \]

where \( \theta \) is the vector of model parameters (Table 5.5), \( Y = [y_1, y_2, \ldots, y_m]^T \) is the vector of mean gene expression values at each time point, \( m \) is the number of time points, \( \sigma_i \) is the sample standard deviation estimate at time point \( i \) (\( i = 1, 2, \ldots, m \)), \( x_i \) is the expression value predicted by the model at time point \( i \) based on the set of parameters \( \theta \).

### Table 5.5 List of model parameters for gene \( x_n \) expression. Units are presented in expression units \([ex]\), hours \([hr]\), or unitless \([1]\).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_{0n} )</td>
<td>Initial expression value</td>
<td>( ex )</td>
</tr>
<tr>
<td>( a_n )</td>
<td>Transcription rate scaling factor</td>
<td>( ex \cdot hr^{-1} )</td>
</tr>
<tr>
<td>( b_n )</td>
<td>Decay rate for the regulators</td>
<td>( hr^{-1} )</td>
</tr>
<tr>
<td>( b_{Fe+}^{n} )</td>
<td>Decay rate at iron sufficiency for the targets</td>
<td>( hr^{-1} )</td>
</tr>
<tr>
<td>( m_n )</td>
<td>Decay rate scaling factor at iron deficiency for the targets</td>
<td>( hr^{-1} )</td>
</tr>
<tr>
<td>( \tau_{rn} )</td>
<td>Unknown factors signal magnitude</td>
<td>1</td>
</tr>
<tr>
<td>( \tau_{rn} )</td>
<td>Unknown factors signal delay</td>
<td>( hr )</td>
</tr>
<tr>
<td>( r_n )</td>
<td>Unknown factors signal rate</td>
<td>1</td>
</tr>
<tr>
<td>( c_{rn} )</td>
<td>Regulator’s impact factor</td>
<td>( ex^{-1} )</td>
</tr>
</tbody>
</table>

By default DREAM software makes a typical statistical assumption about equal measurement variance. We tested this assumption by plotting gene expression values magnitudes against the corresponding standard error estimates and detected heteroskedasticity in the time course data (Figure 5.24A). We approximated the dependence of variance on the measured value with a log-linear function (5.16) and specified the corresponding setting in the DREAM parameter optimization software to account for unequal variance according to the Equation 5.17.

\[ \log_2(\sigma_i) = A + B \log_2(y_i), \quad (5.16) \]
\[ \sigma_i = 2^A y_i^B \quad (5.17) \]

We constructed prior parameter distributions based on the available experimental data. We used the results of decay rate experiments to calculate mean and standard error values for the mRNA decay rates \( b_n \) (subsection 5.3.2.5). We constructed a prior for each \( b_n \) assuming log-normal distribution
for this parameter to account for the fact that these values cannot be negative while preserving other properties of Gaussian distribution. The same assumption was used to calculate prior distributions for the initial conditions $x_n(t = 0)$ from mean and standard deviation values of the time course data at 0 hours of exposure to iron deficiency. Similarly, we constructed Gaussian priors for the regulator’s effect parameters $c_n_i$ based on the regulators single mutant data (subsection 5.3.2.6). Priors for the heteroscedastic noise parameters A and B were obtained by fitting equation (5.16) to measured data as depicted in Figure 5.24B. We used uniform priors for the rest of the parameters because no additional information to shape the corresponding distributions were available. However, we limited the prior for mRNA production rate coefficient $a_n$ to be greater than 0 to preserve a positive sign of the transcription part of the equation. For the same reason we set a lower bound of $-1$ to the magnitude $u_{f_n}$ of the unknown effect function $f_{u_n}(t)$. We also bounded the delay term $\tau_n$ by the span of time course data (i.e. between 0 and 72 hours). The resulting prior distributions were used by DREAM while minimizing the cost function (5.15).

