A Machine Learning Approach to Predict Gene Regulatory Networks in Seed Development in Arabidopsis Using Time Series Gene Expression Data

Ying Ni

Thesis submitted to the faculty of the
Virginia Polytechnic Institute and State University.
in partial fulfillment of the requirements for the degree of

Master of Science
in
Computer Science and Applications

Lenwood S. Heath, Chair
Ruth Grene, Co-Chair
Song Li

June 22, 2016
Blacksburg, Virginia

Keywords: Network inference, signal transduction pathways, gene expression, support vector machines

A Machine Learning Approach to Predict Gene Regulatory Networks in Seed Development in Arabidopsis Using Time Series Gene Expression Data

Ying Ni

(ABSTRACT)

Gene regulatory networks (GRNs) provide a natural representation of relationships between regulators and target genes. Though inferring GRN is a challenging task, many methods, including unsupervised and supervised approaches, have been developed in the literature. However, most of these methods target non-context-specific GRNs. Because the regulatory relationships consistently reprogram under different tissues or biological processes, non-context-specific GRNs may not fit some specific conditions. In addition, a detailed investigation of the prediction results has remained elusive. In this study, I propose to use a machine learning approach to predict GRNs that occur in developmental stage-specific networks and to show how it improves our understanding of the GRN in seed development.

I developed a Beacon GRN inference tool to predict a GRN in seed development in Arabidopsis based on a support vector machine (SVM) local model. Using the time series gene expression levels in seed development and prior known regulatory relationships, I evaluated and predicted the GRN at this specific biological process. The prediction results show that one gene may be controlled by multiple regulators. The targets that are strongly positively correlated with their regulators are mostly expressed at the beginning of seed development. The direct targets were detected when I found a match between the promoter regions of the targets and the regulator's binding sequence. Our prediction provides a novel testable hypotheses of a GRN in seed development in Arabidopsis, and the Beacon GRN inference tool provides a valuable model system for context-specific GRN inference.

A Machine Learning Approach to Predict Gene Regulatory Networks in Seed Development in Arabidopsis Using Time Series Gene Expression Data

Ying Ni

(GENERAL AUDIENCE ABSTRACT)

Deoxyribonucleic acid, DNA, is well known genetic material that stores the information necessary for most living organisms. A segment of DNA encodes a gene. Generally, gene expression process is composed by DNA transcription and translation and this process is well regulated by the organisms. In this thesis, I particularly focus on the regulatory relationships in transcription step. The gene expression in different plants and different biological process is controlled by different regulatory mechanisms. To make the study specific, I present my work based on Arabidopsis in seed development. In a regulatory relationship, the gene that regulates another gene is known as regulator, and the other gene is called target gene. The target gene, in turn, can regulate many other genes. As a result, the regulators and their targets form a complex network. Reveal the structure of the regulatory network will help the researchers get better understanding of how the regulators work as a network and the complexity of the interdependencies among genes, and in this thesis, I use computational approaches to elucidate the topology.

There are four key regulators involved in Arabidopsis seed development, they are ABI3, FUS3, LEC1 and LEC2. The computational tool I developed, called Beacon inference tool, is a machine learning approach makes use of support vector machines (SVMs) that can infer the potential targets of the four regulators. This method predicted 1064, 2569 and 3836 targets for ABI3, FUS3 and LEC1, respectively. Among these targets, I searched for the ones that have their expression levels strongly positively correlated with their regulators. Because we assume that the expression of regulator influences the expression of targets, so I have more confidence of these positive correlations. Further, as it is known that the regulator

binds to the upstream sequence of the gene to regulate the expression level, therefore, if the upstream sequence of the target gene matches the binding site of the regulator, this target is classified as direct target. As a result, 24 out of 65 and 173 out of 1759 are direct targets among the positive correlations for ABI3 and FUS3, respectively.

Acknowledgments

I would like to thank my advisors, Prof. Lenwood S. Heath and Ruth Grene for their guidance and support during my studies at Virginia Tech. Their patience, encouragement and insightful suggestions greatly aided my research. I would also like to thank my committee member, Prof. Song Li, for his valuable time and helpful comments. In particular, I am also grateful to Prof. Eva Collakova for her help in my study.

I am thankful to all of my group members and former colleagues, Mostafa Arefiyan and Elijah Myers for their Beacon editor, Deepti Aggarwal for her advise at the beginning of my research, and Delasa Aghamirzaie and Doaa Altarawy for their cooperation and valuable ideas. I also appreciate Wei Wang, Gustavo Arango Argoty, and Yuzhong Wen, who are good friends that have been great companions and encouraged me during my research. Special thanks to Delasa Aghamirzaie, who helped me collect some data presented in this thesis.

I also want to thank the Department of Computer Science at Virginia Tech for giving me an opportunity to work in Bioinformatics area. I am also grateful to the National Science Foundation (DBI-1062472) for financial support. As always, I would like to thank my family for their unconditional love and support over the years.

Contents

1 INTRODUCTION **1**

2 PRELIMINARIES **6**

2.1 Biological Concepts 6

2.2 Computational Terms and Concepts 8

2.2.1 Types of Graphs 8

2.2.2 Euclidean Distance 9

3 PROBLEM DEFINITION **10**

3.1 Background 10

3.2 Problem Definition 12

4 LITERATURE REVIEW **14**

4.1 Unsupervised Learning Methods 14

4.2 Supervised Learning Methods 16

4.3 Discussion 17

5 DATA MODEL **19**

5.1 Prior Knowledge 19

5.2 Experimental Data 23

5.2.1 Data Set 1 (Gene Time Course Data) 23

5.2.2 Data Set 2 (Differentially Expressed Gene Time Course Data) 26

6 METHODOLOGY **27**

6.1 Data Analysis 27

6.2 ROC and AUC 28

6.3 Unsupervised Learning - CLR Algorithm 29

6.4 Supervised Learning - SVM Algorithm 30

6.4.1 Feature Vector 30

6.4.2 Kernel Function 31

6.4.3 Cross Validation 33

6.4.4 Ranking 34

6.5 Clustering 34

6.6 Experimental Procedure 35

7 RESULTS AND DISCUSSION **40**

7.1 Results 40

7.1.1 Algorithm Evaluation and Comparison 40

7.1.2 Network Prediction 44

7.1.3 Biological Validation 45

7.2 Discussion 51

8 CONCLUSIONS **53**

9 BIBLIOGRAPHY **56**

Glossary

ABA	ABSCISIC ACID.
ABI3	ABSCISIC ACID-INSENSITIVE3.
ABI4	ABSCISIC ACID-INSENSITIVE4.
ABI5	ABSCISIC ACID-INSENSITIVE5.
C3NET	Conservative causal core.
cDNA	Complementary DNA.
CIS-BP	Catalog of inferred sequence binding preferences.
CLR	Context likelihood of relatedness algorithm.
DAP	Days after pollination.
DNA	Deoxyribonucleic acid.
FIMO	Find individual motif occurrences.
FPKM	Fragments per kilobase of transcript per million mapped reads.
FUS3	FUSCA3.
GENIE	Gene network inference with ensemble of trees.
GRN	Gene regulatory network.

iRafNet	Integrative random forest for gene regulatory network inference.
LARS	Least angle regression.
LEC1	LEAFY COTYLEDON1.
LEC2	LEAFY COTYLEDON2.
MI	Mutual information.
MRMR	Minimum redundancy maximum relevance.
mRNA	Messenger RNA.
MRNET	Maximum relevance minimum redundancy.
RN	Relevance network.
RNA	Ribonucleic acid.
RNA-seq	RNA sequencing.
SBGN	Systems Biology Graphical Notation.
SIRENE	Supervised inference of regulatory networks.
SVM	Support vector machine.
TF	Transcription factor.
TIGRESS	Trustful inference of gene regulation using stability selection.
WGCNA	Weighted correlation network analysis.

List of Figures

2.1 Gene expression at the molecular level. 7

3.1 Schematic description of signal transduction pathways in plants. Figure adapted from Shinozaki and Dennis [71]. 11

5.1 Interactions among some of the hormonal and developmental signals and regulatory elements controlling seed maturation. Arrows represent positive regulation and red bars indicate repression. Figure drawn by Beacon editor and it is adapted from Finkelstein. [16]. 22

6.1 Two dimensional representation of SVM using maximum margin with support vectors to classify genes. 32

6.2 Beacon GRN inference and validation workflow. Five phases: method comparison (A), prediction (B), k-means clustering (C), identify the targets contain binding motifs (D), and identify targets containing the downstream TF binding motifs (E). K-means clustering is done by combining strongly correlated known and predicted targets. 38

6.3 The proposed network. The diagram is drawn in Systems Biology Graphical Notation (SBGN) format using the Beacon editor [39]. LEC1, FUS3 and ABI3 represent three master regulators, with ABI3 directly controlled by LEC1 and ABI3 and FUS3 mutually regulated. 39

7.1 Comparison of performance between SVM local models and global model. ABI3, FUS3 and LEC1 represent local models with each of them as a separate SVM. Global model trains one SVM for all the TF-target pairs. 42

7.2 Comparison of performance between SVM local models and CLR algorithm. 43

7.3 A Venn diagram depicting the overlap between the predicted targets among the three regulators. 45

7.4 K-means clusters of (A) ABI3, (B) FUS3, and (C) LEC1 target genes, and the expression profiles for the three regulators (D). The results are organized by developmental stage. Three stages of seed development are involved in the gene expression: early (7 and 8 DAP), middle (10, 12 and 13 DAP), and late (15 and 17 DAP). The color scale indicates the gene expression level: red color represents high expression level, and blue color represents low expression level. A horizontal line is in each cluster, above which are the prior known targets and the remaining are predicted targets. The difference in expression profiles of the regulators may lead to different expression patterns of the target genes. 48

List of Tables

5.1 Number of target genes of LEC1, LEC2, FUS3, and ABI3, number of samples, techniques and tissues used in experiments. 20

5.2 *Arabidopsis thaliana* gene name data set of Data set 1. 24

5.3 *Arabidopsis thaliana* time series data set of Data set 1. 25

7.1 Summary of the number of prior known regulations in Arabidopsis seed development gene and differentially expressed gene data sets. 43

7.2 A summary of the number of predicted and unique targets for each regulator. 44

7.3 A comparison of the total number of targets and the number of strongly positively correlated targets of each regulator. Less than half of the ABI3 and LEC1's targets are strongly positively correlated, while more FUS3 targets are strongly correlated. 46

7.4 The number of direct and indirect targets discovered by FIMO in each cluster. LEC1 does not have known binding site in the CIS-BP database, so only ABI3 and FUS3 binding sites could be identified. For each regulator, the table shows the number of targets that have the binding sites in known and predicted connections, respectively. The last row of each regulator shows the number of targets exist both in our prediction and in GeneMANIA (GM). . 50

Chapter 1

INTRODUCTION

Elucidating and understanding the topology of gene regulatory networks (GRNs) is fundamental to understand how transcription factors (TFs) regulate gene expression and the complexity of interdependencies among genes. The structure of a network can, in theory, be investigated through experiments by using chromatin immunoprecipitation with DNA microarray (ChIP-chip), ChIP-sequencing [57] or protein-binding microarrays [8]. However, wet-lab experiments are technically challenging, financially demanding, and time consuming [60]. Because genes are expressed at certain levels and under certain conditions, the expression profiles are the outcome of the regulation from the GRN. Therefore, many computational methods have been proposed to infer GRNs using gene expression levels. With the advent of high-throughput transcriptome analysis, such as microarray and RNA sequencing (RNA-seq), the computational inference of a regulatory network on a genome scale has been made feasible. Moreover, inference through computational methods is convenient, the results can be easily reproduced, and there are various ways to validate [68, 58]. From a computational view, such biological networks can be depicted as directed graphs, where TFs and genes are nodes and interactions or regulations are edges.

Several approaches have been proposed to discover gene interactions. Most earlier works focused on unsupervised approaches, such as weighted correlation network analysis (WGCNA) [37], the context likelihood of relatedness algorithm (CLR) [18], and trustful inference of gene regulation using stability selection (TIGRESS) [26]. Because these methods predict a network exclusively from expression data, they have an advantage when gene regulation information is limited. However, with the identification of large numbers of TF-target interactions, their failure to utilize these prior known interactions limits the prediction accuracy. The most recent and largest comparison made by [44] compared 17 unsupervised methods with the supervised method support vector machine (SVM) in three different experimental conditions using both simulated and experimental data sets and found that the supervised method performed best except in knockout experiments. Similar results have been published by [53]. They compared supervised inference of regulatory networks (SIRENE) with CLR, the algorithm for the reconstruction of accurate cellular networks (SIRENE), relevance networks (RN), and a Bayesian network on an *Escherichia coli* benchmark data set by [18] and concluded that the supervised method significantly outperformed unsupervised methods. A more recent publication [23] compared the performance of four kernel functions based on SVM with CLR on simulated *E. coli* microarray data sets, and they concluded that not only experimental conditions, but also network size, played an important role in inference accuracy. SVM with Gaussian kernel inferred small networks (<200 nodes) with the highest prediction accuracy, but, with a larger number of nodes (∼500), CLR outperformed all other methods.

The methods cited above are "non-targeted" [1] or condition-independent approaches, because they provide an overall network structure across many conditions. The drawback of these methods, as reviewed by [70], is that the interactions that occur under specific conditions or biological processes are easily missed. One way to solve this problem is to use the data from experiments that are relevant to the biological question [70]. Here, I focus on the GRNs in the model plant *Arabidopsis thaliana* that occur during specific stages of in seed development.

