A combinatorial approach to determine the context-dependent role in transcriptional and posttranscriptional regulation in Arabidopsis Thaliana

Lu, Le

2008

https://hdl.handle.net/10356/6569

https://doi.org/10.32657/10356/6569

Nanyang Technological University

A Combinatorial Approach to Determine the Context-dependent Role in Transcriptional and Posttranscriptional Regulation in *Arabidopsis Thaliana*

Lu Le

School of Biological Sciences

A thesis submitted to the Nanyang Technological University
in fulfillment of the requirements for the degree of
Doctor of Philosophy

2008

ACKNOWLEDGEMENTS

First of all, I would like to express my sincere gratitude to my supervisor, Dr. Li Jinming, to whom I really owe a debt, for his care and encouragement, his patience and generosity, his creative ideas and invaluable guidance, and so on and so forth. It is his support and help that made this thesis possible.

I would like to thank my co-supervisor Dr. Peter Dröge for his great help and valuable suggestions during my four-year's postgraduate study.

A special thank to my labmate Jia Hui and Xiao Jie, for their friendship, good spirits and various helps. Thanks to my dear friend Qu Yi and Zhang Xiaohong for everything they did for me in the past years.

I acknowledge the following people for the permission to reproduce their figures in this thesis: Dr. Phillip D. Zamore at University of Massachusetts Medical School (Figure 1.1.1), Dr. Alexander Schliep at Max Planck Institute for Molecular Genetics in Germany (Figure 1.2.1), Dr. Timothy Gardner at Boston University (Figure 1.4.1 and 1.4.2), and Dr. David Bartel at Whitehead Institute for Biomedical Research in USA (Figure 3.1.1).

Last, but not least, I thank my parents and grandparents for their love, which will never end.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

SUMMARY

We conducted a systematic study on the transcriptional and posttranscriptional regulatory roles at sequence level in *Arabidopsis*. Both miRNA target motifs (miRNA-mediated posttranscriptional regulatory sites) and TFBSs (transcription factor binding motifs) were incorporated with microarray time course gene expression profiles to determine their probabilistic dependences. A novel method based on an inhomogeneous hidden Markov model (HMM) was developed to predict plant miRNA targets without additional conservation constraint. The model was trained using the target information of about one third of the confirmed miRNAs, whereas it was capable of finding all the experimentally validated targets for all the known miRNAs. Bayesian network model was introduced to deduce the conditional dependences between expression profiles and the combinations of two types of sequence motifs. Based solely on the sequence motifs adopted in the network models, we could correctly predict expression patterns for more than 50% of 1,132 genes, which was statistically significant. Furthermore, 20 genes' expression patterns could only be correctly predicted with the involvement of miRNA target motifs. Among the 20 genes, one was experimentally validated as miRNA target and the other two had GO annotation term as RNA binding. The result suggested that microarray time course dataset could be used to detect the change of mRNA steady-state level which might be

affected by miRNA regulation, and the combinatorial approach was efficient to determine the underlying context-dependent roles in transcriptional and posttranscriptional regulation.

The biological background, concepts and strategies for the construction of gene regulatory networks are reviewed in Chapter 1, and the two statistical methods used in this study, namely hidden Markov model (HMM) and Bayesian network model, are introduced in Chapter 2. In Chapter 3, a novel inhomogeneous HMM based approach for plant miRNA target prediction is described. Chapter 4 is a presentation on our systematic study on the transcriptional and posttranscriptional roles at sequence level in *Arabidopsis*. Chapter 5 gives the conclusion of this study.

Chapter 1:Introduction

The completion of various genome sequencing projects and large-scale genomic studies has led to a wealth of available biological data. In the era of post-genomic biology, a key aim is to systematically catalogue all genes and their interactions within a living cell in order to understand the cell's functional organization. Current molecular biological, analytical and computational technologies enable us to systematically investigate the complex molecular processes underlying biological systems [1].

Biological networks represent multiple interactions within a cell, and provide a global view to understand the relationships between molecules and the dictated cellular behavior. Rapid advances in network biology indicate that gene regulatory networks are governed by universal laws and offer a new conceptual framework that can potentially revolutionize our view of biology and disease pathologies in the twenty-first century [2]. Recent progresses in molecular and computational biology have made the study of intricate transcriptional regulatory network possible through describing gene expression as a function of regulatory inputs specified by interactions between protein and DNA [3]. To define the network structure of a cell, network biology researchers use data generated by experimental methods, including high-throughput genomic, proteomic, and metabolic data, as well as computational approaches to identify and map cell signaling networks. High throughput genomic work, such as microarray technology, chromatin immunoprecipitation (CHIP), has now yielded relatively unbiased genome-wide data sets that comprise known metabolic, regulatory, functional, and physical interactions. Application of these technologies will shift the emphasis in biological research from primary data generation to complex quantitative data analysis [4]. It has become obvious that the rate-limiting step in the studies of functional genomics is not the handling of the biological samples, but the post-analytical work in determining what the results actually mean instead [5].

Viewing genes in terms of their underlying network structure is a powerful concept. All networks share common characteristics, and mathematical treatments have been developed to understand their structure and how they can be regulated. Thus, organizing biological information in the context of networks is fundamental to understanding biological function in system level.

Many studies have been done to relate the gene expression data to the combination of *cis*-regulatory elements [6,7]. However, these approaches have not been broadly applied in multi-cellular organisms in spite of the reported success in model organisms. A key limitation of such approaches is that many regulators are regulated posttranscriptionally [8]. While progress is made in mapping transcriptional regulatory networks, posttranscriptional regulatory networks just begin to be uncovered. In this work, we conducted a systematic study on the transcriptional and posttranscriptional regulatory roles at sequence level in *Arabidopsis*. Both miRNA target motifs (miRNA-mediated posttranscriptional regulatory sites) and TFBSs (transcription factor binding motifs) were incorporated with microarray time course gene expression profiles to determine their probabilistic dependences. To facilitate our transcription regulatory network construction, a novel method based on an inhomogeneous Hidden Markov Model (HMM) was developed to predict plant miRNA targets without additional conservation constraint. Bayesian network model was introduced to deduce the conditional dependences between expression profiles and the combinations of two types of sequence motifs.

Before describing the methods proposed by us for transcriptional regulatory network construction, I would like to briefly review the general biological background and the common strategies of building transcriptional and posttranscriptional regulatory networks. Basically, there are three steps, namely classifying co-regulated genes into clusters, detecting transcriptional and posttranscriptional regulatory sequence motifs and building regulatory networks which integrate both gene expression data and sequence motifs.

1.1 Biological background of the transcriptional and posttranscriptional regulation

A large number of experimental and computational studies have been done on locating transcriptional regulator binding DNA sequences and understanding their functions [9-11]. TFs regulate gene expression by binding selectively to sequence sites in promoters of genes, and genes regulated by the same TFs have been assumed to share the common binding sites in their promoter regions and exhibit similar expression profiles [12].

1.1.1 Transcriptional regulation

Transcription in molecular biology is the copying from a DNA pattern to create a RNA molecule. By interacting with RNA polymerase or recruiting chromatin-modifying machinery, transcriptional regulators increase or decrease the transcriptional rate of genes, through transcriptional regulator binding DNA in the neighborhood of protein-coding and RNA genes.

The transcription of genes is regulated by transcription factors (TFs), which bind to DNA regulatory elements near the coding sequences. The "promoter" of a gene includes the DNA sequences which regulate basal levels of gene transcription, and control the exact position where transcription begins. These sequences are generally very near the transcriptional start site because of the functional need of Pol II enzyme. There are proteins involved in this process, such as the "basal transcription factors", which assemble on the promoter in a defined order (Figure 1.1.1). This process of initiation complex formation is followed by other protein-protein interactions that modulate the level of transcription, the cell-specificity, and the timing of gene expression.

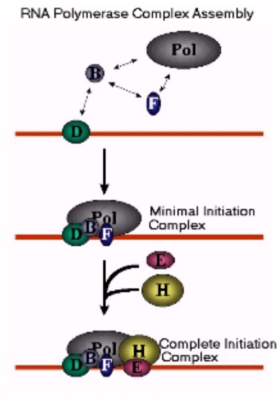

RNA Polymerase Complex Assembly
Pol
B
F
D
Minimal Initiation Complex
E
H
Complete Initiation Complex

Figure 1.1.1 Transcription initiation.

Pol: RNA polymerase II complex. B-H: transcription factors [1]

TFs are usually defined as proteins which show sequence-specific DNA binding and are capable of activating and/or repressing transcription of their target genes [13]. These transcription factor binding sites (TFBSs) contain 6~25 base pairs and are usually located in the upstream regions of the genes being regulated. They are important in facilitating the binding of TFs that controls the transcription of the genes.

1.1.2 Posttranscriptional regulation

microRNAs (miRNAs) was first discovered in nematodes in 1993. They are non-coding RNAs and sculpt gene expression profiles during plant and animal development. In fact, miRNAs may regulate as many as one-third of human genes [14].

miRNAs are transcribed by RNA polymerase II as primary miRNAs

[1] Reproduce from http://homer.ornl.gov/cbps/tfpage.htm

(pri-miRNAs), the size of which ranges from hundreds to thousands of nucleotides in length [15-18]. Most miRNAs are transcribed from the genome regions distinct from previously annotated protein-coding sequences. About half of the known mammalian miRNAs are within the introns of protein-coding genes, or within either the introns or exons of non-coding RNAs, rather than in unique transcription units [19-23].

1.1.2.1 The biogenesis and function of miRNAs in animals

In animals, two processing steps are needed to yield mature miRNAs (Figure1.1.2), each of which is catalyzed by a ribonuclease III (RNase III) endonuclease together with a double-stranded RNA-binding domain (dsRBD) protein partner. First, Drosha, a nuclear RNase III, cleaves the flanks of pri-miRNA to liberate an ~70 nucleotide stem loop, the pre-miRNA [10,19-22,24-28]. The resulting pre-miRNAs have 5’ phosphate and 3’ hydroxyl termini, and two- or three- nucleotide 3’ overhangs, due to the characteristics of RNase III cleavage of dsRNA. Then pre-miRNA is exported from nucleus to cytoplasm by Exportin 5 [14,29-32]. Second, in the cytoplasm, another RNase III, Dicer, together with its dsRBD protein partner makes a pair cut that defines the other end of the mature miRNA, liberating an ~21 nucleotides RNA duplex [14,33-40]. The mature miRNA enters the RNA-induced silencing complex (RISC), the protein complex that represses target gene expression [41,42].