We started model fitting with parameter training for each regulator and then used the best fit to train parameters for each individual target. We used 12 chains of DREAM algorithm to find a set of parameter values that provided the best fit to prior distributions and the time course data. Each chain returned a collection of 10,000 sets of parameter values, which were used to construct parameter posterior distributions. We applied $\hat{R}$ diagnostic of Gelman and Rubin [GR92; Vru16] as
the algorithm convergence criterion.

5.3.4 Model prediction uncertainty quantification

Model prediction uncertainty arises from a host of factors including sample variation, measurement errors, and the discrepancy between an actual biological process dynamics and its mathematical representation. Quantification of this uncertainty is essential to access descriptive and predictive capabilities of the model. We used uncertainty quantification analysis for model predictions to produce credible intervals that cover the true gene expression at a given time point with 95% confidence. To produce such intervals, we randomly sampled 500 parameter sets from the obtained 10,000 chains to simulate the corresponding gene expression trajectories. We sampled those trajectories at each simulated time point and discarded 5% of the furthest from the mean values to obtain a collection of 95% credible intervals [Smi13]. These credible intervals were then combined into a credible region for a given gene expression pattern. We used similar procedure to obtain credible intervals for evaluating the results of double knockout mutant experiments after setting the corresponding regulator gene expression to 0 in model simulations.

5.3.5 Biological experimentation

Plant materials and growth conditions

*Arabidopsis thaliana* ecotype Columbia (Col-0) was used as the wild type accession for all experiments. Mutant lines used are *col-4* (SALK_092012), *etf9-1* (SALK_025328), *asil2-1* (SAIL_258_F06), *myb55-1* (GK-460G09), *bhlh34* (CS411089), *bhlh104-1* (SALK_099496), and *ilr3-3* (SALK_-043690). Combinatorial mutants were either generated *etf9asil2* or obtained from previous studies (*bhlh34bhlh104*, *bhlh34ilr3*, and *bhlh104ilr3*) [Li16]. All mutants were confirmed using qRT-PCR. Seed sterilization was performed as previously described in [Lon10]. For all assays, plants were plated on 100 µm Nite nylon mesh (Genese Scientific) on iron sufficient (+Fe) media for 4 days and transferred to iron sufficient (+Fe) or iron deficient (–Fe) media (0-72h for time course data, 2 days for decay rate data, and 3 days for mutant data. Iron sufficient (+Fe) media is standard Murashige and Skoog (MS) media with 0.05% (w/v) MES, 1% (w/v) Sucrose, 1% (w/v) agar, and 0.1 mM Fe-EDTA substituted for iron sulfate and iron deficient (–Fe) media has no added Fe-EDTA and the addition of the iron chelator 300 µM ferrozine. Plants were grown vertically at 22°C under 16-h-light/8-h-dark in environmentally controlled plant growth chambers.

Cordycepin treatment

Seedlings grown 4 days on +Fe media were transferred to +Fe or –Fe media for 2 days, transferred to liquid media to preincubate for 30 minutes at 50 rpm in the dark, and finally treated with addition
of 400 μM Cordycepin (Sigma-Aldrich) or DMSO (protocol modified from Ravet et al. [Rav12]).

RNA processing

Total RNA was extracted using the GeneJET Plant RNA Purification Kit (Thermo Scientific) and first-strand complementary DNA was synthesized using the Superscript III cDNA synthesis kit (Life Technologies). qRT-PCR was performed using iTaq Universal SYBR Green Supermix (Bio-Rad) and the StepOnePlus Real-Time PCR System (Applied Biosystems). Relative expression was calculated by the $2^{-\Delta C_T}$ method [SL08] with all expression values relative to (β)-tubulin in order to compare between experiments.
6.1 Conclusions