Seed development is an important process in the life cycle of flowering plants and can be divided into three major stages [4, 47]. First is embryogenesis, where the basic body plan of a plant is established. Second is maturation, where storage compounds are synthesized and accumulated. Third is the acquisition of desiccation tolerance and dormancy. Seed development is a tightly controlled complex process regulated by a variety of endogenous factors including plant growth regulators and ambient conditions such as light, temperature, and water availability. In Arabidopsis, genetic studies have identified some key regulators that globally regulate distinct aspects of seed development [32]. The LEC1/AFL (LAFL) TF network is composed of TFs including B3 domain TFs ABSCISIC ACID (ABA)-INSENSITIVE3 (ABI3), FUSCA3 (FUS3), and LEAFY COTYLEDON2 (LEC2, AFL), and two LEC1-type HAP3 family CCAAT-binding factors, LEC1 and LEC1-LIKE [31]. The TFs in the LAFL network, together with many overlapping and unique downstream targets, constitute a complex transcriptional regulatory network that regulates seed development [49]. Previous efforts to infer the GRN in Arabidopsis seeds, such as the seed specific network associated with dormancy or germination established by [3] used the WGCNA algorithm and 138 samples from mature imbibed Arabidopsis seeds, already made progress on understanding the gene interactions in seeds. However, in seed development, the downstream targets of the well-known core TFs of the LAFL TF network and the TFs regulate them are still poorly understood. Here, I propose to use the condition-specific concept to investigate the transcriptional regulation that occurs during seed development using the expression data of the genes expressed at this particular stage, aiming to reveal the regulations during this biological process.

For the inference algorithm, I developed a Beacon GRN inference tool that uses the supervised method SVM as good results have been produced by SVM in both previous studies and our experiments. In the context of supervised methods, global or pairwise approaches and local approaches are two main categories that have been used in the literature to transform the network inference problem to a classification problem [80]. Global or pairwise approaches consider each pair of genes as a single object, and the classification is performed on these objects [5, 44]. Therefore, the feature vector has to be constructed to define the gene pairs.

Instead of focusing on gene pairs, local approaches divide the inference problem into several smaller classification problems. Each sub-classification problem corresponds to a TF of interest, aiming to infer all the target genes that are connected to this TF [23, 53]. The combination of small networks forms the complete network. I used these two concepts to estimate a global model for all gene pairs and local models for each TF and its target genes in the seed development data set. I evaluate the prediction accuracy of the SVM using two widely used kernel functions in comparison to an unsupervised method, trying to find out a suitable method for inferring the regulatory network with respect to seed development. As a supervised method, SVM requires a list of known regulation relationships between TFs and targets as a training set, which is then used to predict unknown connections. For the TFs, I considered ABI3, FUS3, LEC2 and LEC1, as they are at the core of the LAFL regulatory network as described previously [31]. On the other hand, many previous studies have been dedicated to developing suitable and accurate approaches for predicting, but most of them lack adequate investigation and explanation of the prediction results. Thus, analyzing the predicted network is another key part of our work. I clustered the target expression profiles to analyze co-expressed genes, scanned promoter regions of the targets to search for the ones that contain the binding motifs of the relevant TFs, and studied the functional categories that were enriched in each cluster, all of which gives more meaningful insight into how the TFs regulate Arabidopsis seed development.

Systems Biology Graphical Notation (SBGN) provides a standard for the visual representation of biological processes and networks [39]. It incorporates some easily recognizable glyphs, and can incorporate structural, dynamic information, and thereby has the ability to represent a broad range of biological networks. Here I used the SBGN scheme to present the proposed GRN in seed development in Arabidopsis, and it is drawn by the Beacon editor.

The thesis is structured as follows. Chapter 2 describes some preliminary knowledge, including basic biological and computational concepts used in this research. Chapter 3 gives a brief summary of the signal transduction pathway and formulates the actual problem in a

computational way. In Chapter 4, previous related studies are reviewed and their advantages and disadvantages are discussed. Chapter 5 discusses the data that are used in this research. Chapter 6 is the methodology used to solve the problem raised in Chapter 3. Chapter 7 shows the results and a discussion of the results. Chapter 8 concludes this study.

Chapter 2

PRELIMINARIES

2.1 Biological Concepts

Deoxyribonucleic acid, known as DNA, is the genetic material that stores the information necessary for synthesis, functioning, and development of most living organisms. The building blocks of DNA are nucleic acids. A double helix structure is formed with two polynucleotide chains twisting around each other. Each nucleotide comprises one of the four nitrogenous bases, adenine (A), guanine (G), cytosine (C), or thymine (T), along with a deoxyribose and a phosphate group. It is the specific sequences of the nucleobases that support the DNA's ability to store genetic information. Ribonucleic acid (RNA) is also assembled as a linear chain of nucleotides, but it is more often found as a single stranded molecule. The sugar components of a nucleotide in RNA are ribose, and one nitrogenous base is uracil (U) instead of thymine [72].

A gene is a unit of heredity that is encoded by a segment of DNA and produces a functional product. The functional product of a gene is a protein or a functional RNA. The process of synthesizing proteins is called translation. There are two major steps to access the DNA

sequence of a gene and to produce the corresponding protein. First is transcription. The DNA sequence within a gene is copied to messenger RNA (mRNA). In eukaryotes, the initial product of transcription is pre-mRNA. Pre-mRNA consists of exons and introns, and mature mRNA is formed after introns are spliced out. Second, mRNA is translated into the amino acid sequence of a polypeptide, and then the polypeptide is assembled into a protein (Figure 2.1) [11].

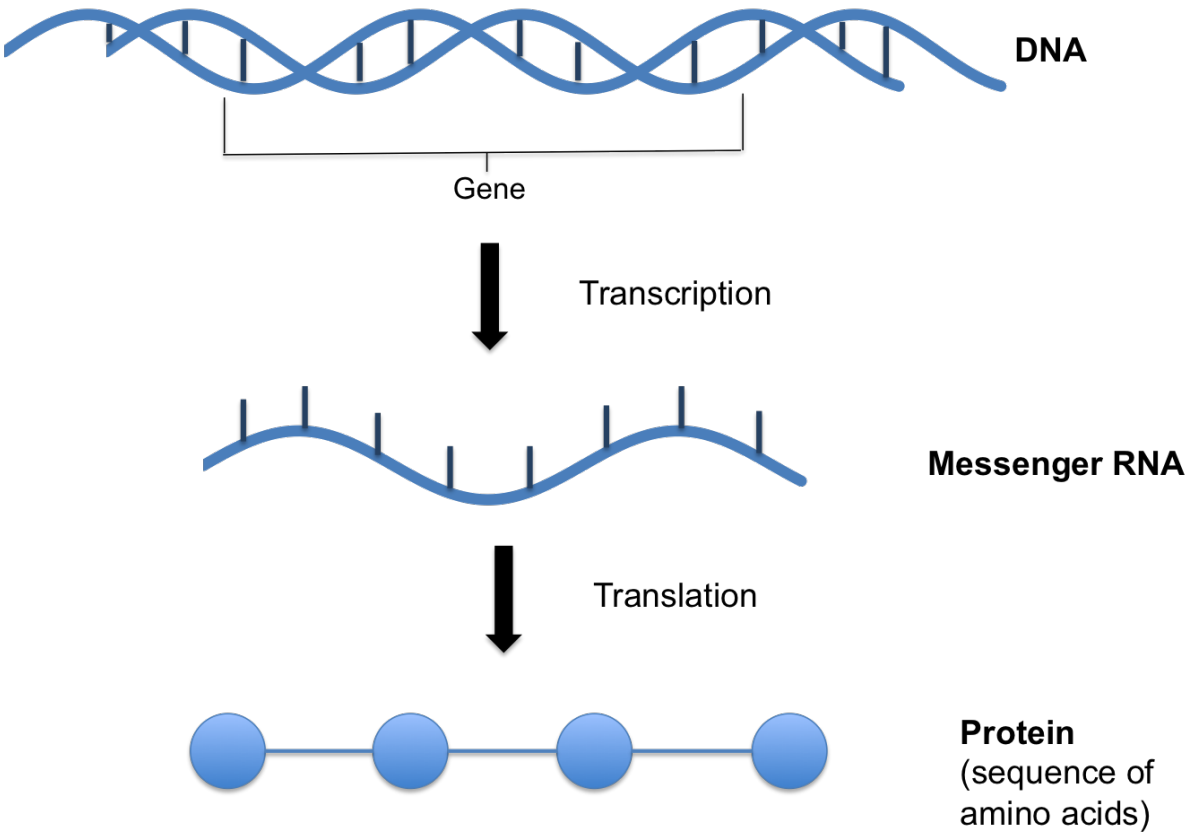


Figure 2.1: Gene expression at the molecular level.

The transcription step plays a crucial role in regulating gene expression. Transcription factors are also known as regulators. They are proteins that can bind to a specific sequence of DNA

to control transcription, either positively or negatively. Transcription factors alone or with other proteins as a complex perform functions of promoting or inhibiting transcription [11].

Genes are expressed at specific times and in specific amounts under different conditions, under the regulation from a wide range of mechanisms [11]. In this work, I obtained and analyzed an *Arabidopsis thaliana* gene expression dataset from developing seeds, trying to infer what genes are regulated by some certain regulators.

Signal transduction pathways are a type of regulatory network that contains a collection of regulators and their targets. The nodes in the network are the genes or transcripts and the edges are interactions. Interactions between nodes are important in controlling gene expression levels and therefore controlling cell functions.

2.2 Computational Terms and Concepts

2.2.1 Types of Graphs

A directed graph is a set of vertices V that are connected by edges and the edges are directed from one vertex to another. The graph is often depicted as $G = (V, E)$ where $E \subseteq \{(u, v)|u, v \in V\}$ [2].

A weighted graph $G = (V, E)$ is a directed graph that has numeric values (weights) assigned to its edges through a weight function $W : E \to \mathbb{R}$.

2.2.2 Euclidean Distance

Euclidean distance [17] is the common distance between two multidimensional points. Given two n-dimension points $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_n)$, the Euclidean distance (d) between $\mathbf{x}$ and $\mathbf{y}$ is defined as:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}.$$

Euclidean distance is computed in clustering target genes (Section 6.5).

Chapter 3

PROBLEM DEFINITION

3.1 Background

A signal transduction pathway is a network of interacting cellular components that mediate the sensing and processing of external signals, such as drought, flooding, heat, cold, ozone, and salt. These pathways coordinate gene expression, enzyme activity, or ion-channel activity by detecting, amplifying, and integrating diverse external stimuli. There are generally four phases for transduction pathways as described below [71] (Figure 3.1).

First, environmental stimuli are transferred to the interior of the cell through the membrane. Some of the signaling molecules diffuse through the membrane and some others transfer information across membrane via membrane-associated receptor protein. Such receptors contain both intracellular and extracellular domains, and the extracellular binding site can specifically recognize the signaling molecule [7].

Second, a received signal is amplified and transduced. The signal information can activate enzymes or membrane channels to produce many small molecules called second messengers.

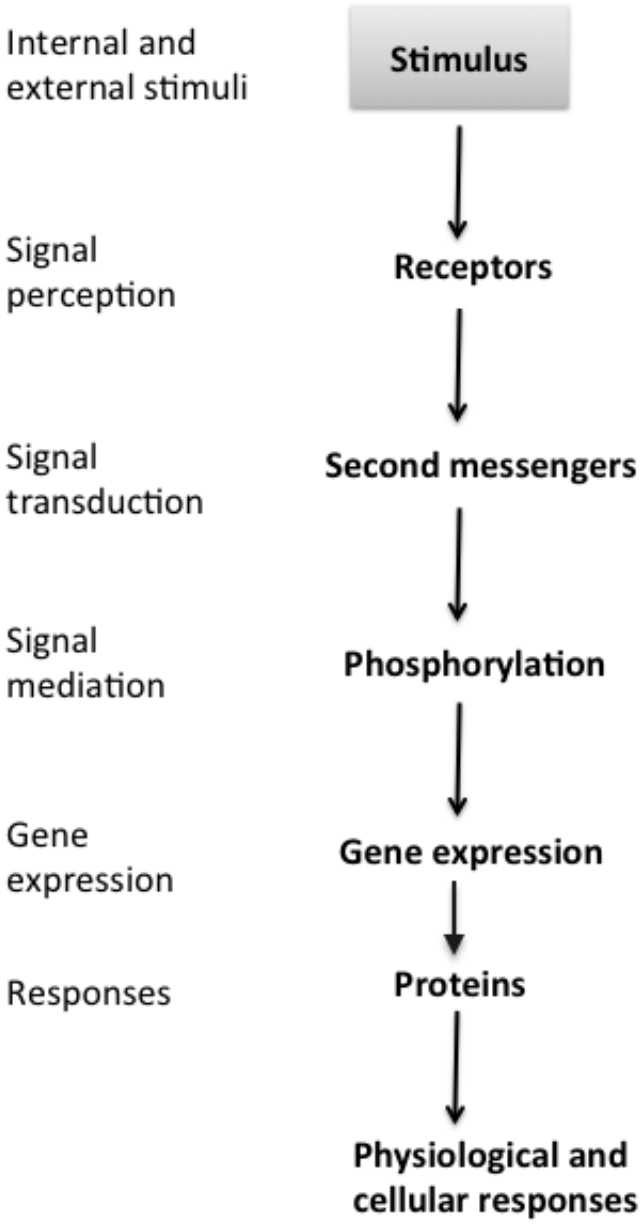


Figure 3.1: Schematic description of signal transduction pathways in plants. Figure adapted from Shinozaki and Dennis [71].

The increased number of second messengers, therefore, is able to significantly amplify even a tiny environmental signal [7].

Third, protein phosphorylation mediates the signal. Second messengers can activate protein kinases to phosphorylate proteins, hence transducing concentration changes in second messengers to protein structure changes. Protein changes can affect cell function and lead to gene activation or repression [7].

Fourth, the signal is terminated. The signaling process is terminated once the information has been transduced to affect other cellular processes [7].

Responses and accommodations to environmental stress are at the heart of many plant activities, and signal transduction pathways are keys to understanding the dynamic logic of those responses. As a result, the elucidation of signal transduction pathways is a big task for many researchers. Though more and more biological and computational approaches are available to study the pathways, the full details of most stress signaling pathway are still not clear. In this thesis, I propose to make use of existing signaling networks in seed development in *Arabidopsis thaliana* to infer potential regulations.

3.2 Problem Definition

Let Y be a set with q genes $Y = \{y_1, y_2, \ldots, y_q\}$, and let T be a set of n time points $T = \{1, 2, \ldots, n\}$. Let $G = (V, E)$ be a directed graph, where $V \subseteq Y$. The expression value for gene y_i at time j is u_{ij} and the gene expression matrix is $\mathbf{U} = (u_{ij})$, where $i \in \{1, 2, \ldots, q\}$ and $j \in T$. In general, gene expression values $u_{ij} \in \mathbb{R}$ and are ≥ 0.