The RISC carries out small RNA-directed gene silencing in the miRNA pathways in both plants and animals [43-47]. When miRNA guides in the RISC pairs extensively to a target mRNA, the RISC functions as an endonuclease, cleaving the mRNA between the target nucleotides paried to bases 10 and 11 of the miRNA. The core component of the RISC is a member of the Argonaute (Ago) protein family, whose members all contain a central PAZ domain and a carboxy terminal PIWI domain. The PIWI and PAZ domains bind to the 5' and 3' ends of the miRNA, respectively [48-53]. And Argonaute is the target-cleaving endonuclease of the RISC [54-59].

In animals, the complementarity between animal miRNAs and their targets is usually restricted to the 5' region (nucleotides 2-8 or 2-7) of the miRNA to the 3' region of the target site [60-63]. In the absence of extensive complementarity between the miRNAs and the targets, binding of the RISC blocks translation of the target mRNA [64].

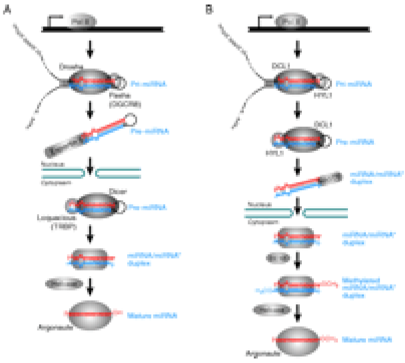


Figure 1.1.2 The miRNA biogenesis pathway. (A) Animal and (B) plant miRNA biogenesis[2].

[2] Reproduced from http://dev.biologists.org/cgi/content/full/132/21/4645 with permission.

1.1.2.2 The biogenesis and function of miRNAs in plants

miRNA maturation in plants differs from the pathway in animals because plants lack a Drosha homolog (Figure1.1.2). The RNase III enzyme DICER-LIKE 1 (DCL1), which is homologous to animal Dicer, is required for miRNA maturation [65-68]. In plants, DCL1 is localized in the nucleus and can make both the first pair cuts made by Drosha as well as the second pair of cuts made by animal Dicer. As for animal Dicer, a dsRNA-binding domain protein partner, HYL1, has been implicated in DCL1 function in plant miRNA maturation [66,69]. HASTY (HST), the plant ortholog of Exportin 5, exports the miRNA/mRNA* duplex and completes its assembly into the RISC in the cytoplasm [70,71]. Plant miRNAs have a methyl group on the ribose of the last nucleotide. The terminal methyl group is added by the methyltransferase HEN1, and the modification of the miRNA by HEN1 either protects the miRNA from further modification or degradation, or may facilitate its assembly into the RISC [14,72,73].

In plants, most miRNAs have perfect or near perfect complementarity to their targets [68]. Upon binding to their targets, the miRNA-containing RISCs function as endonucleases, cleaving the mRNA [74,75]. miRNA-binding motifs are found both in the coding regions and in the untranslated regions of miRNA-regulated plant mRNAs.

miRNAs function in a wide rage of biological processes in plants and animals [76,77]. The first insight into their function came from phenotypic studies of mutations that disrupt core components of the miRNA pathway. *dicer* mutants show diverse developmental defects, such as abnormal embryogenesis in *Arabidopsis*. Similarly, the disruption of Argonaute function causes widespread maintenance and failure to form axillary meristem in an *Arabidopsis* mutant for *PINHEAD/ZWILLE (PNH/ZLL)* or *ARGONAUTE 1* (*AGO1*). *Arabidopsis* mutant for *ZIPPY* (*ZIP*), an Argonaute gene, and *HASTY* (*HST*), which encodes the miRNA export receptor, exhibit a precocious vegetative phenotype and produce abnormal flowers [70]. Overall, these phenotypes suggest that certain miRNAs play important roles in early development [14].

1.2 Gene regulation analysis through clustering of gene expression data

DNA microarrays provide rapid and parallel surveys of gene-expression levels for hundreds or thousands of genes in a single assay. Based on our understanding of cellular processes, genes contained in a particular pathway, or respond to a common environmental challenge, should be co-regulated and consequently show similar patterns of expression. The hypothesis behind using clustering techniques is that genes in a cluster share some common function or regulatory elements. In other words, genes in the same cluster are more likely to have a known interaction or a similar cellular role [78]. However, there are quite a few clustering methods available for grouping genes in terms of their expression levels, differing in how data is normalized within and across experiments and how the similarity is measured and so on and so forth. All of these differences may have substantial effect on the outcome of clustering analysis [79].

1.2.1 Clustering analysis of microarray data: heuristic clustering methods

We can cluster genes into different groups based on the similarity of gene expression profiles. Genes with similar expression patterns across multiple conditions may share the same biological pathway or work coordinately, such that gene clustering can be used to predict the function of unknown genes [80,81]. Basically, there are two kinds of clustering methods: (a) the heuristic clustering methods; and (b) the model-based clustering methods. The former clusters genes based on the expression data only, whereas the latter makes use of prior biological knowledge.

Many clustering methods have been applied to microarray data analysis [82]. Three commonly used methods are hierarchical clustering, *K*-means clustering and self-organizing map. They are all based on the measures of the distance or similarity between the objects of interest (gene expression vectors). The commonly used distance measures include Euclidean distance and the complementary Pearson correlation. The expression vectors of two genes g_1 and

g_2 are $v_{g_1} = (v_{g_{11}}, ..., v_{g_{1N}})$ and $v_{g_2} = (v_{g_{21}}, ..., v_{g_{2N}})$, respectively, where N is the dimension of the vectors. The Euclidean distance can be calculated as:

$$d(v_{g_1}, v_{g_2}) = \sqrt{\sum_{n=1}^{N} (v_{g_{1n}} - v_{g_{2n}})^2} .$$

Another popular distance measure is defined in terms of the complementary Pearson correlation coefficient between two objects:

$$d(v_{g_1}, v_{g_2}) = 1 - r(v_{g_1}, v_{g_2}) ,$$

where $r(v_{g_1}, v_{g_2})$ is the Pearson correlation coefficient:

$$r(v_{g_1}, v_{g_2}) = \frac{\sum_{n=1}^{N} (v_{g_{1n}} - \overline{v}_{g_{1.}})(v_{g_{2n}} - \overline{v}_{g_{2.}})}{\sqrt{\sum_{n=1}^{N} (v_{g_{1n}} - \overline{v}_{g_{1.}})^2 \sum_{n=1}^{N} (v_{g_{2n}} - \overline{v}_{g_{2.}})^2}} .$$

Two advantages of this complementary Pearson correlation coefficient as the distance measure over Euclidean distance are: (a) it is not scale dependent; and (b) it can reflect negative association.

After a distance measure is specified, one can use the following algorithms to cluster genes.

(1) Hierarchical clustering

Agglomerative hierarchical clustering is a commonly used method in clustering genes in terms of their expression profiles. Given a set of N genes to be clustered, the hierarchical clustering starts by assigning each gene to a cluster; and the two "closest" clusters are merged together in each step. This process continues until all the genes are merged into one single cluster of size N . Besides the distance measure, "linkage method", which defines the distance between the two clusters to be merged, needs to be specified. Linkage methods include: (a) average linkage, which uses the average distance between all pairs of genes (one in each cluster) as the distance between clusters; (b) complete linkage (also called the diameter or maximum method), which uses the greatest distance between pairs of genes; and (c) single linkage (also called the connectedness or minimum method), which uses the shortest distance between pairs of genes. In general, complete linkage and average linkage are preferred over single linkage because they tend to produce relatively compact clusters. Hierarchical clustering has been widely applied to microarray data analysis since Eisen et al. and Bittner et al. [80,83].

(2) K-means clustering

The objective of K-means clustering is to segregate the genes into *k* clusters. The number of clusters *k* is predetermined and the main procedure is to define *k* centroids, one for each cluster. Usually the algorithm begins with a random initialization of the centroid of each cluster. Because different initial centroids will bring different result, the centroids should be placed far away as much as possible from each other. Each gene is assigned to the cluster whose centroid is nearest, and the centroid of that cluster is updated accordingly. After all the genes are assigned, the process starts over to reallocate the genes among clusters and update the centroids of both the donor cluster and the recipient cluster. The iteration continues until no more allocations take place.

The major advantages of the K-means method over hierarchical clustering are (a) the “clusters” are more clearly defined; and (b) it does not request for pair-wise distance between genes. This feature substantially improves the ability of the algorithm in dealing with large datasets. A disadvantage of this method is that the number of clusters *k* is to be chosen arbitrarily, and the outcomes could be very sensitive to the choice of *k* and the initial cluster centroids. The K-means clustering was first proposed by MacQueen and it is widely used in gene expression profile analysis [84-86]. Because of the instability of the outcomes, Rahnenfuehrer suggested clustering the data for a number of times with different random initial centroids and then choosing the best classification, which minimized the within-cluster sum of squares. This method performed very well on his test datasets [86].

(3) Self-organizing maps (SOM)

SOM is a data visualization technique which reduces the dimension of data using self-organizing neural networks. It produces a map of 1 or 2 dimensions which plots the similarities of the data.

SOM is similar to the K-means method in that it also attempts to classify genes into a predefined fixed number of clusters, and it does so by comparing the distance between the gene and the centroid of each cluster. The difference between the two is as follows. For SOM, when a gene is classified, the centroid of the recipient cluster and the centroids of some defined “neighborhood” clusters are updated simultaneously. This algorithm builds up a “map” for the clusters in a lower-dimensional space (for example, a two-dimensional plane). Each cluster is represented by a node in the map. The configuration of the map is arbitrarily determined and may be irrelevant of the real data structure. When using a two-dimensional plane, people usually arrange the nodes in a rectangle grid. Each node is associated with a “codebook” vector, which is of the same dimension as the data vectors. At the beginning of the algorithm, the codebook vector for each cluster is randomly chosen. When a gene is presented to the algorithm, the distance between this gene and each of the codebook vectors are computed to determine which codebook vector is the closest one. Then both the closest codebook vector and other codebook vectors whose corresponding nodes are within some neighborhood of the closest node are updated (moved toward to the gene). At the end of the algorithm, each gene is assigned to the cluster whose codebook vector is closest to it. However, the outcome is sensitive to the number of clusters, the configuration of the map, the choice of the learning rate and the neighborhood function. Tamayo et al. used this algorithm to classify 585 genes in a time-course experiment into 12 clusters arranged in a 4×3 rectangle grid [87].

1.2.2 Clustering analysis of microarray data: model-based clustering methods

There is another way to deal with clustering problems: model-based methods, which use certain models for clusters and attempt to optimize the fit between the data and the model.