The goal of the work presented in this dissertation was to develop computational tools for transcriptome data processing allowing to gain a deeper understanding of molecular processes underlying the iron deficiency response in the model plant *Arabidopsis thaliana* on a genetic level. To do so, collaborators and I (1) explored the existing experimental methods for gene activity/expression quantification, (2) surveyed state of the art computational tools for gene expression data processing and the inference capabilities of those tools, (3) organized the computational tools based on complexity and the corresponding inference power, (4) proposed a transition path through the complexity levels starting from a test for significant change in gene expression and finishing with a dynamic model of gene regulations, (5) identified novel gene regulators affecting known iron response genes under iron deficiency in *A. thaliana*, (6) developed a computational framework for capturing temporal aspects of gene interactions with a collection of ordinary differential equations structures for different sets of available data, (7) formulated and trained the dynamic model of iron deficiency response between the novel gene regulators and the known iron response genes by applying the developed framework to a collection of experimental data that describe different aspects of iron deficiency related gene response.
6.1.1 Computational tools for transcriptome analysis

We explored inference capabilities of the state of the art computational tools used for analyzing time course gene expression data and organized these tools in the order of increasing complexity. We hypothesized how the results of a less complex inference procedure could be used as an input to a more elaborated computational tool when supplied with additional information. We proposed a possible progression path through the complexity levels starting from the differential expression analysis, proceeding towards gene correlation graphs, causality networks, and Boolean networks towards a continuous model of gene regulatory interactions over time represented with a system of ordinary differential equations. We presented application examples for each complexity level, so a researcher would be able to relate the data at hand with a possible inference procedure or to decide on additional experimentation to answer the question of interest.

6.1.2 Causality inference for iron deficiency related genes

We analyzed gene expression patterns from the available time course gene expression data on iron deficiency response in *A. thaliana* root and discovered that the majority of differentially expressed genes have a significant change in activity only after 24 hours of iron deprivation. All known iron response gene mediators, namely PYE, BTS, MYB10, MYB72, bHLH39, bHLH101, and bHLH115, appeared that group. We hypothesized that these known genes should be affected by gene regulators that get activated at the earlier stages. We developed and applied an algorithm for gene-wise causality inference to identify candidates for the early activity regulators. The algorithm predicted a set of 7 early activity gene regulators for the known iron response gene mediators. None of the predicted regulators have been previously linked to iron deficiency. 4 of the identified gene regulators, namely ETF9, COL4, ASIL2, and MYB55, were experimentally confirmed to influence the known iron deficiency mediators. The obtained regulatory relationships between the newly identified and known iron related gene mediators extended the body of knowledge about the early stages of iron deficiency response and provided avenues for gene response manipulation in plants exposed to iron deficiency.

6.1.3 Computational framework for building a dynamic model

We explored ways to gain a deeper understanding of gene interactions in a given gene regulatory network through modeling gene dynamics over time. A continuous ODE-based model of gene interactions can provide a quantitative assessment of gene activity and predict the effects of genetic manipulations leading to a more efficient gene engineering. Given a wide variety of proposed structures of ordinary differential equations for dynamic modeling of gene regulation over time, we had to find criteria for relating an equation structure with a collection of experimental data that can support it. For the cases when data was not sufficient for implementing a model structure of interest,
we needed to identify additional experiments that can efficiently increase the information content for model parameter training and an approach to aggregate the information from different types of experimental data. We developed a computational framework for building a dynamical model of gene interactions which addresses those questions. We proposed (1) parameter identifiability analysis to determine whether a given model structure is supported by the available data, (2) experiments such as loss of function, overexpression, dynamic parameter inference, and phenotypic parameter measurements that can be used to efficiently enrich the amount of available information for more descriptive and predictive model representations, and (3) a Bayesian approach for aggregating the diverse dataset types for model training. This framework provides guidance for quantitative modeling of stress response dynamics on the gene level in plants and can be extended to other organisms if similar types of the proposed experiments are applicable.