I start with a given time series gene data set of q genes as gene expression matrix $\mathbf{U} = (u_{ij})$, and a directed graph $G = (V, E)$, where regulator $y_r \in V$ and target gene $y_t \in V$ and

$(y_r, y_t) \in E$. Our goal is to predict new edges $(y_r, v) \in E'$ and $(y_r, v) \notin E$, where $v \in V$, to form a new directed graph $G' = (V, E')$.

Chapter 4

LITERATURE REVIEW

Many computational approaches have been developed in the literature from gene expression data to infer regulatory networks. Since the gene expression level is the consequence of regulation, this class of methods is known as the reverse engineering approach [28]. These methods can be classified into two major categories, supervised and unsupervised, according to the classification type.

4.1 Unsupervised Learning Methods

Unsupervised inference methods usually compute a score for the interaction between a pair of genes, based on analysis of their gene expression data.

Correlation-based

Some early models are based on correlation coefficients between expression patterns of all pairs of genes and the pairs with correlated expression profiles are indicative of regulatory interactions. The most commonly used correlation measures are Pearson and Spearman cor-

relations [12]. Later modifications include WGCNA (Weighted correlation network analysis) [37] that amplifies high correlation coefficients by raising the correlation coefficients to the power of β ($\beta \geq 1$).

Mutual information-based

Approaches based on mutual information (MI) can capture both linear and non-linear correlations. The basic idea of these approaches is to compute MI values for all pairs of genes using their expression levels and to infer the regulatory interactions when MI is larger than a given threshold [12]. The network is constructed based on this threshold by including the inferred interactions and a score as the weight of each interaction [36]. Various modifications have been proposed to compute the scores. The ARACNE (Algorithm for the Reconstruction of Accurate Cellular Networks) algorithm [46, 65] filters out the weakest interaction from triplets of genes. The CLR (Contxt Likelihood or Relatedness Network) algorithm [18] modifies the MI score based on the backgroud distribution of all MI scores (See Section 6.3 for more details). MRNET (Maximum Relevance Minimum Redundancy) [50, 51] infers a network of interactions between genes using MI and a feature selection algorithm minimum-redundancy-maximum-relevance (MRMR). Finally, C3NET (conservative causal core) iterates through every gene and considers one edge per gene such that the MI value between this gene and its neighbor is maximal.

Regression-based

The basic assumption of regression based methods is that the putative regulators of a target gene are the ones that are the most informative to predict the expression level of the target gene. The importance of the putative regulators are ranked through regression coefficients. TIGRESS (trustful inference of gene regulation using stability selection) [26] uses the least angle regression (LARS) approach [76] combined with stability selection [48] to assess the significance of candidate regulators. GENIE (gene network inference with ensemble of trees) [29] infers gene network by using a tree-based regression method. It splits the problem of predicting a regulatory network between p genes into p sub-regression problems. In each

of the regression problems, the tree-based method random forest [10] or extra trees [22] is used to predict the expression level of one of the genes from the expression levels of all the other genes. A recent improvement of GENIE is iRafNet (integrative random forest for gene regulatory network inference) [61], which allows information from heterogeneous data to be jointly considered for gene network inference.

4.2 Supervised Learning Methods

In contrast to unsupervised learning, supervised methods exploit some supervised algorithms to classify the unknown gene pairs based on knowledge of part of the network. Global or pairwise and local approaches are two main methods that have been stated in the literature to transform the network inference problem to classification problems [80].

Global or pairwise approaches

A global or pairwise approach considers each pair of genes as a single object, and the classification is performed on these objects [75]. Therefore, the feature vector has to be constructed to define the gene pairs. A simple way to achieve this is to concatenate or add the features from each of the nodes in the pair [14, 56]. Some more complex combination approaches, such as computing outer product of two gene expression profiles [44], have also been proposed. In addition to constructing feature vectors, many classification mathods have been investigated too, including support vector machines (SVMs) [23], logistic regression [45], and tree-based methods [69].

Local approaches

Instead of being interested in gene pairs, local approaches divide the inference problem to several smaller classification problems. Each small classification problem is corresponding to a gene or regulator of interest, aiming to infer all the target genes that are connected to this gene. In principle, any classification method can be used to train each of the small

classification problems, but many recent publications focus on SVM in this context [23, 68, 80].

Except for the methods mentioned above, the network inference problem can also be solved by using the network itself or the network in combination with classification algorithms. For instance, Cheng et al. [15] exploit network topology to derive a similarity measure between nodes and infer new relations using this similarity, and Turki and Wang [79] include network topology features in the supervised learning approach to improve network inference.

4.3 Discussion

The advantage of unsupervised methods is that they can be applied when prior knowledge on gene interaction is limited because new interactions in the network are predicted exclusively from gene expression data. Faith et al. [18] compiled an *Escherichia coli* benchmark data set and found CLR was the top performing method in recovering known interactions when compared with ARACNE, Bayesian networks [21], and linear regression networks. In this work, I consider this state-of-art unsupervised learning method on our Arabidopsis data set. However, the unsupervised methods do not take advantage of known interactions that may improve the prediction accuracy. Numerous papers have been published to discuss and compare the performance of unsupervised and supervised algorithms. The most recent and largest comparison has been conducted by Maetschke et al. [44]. They compared the prediction accuracy of 17 unsupervised methods and one supervised method on both simulated and experimental data sets. Another comprehensive comparison has been done by Madhamshettiwar et al. [43], in which they compared eight unsupervised and one supervised method on 38 simulated data sets. Both of the works show that the methods performed differently on different data sets, but the supervised method was found to be the best across all the experiments. Evaluations from some other earlier studies are in agreement with

these results. For example, Mordelet and Vert [53] compared supervised learning method SVM with CLR, ARACNE, relevance networks, and a Bayesian network on the *Escherichia coli* benchmark data set, and Cerulo et al. [13] made a comparison between SVM, CLR, and ARACNE on the *Escherichia coli* benchmark data set and simulated data set. All their results indicate that the supervised method outperforms unsupervised approaches. As better performance has been reported for supervised methods in all these publications, in this thesis I focus on a supervised learning algorithm.

In the context of supervised learning, the inference approaches that exploit only a part of the network structure does not fit my problem. With only three star sub-networks as prior knowledge, the network topological features are limited. The local model, in my case, is more suitable because I am only interested in predicting the target genes for three particular regulators. I train the local model for each regulator based on its expression profile and the genes it regulates. The most widely used algorithm in local supvervised learning model is SVM [68]. The SVM technique is popular due to its robust performance in classification and ability to tolerate noise [6]. For example, one of the most recent papers by Schrynemackers et al. [69] proposed a tree-based ensemble method and compared its performance with SVM. The comparison was also done by using the *Escherichia coli* benchmark data set, and the result showed that their new method performed very close to, but slightly worse than, SVM.

Chapter 5

DATA MODEL

5.1 Prior Knowledge

It is known that seed development is regulated by many transcription factors and these transcription factors are part of networks that control downstream target genes. In this study, we picked LEAFY COTYLEDON1 and 2 (LEC1 and LEC2), FUSCA3 (FUS3) and ABSCISIC ACID INSENSITIVE3 (ABI3), as they are known to be major regulators of seed development in *Arabidopsis thaliana* [25, 66, 81]. LEC1, which contains a CCAAT-BOX BINDING FACTOR, is required and predominantly expressed during embryo development and seed maturation [41]. In contrast, transcription factors LEC2, FUS3, and ABI3 are more related, since they each contain a DNA-binding B3 domain, and they play critical roles in seed maturation [73].

Previous studies have been carried out to identify the genes regulated by LEC1, LEC2, FUS3, and ABI3. Lists of target genes have been obtained from Junker et al. [34] for LEC1, from Braybrook et al. [9] for LEC2, from Wang and Perry [82] for FUS3, and from Mönke et al. [52] for ABI3. Information including experimental design and number of target genes

are summarized in Table 5.1

Table 5.1: Number of target genes of LEC1, LEC2, FUS3, and ABI3, number of samples, techniques and tissues used in experiments.

Datasets	Number of Samples	Number of Target Genes	Techniques	Tissues
LEC1	16	356	ChIP-chip	Two-week old seedlings
LEC2	8	14	Microarray	8-day old seedlings
FUS3	1	1218	ChIP-chip	Embryonic culture expressing FUS3
ABI3	40	98	ChIP-chip	Two-week old seedlings

Different experimental techniques were used for identifying target genes reported in these publications. DNA microarrays are often used to detect and measure expression levels of genes [78]. Two key concepts are applied in the DNA microarray technique. First is complementary DNA (cDNA), which is a double-stranded DNA synthesized from single strand mRNA. Second is hybridization. Hybridization is the phenomenon that single strand DNA molecules anneal to cDNA with sequence complementarity. A DNA microarray is an orderly arrangement of tens to hundreds of thousands of DNA fragments (probes) of known sequence immobilized to a solid surface (array), such as a small glass, plastic, or nylon membrane [40]. It provides a platform for the probes to hybridize to a cDNA sample (target). After hybridization, radioactive or fluorescent labeled cDNAs are detected and quantified. The intensity of the radioactive or fluorescent signals reveals the level of cDNAs in the samples under study. To discover the target genes for LEC1 [34], LEC2 [9], FUS3 [82], and ABI3 [52], only Braybrook et al. [9] made use of microarray data. Since LEC2 activity can be induced by treating Arabidopsis containing the *35S:LEC2-GR* chimeric gene with the steroid-hormone

analogue dexamethasone (Dex), they isolated RNA from *35S:LEC2-GR* seedlings that were treated with Dex for either 1 or 4 hours (1-h or 4-h Dex). Microarray analysis was conducted and the RNAs present in only 1-h Dex or only 4-h Dex or in both treatments were reported as being induced by LEC2. Because only 14 target genes were reported in this data set, no statistically significant result can be inferred from such a small number of relations, so I eliminated the use of this data set as prior knowledge for training.

On the other hand, the ChIP-chip technique utilizes a combination of chromatin immunoprecipitation (ChIP) and whole-genome DNA microarrays (chip) and is often used to investigate interactions between proteins and DNA [30, 63]. ChIP is one of the techniques that investigates protein-DNA interactions. Transient protein-DNA interactions in living cells can be captured using crosslinking agents. By using specific antibodies to a putative DNA binding protein, a protein-DNA complex is immunoprecipitated. Reversing the cross-linking of the complex allows the DNA to be separated from the proteins. The DNA microarray technique is then used to identify the separated DNA fragments that interact with the protein. ChIP-on-chip allows for high resolution of genome-wide maps and can determine binding DNAs. However, the binding DNAs are not necessarily true downstream genes of a transcription factor because the association may be due to a pure physical interaction other than gene regulation.

In addition to the four major regulators, phytohormone abscisic acid (ABA), a ubiquitous plant hormone, plays an important role in development processes such as seed dormancy, germination, and embryo maturation. In seed development, ABA not only regulates maturation, dormancy, and germination but also mediates responses to abiotic stresses such as drought, cold, and salinity [16, 67]. As was reviewed in Finkelstein [20], ABA and regulatory elements interact to control the seed maturation and germination processes (Figure 5.1). For example, LEC1 promotes expression of LEC2 and FUS3, which are active in promoting expression of ABI3, ABI4, and ABI5, which in turn regulates the ABA response [33, 74]. Interactions shown in Figure 5.1 were taken as part of the prior knowledge.

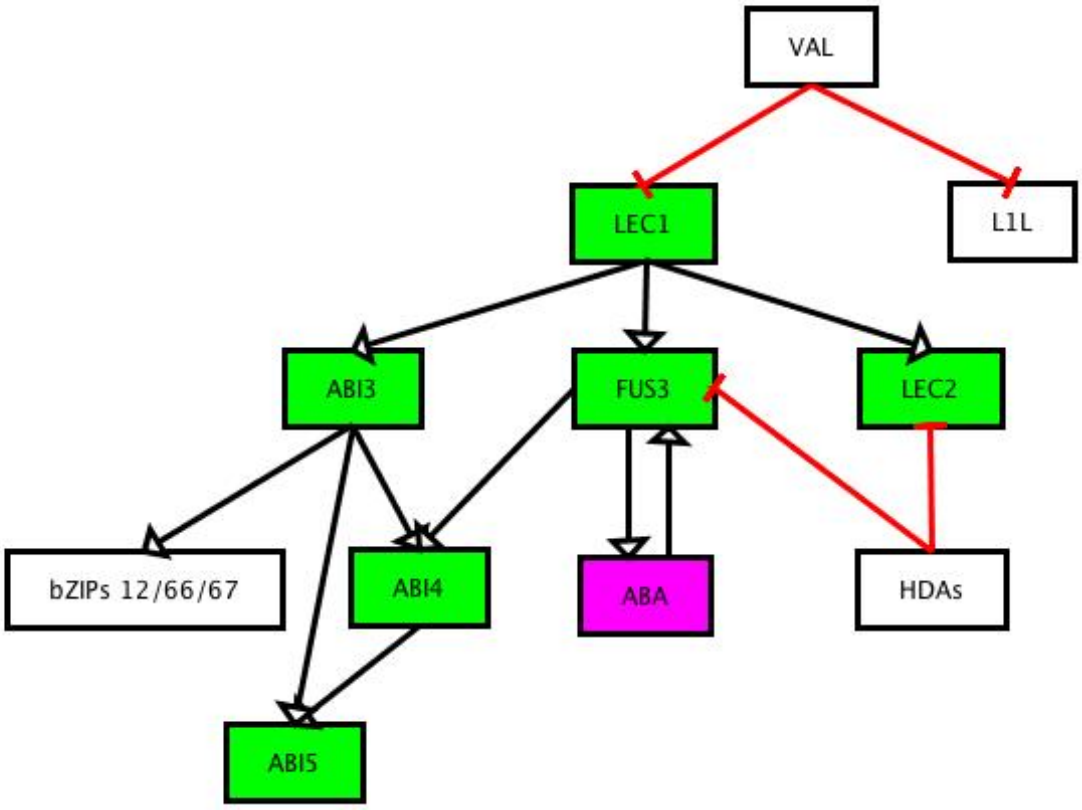


Figure 5.1: Interactions among some of the hormonal and developmental signals and regulatory elements controlling seed maturation. Arrows represent positive regulation and red bars indicate repression.