An important question in microarray data analysis is whether to use a time-series (dynamic) design or a static design. The steady-state design may miss dynamic events that are critical for correctly inferring the structure of a gene network, but it enables one to observe gene expression under more diverse experimental conditions. On the other hand, time-series experiments can capture dynamics, but many of the data points may contain redundant information leading to inefficient use of experimental resources [88,89]. There are many approaches to analyze time course data, including the three aforementioned "heuristic" methods. These methods assume the different experiments to be independent and do not consider any dependencies between profiles belonging to subsequent time-points, so that permuting time points arbitrarily does not change the result of the clustering [90]. As an alternative to these methods, model-based clustering methods have also been proposed and applied to microarray data, which are especially suitable for time-series expression data [91-96]. The advantage of model-based clustering is that, usually the assumed underlying distributions are well defined and extensively studied in statistics.

Two of the recently proposed model-based clustering methods, i.e. cubic spline clustering and hidden Markov model (HMM), are briefly reviewed in the following.

(1) Cubic spline clustering model

Cubic splines can be used to represent gene expression curves. They are a set of piecewise cubic polynomials and are frequently used for fitting time series and other noisy data [97]. Ma et al. modeled the “mean curve” for each cluster of genes by smoothing spline [91]. The gene expression level in a time series experiment in a given cluster is assumed to follow the shape of the mean curve, with an additional gene-specific “random effect”. Given the gene i in cluster k, the expression at time t_{ij} (j is the time point) can be written as:

$$y_{ij} = \mu_k(t_{ij}) + b_i + \varepsilon_{ij},$$

where μ_k is the mean curve, $b_i \sim N(0,\sigma_{bk}^2)$ explains the gene specific deviation from μ_k and $\varepsilon_{ij} \sim N(0,\sigma^2)$ is the Gaussian measurement error.

$y_k \sim N(\mu_k, \sum_k)$, where y_k and μ_k are the vector representations of the expression and the mean curve, respectively. And the time series expression vector of y_i can be modeled by a mixture Gaussian distribution:

$$y_i \sim p_1 N(\mu_1, \sum_1) + p_2 N(\mu_2, \sum_2) + \cdots + p_K N(\mu_K, \sum_K),$$

where K is the total number of clusters and p_k (a prior) is the relative size (proportion) of cluster k. The maximum likelihood of the entire mixed-effect model can be calculated by EM algorithm and Bayesian Information Criterion (BIC) can be used to estimate the number of clusters.

Luan et al. used a mixed-effects model based on B-spline to account for time dependency of the gene expression measurements over time and the noisy nature of the microarray data [96]. For gene expression, a B-spline can be described as a polynomial spline and represents each point as a linear combination of a set of basis polynomials; therefore it fits different curves in different regions instead of fitting one polynomial curve across the data points [95]. B-spline requires specification of series of internal knots (break points). Once the value of these splines at a set of control points is known, the entire set of polynomials can be generated using these basis functions. The parameters of these models are usually estimated by an EM algorithm [95].

The cubic splines model provides a superior fit for time series expression data. However, it is only appropriate for relatively long time-series experiments (> 10 time points), and for practical application the number and location of the knots for the splines corresponding to the mean function and the random effects have to be specified.

(2) Hidden Markov models (HMM) for clustering time course microarray data

HMM takes into account the temporal nature between the time points and the duration each time point represents. HMM can be defined by the following parameters: the hidden states S_i, the initial probability π_i at a given state, the transition probability a_{ij} from state i to state j, and the emission probability $b_i(w)$ of symbol w at state S_i.

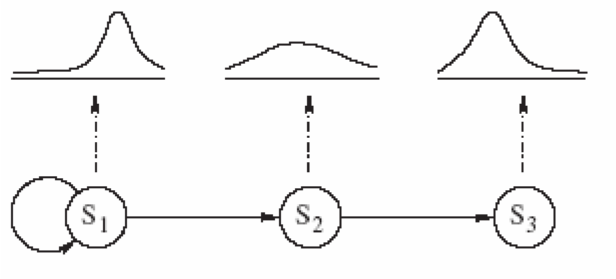


Figure 1.2.1 A hidden Markov model visualized as a directed graph.

The emission probability density functions are attached to the nodes. The model depicted is a prototype for down-regulation[3].

Given gene expression data for n genes, the objective is to classify the genes into K clusters (K HMMs $\Lambda_1, \Lambda_2, ..., \Lambda_K$), which gives the maximum likelihood of the observed gene expression profiles (illustrated in Figure 1.2.1). Schliep et al. presented a HMM based clustering algorithm for time series expression data [90,97]. The emission probabilities were assumed to follow Gaussian distribution with fixed variance. An EM algorithm was used to estimate the parameters: each gene is assigned to the most likely HMM (E-step) and the parameters of each HMM are determined using the genes that are assigned to it (M-step).

1.3 Detection of transcriptional and posttranscriptional regulatory motifs

A flood of genome sequences and large scale expression data have combined to stimulate the maturation of bioinformatics methods for the analysis of sequences that mediate the regulation of gene transcription [98]. By interacting with RNA

[3] Reproduced from http://bioinformatics.oxfordjournals.org/cgi/reprint/19/suppl_1/i255 with permission.

polymerase or recruiting chromatin-modifying machinery, transcriptional regulators increase or decrease the transcriptional rate of genes, through transcriptional regulator binding DNA in the neighborhood of protein-coding and RNA genes. On the other hand, other level of transcriptional regulation besides that through transcription factors has recently arrested much attention due to the discovery of microRNAs (miRNAs) [99,100]. A large number of experimental and computational studies are aimed at locating the specific sequence motifs which can bind to either transcription factors (transcriptional regulatory motifs) or miRNAs (posttranscriptional regulatory motifs). Some prevailing algorithms for detecting these motifs are briefly reviewed in the following sections.

1.3.1 Identifying the *cis*-regulatory control elements shared by the co-expressed genes

Transcription in molecular biology is the copying from a DNA pattern to create a RNA molecule. The creation of new RNA from DNA cannot occur without transcription factors (TFs) and TFs can be activated or deactivated by other proteins. TFs are primarily involved in the initiation stage of RNA transcription -- they are the key to determining the position at which the DNA chain will be "unzipped". TFs are usually defined as proteins which show sequence-specific DNA binding and are capable of activating and/or repressing transcription of their target genes [13]. These transcription factor binding sites (TFBSs) contain 6~25 base pairs and are usually located in the upstream regions of the genes being regulated. They are important in facilitating the binding of TFs that control the transcription of the genes. Therefore it is reasonable to assume that genes regulated by the same TF should all contain the corresponding binding sites in their regulatory regions and exhibit similar expression profiles as measured by, for example, microarray technology [12]. Based on this assumption, a considerable number of algorithms have been developed that group genes into coexpressed classes and then search their upstream sequences for common motifs [12,101,102]. It is important to note that the only assumption required for these approaches is that genes regulated by the same TF should contain common binding sites and exhibit similar expression profiles. However, this is a quite strong assumption that allows investigation of coregulation through genome-wide sequence and expression data analysis [12]. Recognition of TFBS in the upstream regions of co-expressed (co-regulated) genes is crucial for elucidating gene regulatory networks. Among the numerous methods available for detecting motifs, three commonly used ones are briefly introduced here, namely phylogenetic footprinting, EM (expectation maximization) and Gibbs sampling.

(1) Phylogenetic footprinting

The method phylogenetic footprinting is based on the observation that regulatory elements are under selective pressure, which result in a slower evolution rate than that of the surrounding nonfunctional sequences [103]. The phylogenetic footprinting approach is aimed at detecting conserved sequences cross multiple closely related species. There are three components in the existing phylogenetic footprinting algorithms: comparison of suitable orthologous gene sequences, promoter sequence alignment of orthologous gene sequences and identification of significantly conserved segments [98]. Phylogenetic footprinting predicts highly conserved sequence elements that could function as potential regulatory motifs by comparing orthologous regulatory regions from multiple related species [103]. Prakash and Tompa have analyzed issues of reliability in studies in which comparative genomic approaches have been applied to the discovery of regulatory elements at a genome-scale level in vertebrates [103]. A key assumption in the application of phylogenetic footprinting is the implicit hypothesis that the regulation of orthologous genes are under the same regulatory mechanisms in different species [98]. Although straightforward and powerful, this approach can not be used when upstream regions are very divergent or have undergone genomic rearrangements. Furthermore, multiple closely related genome sequences are not always available [9].

(2) Expectation maximization (EM)

The EM algorithm is a general iterative algorithm for parameter estimation based on maximum likelihood when some random variables involved are not observed, i.e. missing or incomplete. In each repeat, the EM algorithm contains two steps, namely E-step (Estimation step) and M-step (Maximization step). The E-step replaces missing values by estimated values and computes the expected value of the log likelihood. The M-step maximized the expected log likelihood computed in the E-step in order to find the next estimates of the parameters.

Lawrence and Reilly handled the uncertainty in the location of binding sites by employing the missing information principle to develop an EM algorithm [104]. They assumed that each TFBS had a fixed length L and each input sequence needed to contain and only contained one copy of each TFBS. And these assumptions were also the limitation of their model. Each DNA binding motif was represented by a $4 \times L$ matrix, whose L-column elements represented the probabilities of A, C, G, T occurred at position l and$1 \le l \le L$.

Bailey and Elkan developed a motif finding approach – MEME, based on an extended EM algorithm [105], which could be used for discovering multiple motifs.

The major disadvantage of EM based algorithm is that it may end up with local optimum, such that different runs will give different results. Other limitations of EM algorithm based motif finding approaches are that the motif length L and expected number of motifs are both required to be pre-defined.

(3) Gibbs sampling

Gibbs sampling is a general method for probabilistic inference, which is well suited for coping with incomplete information. The algorithm generates a sequence of samples from the joint probability distribution of two or more random variables. The purpose is to approximate the joint distribution, or to compute an integral. In other words, Gibbs sampling approach is essentially a special numerical approximation method Markov Chain Monte Carlo (MCMC), which enables one to draw samples of high dimensional random variables in an iterative fashion. If $\pi(x)$ is a multivariate target distribution and the entire vector x will be updated by generating y from a density $q(x,y)$. At the i^{th} step, y_i is generated from the density $q_i(x_i, y_i)$, where q_i depends on the current value of x_i and $\pi(x)$ is uniquely determined by the set of full conditionals $\pi_i(x_i)$, where $i = 1,...,k$. The distribution for updating the i^{th} component of x is:

$$q_i(x_i, y_i) = \pi_i(y_i),$$

where $\pi_i(y_i)$ is the full conditional distribution of y_i. Thus Gibbs sampling consists of sampling from full conditionals of the target distribution. Given a set of sequences, Gibbs sampling algorithm can be used to find the motif shared by all or most sequences, while the motif's starting position in each sequence is unknown.