6.1.4 Dynamic modeling of iron deficiency response

We used the developed framework to formulate a dynamic model of iron deficiency regulation between the known and newly identified iron deficiency related genes. We added three more regulators of the known mediators, namely ILR3, bHLH104, and bHLH34, which were identified by other researchers, to extend the gene regulatory network. Time course data alone was not sufficient to support a system of ordinary differential equations that could describe gene expression dynamics, so we designed experiments and collected the data that solved the model parameter identifiability issues. Loss of function experiments provided information for stabilizing individual regulatory effects parameters. mRNA decay rate measurements provided the corresponding parameter estimates and revealed a dependence of this parameter on iron condition for most of the target genes (i.e. known iron mediators). We captured this dependence by introducing a correction factor to the model equations’ structure. Bayesian inference based approach allowed to account for measurement variation in each type of experimental data and to estimate the uncertainty associated with model predictions. The model was able to fit the training data within the range of estimated uncertainty and predict the results of 4 out of 6 regulator double mutant experiments which were not used for model training. Further analysis of the data from double mutant experiments revealed possible synergistic effects between bHLH34 and bHLH104 when regulating BTS and between bHLH104 and ILR3 when regulating bHLH115. Overall, the developed model provided insights into aspects of gene regulation dynamics under iron deficiency and showed promising results for replacing greenhouse experiments with computational simulations.
6.2 Future work

The computational framework for the whole genome temporal gene expression data analysis described in the dissertation provided a basis for deep exploration of gene regulation dynamics in a plant exposed to iron deficient media. Further exploratory process in this direction will include refinements at each stage of the framework, design of experiments capable of capturing the missing information efficiently, and aggregation of new relevant data in an effort to further increase descriptive and predictive power of the developed dynamic model. Possible directions for advancing the inference process and extending the dynamic model capabilities include the following topics.

6.2.1 Time course experiment design optimization

The developed dynamic model of iron deficiency response relies on the time course data to train the parameter values. The information capacity of that data depends on the number and choice of the sampling time points. Time, cost, and effort required for collecting each sample limit the feasible number of time points for a given experimental design, so the researchers use intuition to get the most information from a typical number of 2 to 8 time points per time course [Ros12]. Dinneny et al. [Din08] attempted to increase the informational content of the iron deficiency dataset by sampling more densely at the first day where the most activity was anticipated, yet the dataset analysis showed that the major known iron regulators got turned on only after 24 hours of *A. thaliana* exposure to iron deficiency. Hence, rules of thumb can be misleading in experimental design. A computational approach for optimal time point selection is needed to increase the information content of time course data for model training. Rosa et al. [Ros12] proposed to use the results of previous experiments under the same condition and time course data analysis under similar conditions as the prior knowledge for time point selection. Applying those concepts for designing new time course experiments will provide a higher informational content for inferring novel gene regulatory interactions and improving the parameter estimates of the developed dynamical model. Other improvements in the future time course experimental design can include the use of a new gene expression measurements platforms such as RNA-Seq for improved measurements precision, control samples at each measurement time point to avoid development and other time dependent effects that confound the differential expression analysis, and additional sampling time points for a higher temporal pattern resolution.

6.2.2 Cell specific inference of genetic regulations

*A. thaliana* is comprised of cell layers with specific functions and spatial patterns within the plant. The experimental data used in this dissertation was collected from the whole root samples. The idea was to capture and characterize the strongest iron deficiency drivers that would stand out even
if averaged across the cell types. This approach, however, does not guarantee that those drivers have a significant activity in each cell type and can miss a significant regulator that acts in one cell type only. Moreover, temporal gene activity patterns can vary across cell types because iron deficiency sensing takes time to propagate from the outer layer of the root to the inner layers. So, cell specific gene expression data analysis would provide a more refined understanding of gene activity regulations within *A. thaliana* root. The insights obtained from the whole root gene expression analysis coupled with the information about cell types’ sizes and spatial locations would provide information for efficient experimental design. Experimental techniques such as cell sorting [Bir03; BB09] allow separating different cell types before performing gene activity measurements. The data analysis framework presented in this dissertation can be applied to the cell specific data to obtain cell specific gene regulatory models which can lead to a compartmental model and, possibly, to a spatiotemporal continuous model in terms of partial differential equations.