Figure drawn by Beacon editor and it is adapted from Finkelstein. [16].

5.2 Experimental Data

In this thesis, expression data from the model organism *Arabidopsis thaliana* in seed development has been analyzed to infer possible gene networks. Two data sets, genes, differentially expressed genes, and differentially expressed transcripts in time course data, were provided by Dr. Eva Collakova and Dr. Ruth Grene, professors at Virginia Tech. Further details of the experiments can be found in [67]. In their work, high-throughput RNA sequencing (RNA-Seq) experiments were conducted on developing Arabidopsis seeds and the Tuxedo Suite [77] was used for RNA-Seq data analysis. Each expression data set can be divided into a gene/transcript name data set and a time course gene expression data set. The gene expression data is already normalized to fragments per kilobase of transcript per million mapped read (FPKM) values.

5.2.1 Data Set 1 (Gene Time Course Data)

This data set contains the expression levels of all detectable known transcripts in the developing embryo of *Arabidopsis thaliana*. The gene name data set contains six fields described as follows. First, gene_id is the unique identifier of the gene, and it is assigned by Tuxedo Suite [77]. Second, gene_name is the gene annotation in TAIR10. Third, AGI index is the gene identifier as per the *Arabidopsis thaliana* initiative. Fourth, tss_id is the transcription start site of the gene. Fifth, locus_id shows the specific location of the gene located on the chromosome. Sixth, the description field briefly describes the gene function (Table 5.2).

Table 5.3 shows an example of time course gene expression data, and it is related to Table 5.2 via unique gene identifier gene_id. The expression matrix only contains transcripts whose expressions were detectable at one or more time points throughout the time course of embryo development. There are 23866 genes contained in this data set, and 7 time points are available for each gene: day_7, day_8, day_10, day_12, day_13, day_15 and day_17.

Table 5.2: *Arabidopsis thaliana* gene name data set of Data set 1.

gene_id	gene_short_name	AGI	description	tss_id	locus
XLOC_000001	ANAC001	AT1G01010	NAC domain containing protein...	TSS1	1:3630-5899
XLOC_000002	DCL1, MIR838A	AT1G01046; AT1G01040	dicer-like 1 (DCL1); CONTAINS...	TSS2, TSS3, TSS4	1:23145-33153
XLOC_000004	IQD18	AT1G01110	IQ-domain 18 (IQD18);FUCTION...	TSS6, TSS7, TSS8	1:52047-54692
XLOC_000006	GIF2	AT1G01160	GRF1-interating factor 2 (GIF2)...	TSS10	1:72338-74967
XLOC_000007	AT1G01180	AT1G01180	S-adenosyl-L-methionine-dependent...	TSS11	1:75582-76758

Table 5.3: *Arabidopsis thaliana* time series data set of Data set 1.

gene_id	d7_FPKM	d8_FPKM	d10_FPKM	d12_FPKM	d13_FPKM	d15_FPKM	d17_FPKM
XLOC_000001	0.87684	0.555093	0.133433	0.418929	0.504827	1.13617	1.34929
XLOC_000002	1.62084	2.02802	5.26349	6.25829	4.89964	3.41851	3.59797
XLOC_000004	5.94789	3.87047	2.2977	2.63623	1.36597	0.434363	0.17935
XLOC_000006	33.0851	30.3209	27.2214	30.3298	37.7312	61.1768	64.1005
XLOC_000007	0.0447507	0.0552952	0.193438	0.814342	1.62088	14.3186	15.4317

5.2.2 Data Set 2 (Differentially Expressed Gene Time Course Data)

Limma analysis [64] was done on Data Set 1 in this thesis, and genes that are differentially expressed at least at one time point with respect to its following time point are recorded in Data Set 2. Information about Limma differential analysis is described in Section 5.1. The columns in Data Set 2 are the same as in Data Set 1. The 7376 entries in this data set represented 7376 genes that are differentially expressed in seed development.

Chapter 6

METHODOLOGY

Many computational methods have been developed to infer gene networks, using both unsupervised and supervised approaches. In this thesis, methods from these two categories were applied, and the results were compared. Unsupervised learning does not require any labeled data, while supervised methods exploit labeled training data to find the optimal model parameters. The following sections describe my methodology of evaluating classifiers and inference methods in detail.

6.1 Data Analysis

The expression of all transcripts data set was directly obtained from Dr. Eva Collakova and Dr. Ruth Grene. This data set contains 53,989 entries. Gene level FPKM values can be calculated by adding up the expression values from all the transcripts detected for each gene. This is the gene data set presented in Tables 5.2 and 5.3.

Limma analysis [64] was then applied to find differentially expressed genes as [67] did for finding differential expressed transcripts between wild type and *val*1 mutant embryos. Instead of using FPKM values, Limma requires raw counts as input data, and the raw counts are the number of reads overlapping a given gene. This data set was also obtained from Dr. Eva Collakova and Dr. Ruth Grene. In the Limma pipeline, the VOOM package [38] was first used to normalize the counts. Empirical Bayes, moderated t-statistics, and their associated p-values were then used to assess the significance of the observed expression changes between one time point with respect to its following time point. Genes with adjusted p-value < 0.05 were declared to be differentially expressed. The differentially expressed genes belong to the data set presented in Section 5.2.2.

6.2 ROC and AUC

To evaluate the performance of the inference algorithm, I drew receiver operator characteristic (ROC) curves and computed the area under the receiver operator characteristic curve (AUC) as has been used by much previous work [27, 35, 53, 55]. ROC curves show the true positive rates over the full range of false positive rates at different thresholds, and AUC quantifies the quality of the classifier. The AUC value represents the probability that the classifier ranks a randomly chosen positive instance higher than a randomly chosen negative instance. AUC is a portion of a unit square and hence its value will always be between 0 and 1. An AUC above 0.5 is expected for a realistic classifier since it should perform better than random guessing, while an AUC of 1 indicates perfect performance [19].

An unsupervised method does not require any parameter optimization. For supervised methods, on the other hand, 3-fold cross validation was applied and parameters were optimized on the training data only (See Section 6.4.3).

6.3 Unsupervised Learning - CLR Algorithm

The CLR (Context Likelihood of Relatedness) method is a widely used unsupervised learning method for gene network inference, which was first introduced by Faith et al. [18]. In this work, the CLR method was implemented according to the publication and was called with its default parameters.

CLR extends the relevance network method [12] and makes use of mutual information (MI) values. MI between two discrete random variables X_i and X_j is defined as

$$I(X_i, X_j) = \sum_{x_i \in X_i} \sum_{x_j \in X_j} p(x_i, x_j) \log \frac{p(x_i, x_j)}{p(x_i)p(x_j)},$$

where $p(x_i)$ and $p(x_j)$ are marginal probabilities, and $p(x_i, x_j)$ is the joint probability distribution of X_i and X_j.

CLR calculates the MI values between all gene pairs and produces an MI matrix $\mathbf{M}$, where $\mathbf{M}_{ij}$ is the MI value between gene i and gene j. The background MI distribution is then taken into account to estimate the interaction between genes i and j. The background distribution consists of two sets of MI values: all the MI values for gene i, $\mathbf{M}_{ik}$, $k = 1, \ldots, n$, and all the MI values for gene j, $\mathbf{M}_{kj}$, $k = 1, \ldots, n$. The CLR technique assumes that the interactions with MI that deviate most from the background distribution are the most probable interactions. Thus, a maximum z-score is computed for each gene i as

$$z_i = \max_j(0, \frac{\mathbf{M}_{ij} - \mu_i}{\sigma_i}),$$

where μ_i and σ_i are the mean value and standard deviation, respectively, of the MI values $\mathbf{M}_{ik}$. The final form of the CLR likelihood estimation between gene pair i and j is

$$w_{ij} = \sqrt{z_i^2 + z_j^2}.$$

Putative regulator-gene interactions are then ranked by decreasing w_{ij}.

6.4 Supervised Learning - SVM Algorithm

A variety of different supervised machine learning approaches are available, but I limited my inference to support vector machines (SVMs), as good results have been produced using this method in previous studies [44, 53]. I used the Python implementation of SVM called sklearn.svm, published by Pedregosa et al. [59].

6.4.1 Feature Vector

An SVM algorithm operates on column feature vectors. Various ways can be used to construct the feature vector. In this thesis, I compared the performance of two of them. Let t be the target gene, r be the regulator, $i = 1, \ldots, k$ be the time point, and $e(t_i)$ and $e(r_i)$ be the expression levels of genes t and r at time point i, respectively; feature vector of the gene pair (r, t) is defined as $\mathbf{x}$. The first way of constructing $\mathbf{x}$ is to directly concatenate the expression data of regulator and target: $\mathbf{x} = (e(r_1), \ldots, e(r_k), e(t_1), \ldots, e(t_k))^T$. This belongs to global approach because each gene pair is treated as a single object and only one SVM is used to train and predict. The second way is $\mathbf{x} = (\log \frac{e(t_2)}{e(t_1)}, \ldots, \log \frac{e(t_k)}{e(t_{k-1})})^T$, which belongs to the local approach because each regulator is treated as a separate SVM.

6.4.2 Kernel Function

The kernel function is a fundamental component for an SVM algorithm. Given r as the regulator and n target genes $t_1, \ldots, t_n$, the gene pairs $(r, t_1), (r, t_2), \ldots, (r, t_n)$ belong to two classes +1 and -1. Class +1 means gene r regulates gene t; class -1, in contrast, means gene r does not regulate gene t. The basic idea is to construct a hyperplane to separate these two classes, and the optimal hyperplane maximizes the distance of the closest point to the hyperplane. Let $\mathbf{x}_i$, l_i, and $\mathbf{x}_j$, l_i denote the feature vectors and labels of gene pairs (r, t_i) and (r, t_j), respectively, and the kernel function between $\mathbf{x}_i$ and $\mathbf{x}_j$ is $k(\mathbf{x}_i, \mathbf{x}_j)$. The SVM is trained through maximizing a constrained, quadratic optimization problem over Lagrange multipliers α:

$$\max_{\alpha} L(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j l_i l_j k(\mathbf{x}_i, \mathbf{x}_j)$$

subject to

$$\begin{cases} \sum_{i=1}^{n} \alpha_i l_i = 0 \\ 0 \leq \alpha_i \leq C \text{ for } \forall_i. \end{cases}$$

C is the complexity parameter that needs to be tuned for optimal prediction performance. With a very high value of C, the training mistakes have very high cost. In here, I chose $C = 1000$ to train all SVMs. This choice was also used by SIRENE [53].

The prediction makes use of the optimized α_i. Let $\mathbf{x}'_j$ denote the feature vector of a new gene pair (r, t_j), the kernel function between $\mathbf{x}_i$ and $\mathbf{x}'_j$ is $k(\mathbf{x}_i, \mathbf{x}'_j)$. An SVM estimates a scoring function for any new gene pair (r, t_j) of the form:

$$f(\mathbf{x}'_j) = \sum_{i=1}^{n} \alpha_i k(\mathbf{x}_i, \mathbf{x}'_j).$$

In this way, the scoring function $f(\mathbf{x}'_j)$ classifies gene pairs from unknown classes in the test set (Figure 6.1).

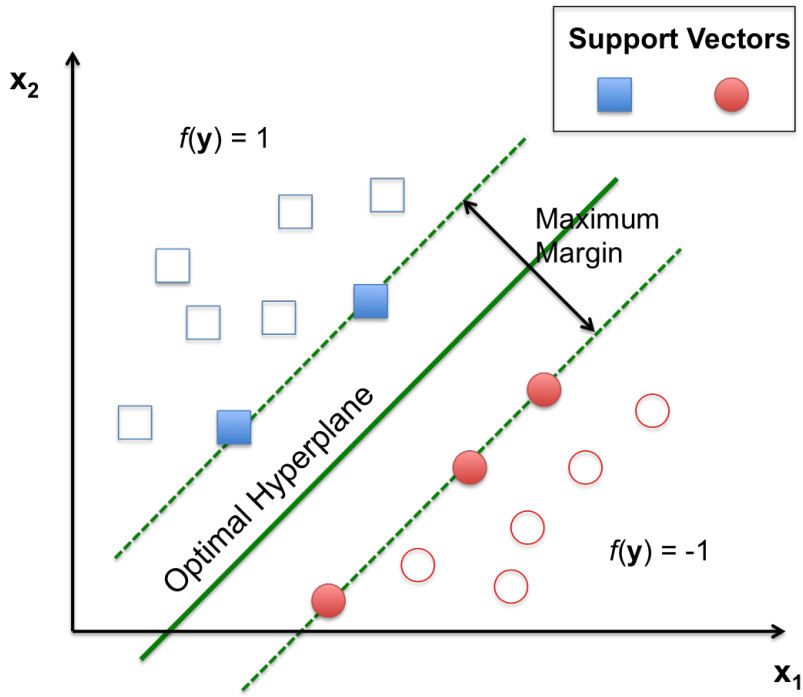


Figure 6.1: Two dimensional representation of SVM using maximum margin with support vectors to classify genes.

To find out the SVM kernel with the best performance, I did experiments to evaluate the following two kernel functions. Though there are many kernel functions available, these two are mostly used in gene network inference and have proved to perform well in previous studies [13, 44, 53].

1. Linear Kernel

   The linear kernel is the simplest kernel function for an SVM. The linear kernel is defined as the dot product of two vectors $\mathbf{x}$ and $\mathbf{x}'$ with addition of c constant:

   $$k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T\mathbf{x}' + c.$$

2. Gaussian Kernel

   The gaussian kernel is a radial basis kernel function or RBF kernel defined by

   $$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\gamma||\mathbf{x} - \mathbf{x}'||^2\right),$$

   where $\gamma = \frac{1}{2\sigma^2}$ and $\sigma > 0$. σ is a parameter that controls the width of the Gaussian. If σ is underestimated, the kernel becomes more local and forms greater curvature of the decision surface, which makes the radius of the area of influence of the support vectors too small so that it only includes the support vector itself. If overestimated, the model behaves similarly to the linear model, resulting in a failure to capture the shape of the data. In this thesis, I used the default choice $\gamma = \frac{1}{\text{number of features}}$.