After the original Gibbs sampling algorithm for motif finding was reported in 1993 [106], Liu et al. developed a full Bayesian foundation of this algorithm and presented a rank test for the assessment of the significance of multiple sequence alignment [107]. Till now, various Gibbs sampling based motif finding approaches have been developed, such as Gibbs motif sampler [106,107], BioProspector [108] and AlignACE [109,110]. In this study, we used AlignACE to detect motifs located upstream of co-regulated genes, that correspond to the DNA binding preferences of TFs (Chapter 4).

1.3.2 Identifying the miRNA target motifs

microRNAs (miRNAs) are endogenous 20~24 nt RNAs that can play important roles by targeting mRNAs for cleavage or translational repression. Although they captured attention only recently, there are evidence that miRNAs are parts of ancient, conserved regulatory modules underlying developmental outcomes and comprise one of the abundant classes of gene regulatory molecules in multicellular organisms and likely influence the expression of many protein-coding genes [99,100,111].

Substantial differences have been reported between the animal and plant kingdoms in regard to the mechanisms and scope of miRNA-mediated gene regulation, including the biogenesis of miRNAs and the recognition between miRNA-mRNA duplex. Before describing our own approach to detect miRNA target motifs in plants (in Chapter 3), I will briefly review several existing algorithms of predicting either animal or plant miRNA targets.

1.3.2.1 Animal miRNA targets prediction

miRNAs are among the major class of regulatory genes, present in most metazoans and play important roles for a range of biological functions [112]. In animals, most of the miRNAs bind to the target 3’ untranslated region (UTR) with imperfect complementarity and function as translational repressors. The limited complementarity between miRNA-mRNA duplex are basis of all the methods to computationally predict animal miRNA targets. Recently, Rajewsky reviewed some of the computational approaches developed to predict miRNA targets in animals [112].

In 2003, the fly pro-apoptotic gene *hid* was the first miRNA (miRNA *bantam*) target identified by performing a genome-wide, sequence-based bioinformatics screening for targets of a miRNA [112-114]. Thereafter, different algorithms have been published in predicting targets for miRNAs in *Drosophila* and vertebrates [60,112,115-118]. Most of these methods detected target candidates generally based on two natures: (a) the 5’ end of miRNA (6-8 bp, called “seed sites”) needs to perfectly match to its target site, and the term “nucleus” was used for these seed sites which were found to be the key component of target recognition [111,112,117]; (b) The multiplicity of complementary sites in a 3’ UTR could exponentially boost the efficacy of target repression [111,112]. Additional constraints were also required in almost all the methods in order to increase the specificity of the algorithms, including conservation of the seed sites, free energy of the miRNA-mRNA duplex lower than a cut-off value, and the requirement of more complementary sites for miRNAs than their shuffled sequences (control). Although the additional constraints could dramatically reduce the number of target candidates which might lead to the increased specificity, there were some arguments about the application of these additional constrains. Firstly, the conservation constraint will discard species-specific candidates. Secondly, there was evidence that the nuclei initiates a rapid zip up of the miRNA-mRNA duplex to overcome thermal diffusion, followed by a stabilizing thermodynamic step of further annealing of the miRNA to the target site [117]. The study indicated that the free energy requirement of the entire miRNA-mRNA duplex was generally a bad predictor for miRNA target sites [112]. Thirdly, in order to assess the statistical significance of targets, the abundance of perfect matches to the 5' end of the shuffled sequences and miRNAs was compared. Since the shuffled sequences were unlikely to be biologically relevant, so the observation of more "hit" for miRNA than shuffled sequence could serve as an indicator that many of the target candidates were indeed biological relevant. The ratio of "real" versus "shuffled" hits provided an estimate of the signal to noise ratio of the target predictions. This logic seemed natural, however, Rajewsky suggested that if a miRNA had only very few targets because it might be functional disadvantageous for a mRNA to be targeted, then the number of its target candidates could be fewer than the number of hits produced by shuffled sequences [112].

New experimental evidence emerged in 2005, which showed that there are at least two classes of miRNA targets in animals. One class is as the aforementioned, the other class has imperfect 5' matches but compensates via additional base pairing in the 3' end of miRNAs [112]. The number of predicted targets for each miRNA varies from a handful to more than 800 unique genes. Lall et al. suggested that a high fraction of all the conserved 3’ UTR motifs among vertebrates are complementary to the 5’ end of known miRNAs, which indicated that these motifs were likely under functional constraints mediated by miRNAs [112,119]. Chan et al. used network-level conservation between pairs of fly (*Drosophila melanogaster*/*D. pseudoobscura*) and worm (*Caenorhabditis elegans*/*C. briggsae*) genomes to detect highly conserved mRNA motifs in 3’ UTRs. Many of these elements were complementary to the 5’ end of known miRNAs and likely to be their target sites [100].

A systematic comparison of miRNA target prediction algorithms was recently carried out versus ~ 130 experimentally assayed miRNA-mRNA regulatory relationships in *D. melanogaster* and the result showed that the algorithm PicTar ranked relatively high in the comparisons of both accuracy and sensitivity [112,120].

PicTar computed a maximum likelihood score that a certain 3’ UTR was targeted by a fixed set of miRNAs. At each step one of the states, i ($i \in \{0...M\}$) was chosen with prior probability ρ_i, where ρ_0 was the prior probability for the background, and M was the total number of different miRNAs whose combinatorial effects were assessed. Depending on the nature of the state, a certain sequence would be emitted, namely one nucleotide would be emitted in the background state and a 7~8 mer would be emitted in miRNA target site state.

The emitted sequence was appended to the previously generated sequence and the process repeated until the length of 3'UTR was reached. The likelihood was calculated as:

$$P(S \mid \theta, T) = \prod_{i=1}^{N(T)} \rho_i \cdot m_i(s),$$

where T was the state path and $N(T)$ was the total number of states in the path. $m(s)$ was the emission probabilities for the miRNA binding sites (the probability that a certain mRNA subsequence would be the miRNA binding site), which was estimated from the experimental data [121].

1.3.2.2 Plant miRNA targets prediction

One of the major differences between plant miRNAs and animal miRNAs is that for plant miRNAs, base paring with the corresponding targets is near-perfect and their complementary sites are located in coding regions of the target genes instead of being limited to the 3' UTRs [122]. The first group which used computational approach to predict *Arabidopsis* miRNA targets was Bartel's lab at the Whitehead Institute for Biomedical Research and the Massachusetts Institute of Technology [122]. They used PatScan [123] to search complementary sites of miRNAs. PatScan is based on "fuzzy matching" algorithm, which might result in incorrect output of the counts of mismatches. In 2004, Jones-Rhoades and Bartel as well as Wang et al. both developed comparative genomic approaches to systematically identify miRNAs ant their targets that were conserved in *Arabidopsis* and rice

[124].

1.4 Recent approaches aimed at elucidating gene regulatory networks

A comprehensive study should be at the network level which focuses on the interactions between genes, and attempts to build descriptive and predictive models for different systems in a cell. Microarray technology, which is capable of the simultaneous measurement of all RNA transcripts in a cell, has spawned the development of algorithms for reverse-engineering transcriptional networks. DNA microarray technology offered the possibility to infer, or “reverse-engineer”, a model of cell’s underlying transcription control systems (Figure 1.4.1). The development of reverse-engineering methods remains a challenge because of the nature of data, which are typically noisy, high dimensional and substantially under-sampled. Moreover, well-understood and standardized benchmark systems are not available which can evaluate algorithm performance without any bias. There is still an open question regarding to experimental design, the reliability of the predicted networks, and the utility of various approaches for particular applications [89].

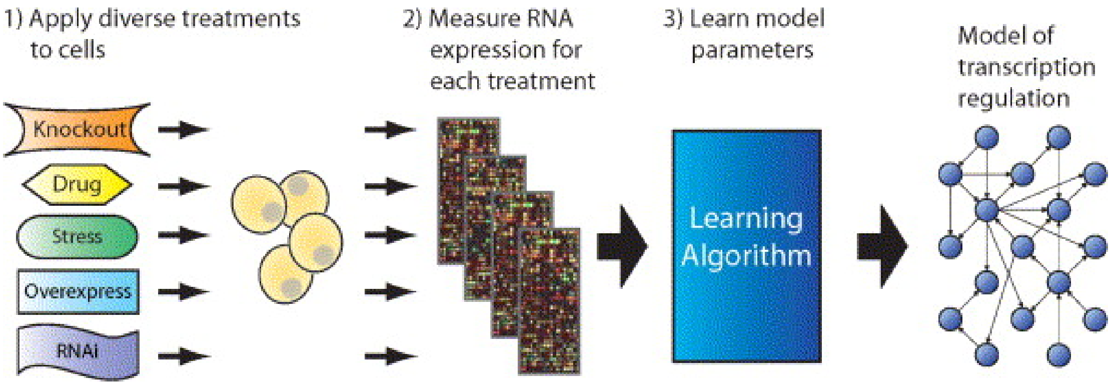

1) Apply diverse treatments to cells
2) Measure RNA expression for each treatment
3) Learn model parameters
Model of transcription regulation
Knockout
Drug
Stress
Overexpress
RNAi
Learning Algorithm

Figure 1.4.1 The general strategy for reverse-engineering transcription control systems.

(1) The cells are perturbed with various treatments to elicit distinct responses. (2) The expression level (concentration) of many or all RNA transcripts in the cells is measured. (3) Parameter learning of the model that describes the transcription control system underlying the observed responses. The resulting model may then be used in the analysis and prediction of the control system function[4].

Reverse-engineering techniques have principally focused on decoding the mechanisms that control gene transcription and seek to model causal relationship between RNA transcripts and the causal relationships may or may not correspond to true molecular interaction.

There are many methods which could be used to reconstruct the biological networks, and each has its own advantage and disadvantage, respectively (Table 1-1).

Methods	**Advantage**	**Disadvantage**
Clustering		inconsistent
Linear modeling		rough network models
Boolean networks	logic, general	rough, determinism
Differential equations	Exactitude	less of training data, time delay, high computational cost

[4] Reproduced from http://gardnerlab.bu.edu/publications/Gardner with permission.

Bayesian networks	local	limitation in network structure (e.g. self-feedback)

Table 1-1 The comparison among different methods for the reconstruction of biological networks

Three statistical approaches are further reviewed, i.e. Boolean network, differential regression and Bayesian network [89].

1.4.1 Boolean network models

In Boolean network models, a gene takes one of the two states from binary space $\{0,1\}$, and a gene regulation rule is given as a Boolean function [125]. A "0" represents that a gene is not expressed or its expression level remains unchanged relative to control sample. A "1" represents that a gene is expressed or its expression level is changed under certain condition. The state of a gene is determined as a Boolean function of the state of the input genes [89]. A Boolean network $G(V,F)$ consists of a set $V=\{v_1,...,v_N\}$ of nodes representing genes and a list $F=(f_1,...,f_N)$ of Boolean functions, where a Boolean function $f_i(v_{i1},...,v_{ik})$ with inputs from specified nodes $v_{i1},...,v_{ik}$ is assigned to each node v_i. An expression pattern ψ is an index function from V to $\{0,1\}$. That is, ψ assumed to take either 0 or 1 as its state value. Expression ψ_{t+1} at time $t+1$ is determined by Boolean functions F from expression pattern ψ_t at time t, i.e.

$$\psi_{t+1}(v_i)=f_i(\psi_t(v_{i1}),...,\psi_t(v_{ik}))$$ [125].