### 6.2.3 Combinatorial gene regulatory effect inference

Experimental validation of the dynamic model presented in this dissertation uncovered regulatory effects that could not be explained with the assumption of the independent action of multiple regulators on the same target. We made this simplifying assumption because such synergistic, or gene interaction, effects were shown to constitute only about 3% of all pairwise regulatory influences in yeast [Cos10] while the experimental estimation of the same quantity in higher organisms was almost intractable [Van14]. Moreover, a high level of experimental noise coupled with a small number of samples per time point did not allow for inferring combinatorial effects computationally based on the data that we used for model training. However, non-independent regulatory effects should be characterized and included into the developed mathematical model to increase its descriptive and predictive power. Future steps in that direction would be to extend the existing model to account for the identified synergistic effects, develop recommendations for an experimental design that would allow for differentiating between models with different synergistic effects assumptions, and collect the corresponding data to further improve the understanding of regulator genes interactions.

### 6.2.4 Other model extension strategies

The model for gene regulations in response to iron deficiency in *A. thaliana* presented in this dissertation was developed based on the current knowledge and experimental data on that topic. However, new experimental data and the associated insights appear in the literature due to the overall interest in iron regulation in plants. This opens potential ways for extending the developed model in a variety of directions based on specific research questions. First, the model's gene regulatory network structure can be extended with newly identified iron related gene regulators and/or non-regulator genes with a specific function of interest. Second, regulator effect functions $f_r(x_r)$ can
be improved to capture the nonlinear nature of gene interactions as described in subsection 4.2.2 when the corresponding data becomes available. Third, the effect of iron concentration on gene activity levels can be included in the model by finding the most informative sampling concentrations based on phenotypic data as suggested in subsection 4.2.3 and collecting the corresponding gene expression measurements. Fourth, gene activity effects on plant physiology can be captured with additional model constructs describing a physiological parameter of interest based on the relevant phenotypic data. Finally, gene product dynamics can be captured by collecting and incorporating proteomic and metabolomic data as it is suggested in [Bas12].
BIBLIOGRAPHY


A.1 Dissimilarity tables for the candidate regulators of the known iron response mediators
A.2 Posterior probability distributions for the model parameters

![Posterior probability distributions for the model parameters](image-url)
Posterior of parameters of ASIL2

- **a**: Distribution of parameter a with a best fit at approximately 0.5.
- **u_T**: Distribution of parameter u_T with a best fit at approximately 0.2.
- **b**: Distribution of parameter b with a best fit at approximately 1.2.
- **x_0**: Distribution of parameter x_0 with a best fit at approximately 0.7.
- **\( \tau \)**: Distribution of parameter \( \tau \) with a best fit at approximately 60.
- **A**: Distribution of parameter A with a best fit at approximately 0.
- **B**: Distribution of parameter B with a best fit at approximately 0.

The plots show the posterior distribution of each parameter along with the best fit values.
Posterior of parameters of MYB55

- Parameter $a$
  - Range: 0.05 to 0.2
  - Posterior distribution

- Parameter $u_T$
  - Range: -3 to 0.5
  - Posterior distribution

- Parameter $b$
  - Range: 0.4 to 0.9
  - Posterior distribution

- Parameter $x_0$
  - Range: 0.1 to 0.45
  - Posterior distribution

- Parameter $\tau$
  - Range: 0 to 80
  - Posterior distribution

- Parameter $A$
  - Range: -4 to 0
  - Posterior distribution

- Parameter $B$
  - Range: 0 to 2
  - Posterior distribution

**Legend:**
- Blue: Posterior
- Red cross: Best Fit
Posterior of parameters of bHLH101

- a
- u_T
- b
- x_0
- \tau
- m
- C_{MYB55}
- C_{ILR3}
- A
- B

- Posterior
- Best Fit
A.3 Model predictions for single regulator mutant dynamics
A.4 Uncertainty regions for the currently unknown regulators' expression patterns time course trajectories
A.5 Credible region of the modeled bHLH115 trajectory overlaid with credible regions of its known targets’ unknown effect functions
A.6 Initial and bias compensated time course expression patterns