6.4.3 Cross Validation

As a supervised learning method, SVM needs both positive and negative examples in the training sets. Positive examples are known relationships between well-studied regulators and their targets as described in Section 5.1. However, there is little information about a regulator not regulating some certain target genes. In this thesis, I assigned regulator-target gene pairs not reported in the prior knowledge to negative examples. All genes known to be regulated by this regulator form a set of positive examples, and the same number of genes were randomly chosen from the remaining genes to form the set of negative examples. A

3-fold cross validation is done by randomly splitting the positive and negative example sets into three parts, training the SVM on two of the subsets, and evaluating the prediction on the third subset. This process was repeated three times, testing successively on each subset. The prediction quality is averaged over all three iterations.

6.4.4 Ranking

The output of an SVM prediction is a label $l = +1$ or $l = -1$ in my case. However, except for knowing which class the gene belongs to, we sometimes are also interested in the degree of certainty about the classifier. Platt Scaling [62] can transform SVM predictions to posterior probabilities by passing them through a sigmoid. The output of SVM is $f(\mathbf{y})$ as described in Section 6.4.2, Platt scaling passes $f(\mathbf{y})$ through a sigmoid to get calibrated probabilities:

$$P(l = 1|f) = \frac{1}{1 + \exp(Af + B)},$$

where the parameters A and B are fitted using maximum likelihood estimation from a fitting training set (f_i, l_i).

Platt scaling has been shown to be effective for SVMs [54], and the implementation of SVM I used in this thesis incorporated Platt scaling to output probabilities [59].

6.5 Clustering

To analyze target genes and visualize their expression patterns, I grouped these genes by similar expression profiles using the k-means clustering algorithm [42], as implemented in

Python [59]. It is a partition-based clustering method that can automatically partition a data set into k groups. Given a predetermined number k, and a set of gene expressions $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$, where each gene expression is a k-dimensional vector, the goal is to minimize the objective function

$$E = \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} |\mathbf{x} - \boldsymbol{\mu}_i|^2,$$

where $\boldsymbol{\mu}_i$ is the centroid of cluster S_i. Thus, E is to minimize the sum of squared distances (euclidean distance) of gene expressions from their cluster centers. It proceeds by randomly choosing k cluster centers and then iteratively updating them as follows:

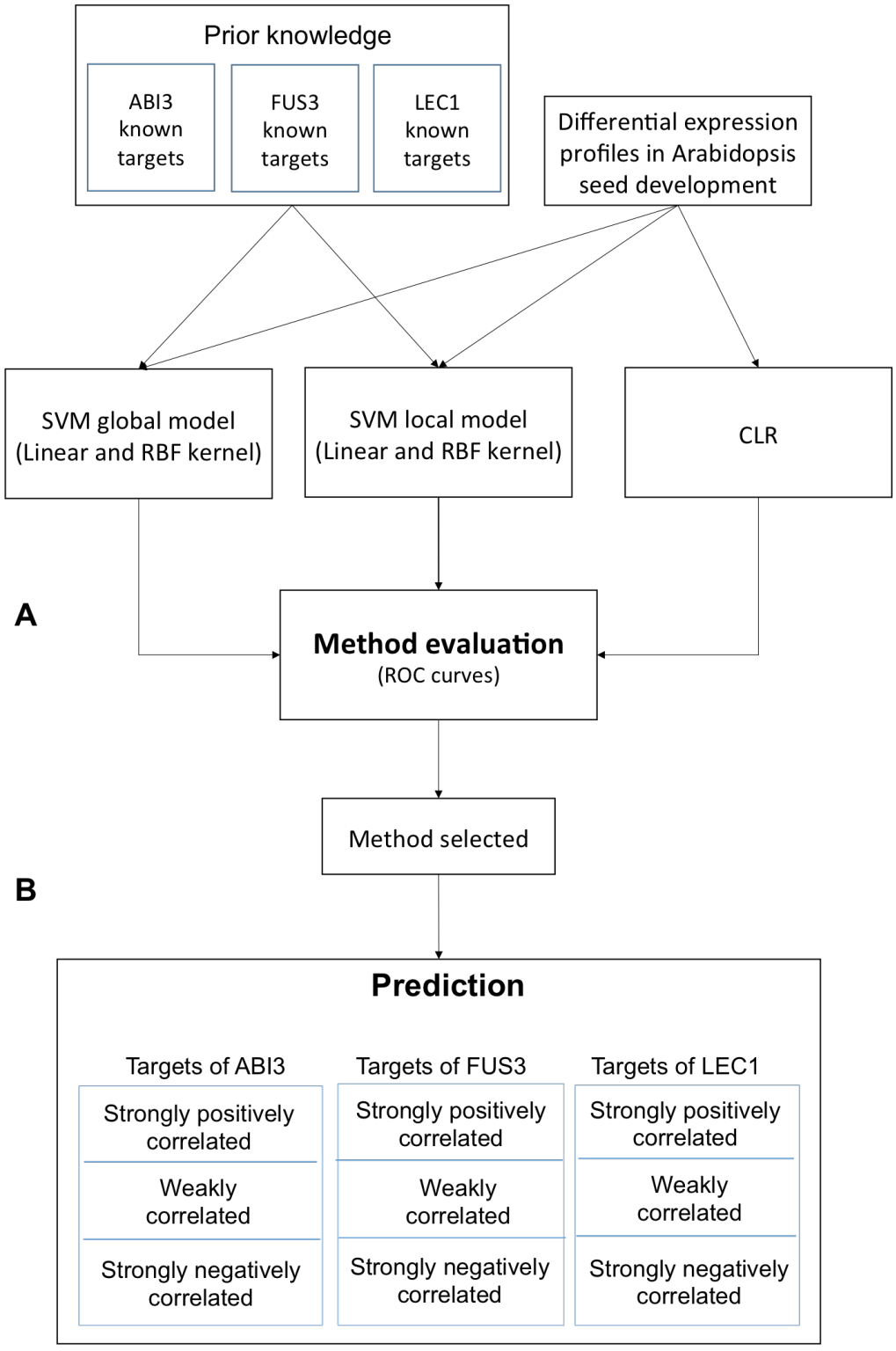
1. Each gene is assigned to its closest cluster center.
2. Each cluster center is updated to the mean of its constituent genes.

The algorithm converges when there is no further change in assignment of genes to clusters.

6.6 Experimental Procedure

The workflow for the Beacon GRN inference tool contains five phases, namely, comparison, prediction, clustering, searching for direct and indirect targets of regulators, and searching for direct and indirect targets of secondary TFs; they are shown in Figure 6.2. The purpose of the comparison phase is to generate the ROC curve using the supervised method with global and local SVMs and the unsupervised method CLR (Figure 6.2A). To train the SVM, two types of inputs are required. First is a list of gene names to be trained and tested and their expression levels. Second is a list of positive examples and negative examples. With the prepared lists, the problem is divided into three subproblems in the local model, and

each subproblem is related to one of ABI3, FUS3, and LEC1. An SVM classifier is trained for each regulator based on its known target genes and non-target genes. For the global model, the three subproblems are combined to obtain one problem, where a global SVM classifier is trained based on all known regulations and non-regulations. The list of testing relationships can then be assigned into different classes according to the trained SVM. This process is repeated for each kernel. Since the CLR algorithm does not require a training data set, the final ROC curve is generated on all genes simultaneously. The approach with the higher accuracy are used to predict new target genes of the regulators (Figure 6.2B). The result of prediction is three networks with the center node of each being ABI3, FUS3, or LEC1. The target genes controlled by single or multiple regulators are identified. The following procedures are all based on individual networks. Pearson correlation is performed to determine how the expression levels of the targets are correlated with the expression levels of their corresponding regulator. A threshold of 0.6 is chosen to filter the strongly correlated targets. The third phase is to group the known and predicted strongly positively correlated target genes according to their expression patterns (Figure 6.2C). With co-expressed targets in each regulator, the FIMO (Find Individual Motif Occurrences) algorithm was used to search for the direct targets of each cluster (Figure 6.2D) [24]. Finally, in each cluster, secondary TFs and their binding motifs among the direct targets were identified, and FIMO was run again on indirect targets in this cluster to identify the direct targets of these secondary TFs (Figure 6.2E). As reviewed by [32], LEC1 influences ABI3, and ABI3 and FUS3 are mutually regulated. Combining these relationships with our predicted three sub-networks, I obtain the entire network as shown in Figure 6.3.


Prior knowledge
ABI3 known targets
FUS3 known targets
LEC1 known targets
Differential expression profiles in Arabidopsis seed development
SVM global model (Linear and RBF kernel)
SVM local model (Linear and RBF kernel)
CLR
A
Method evaluation
(ROC curves)
Method selected
B
Prediction
Targets of ABI3
Targets of FUS3
Targets of LEC1
Strongly positively correlated
Weakly correlated
Strongly negatively correlated
Strongly positively correlated
Weakly correlated
Strongly negatively correlated
Strongly positively correlated
Weakly correlated
Strongly negatively correlated

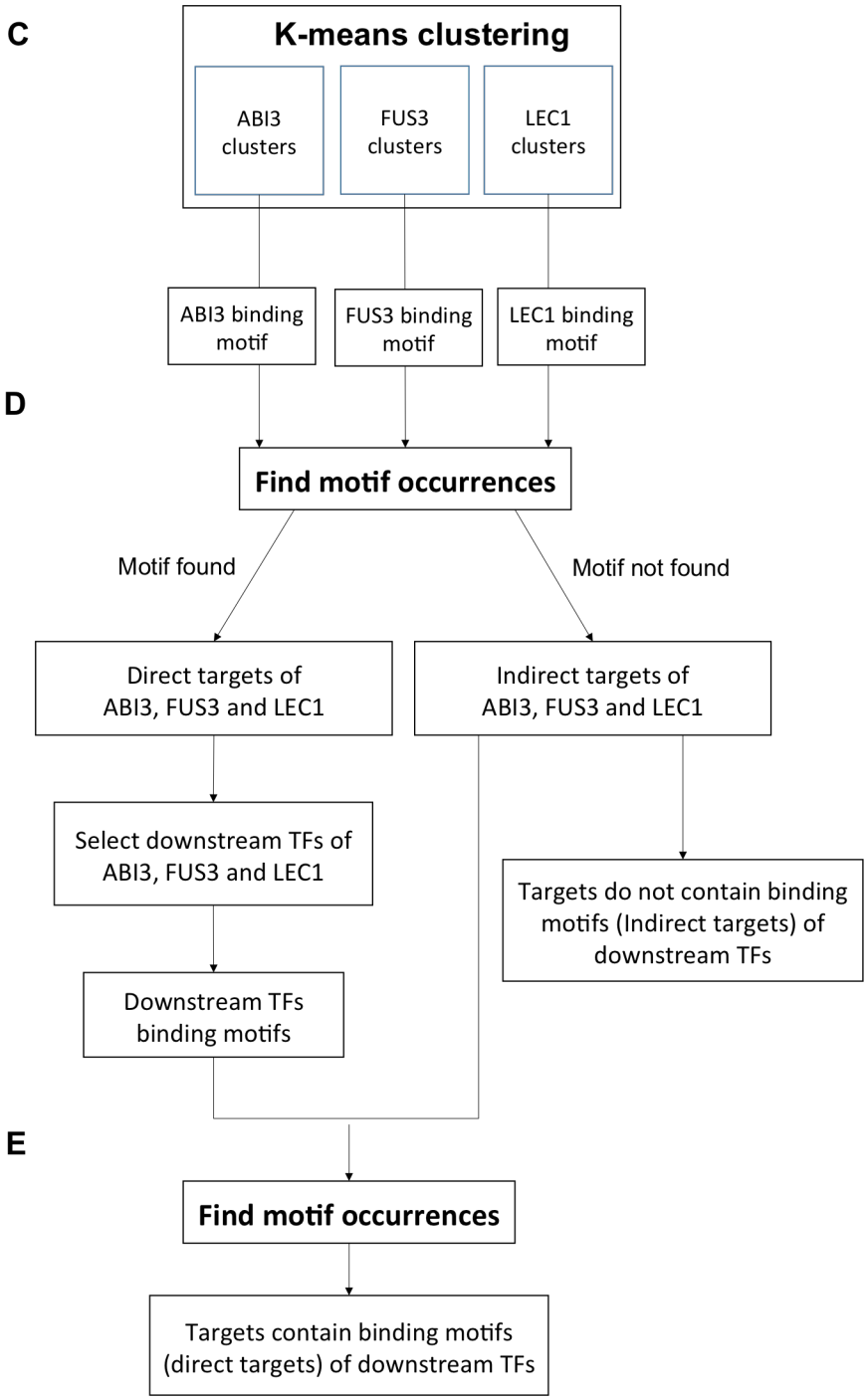


Figure 6.2: Beacon GRN inference and validation workflow. Five phases: method comparison (A), prediction (B), k-means clustering (C), identify the targets contain binding motifs (D), and identify targets containing the downstream TF binding motifs (E). K-means clustering is done by combining strongly correlated known and predicted targets.