Learning Boolean network requires large amount of experimental data since it does not place constraints on the form of Boolean function f_i. For a fully connected Boolean network, an algorithm needs to require approximately 2^N data points to infer all interaction functions [1]. In practice, sparsely connected networks are usually assumed in order to reduce the demand for data. However, the data requirements are still considerable. Most studies based on Boolean networks have examined only simulated data sets. Hence it is difficult to assess their practical utility. Although Boolean networks had offered promise, such as Yuh et al. provided a Boolean network model to reveal the logic interrelations between a sea urchin control element and gene expression in the endoderm during development [126], the data requirements may impede their practical use in reverse-engineering [89].

1.4.2 Differential equation models

A gene network can be described as a system of differential equations. The rate of change in expression level of a particular gene x_i, is given by nonlinear influence function f_i of the expression levels of the genes in the network:

$$\frac{dx_i}{dt} = f_i(x_1, \ldots, x_N),$$

where N is the number of genes in the network. The influence functions f_i are usually presupposed, including sigmoidal functions and linear functions [88,89]. The linear function has simplifying power, which can dramatically reduce the number of parameters needed to describe the conditional dependence between gene expression levels and avoid overfitting problem. Therefore, the amount of data to learn a linear model is much less than that required by more complex nonlinear models, such as Boolean network and Bayesian network models. This advantage is significant considering the high cost of experimental data. However, this gain in experimental efficiency is acquired at the cost of placing strong constraints on the nature of regulatory mechanisms in the cell, which may lead to errors in the network model and end up with models that can only describe the regulatory mechanism under certain conditions [89]. Akutsu et al. combined Boolean network model and nonlinear differential equation model to infer genetic network architecture from time series data of gene expression profiles [125]. Bussemaker et al. provided a method for discovering *cis*-regulatory elements based on a multiple linear regression model in which upstream motifs contribute additively to the log-expression level, and the authors pointed out that the log-linear model captured about 30% of the signal given a test set of yeast expression experiments [127]. However, the success of their approach depended on the assumption that the combinations of TFs act as a log-linear function of gene expression level which may lead to errors in predictions. Hence, more complex nonlinear models of gene regulatory network may be more suitable, such as Bayesian network model which could also capture location and orientation features of binding motifs, as described below [7].

1.4.3 Bayesian network models

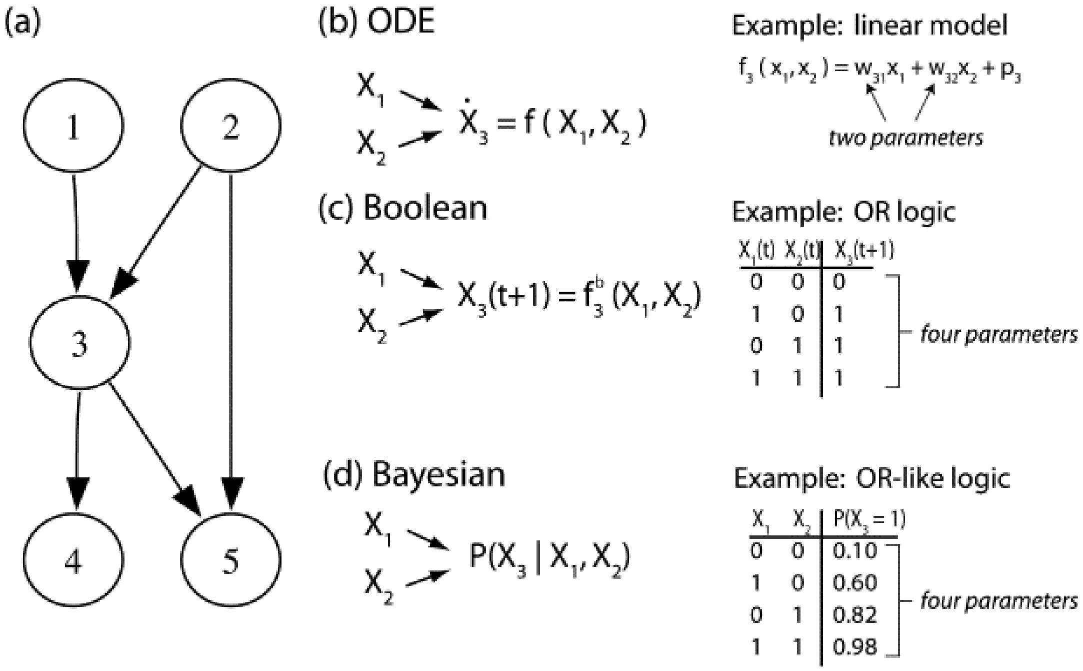
Data are treated as random variables in a probabilistic model, and the probability distribution of a random variable depends on parameter values. Bayesian models are sometimes called fully probabilistic for the parameter values are also treated as random variables [128]. A Bayesian belief-network structure B_S is a directed acyclic graph in which nodes represent domain variables and arcs between nodes represent probabilistic dependencies [129,130]. An attractive feature of Bayesian analysis is its ability to incorporate background information into the specification of the model [128].

Let X_i denote the state of a particular gene, which can be specified by a probability distribution function f_i and depends on a set of regulatory genes X_j :

$$P(X_i = x_i \mid X_j = x_j) = f_i(x_i \mid x_j),$$

where $j = 1,...,N$ (N is the number of genes), $j \neq i$, and x is a unique instantiation of X. As for differential equation models, the algorithm used to learn the model usually presupposes the form of conditional probability function f_i. Any function may be used, such as Boolean and linear functions. However, there will be a trade-off between model realism and model simplicity. More realistic models contains more parameters, which will require more experimental data and greater computational effort [89].

Two sets of parameters need to be determined in order to reverse-engineer a gene regulation network using Bayesian analysis: the model topology and the conditional probability functions relating the state of regulatory genes to the state of regulated gene.



$X_1(t)$	$X_2(t)$	$X_3(t+1)$
0	0	0
1	0	1
0	1	1
1	1	1

X_1	X_2	$P(X_3 = 1)$
0	0	0.10
1	0	0.60
0	1	0.82
1	1	0.98

Figure 1.4.2 An example for gene regulatory network.

(**a**) An example for gene regulatory network with five genes and five nodes. Three different models are used to represent the conditional dependencies among gene expression levels, as illustrated in (**b**), (**c**), and (**d**). (**b**) Relationship obtained by differential equation model. For a linear relationship, the model requires two parameters to relate gene 1 and 2 to gene 3, respectively. (**c**) Relationship obtained by Boolean network model. The model requires four parameters to relate gene 1, 2 and 3. "OR" logic is illustrated. (**d**) Relationship obtained by Bayesian network model. The model is similar to the Boolean model; it requires four parameters to relate genes 1, 2 and 3, though the parameters specify probability distributions rather than deterministic relationships. "OR-like" logic is illustrated[5].

Figure 1.4.2 illustrated an example for gene regulatory network based on three different models. When the expression level of a gene is discretized into binary random variables, the algorithm needs to learn four parameters to

[5] Reproduced from http://gardnerlab.bu.edu/publications/Gardner with permission.

specify $P(X_3 | X_2, X_1)$, i.e. the conditional distribution for gene 3. For each of the 2^k combinations of the states of the k parents, there is a parameter. Hence, to fully specify the distribution for a particular gene, each of the 2^k state combinations must be experimentally observed at least once (as is the case for Boolean networks), which may result in an impractically large number of experiments. Training data for such models are often incomplete. However, the Bayesian network structure enables an algorithm to partially specify the conditional distribution function for those states that are observed and the set of partially complete distribution functions can be sufficient to determine the topology of the network. The network structure is usually determined by a heuristic search, such as greedy-hill climbing approach. When the training data are not complete, the learning problem is underdetermined and several high-scoring networks are found. So we can average model or do bootstrapping to select the most probable regulatory interactions [131]. Furthermore, the probabilistic structure of a Bayesian network enables incorporation of the prior knowledge via application of Bayes rule [89,132]. More details on Bayesian network model is provided below in Chapter 2.

Bayesian network model is a promising tool for analyzing gene expression patterns. Firstly, it is particularly useful for describing processes composed of locally interacting components, that is, the value of each component directly depends on the values of a relatively small number of components. Secondly, statistical foundations for learning Bayesian networks from observations and computational algorithms are well understood and have been used successfully in many applications [133]. Thirdly, Bayesian network models provide a way to account for the noise and data limitations inherent in expression studies, as well as retain the combinatorial logic of transcription regulation [89]. The main limitation of Bayesian network models is that the network structure does not allow cycles, i.e., feedback loops.

Researchers have devoted considerable attention in recent years to the use of Bayesian network approaches for reverse-engineering gene regulatory networks [89,133-136]. Using Bayesian networks to represent statistical dependencies, Friedman et al. proposed a framework for discovering interactions between genes based on multiple expression measurements [133]. Segal et al. used a Naïve Bayes model [137] to identify transcriptional modules—sets of genes that are co-regulated in a set of experiments, through a common motif profile [6]. Instead of relying on a linear model for gene expression, Beer and Tavazoie used a Bayesian network model to capture non-linear effects, and could correctly predict the expression patterns for 73% of the 2587 genes examined [7]. Their results suggested the presence of non-linear regulatory control and showed a marked improvement over the predictive capabilities of the Bussemaker algorithm. Sabatti and James used a Bayesian hidden component model which integrated literature information, DNA sequences and expression arrays to identify the potential binding sites actually used by the regulatory proteins in the studied cell conditions, the strength of their control, and their activation profile in a series of experiments [138].

In our own work, we have applied a Bayesian network model to reverse-engineer the gene regulatory network at sequence level (see Chapter 4).

Chapter 2: Statistic Methods – Hidden Markov Model and Bayesian Network

In this chapter, two statistical modeling methods used in this study are introduced, namely hidden Markov model and Bayesian network model. HMM and Bayesian network were used to predict plant miRNA targets and model the transcriptional and posttranscriptional regulatory networks, respectively.

2.1 Hidden Markov model

Hidden Markov model is an extension to the classic Markov chain model, in which the state of each observation is drawn randomly from a distribution, the state transition of which follows a Markov chain.