Figures show initial and scaled (sc) qRT-PCR expression patterns for the genes in the GRN. First half of the samples (0 to 36 hours) was initially processed in 2015 and the rest was processed in 2017. Digits on the sc graphs correspond to replicates numbers.
A.7 Profile likelihood results for the regulators

**COL4**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Range</th>
<th>Profile Likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pl(a)</td>
<td>0.2 - 0.4</td>
<td></td>
</tr>
<tr>
<td>Pl(uₜ)</td>
<td>0 - 4</td>
<td></td>
</tr>
<tr>
<td>Pl(ᵦ)</td>
<td>0 - 0.8</td>
<td></td>
</tr>
</tbody>
</table>

**ETF9**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Range</th>
<th>Profile Likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pl(a)</td>
<td>0.02 - 0.06</td>
<td></td>
</tr>
<tr>
<td>Pl(uₜ)</td>
<td>0 - 0.2</td>
<td></td>
</tr>
<tr>
<td>Pl(ᵦ)</td>
<td>-0.1 - 0.2</td>
<td></td>
</tr>
</tbody>
</table>

159
A.8 Decay rate estimation data fitting

Fe+: $b = 0.59$

$p_\gamma = 0.4, p_b = 2\times10^{-27}, p_\phi = 0.3$

Fe+: $b = 0.59$

$p_\gamma = 8\times10^{-05}, p_b = 0.5, p_\phi = 0.5$

Fe+: $b = 0.019$

Fe+: $b = 0.019$
Fe+: $b = 0.35$

Fe−: $b = 1$

$\rho = 1 \times 10^{-71}$, $p_b = 3 \times 10^{-22}$, $p_\gamma = 4 \times 10^{-07}$

Fe+: $b = -0.0072$

Fe−: $b = -0.0072$

$\rho = 1 \times 10^{-13}$, $p_b = 0.8$, $p_\gamma = 0.6$
\[ p_γ = 4 \times 10^{-54}, \ p_b = 5 \times 10^{-19}, \ p_ϕ = 1 \times 10^{-14} \]

\[ p_γ = 8 \times 10^{-30}, \ p_b = 5 \times 10^{-16}, \ p_ϕ = 1 \times 10^{-09} \]
Fe⁺: $b = -0.033$

$p_{\gamma} = 2 \times 10^{-07}$, $p_{b} = 2 \times 10^{-08}$, $p_{\phi} = 0.0002$

Fe⁻: $b = 0.33$

$p_{\gamma} = 1 \times 10^{-14}$, $p_{b} = 5 \times 10^{-17}$, $p_{\phi} = 0.0004$

Fe⁺: $b = 0.39$

Fe⁻: $b = 0.87$

Means
Estimate
Samples
\[ p_\tau = 3e^{-37}, p_b = 1e^{-18}, p_\phi = 4e^{-05} \]

**Fe+:** \( b = 0.52 \)

**Fe-:** \( b = 1.1 \)
A.9  Regulator-target influence parameters’ samples histograms

![Histograms of regulator-target influence parameters](image-url)
Histogram
Gaussian approximation
Mean $c_n$

Histogram
Gaussian approximation
Mean $c_n$

Histogram
Gaussian approximation
Mean $c_n$

Histogram
Gaussian approximation
Mean $c_n$
Density

ASIL2 (i) -> bHLH39 (n)

BHLH104 (i) -> bHLH39 (n)

BHLH34 (i) -> bHLH39 (n)

COL4 (i) -> bHLH39 (n)

ETF9 (i) -> bHLH39 (n)

ILR3 (i) -> bHLH39 (n)

MYB55 (i) -> bHLH39 (n)