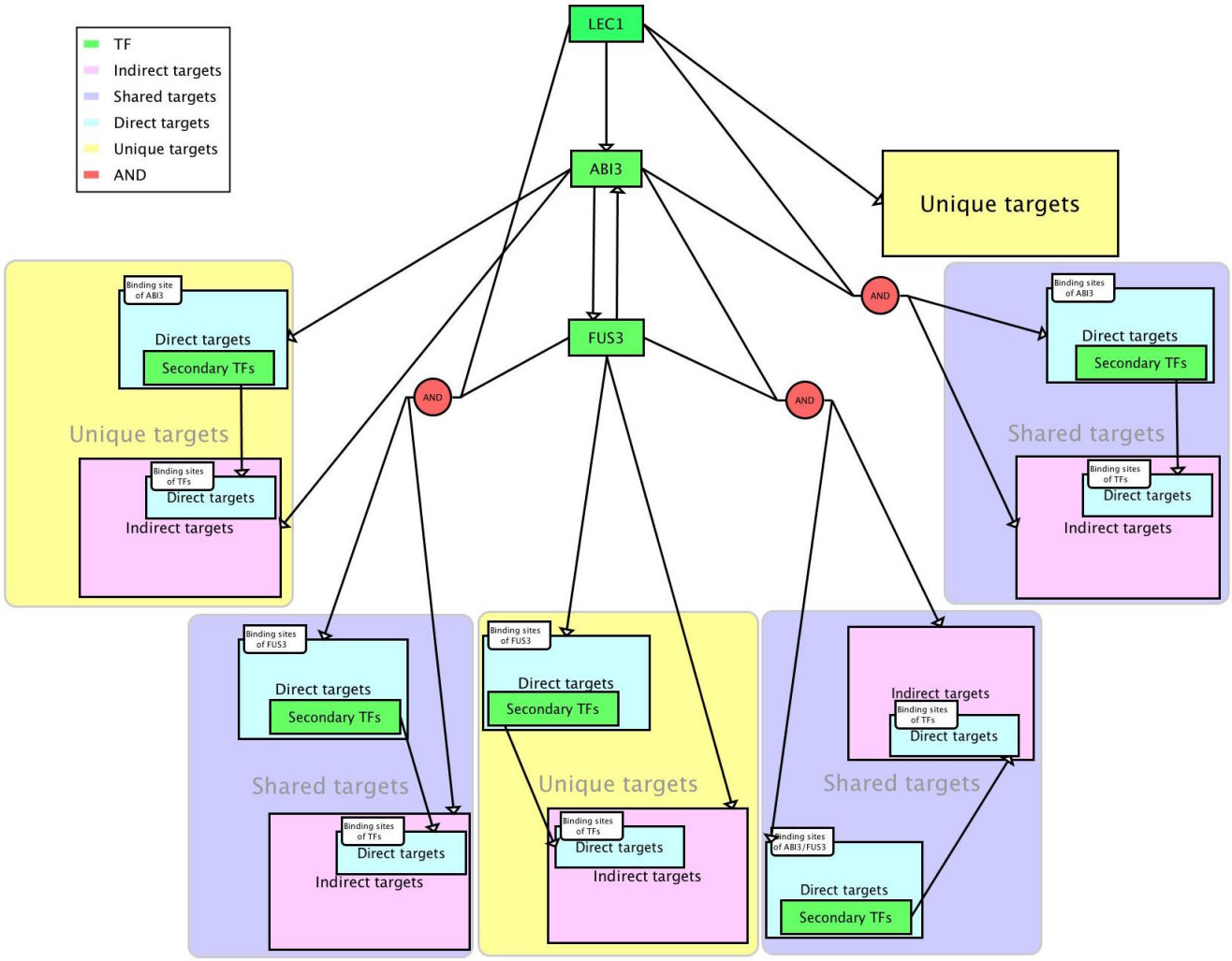


Figure 6.3: The proposed network. The diagram is drawn in Systems Biology Graphical Notation (SBGN) format using the Beacon editor [39]. LEC1, FUS3 and ABI3 represent three master regulators, with ABI3 directly controlled by LEC1 and ABI3 and FUS3 mutually regulated.

Chapter 7

RESULTS AND DISCUSSION

7.1 Results

7.1.1 Algorithm Evaluation and Comparison

In the following, I first evaluate the performance of the SVM before comparing CLR with the best performing SVM model. Figure 7.1 shows the comparison between the prediction accuracies measured by AUC for linear and RBF SVMs. Figures 7.1A through 7.1C are the results of local models. Among all the three regulators, the SVM of ABI3 with AUC approximately 0.9 performs the best. Figure 7.1D is the result of the global model. The global model performs worse than ABI3 but is comparable with FUS3 and LEC1. To compare the performance between the two kernels, they result in similar AUC values with the RBF kernel somewhat better for all four cases. The reason why the global approach performs not as well as the local approaches is due to its failure to capture the unique characteristics of different regulators. Different regulators may have different regulation mechanisms, and thus it is hard to learn all different features in one SVM. Furthermore, as summarized in Table

7.1, FUS3 has 1045 known target genes, which is much greater than the known targets of the other two. Hence, the majority of the positive examples consist of FUS3 regulations, while FUS3 relations are the minority in the negative example set. The consequence is the SVM may simply capture the feature of FUS3 regulation as positives and considers all features different from FUS3 regulations as negatives. Since the local model is more meaningful and powerful than the global model in our case, I focused our study on the local model with the RBF kernel. The SVM local RBF model was then compared to the CLR algorithm, and Figure 7.2 shows the ROC curves. The CLR prediction accuracy approaches 55%, which performs much worse than the supervised method.

The evaluation of the methods indicate that a local SVM model with RBF kernel is the most suitable method for predicting regulatory network among the three regulators using the differentially expressed gene data in Arabidopsis seed development. I call this approach the Beacon GRN inference tool.

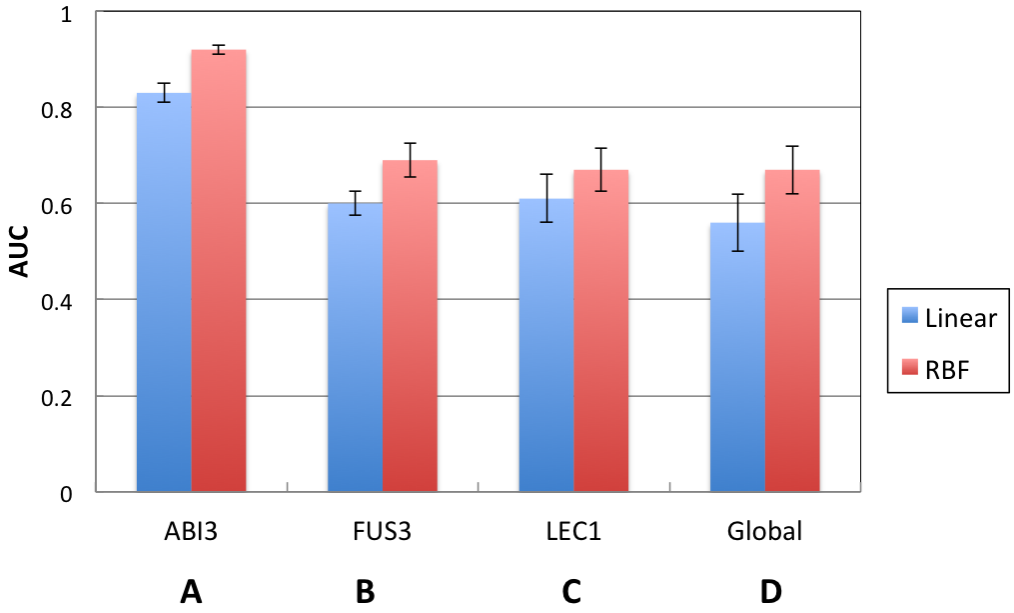


Figure 7.1: Comparison of performance between SVM local models and global model. ABI3, FUS3 and LEC1 represent local models with each of them as a separate SVM. Global model trains one SVM for all the TF-target pairs.

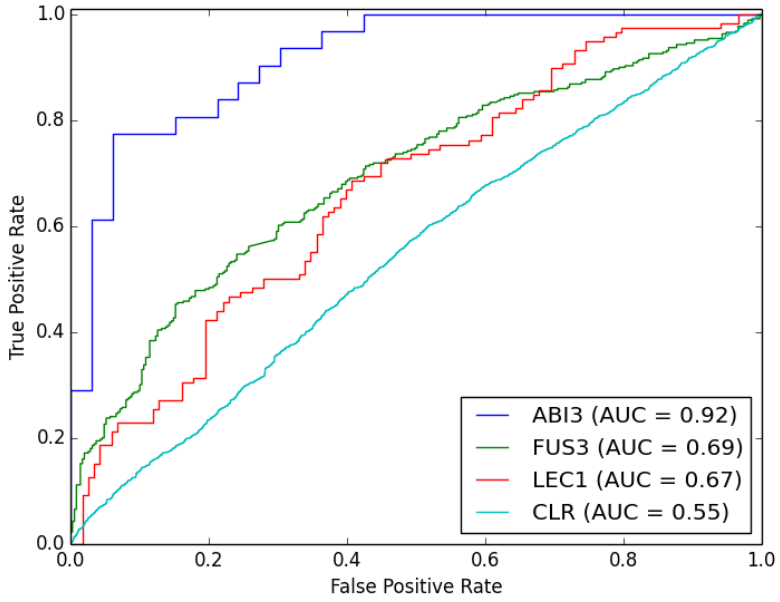


Figure 7.2: Comparison of performance between SVM local models and CLR algorithm.

Table 7.1: Summary of the number of prior known regulations in Arabidopsis seed development gene and differentially expressed gene data sets.

Regulator	Number of Target Genes in the Gene Data Set	Number of Differentially Expressed Target Genes	Number of Not Differentially Expressed Target Genes
LEC1	353	174	179
LEC2	14	14	0
FUS3	1045	508	537
ABI3	98	94	4

7.1.2 Network Prediction

As described in Section 7.1, I treated ABI3, FUS3, and LEC1 as separate SVMs to predict networks based on all the differentially expressed genes. The predicted networks were then combined to make one network.

I used all 98, 1045, and 353 positive examples and the same number of negative examples as the training sets for ABI3, FUS3 and LEC1, respectively. Overall statistics for the predictions are presented in Table 7.2 and Figure 7.3. Table 7.2 lists these three regulators along with the number of targets they regulate. Figure 7.3 presents the portion of unique and shared target genes controlled by two or three of the regulators.

Table 7.2: A summary of the number of predicted and unique targets for each regulator.

Regulator	Number of Predicted Targets	Number of Unique Targets
ABI3	1064	275
FUS3	2569	862
LEC1	3836	1732

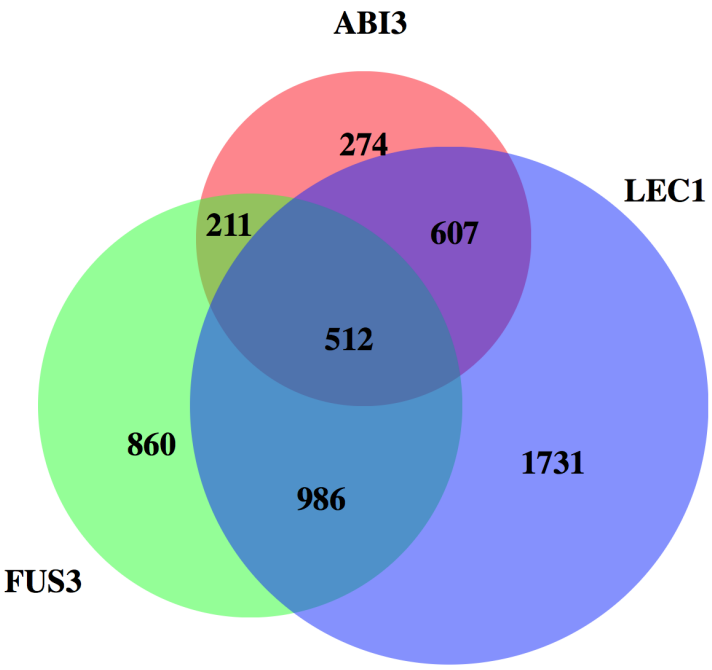


Figure 7.3: A Venn diagram depicting the overlap between the predicted targets among the three regulators.

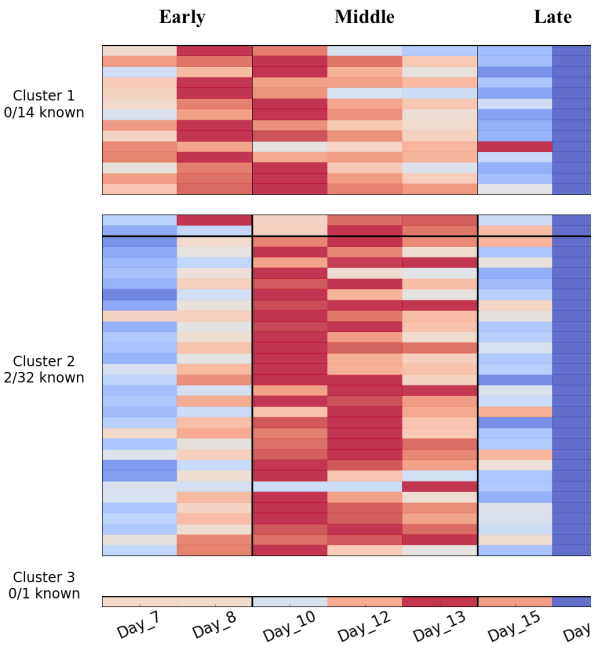
7.1.3 Biological Validation

To further filter the prediction, I identified the targets whose expression levels are strongly positively correlated with the expression level of their related regulator (Table 7.3). Approximately half of the FUS3 and LEC1's targets were discarded according to the correlation coefficient threshold setting of 0.6. Only 47/1698 ABI3 targets were found strongly positively correlated, which is because ABI3 is not differentially expressed over the time course but its potential targets are. Our analysis in the following is based on these strongly positively correlated targets.

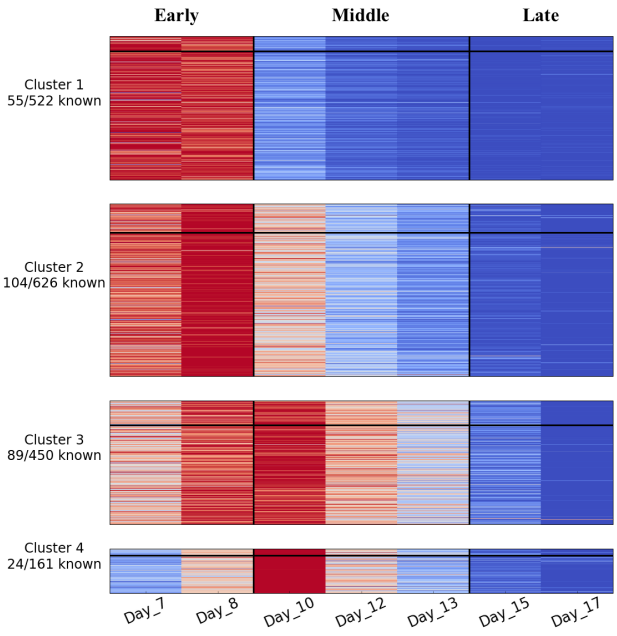
Table 7.3: A comparison of the total number of targets and the number of strongly positively correlated targets of each regulator. Less than half of the ABI3 and LEC1's targets are strongly positively correlated, while more FUS3 targets are strongly correlated.