2.1.1 Markov chains

A classical Markov chain model generates sequences in which the probability of a symbol depends on the previous symbol [139]. The key property of a Markov chain (first order) is that the probability of each symbol o_t depends only on the value of the preceding symbol o_{t-1}, not on the rest previous symbols:

$$P(o) = P(o_T \mid o_{T-1})P(o_{T-1} \mid o_{T-2}) \cdots P(o_2 \mid o_1)P(o_1) = P(o_1)\prod_{t=2}^{T} a_{o_{t-1}o_t} ,$$

where the $a_{o_{t-1}o_t}$ is the transition probability.

2.1.2 Elements of an HMM

A HMM contains two chains. One is the state chain (hidden) and the other is the symbol chain (observed). The state sequence is called the path I . The path itself follows a Markov chain, thus the probability of a state only depends on the immediate previous state. The state in the path is called I_t

and $a_{ij} = P(I_t = j \mid I_{t-1} = i)$ is the state transition probability and $A = \{a_{ij}\}$ is the state transition probability matrix.

In general, a state can produce a symbol from a distribution over all possible symbols and:

$$e_j(b) = P(o_t = b \mid I_t = j),$$

where $e_j(b)$ is defined as the emission probability that symbol b is observed when in state j and $E = \{e_j(b)\}$ is the emission probability matrix. Then the joint probability of an observed sequence o and a state path I is [139]:

$$P(o, I) = a_{0I_1} \prod_{t=1}^{T} e_{I_t}(o_t) a_{I_t I_{t+1}} \;,$$

where a_{0I_1} is the initial probability that the state path begins at state I_1 and $Z = \{a_{0I_1}\}$ is the initial probability matrix.

$\lambda = (A, E, Z)$ will be used to denote an HMM as a compact notation.

Most applications of HMMs mainly focus on three problems:

Problem 1: Given the model $\lambda = (A, E, Z)$, how to compute $P(o \mid \lambda)$, the probability of observing sequence o ?

Problem 2: Given the model $\lambda = (A, E, Z)$, how to choose a state path $I*$ in order to maximize $P(o, I \mid \lambda)$, the joint probability of the observation o and the state path I ?

Problem 3: How to obtain the HMM model parameters $\lambda = (A, E, Z)$ which can maximize $P(o \mid \lambda)$ (or $P(o, I \mid \lambda)$)?

Problem 1 and *Problem 2* can be viewed as analysis problems while *Problem 3* is a typical model identification or training problem.

2.1.3 Forward algorithm

Forward and backward algorithm both can solve the *problem 1*. The forward algorithm is a dynamic programming algorithm which can efficiently enumerate all paths to calculate the probability of a sequence, as $P(o) = \sum_I P(o, I)$. To calculate $P(o)$, we first define:

$$f_i(t) = P(o_1 ... o_t, I_t = i),$$

where $f_i(t)$ (forward variable) is the probability of the observed sequence up to and including o_t, and $I_t = i$ [139]. And the recursion equation for updating the forward variable is:

$$f_j(t+1) = e_j(o_{t+1}) \sum_i f_i(t) a_{ij} .$$

Following is the pseudo-code for forward algorithm.

Input: observed sequence o, transition probabilities a_{ij}, and emission probabilities $e_j(b)$.

Output: $P(o)$, the probability of observing a sequence o.

```
$f_0(0) = 1;$          # initialization
$f_i(0) = 0;$          # for $i > 0$
for $t = 1$ to $T$ do                    # recursion
    $f_j(t) = e_j(o_t)\sum_i f_i(t-1)a_{ij}$ ;
end
$P(o) = \sum_i f_i(T)a_{i0}$ ;
```

2.1.4 Backward algorithm

Analogous to the forward variable, the backward variable $b_i(t)$ is the probability of the observed sequence back to and including o_t [139], which is defined as:

$$b_i(t) = P(o_{t+1} \ldots o_T \mid I_t = i),$$

where $b_i(t)$ is obtained by a backward recursion starting at the end of the sequence, and the pseudo-code for backward algorithm is given below.

Input: observed sequence o, transition probabilities a_{ij}, and emission probabilities $e_j(b)$.

Output: $P(o)$, the probability of observing a sequence o.

```
$b_i(T) = a_{i0}$;                # initialization
for $t = T-1$ to 1 do             # recursion
    $b_i(t) = \sum_j a_{ij} e_j(o_{t+1}) b_j(t+1)$;
end
$P(o) = \sum_j a_{0j} e_j(o_1) b_j(1)$;
```

The posterior probability of state i at time t, given the observed sequence o, can be calculated as [140] :

$$P(I_t = i \mid o) = \frac{f_i(t) b_i(t)}{P(o)},$$

where $P(o)$ can be calculated by either forward or backward algorithm.

2.1.5 Viterbi algorithm

The Viterbi algorithm is a common method for finding the most probable state transition path and its probability in HMMs [139,141], so it can be used to solve the *problem 2*. The path I^* with the highest probability should be chosen as follows:

$$I^* = \arg\max_I P(o, I).$$

Let $v_i(t)$ (the Viterbi variable) be the probability of the most probable path ending in state i with observation t and:

$$v_j(t+1) = e_j(o_{t+1})\max_i(v_i(t)a_{ij}).$$

Each state path starts from the initial state 0, thus $v_0(0) = 1$. The most probable path $I*$ can be found by backtracking recursively [139]. Following is the pseudo-code for Viterbi algorithm.

Input: observed sequence o, transition probabilities a_{ij}, and emission probabilities $e_j(b)$.

Output: the most probable state path $I*$.

```
$v_0(0) = 1$;          # initialization
$v_i(0) = 0$;          # for $i > 0$
for $t = 1$ to $T$ do                   # recursion
    $v_j(t) = e_j(o_t)\max_i(v_i(t-1)a_{ij})$;
    $\psi_t(j) = \arg\max_i(v_i(t-1)a_{ij})$;
end
$P(o, I*) = \max_i(v_i(T)a_{i0})$;          # $P(o, I*)$ gives the required state-optimized
                                            probability
$I_T* = \arg\max_i(v_i(T)a_{i0})$;        # termination
for $t = T$ to $1$ do               # tracing back
    $I_{t-1}* = \psi_t(I_t*)$;
end
# $I* = \{I_1*, I_2*, ..., I_T*\}$ is the optimal state path
```

2.1.6 Parameter estimation for HMMs: Baum-Welch algorithm

Baum-Welch algorithm [140], which can be used to solve *problem 3*, is a particular iteration method that is used to estimate parameter values. This algorithm maximizes $P(o \mid \lambda)$ by adjusting the parameters of λ. The optimization criterion is called the maximum likelihood criterion and the function $P(o \mid \lambda)$ is called the likelihood function.

The maximum likelihood estimators for a_{ij} and $e_j(b)$ are given by [139]:

$$a_{ij} = \frac{A_{ij}}{\sum_{j'} A_{ij'}} \quad \text{and} \quad e_j(b) = \frac{E_j(b)}{\sum_{b'} E_j(b')},$$

where A_{ij} and $E_j(b)$ are the counts of particular transition (state i to state j) and emissions (state i to observed symbol b), respectively.

The Baum-Welch algorithm calculates the A_{ij} and $E_j(b)$ as the expected number of times each transition and emission is used. And the probability that state i to state j transition is used at position t in sequence o can be calculated as:

$$P(I_t = i, I_{t+1} = j \mid o, \lambda) = \frac{f_i(t) a_{ij} e_j(o_{t+1}) b_j(t+1)}{P(o)},$$

then the A_{ij} can be derived by summing over all positions and over all training sequences:

$$A_{ij} = \sum_k \frac{1}{P(o^k)} \sum_t f_i^k(t) a_{ij} e_j(o_{t+1}^k) b_j^k(t+1),$$

where $f_i^k(t)$ is the forward variable calculated for sequence k and $b_j^k(t)$ is the backward variable calculated for sequence k. The $E_j(b)$ can be calculated as:

$$E_j(b) = \sum_k \frac{1}{P(o^k)} \sum_{\{t \mid o_t^k = b\}} f_i^k(t) b_i^k(t),$$

where the inner sum is only over those positions at which the symbol is b [139].

Baum-Welch algorithm is a special case of EM algorithm, which is a very powerful approach for probabilistic parameter estimation. The E-step calculates the expectations A_{ij} and $E_j(b)$, which is done by the forward and backward algorithm. The M-step plugs A_{ij} and $E_j(b)$ into the re-estimation formulae for a_{ij} and $e_j(b)$, respectively [139]. The pseudo-code for Baum-Welch algorithm is given below.

```
Input: observed sequence $o$.
Output:   transition probabilities $a_{ij}$, and emission probabilities $e_j(b)$.
Arbitrarily assign model parameters;          # initialization
while ($change > e^{-12}$) do
    $A_{ij} = 0$;
    $E_j(b) = 0$;
    $likelihood' = likelihood$;
    for $k = 1$ to $K$ do                # for each sequence $k = 1..K$
        $f_j^k(t) = e_j(o_t)\sum_i f_i^k(t-1)a_{ij}$;
        #calculate $f_i(t)$ for sequence $k$ using the forward algorithm
        $b_i^k(t) = \sum_j a_{ij}e_j(o_{t+1})b_j^k(t+1)$;
        #calculate $b_j(t)$ for sequence $k$ using the backward algorithm
        $A_{ij} = \sum_k \frac{1}{P(o^k)}\sum_t f_i^k(t)a_{ij}e_j(o_{t+1}^k)b_j^k(t+1)$;
        $E_j(b) = \sum_k \frac{1}{P(o^k)}\sum_{\{t|o_t^k=b\}} f_i^k(t)b_i^k(t)$; # add the contribution of sequence $k$ (E-step)
    end
    $a_{ij} = \frac{A_{ij}}{\sum_{j'}A_{ij'}}$;     # calculate the new model parameters (M-step)
    $e_i(b) = \frac{E_j(b)}{\sum_{b'}E_j(b')}$;
    $likelihood = \log(\prod_k P(o^k))$;     # calculate the new log likelihood of the model
    $change = likelihood - likelihood'$;
    # convergence criterion
    # program will stop when the change in total log likelihood is sufficiently small
end
output the parameters;
```

2.2 Bayesian network model

A Bayesian network represents a joint probability distribution [133]. It is a graph-based model of joint multivariate probability distributions which can capture properties of conditional dependence between variables. Such a model provides a clear methodology for learning from observations and is attractive for its ability to describe complex stochastic processes [133].