Regulator	Total Number of Targets	Strongly Positively Correlated Targets
ABI3	1698	47
FUS3	3076	1759
LEC1	4010	1789

The time course gene expression data covers three major stages in seed development: early maturation (7 and 8 days after pollination (DAP)), middle maturation (10, 12 and 13 DAP), and late maturation/early desiccation (15 and 17 DAP). Clustering all the targets (including predicted and prior known targets) according to their expression profiles allows us to understand which of the targets are expressed at particular phases of seed development. Gene expression was normalized to the 0 to 1 range before clustering through $z_i = \frac{x_i - min(x)}{max(x) - min(x)}$, where $x = (x_1, \ldots, x_n)$ and x_i is the gene expression value at time point i. Three clusters are presented for ABI3 and LEC1, and four clusters are presented for FUS3 (Figure 7.4). ABI3 and FUS3 have targets whose maximum expression occurred at early and middle maturation stages but with different distributions. In contrast, in the case of LEC1, cluster 3 showed a high expression levels at the early and late maturation stages. Furthermore, known targets are distributed in each cluster, except for cluster 1 and cluster 3 of ABI3, where there is no known target.


Early
Middle
Late
Cluster 1
0/14 known
Cluster 2
2/32 known
Cluster 3
0/1 known
Day_7
Day_8
Day_10
Day_12
Day_13
Day_15
Day

A. ABI3


Early
Middle
Late
Cluster 1
55/522 known
Cluster 2
104/626 known
Cluster 3
89/450 known
Cluster 4
24/161 known
Day_7
Day_8
Day_10
Day_12
Day_13
Day_15
Day_17

B. FUS3

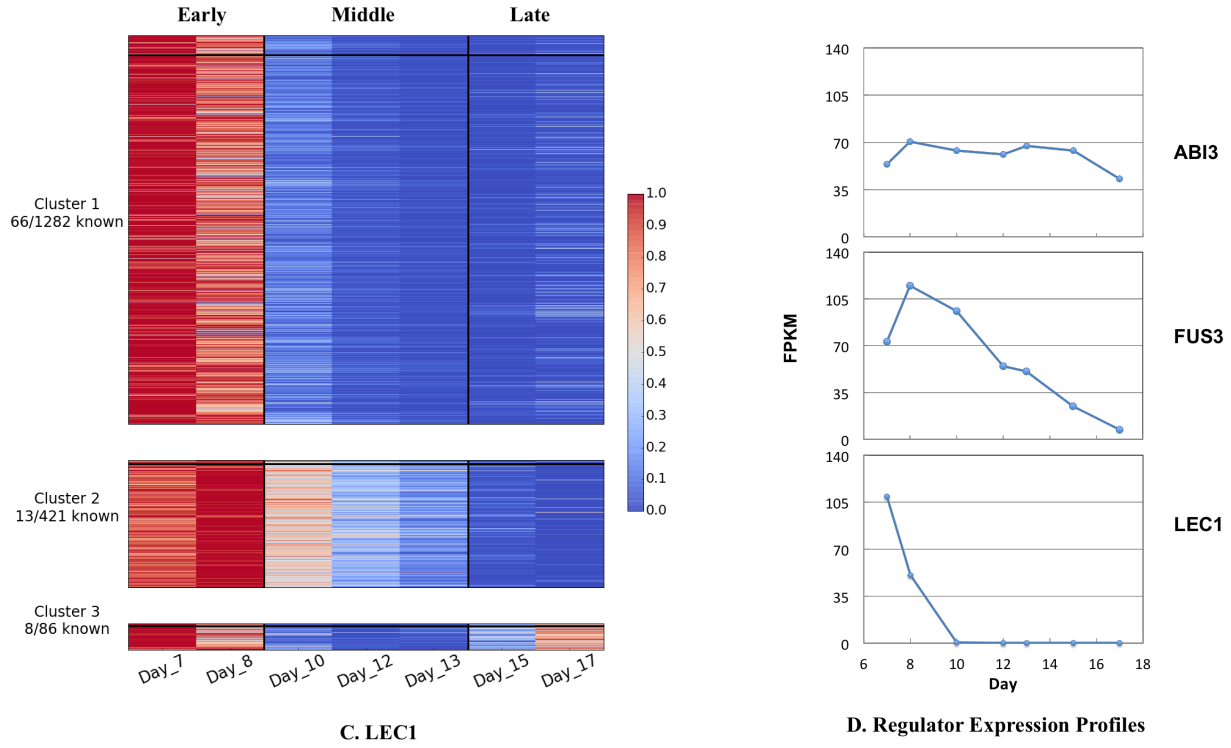


Figure 7.4: K-means clusters of (A) ABI3, (B) FUS3, and (C) LEC1 target genes, and the expression profiles for the three regulators (D). The results are organized by developmental stage. Three stages of seed development are involved in the gene expression: early (7 and 8 DAP), middle (10, 12 and 13 DAP), and late (15 and 17 DAP). The color scale indicates the gene expression level: red color represents high expression level, and blue color represents low expression level. A horizontal line is in each cluster, above which are the prior known targets and the remaining are predicted targets. The difference in expression profiles of the regulators may lead to different expression patterns of the target genes.

To further evaluate the prediction results, I ran the FIMO algorithm to classify all the targets into direct and indirect categories. The binding site study was limited to ABI3 and FUS3, because LEC1 is not in the CIS-BP (Catalog of Inferred Sequence Binding Preferences) database (Table 7.4) [84]. Secondary TFs were found among the direct targets in each

cluster, and their binding motifs were also searched against the CIS-BP database. The result is in cluster 3 of FUS3, secondary TF AT1G01260 has known binding motif, and 60 indirect targets contain the binding site of this TF in this cluster. According to our inference, gene AT1G01260 is only controlled by FUS3.

We then compare our results with the interactions GeneMANIA [83] (http://www.genemania.org) reports as another validation (Table 7.4). Only small portion of the interactions predicted by Beacon tool are in the GeneMANIA database. This is because the prior known interactions are from ChIP-Seq experiments, while GeneMania has curated several resources based on co-expression, physical interactions, and genetic interactions.

Table 7.4: The number of direct and indirect targets discovered by FIMO in each cluster. LEC1 does not have known binding site in the CIS-BP database, so only ABI3 and FUS3 binding sites could be identified. For each regulator, the table shows the number of targets that have the binding sites in known and predicted connections, respectively. The last row of each regulator shows the number of targets exist both in our prediction and in GeneMANIA (GM).

Regulator	Targets	Cluster 1	Cluster 2	Cluster 3	Cluster 4
ABI3	Direct in known	0	2	0	-
	Indirect in known	0	0	0	-
	Direct in predicted	2	2	18	-
	Indirect in predicted	12	28	1	-
	Overlap with GM	2	5	0	-
FUS3	Direct in known	9	16	15	4
	Indirect in known	46	88	74	20
	Direct in predicted	40	37	36	16
	Indirect in predicted	427	485	325	121
	Overlap with GM	3	7	8	4
LEC1	Overlap with GM	30	6	2	-

7.2 Discussion

In this analysis, I have developed a Beacon GRN inference tool, a supervised machine learning method based on SVM local approach, to decipher the complex GRNs that occur in Arabidopsis seed development from gene expression data and prior known regulatory relationships. The SVM local approach with RBF kernel was chosen according to the performance comparison with the SVM global approach and the unsupervised method CLR. CLR does not take into account any known interactions and performs worse than supervised methods. The SVM global approach makes the assumption that all the TFs regulate their downstream targets in the same way, and it performs worse than the SVM local models. A linear SVM kernel generates a linear hyperplane to separate positive and negative examples, which is less flexible than the non-linear kernel RBF. I concluded that the SVM local approach with RBF is the most suitable method to infer GRN in seed development (Figure 7.1 and Figure 7.2). This selected method decomposes the problem of inferring a network into three different subproblems, where the goal is to identify the targets of each of the three regulators.

As with many inference models, there is a limitation based on the initial data set used to make predictions. Our prediction accuracy can be definitely improved with experimentally validated non-regulations and the addition of more known TF-target pairs. Furthermore, the AUC was computed by assuming that the known interactions are accurate and do not include undiscovered relationships.

In Figure 7.3 and Table 7.2, I present the prediction results and show targets controlled by one or more regulators. There are 521 genes regulated by all three regulators, but more shared targets are found between any two of the regulators, which suggests that the regulation mechanism of these three regulators are different but still correlated. This result, in turn, shows that the SVM global model does not fit for our problem. It is also important to note that one TF controls so many targets, which greatly enlarges the regulatory network we

already know and points to the importance of studying these relationships. Though actual relationships remain to be verified by experiments, the TF-target predictions I generated can be preliminary knowledge for the researchers who are interested in gene regulations in Arabidopsis seed development. Moreover, our method of TF target prediction can be easily expanded to infer regulatory network of other plants in other biological processes by replacing the data source.

Chapter 8

CONCLUSIONS

GRNs are key to represent and understand biological activities and their elucidation is one of the main challenges of the researchers. Doing experiments to reveal the network is hard. However, gene expression data embeds much regulatory information. The work presented in this thesis is related to the problem of supervised inference of GRNs using gene expression levels. Many current studies focus on global GRNs, which may overlook the relationships in specific biological process. I addressed this issue by using the gene expression data in specific biological process, i.e., seed development in Arabidopsis. The method I developed in this thesis is the Beacon GRN inference tool. It is a supervised machine learning method based on SVM made use of the gene expression data and applied on a training sample of known interacting and non-interacting pairs, to predict GRN in this specific biological process. The main topic we addressed in this work is to predict the GRN in seed development in Arabidopsis, and interpret the predicted network.

An in-depth and comprehensive examination and analysis of the Beacon GRN inference tool is presented. All the evaluations of the supervised methods are based on 3-fold cross validation. The positive examples are known interactions described in Section 5.1, and

the negative examples were taken randomly from all the remaining pairs. The comparison between the Beacon inference tool with many other methods, including unsupervised method CLR, SVM with global model, SVM with linear kernel, indicates that the tool I developed is superior to all other models in this context. The Beacon inference tool, in detail, is an SVM local model that treats three major regulators, ABI3, FUS3, and LEC1, in seed development separately. Using their own known interactions, the Beacon tool evaluated their performance separately, and the ROC curves are given in Figure 7.2. The AUC of ABI3 is $\sim$0.92, FUS3 is $\sim$0.69, and LEC1 is $\sim$0.67.

In Section 7.1.2, the potential targets inferred from the Beacon inference tool among all the differentially expressed genes are presented. Among 7376 genes differentially expressed in seed development, 1064 of them are predicted as ABI3's targets, 2569 are FUS3's targets, and 3836 are LEC1's targets. There are also large overlaps between the predicted targets, which suggests that many genes are regulated by multiple regulators. In total, 275, 862, and 1732 targets are uniquely regulated by ABI3, FUS3, and LEC1, respectively.

Interpretation of the predicted GRN was done by doing clustering and binding site analysis. From Figure 7.4, I concluded that the targets that strongly positively correlated with their regulators are expressed early in seed development. Binding site analysis shows 24/47 and 173/1759 of the targets have the binding site of ABI3 and FUS3, respectively.

One of the limitations of the Beacon GRN inference tool is its inability to predict regulatory relationships with no prior known relations. The performance of the Beacon inference tool is dependent upon the list of the known target genes, and therefore an incomplete list will produce poor GRN prediction results. A possible future direction to address such challenge to employ semi-supervised models to deal with the unlabeled data.

In spite of the limitations, the predicted network can be preliminary knowledge of the GRN in seed development in Arabidopsis, and the Beacon GRN inference tool can be easily expanded

to infer regulatory networks existing in other plants or other biological processes by replacing the data source.

Bibliography

[1] K. Aoki, Y. Ogata, and D. Shibata, *Approaches for extracting practical information from gene co-expression networks in plant biology*, Plant and Cell Physiology, 48 (2007), pp. 381–390.

[2] J. Bang-Jensen and G. Z. Gutin, *Digraphs: Theory, Algorithms and Applications*, Springer Science & Business Media, 2008.

[3] G. W. Bassel, H. Lan, E. Glaab, D. J. Gibbs, T. Gerjets, N. Krasnogor, A. J. Bonner, M. J. Holdsworth, and N. J. Provart, *Genome-wide network model capturing seed germination reveals coordinated regulation of plant cellular phase transitions*, Proceedings of the National Academy of Sciences, 108 (2011), pp. 9709–9714.

[4] S. Baud, B. Dubreucq, M. Miquel, C. Rochat, and L. Lepiniec, *Storage reserve accumulation in Arabidopsis: Metabolic and developmental control of seed filling*, The Arabidopsis Book, 6 (2008), p. e0113.

[5] A. Ben-Hur and W. S. Noble, *Kernel methods for predicting protein–protein interactions*, Bioinformatics, 21 (2005), pp. i38–i46.

[6] A. Ben-Hur and J. Weston, *A user's guide to support vector machines*, Data Mining Techniques for the Life Sciences, (2010), pp. 223–239.

[7] J. M. Berg, J. L. Tymoczko, and L. Stryer, *Biochemistry*, WH Freeman, 2002.

[8] M. F. Berger and M. L. Bulyk, *Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors*, Nature Protocols, 4 (2009), pp. 393–411.

[9] S. A. Braybrook, S. L. Stone, S. Park, A. Q. Bui, B. H. Le, R. L. Fischer, R. B. Goldberg, and J. J. Harada, *Genes directly regulated by LEAFY COTYLEDON2 provide insight into the control of embryo maturation and somatic embryogenesis*, Proceedings of the National Academy of Sciences of the United States of America, 103 (2006), pp. 3468–3473.

[10] L. Breiman, *Random forests*, Machine Learning, 45 (2001), pp. 5–32.

[11] R. J. Brooker, *Genetics: Analysis and Principles*, Addison-Wesley, 1999.

[12] A. J. Butte and I. S. Kohane, *Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements*, in Pac Symp Biocomput, vol. 5, Citeseer, 2000, pp. 418–429.

[13] L. Cerulo, C. Elkan, and M. Ceccarelli, *Learning gene regulatory networks from positive and unlabeled data*, BMC Bioinformatics, 11 (2010), p. 16.

[14] X.-W. Chen and M. Liu, *Prediction of protein–protein interactions using random decision forest framework*, Bioinformatics, 21 (2005), pp. 4394–4400.

[15] F. Cheng, C. Liu, J. Jiang, W. Lu, W. Li, G. Liu, W. Zhou, J. Huang, and Y. Tang, *Prediction of drug-target interactions and drug repositioning via network-based inference*, PLoS Computational Biology, 8 (2012), p. e1002503.

[16] S. R. Cutler, P. L. Rodriguez, R. R. Finkelstein, and S. R. Abrams, *Abscisic acid: Emergence of a core signaling network*, Annual Reviews Plant Biology, 61 (2010), pp. 651–679.

[17] M. Deza and E. Deza, *Encyclopedia of Distances*, Springer, 2014.

[18] J. J. Faith, B. Hayete, J. T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J. J. Collins, and T. S. Gardner, *Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles*, PLoS Biology, 5 (2007), p. e8.

[19] T. Fawcett, *An introduction to ROC analysis*, Pattern Recognition Letters, 27 (2006), pp. 861–874.

[20] R. Finkelstein, *Abscisic acid synthesis and response*, The Arabidopsis Book, (2013), p. e0166.

[21] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, *Using Bayesian networks to analyze expression data*, Journal of Computational Biology, 7 (2000), pp. 601–620.

[22] P. Geurts, D. Ernst, and L. Wehenkel, *Extremely randomized trees*, Machine Learning, 63 (2006), pp. 3–42.

[23] Z. Gillani, M. S. Akash, M. M. Rahaman, and M. Chen, *CompareSVM: Supervised, Support Vector Machine (SVM) inference of gene regularity networks*, BMC Bioinformatics, 15 (2014), p. 395.

[24] C. E. Grant, T. L. Bailey, and W. S. Noble, *Fimo: Scanning for occurrences of a given motif*, Bioinformatics, 27 (2011), pp. 1017–1018.

[25] L. Gutierrez, O. Van Wuytswinkel, M. Castelain, and C. Bellini, *Combined networks regulating seed maturation*, Trends in Plant Science, 12 (2007), pp. 294–300.

[26] A.-C. Haury, F. Mordelet, P. Vera-Licona, and J.-P. Vert, *TIGRESS: Trustful inference of gene regulation using stability selection*, BMC Systems Biology, 6 (2012), p. 145.

[27] B. C. Haynes and M. R. Brent, *Benchmarking regulatory network reconstruction with GRENDEL*, Bioinformatics, 25 (2009), pp. 801–807.

[28] B. He and K. Tan, *Understanding transcriptional regulatory networks using computational models*, Current Opinion in Genetics & Development, 37 (2016), pp. 101–108.

[29] A. Irrthum, L. Wehenkel, P. Geurts, et al., *Inferring regulatory networks from expression data using tree-based methods*, PloS One, 5 (2010), p. e12776.

[30] V. R. Iyer, C. E. Horak, C. S. Scafe, D. Botstein, M. Snyder, and P. O. Brown, *Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF*, Nature, 409 (2001), pp. 533–538.

[31] H. Jia, D. R. McCarty, and M. Suzuki, *Distinct roles of LAFL network genes in promoting the embryonic seedling fate in the absence of VAL repression*, Plant Physiology, 163 (2013), pp. 1293–1305.

[32] H. Jia, M. Suzuki, and D. R. McCarty, *Regulation of the seed to seedling developmental phase transition by the LAFL and VAL transcription factor networks*, Wiley Interdisciplinary Reviews: Developmental Biology, 3 (2014), pp. 135–145.

[33] A. Junker, A. Hartmann, F. Schreiber, and H. Bäumlein, *An engineer's view on regulation of seed development*, Trends in Plant Science, 15 (2010), pp. 303–307.

[34] A. Junker, G. Mönke, T. Rutten, J. Keilwagen, M. Seifert, T. M. N. Thi, J.-P. Renou, S. Balzergue, P. Viehöver, U. Hähnel, et al., *Elongation-related functions of LEAFY COTYLEDON1 during the development of Arabidopsis thaliana*, The Plant Journal, 71 (2012), pp. 427–442.

[35] N. A. Kiani and L. Kaderali, *Dynamic probabilistic threshold networks to infer signaling pathways from time-course perturbation data*, BMC Bioinformatics, 15 (2014), p. 250.

[36] A. Kraskov, H. Stögbauer, and P. Grassberger, *Estimating mutual information*, Physical Review E, 69 (2004), p. 066138.

[37] P. Langfelder and S. Horvath, *WGCNA: An R package for weighted correlation network analysis*, BMC Bioinformatics, 9 (2008), p. 559.

[38] C. W. Law, Y. Chen, W. Shi, and G. K. Smyth, *VOOM: Precision weights unlock linear model analysis tools for RNA-seq read counts*, Genome Biology, 15 (2014), p. R29.

[39] N. Le Novere, M. Hucka, H. Mi, S. Moodie, F. Schreiber, A. Sorokin, E. Demir, K. Wegner, M. I. Aladjem, S. M. Wimalaratne, et al., *The systems biology graphical notation*, Nature Biotechnology, 27 (2009), pp. 735–741.

[40] D. J. Lockhart, H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, et al., *Expression monitoring by hybridization to high-density oligonucleotide arrays*, Nature Biotechnology, 14 (1996), pp. 1675–1680.

[41] T. Lotan, M.-a. Ohto, K. M. Yee, M. A. West, R. Lo, R. W. Kwong, K. Yamagishi, R. L. Fischer, R. B. Goldberg, and J. J. Harada, *Arabidopsis LEAFY COTYLEDON1 is sufficient to induce embryo development in vegetative cells*, Cell, 93 (1998), pp. 1195–1205.

[42] J. MacQueen et al., *Some methods for classification and analysis of multivariate observations*, in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, Oakland, CA, USA., 1967, pp. 281–297.

[43] P. B. Madhamshettiwar, S. R. Maetschke, M. J. Davis, A. Reverter, and M. A. Ragan, *Gene regulatory network inference: Evaluation and application to ovarian cancer allows the prioritization of drug targets*, Genome Medicine, 4 (2012), pp. 1–16.

[44] S. R. Maetschke, P. B. Madhamshettiwar, M. J. Davis, and M. A. Ragan, *Supervised, semi-supervised and unsupervised inference of gene regulatory networks*, Briefings in Bioinformatics, (2014), p. bbt034.

[45] D. Marbach, S. Roy, F. Ay, P. E. Meyer, R. Candeias, T. Kahveci, C. A. Bristow, and M. Kellis, *Predictive regulatory models in Drosophila melanogaster by integrative inference of transcriptional networks*, Genome Research, 22 (2012), pp. 1334–1349.

[46] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. D. Favera, and A. Califano, *ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context*, BMC Bioinformatics, 7 (2006), p. S7.

[47] D. W. Meinke, *Molecular genetics of plant embryogenesis*, Annual Review of Plant Biology, 46 (1995), pp. 369–394.

[48] N. Meinshausen and P. Bühlmann, *Stability selection*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 72 (2010), pp. 417–473.

[49] A. Mendes, A. A. Kelly, H. van Erp, E. Shaw, S. J. Powers, S. Kurup, and P. J. Eastmond, *bZIP67 regulates the omega-3 fatty acid content of Arabidopsis seed oil by activating fatty acid desaturase3*, The Plant Cell, 25 (2013), pp. 3104–3116.

[50] P. Meyer, D. Marbach, S. Roy, and M. Kellis, *Information-theoretic inference of gene networks using backward elimination.*, in Conference on Bioinformatics & Computational Biology (BIOCOMP'10), 2010, pp. 700–705.

[51] P. E. Meyer, K. Kontos, F. Lafitte, and G. Bontempi, *Information-theoretic inference of large transcriptional regulatory networks*, EURASIP Journal on Bioinformatics and Systems Biology, 2007 (2007), pp. 1–9.

[52] G. Mönke, M. Seifert, J. Keilwagen, M. Mohr, I. Grosse, U. Hähnel, A. Junker, B. Weisshaar, U. Conrad, H. Bäumlein, et al., *Toward the identification and regulation of the Arabidopsis thaliana ABI3 regulon*, Nucleic Acids Research, 40 (2012), pp. 8240–8254.

[53] F. Mordelet and J.-P. Vert, *SIRENE: Supervised inference of regulatory networks*, Bioinformatics, 24 (2008), pp. i76–i82.

[54] A. Niculescu-Mizil and R. Caruana, *Predicting good probabilities with supervised learning*, in Proceedings of the 22nd International Conference on Machine Learning, ACM, 2005, pp. 625–632.

[55] N. Omranian, J. M. Eloundou-Mbebi, B. Mueller-Roeber, and Z. Nikoloski, *Gene regulatory network inference using fused LASSO on multiple data sets*, Scientific Reports, 6 (2016).

[56] B. Palsson, *Systems Biology*, Cambridge University Press, 2015.

[57] P. J. Park, *ChIP–seq: Advantages and challenges of a maturing technology*, Nature Reviews Genetics, 10 (2009), pp. 669–680.

[58] N. Patel and J. T. Wang, *Semi-supervised prediction of gene regulatory networks using machine learning algorithms*, Journal of Biosciences, 40 (2015), pp. 731–740.

[59] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., *Scikit-learn: Machine learning in Python*, The Journal of Machine Learning Research, 12 (2011), pp. 2825–2830.

[60] C. A. Penfold and D. L. Wild, *How to infer gene networks from expression profiles, revisited*, Interface Focus, 1 (2011), pp. 857–870.

[61] F. Petralia, P. Wang, J. Yang, and Z. Tu, *Integrative random forest for gene regulatory network inference*, Bioinformatics, 31 (2015), pp. i197–i205.

[62] J. Platt et al., *Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods*, Advances in Large Margin Classifiers, 10 (1999), pp. 61–74.

[63] B. Ren, F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, et al., *Genome-wide location and function of DNA binding proteins*, Science, 290 (2000), pp. 2306–2309.

[64] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth, *Limma powers differential expression analyses for RNA-sequencing and microarray studies*, Nucleic Acids Research, (2015), p. gkv007.

[65] G. Sales and C. Romualdi, *parmigenea parallel R package for mutual information estimation and gene network reconstruction*, Bioinformatics, 27 (2011), pp. 1876–1877.

[66] M. Santos-Mendoza, B. Dubreucq, S. Baud, F. Parcy, M. Caboche, and L. Lepiniec, *Deciphering gene regulatory networks that control seed development and maturation in Arabidopsis*, The Plant Journal, 54 (2008), pp. 608–620.

[67] A. Schneider, D. Aghamirzaie, H. Elmarakeby, A. Poudel, A. Koo, L. Heath, R. Grene, and E. Collakova, *Potential targets of VIVIPAROUS1/ABI3-LIKE1 (VAL1) repression in developing Arabidopsis thaliana embryos*, The Plant Journal, 85 (2016), pp. 305–319.

[68] M. Schrynemackers, R. Küffner, and P. Geurts, *On protocols and measures for the validation of supervised methods for the inference of biological networks*, Frontiers in Genetics, 4 (2014).

[69] M. Schrynemackers, L. Wehenkel, M. M. Babu, and P. Geurts, *Classifying pairs with trees for supervised biological network inference*, Molecular BioSystems, 11 (2015), pp. 2116–2125.

[70] E. A. Serin, H. Nijveen, H. W. Hilhorst, and W. Ligterink, *Learning from co-expression networks: Possibilities and challenges*, Frontiers in Plant Science, 7 (2016), p. 444.

[71] K. Shinozaki and E. S. Dennis, *Cell signalling and gene regulation: Global analyses of signal transduction and gene expression profiles*, Current Opinion in Plant Biology, 6 (2003), pp. 405–409.

[72] R. R. Sinden, *DNA Structure and Function*, Elsevier, 2012.

[73] S. L. Stone, L. W. Kwong, K. M. Yee, J. Pelletier, L. Lepiniec, R. L. Fischer, R. B. Goldberg, and J. J. Harada, *LEAFY COTYLEDON2 encodes a B3 domain transcription factor that induces embryo development*, Proceedings of the National Academy of Sciences, 98 (2001), pp. 11806–11811.

[74] M. Suzuki and D. R. McCarty, *Functional symmetry of the B3 network controlling seed development*, Current Opinion in Plant Biology, 11 (2008), pp. 548–553.

[75] M. Takarabe, M. Kotera, Y. Nishimura, S. Goto, and Y. Yamanishi, *Drug target prediction using adverse event report systems: A pharmacogenomic approach*, Bioinformatics, 28 (2012), pp. i611–i618.

[76] R. Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society. Series B (Methodological), (1996), pp. 267–288.

[77] C. Trapnell, A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn, and L. Pachter, *Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks*, Nature Protocols, 7 (2012), pp. 562–578.

[78] V. Trevino, F. Falciani, and H. A. Barrera-Saldaña, *DNA microarrays: A powerful genomic tool for biomedical and clinical research*, Molecular Medicine, 13 (2007), pp. 527–541.

[79] T. Turki and J. T. Wang, *A new approach to link prediction in gene regulatory networks*, in Intelligent Data Engineering and Automated Learning–IDEAL 2015, Springer, 2015, pp. 404–415.

[80] J.-P. Vert, *Reconstruction of biological networks by supervised machine learning approaches*, Elements of Computational Systems Biology, (2010), pp. 165–188.

[81] J. Vicente-Carbajosa and P. Carbonero, *Seed maturation: Developing an intrusive phase to accomplish a quiescent state*, International Journal of Developmental Biology, 49 (2005), pp. 645–651.

[82] F. Wang and S. E. Perry, *Identification of direct targets of FUSCA3, a key regulator of Arabidopsis seed development*, Plant Physiology, 161 (2013), pp. 1251–1264.

[83] D. Warde-Farley, S. L. Donaldson, O. Comes, K. Zuberi, R. Badrawi, P. Chao, M. Franz, C. Grouios, F. Kazi, C. T. Lopes, et al., *The genemania prediction server: Biological network integration for gene prioritization and predicting gene function*, Nucleic Acids Research, 38 (2010), pp. W214–W220.

[84] M. T. Weirauch, A. Yang, M. Albu, A. G. Cote, A. Montenegro-Montero, P. Drewe, H. S. Najafabadi, S. A. Lambert, I. Mann, K. Cook, et al., *Determination and inference of eukaryotic transcription factor sequence specificity*, Cell, 158 (2014), pp. 1431–1443.