From the Bayes theorem, we have:

$$P(\phi \mid D) = \frac{P(D \mid \phi)P(\phi)}{P(D)},$$

where the likelihood $P(D \mid \phi)$ is the probability of the data given a particular set of parameter values. The marginal likelihood $P(D)$ is known as the “prior predictive distribution”, namely the probability distribution of the data irrespective of the parameter values. The prior $P(\phi)$ is the probability distribution of all combinations of parameter values before observing the data. And the posterior distribution $P(\phi \mid D)$ is the conditional distribution of the parameters given the observed data [2]. Typically, the likelihood will arise from a statistical model in which it is necessary to consider how the data can be “explained” by the parameter(s). The prior is an assumed distribution of the parameters, and it is obtained from background knowledge [2].

The goal of Bayesian network is not to prove the correlation between two variables but rather to select the variables that are likely to be correlated given the observed data. In the biological context, Bayesian network should be used to understand the network of dependencies among the factors involved, to detect the strongest dependencies and remove independencies [142].

2.2.1 Induction of probabilistic network from data

Let Pa_i denote the parent nodes of variable X_i.

$$P(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} P(x_i \mid pa_i),$$

where x_i denotes the instantiation of X_i and pa_i denotes the instantiation of Pa_i, respectively. Let q_i be the number of different Pa_i values and w_{ij} be the jth unique instantiation of the values in pa_i, relative to the ordering of the cases in D [130]. Let Z be a set of n discrete variables, where a variable X_i in Z has r_i possible value assignments $(v_{i1}, ..., v_{ir_i})$. N_{ijk} denotes the number of cases in D in which variable X_i has the value v_{ik} and pa_i is instantiated as w_{ij}, and

$$N_{ij} = \sum_{k=1}^{r_i} N_{ijk}.$$

When there were unrestricted multinomial distribution, parameter independence, Dirichlet priors and complete data, we have

$$P(D \mid \phi) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(a_{ij})}{\Gamma(a_{ij} + N_{ij})} \cdot \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})},$$

where $a_{ijk} = 1$ and $\alpha_{ij} = \sum_{k=1}^{r_i} \alpha_{ijk}$ under the assumption of uniform priors [130,143].

2.2.2 K2 algorithm

K2 is a greedy-search algorithm, which begins by making the assumption that a node has no parents, and then adds incrementally the parent nodes which can most increase the probability of the resulting structure. When the addition of no parent can further increase the probability, the process will stop [130,143].

Function $g(i, Pa_i)$ is used in K2:

$$g(i, Pa_i) = \prod_{j=1}^{q_i} \frac{\Gamma(a_{ij})}{\Gamma(a_{ij} + N_{ij})} \cdot \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})}.$$

Below is the pseudo-code of K2 algorithm, where u is the maximum number of parents that a node is allowed to have (decided by the user) and we use $\Pr ed(X_i)$ to denote the set of nodes that precede X_i in the node ordering [130].

```
Input: An ordered list of $n$ nodes, an upper bound $u$ for the number of parents a node
may have, and a database $D$ containing $h$ cases.
Output: For each node, a printout of the parents of the node.
for $i = 1$ to $n$ do
    $Pa_i$ = Ø;
    $P_{old} = g(i, Pa_i)$;
    OKToProceed = true;
    while OKToProceed and $|Pa_i| < u$ do
        $z = \arg\max_{\Pr ed(X_i)} g(i, Pa_i \cup \Pr ed(X_i))$;
        # $z$ is the node in $\Pr ed(X_i)$, which can maximize $g(i, Pa_i \cup \{z\})$
        $P_{new} = g(i, Pa_i \cup \{z\})$;
        if $P_{new} > P_{old}$ then
        $P_{new} = P_{old}$;
        $Pa_i = Pa_i \cup \{z\}$;
        else OKToProceed = false;
    end
    output parents of node $i$;
end
```

2.2.3 Predictions of interest

A model with the highest log marginal likelihood (or the highest posterior probability, assuming equal priors on structure) is the best sequential predictor of the data D. We can average the possible configurations of θ_S (vector of parameters) to obtain predictions of interest:

$$P(x \mid D, \phi) = \prod_{i=1}^{n} \frac{\alpha_{ijk} + N_{ijk}}{\alpha_{ij} + N_{ij}},$$

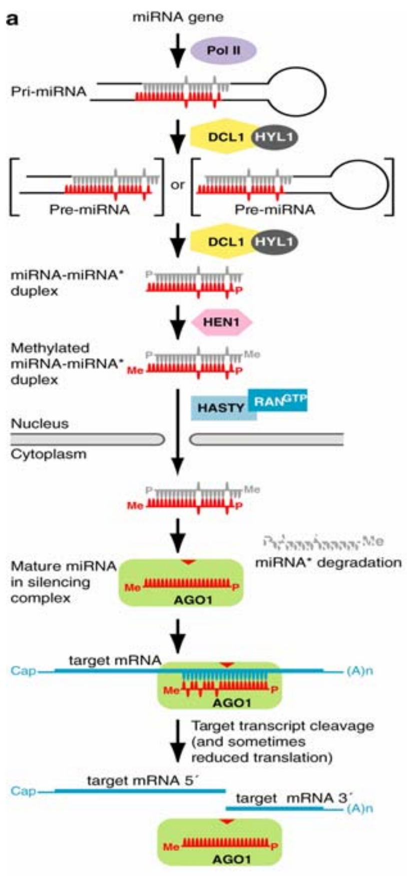
where x is the next case to be seen after D [144].

Chapter 3:Prediction of Plant miRNA Targets

The approaches reviewed in Chapter 1 make use of both genome sequence data and gene expression data to enhance the sensitivity and specificity of predicted regulatory networks. The limitation of these approaches is that they can not detect regulatory control by mechanisms aside from transcription factors, however, many regulators are posttranscriptionally regulated [8]. Another level of regulation beyond that through transcription factors can be considered by including information of miRNA targets. The role of miRNAs in the negative regulation of the expression of their target genes has been demonstrated recently. miRNAs regulate the expression of their target genes, by mRNA cleavage or translational repression. Most plant miRNAs can mediate the destruction of their target mRNAs, and through targeting transcription factors, they can regulate the transcription of a large number of genes, directly or indirectly. The first step in understanding the mechanism of miRNAs is to identify their regulatory targets. In this chapter, we describe a novel method based on an inhomogeneous hidden Markov model (HMM) developed to predict plant miRNA targets without additional conservation constraint. This model was trained by the information about one third of all the confirmed miRNAs, whereas it was capable of finding all the experimentally validated targets for all the known miRNAs.

3.1 Introduction

Recent studies have shown that a number of known functional elements are noncoding sequences, which include regulatory signals, RNA genes and structure elements [63]. Posttranscriptional regulation through RNA-RNA interaction has recently arrested much attention due to the discovery of microRNAs (miRNAs) [99,100]. Figure 3.1.1 illustrated the biogenesis of plant miRNAs (Figure 3.1.1).Transcription regulation is thought to occur primarily through the binding of TFs to *cis*-regulatory motifs, whereas posttranscriptional regulatory mechanism such as miRNA-mediated degradation has also been reported [76,145,146]. miRNAs are a class of endogenous small RNAs that negatively regulate mRNA expression either by inducing degradation of the targeted transcript or by decreasing translational efficiency [9,45,111]. Hutvagner and Zamore suggested that the different regulatory mechanisms are due to the degree of complementarity between miRNAs and their targets [45]. Plant miRNAs are commonly perfectly complementary to their targets and cause the cleavage of the targets by RNA-induced silencing complex (RISC), whereas in animals targets with weaker complementarity appear to have decreased translational efficacy [100,122,147-149]. Recent evidence raised the possibility that miRNAs can also guide transcriptional silencing in plants [150,151].



Jones-Rhoades MW, et al. 2006.
Annu. Rev. Plant Biol. 57:19–53

Figure 3.1.1 A model for miRNA biogenesis in *Arabidopsis*[6].

The pri-miRNA is processed by DCL1, perhaps with the aid of HYL1, to a miRNA-miRNA* duplex with 5' phosphates (P) and 2-nt 3' overhangs. The 3' sugars of the miRNA-mRNA* duplex are methylated by HEN1. The mature, methylated miRNA is incorporated into a silencing complex that can include AGO1 and the miRNA* is degraded. Within the silencing complex, the miRNA is capable of targeting complementary RNAs for cleavage by AGO1, and perhaps also for translational repression.

One of the most important steps to understand the mechanisms of miRNAs is to identify their regulatory targets. Since the miRNAs recognize their regulatory targets through base pairing, computational methods can be applied to predict miRNA targets based on the high complementarity between miRNA-mRNA

[6] Reproduced from http://arjournals.annualreviews.org/doi/full/10.1146/annurev.arplant.57.032905.105218 with permission.

duplex. All the reported prediction approaches detect potential targets with up to 4 mismatches between miRNA-mRNA duplex and many of the predicted targets have been experimentally validated [122,124,152]. Our approach to detect miRNA target motifs consists of two steps. Firstly, we worked out a direct search approach which detected the miRNA potential targets with at most N mismatches. When we set N to 4, the signal (number of potential targets for miRNAs) was significantly greater than the noise (number of potential targets for randomly shuffled sequences), which suggested that the false positive rate under this setting was relatively low. However, there are natural targets with more than 4 mismatches [153]. Although the direct search approach can be used to find potential targets with any number of mismatches, the more the mismatches allowed the higher the false positive rate would be. Therefore how to determine the optimal value of N in plant miRNA target prediction remains an open question. Most published miRNA prediction algorithms have not included the information about position-specific matches/mismatches of miRNA-mRNA duplex, which treated all the mismatches equally regardless of their positions. We assumed that there might be some position-specific rule of particular matches/mismatches and this information was a general rule contained (hidden) in each miRNA-mRNA duplex. We chose an inhomogeneous HMM to detect this hidden information by using a relatively small training dataset and further applied it to each identified miRNAs. This approach dramatically extends our knowledge of underlying miRNA-mediated regulatory role and enlarges the dataset of miRNA target motifs by loosing the requirement for the degree of complementarity between miRNA-mRNA duplex while reserves the information about position-specific mismatches/matches.

3.2 Materials and Methods

3.2.1 Materials

3.2.1.1 Sequences of known mature miRNA

The *Arabidopsis* mature miRNA sequences were downloaded from miRBase (Release 9.1) released in Feb 2007 [154,155]. The 19 miRNA sequences in Release 3.0 (released in Jan 2004) [154,155] were used to generate the training set of potential miRNA targets for the HMM of miRNA target prediction.

3.2.1.2 Coding sequences of *Arabidopsis*

Sequences of all the transcripts were retrieved from the TAIR (ftp://ftp.arabidopsis.org/seq_analysis_updates/), released in Feb 2006.

3.2.2 Methods

3.2.2.1 Training set preparation

The direct search approach detected the mRNA sequences that were complementary, with at most N mismatches, to at least one of the identified miRNAs (released in miRBase 3.0), where N was the number of the mismatches allowed for a miRNA target. Strings of “0” and “1” were generated in order to represent the position specific complementarity (match/mismatch) between miRNA-mRNA duplex, where 0 represented a position specific match and 1 represented a position specific mismatch. For example, 00000000000000000111 and 00100000010000000010 both represented 20-nt sequences that had three mismatches with miRNAs, though the position of matches/mismatches were different. Each string was then converted to sequence patterns which had the corresponding position specific match/mismatch with miRNA. Gap was not allowed in this algorithm and the noncanonical pairs such as G-U were treated as mismatches.

3.2.2.2 Randomly shuffled sequences

Simulation study to assess the statistic significance of the detected miRNA target candidates was performed using randomly shuffled sequences that had identical length and base composition as the mature miRNA sequences. There were four different ways of generating randomly shuffled sequences to evaluate the sequence specific recognition between miRNA and their targets, i.e. *monoshuffled* and *zeroshuffled* (based on the mononucleotide distribution of the miRNA sequences), *firstshuffled* and *dishuffled* methods (based on the dinucleotide distribution of the miRNA sequences) [156]:

- *Monoshuffled* sequences: The counts of the mononucleotide of each miRNA sequence were calculated and bases were drawn iteratively and randomly according to the nucleotide composition.

- *Zeroshuffled* sequences: The mononucleotide frequencies $(P(A), P(C), P(G), P(U))$ for each miRNA sequence were calculated and used to generate a random sequence in which bases were chosen at random with probability $P(i)$, where $i = A, C, G, U$, until the length of the miRNA sequence was reached.
- *Firstshuffled* sequences: From each miRNA sequence, the conditional probability $P(a \mid b)$ for nucleotide a given b was calculated from the frequencies of the 16 possible nucleotide pairs. A random sequence was generated by first choosing a random nucleotide x_i using *zeroshuffled* method, then the rest of the nucleotides were generated iteratively based on the conditional probabilities $P(x_{i+1} \mid x_i)$ (first order Markov process), where x_{i+1} could be any of the four nucleotides. The process was stopped when the sequence had reached the same length as the miRNA sequence.
- *Dishuffled* sequence*:* A random trinucleotide was chosen (e.g. AUU) iteratively and all the non-overlapping trinucleotides that began and ended with the same bases (i.e. AAU, ACU, AGU and AUU) were shuffled at random. The whole process was done R times, where R was 10 times of the length of the miRNA sequence.

3.2.2.3 miRNA target motifs detection based on an inhomogeneous HMM

In our HMM model, hidden states were defined over the binary space $\{P, N\}$, where P meant a conserved matching state, namely an endogenous miRNA needs to consistently match to its target on the specific site. And the hidden symbol is regarding to the conservation level of each position, whether it should be consistently matching or mismatching in the miRNA-target duplex. Each conserved matching state could generate A-U, U-A, G-C or C-G as an emission symbol. N meant a nonconserved matching state, namely a miRNA does not need to consistently match to its target on this specific site. Each nonconserved matching state could emit one of the remaining combinations except the aforementioned four symbols (Figure 3.2.1).

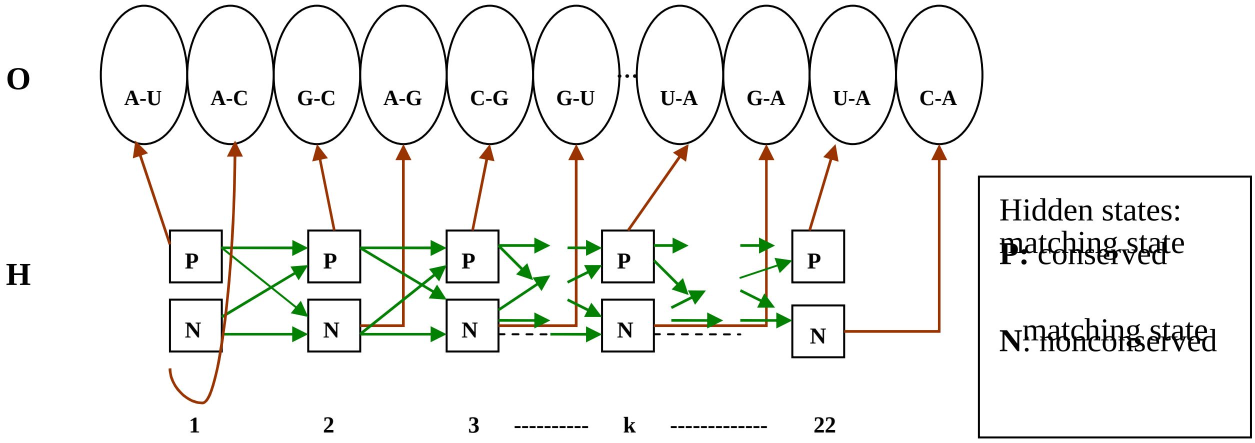


Figure 3.2.1 An exemplar diagram of inhomogeneous HMM.

The position specific transition probabilities and emission probabilities would be estimated using a training set of potential miRNA targets (The transition probabilities and emission probabilities shown in the diagram were arbitrarily assigned). O: Observations; H: Hidden States.

Two types of probabilities needed to be estimated: transition probabilities and emission probabilities. These probabilities were position specific in the inhomogeneous HMM. The parameters were estimated from a training set of the potential targets with up to 4 mismatches to one of the 19 miRNAs. Baum-Welch algorithm was used to update the parameters in the model until it reached (local) maximal log likelihood [139,140,157,158].

The Baum-Welch algorithm calculated A_{ij}^t and $E_j^t(k)$ as the expected counts of each transition or emission at position t for given training sequences [139]:

$$A_{ij}^t = \sum_s \frac{1}{P(x^s)} f_i^s(t) a_{ij}^t e_j^{t+1}(x_{t+1}^s) b_j^s(t+1), \qquad (1)$$

$$E_i^t(k) = \sum_s \frac{1}{P(x^s)} f_i^s(t) b_i^s(t), \qquad (2)$$

$$a_{ij}^t = \frac{A_{ij}^t}{\sum_{j'} A_{ij'}^t}, \qquad (3)$$

$$e_i^t(k) = \frac{E_i^t(k)}{\sum_{k'} E_i^t(k')}, \qquad (4)$$

where a_{ij}^t and $e_j^{t+1}(x_{t+1}^s)$ were the position specific transition probability and emission probability, respectively, and the values of the two parameters were updated iteratively in terms of (3) and (4). The *s*-th observation was denoted by x^s ($s = 1,2,...$) and x_{t+1}^s is the symbol observed at position *t+1*. The forward variable and the backward variable for the *s*-th observation were given by $f_i^s(t)$ and $b_j^s(t)$, respectively. All possible sequence patterns that found by the direct search approach were used as members of the training set of the inhomogeneous

HMM. Convergence of the negative log likelihood was checked up to a precision of 1e-12.

We used I to denote a transition path of the hidden states. For each training sequence, the optimal path $I*$ with the highest probability should be chosen as follows:

$$I* = \arg\max_{I} P(x, I), \qquad (5)$$

where $P(x, I)$ was the probability of an observation x with a state transition path I . The Viterbi algorithm was used to find the most probable (optimal) state transition path in the HMM [139,141]:

$$v_j(t+1) = e_j^t(x_{t+1}) \max_i (v_i(t) a_{ij}^t), \qquad (6)$$

where $v_i(t)$ was the probability of the most probable path ending in state i at position t . We got 103 optimal paths in total after removing the redundant ones. The experimentally verified miRNAs and the optimal state paths obtained above were then used to scan for miRNA target motifs in the *Arabidopsis* genome.

The inhomogeneous HMM was implemented as a Perl script and a genome-scale scanning for miRNA targets took about 10 hrs on a UNIX workstation with 2GHz processor and 2G memory.

3.3 Results

3.3.1 Training set preparation

When the maximum number of mismatches tolerated was set to 4, the direct search approach detected 215 genes whose mRNAs had the complementary site with at least one of the 19 miRNAs. Table 3-1 listed the 215 potential targets that had less than or equal to 4 mismatches (Table 3-1), and 36 of the 215 potential targets had been experimentally validated [124,150,152,153,159-166].

Table 3-1 miRNA potential targets detected by the direct search approach

miRNA	Targets Protein Family	Target Genes (Number of Mismatches)
miR156	SBP (squamosa promoter binding protein)	*At1g27360/SPL11(1), At1g27370/SPL10(1)*[a]*, At1g69170/SPL6(1), At2g42200/SPL9(1), At3g47170(3), At3g57920/SPL9(1), At5g43270/SPL2(1)*[a]*, At5g50570/SPL9(1)* [a]*, At5g50670/SPL9(1)* [a]*, At2g33810/SPL3(2)*[a]*, At1g53160/SPL4(2)*[a]*, At3g15270/SPL5(3)*[a]*, At3g28690(3)*
	F-box family protein	*At1g22000(4), At3g17480(4), At3g58860(4)*
	Expressed proteins	*At3g47170(3), At1g30240(4), At1g48430(4), At1g53240(4), At1g71400(4), At2g21840(4), At3g20840(4), At3g46280(4) , At4g25440(4), At4g27470(4), At4g35170(4), At4g35620(4), At5g11380(4)*
miR157	SBP (squamosa promoter binding protein)	*At1g27360/SPL11(1), At1g27370/SPL10(1)* [a]*, At1g69170/SPL6(2), At2g42200/SPL9(1), At3g57920/SPL9(1), At5g43270/SPL2(1)* [a]*, At5g50570/SPL9(2)* [a]*, At5g 50670/SPL9(2)* [a]*, At1g53160/SPL4(3)*[a]
	F-box family protein	*At1g22000(3), At1g32140(4), At3g58860(4)*
	Expressed proteins	*At3g47170(3), At1g09170(4), At1g30450(4), At5g08620(3), At1g48090(4), At2g45990(4), At3g07160(4), At3g15950(4), At5g18590(4), At5g63060(4)*
miR158	PPR (pentatricopeptide) repeat containing protein	*At1g64100/PPR(2), At3g03580/PPR(3), At1g03540(4), At2g17525(4), At3g15130(4), At4g32430(4)*
	WD-40 repeat family protein	*At1g49910(3), At3g19590(4)*
	Fucosyltransferase	*At2g03210/ FUT2(3), At2g03220/FUT1(3), At1g14070/FUT7(4)*

	Expressed proteins	*At1g09040(3), At1g63770(3), At3g07400(3), At1g04190(4), At1g09050(4), At1g11480(4), At1g23070(4), At1g48460(4), At1g55050(4), At2g27240(4), At2g31620(4), At2g41210(4), At2g45810(4), At3g01590(4), At3g23310(4), At3g28890(4), At3g56620(4), At3g60400(4), At4g17565(4), At4g27180(4), At5g07270(4), At5g07720(4), At5g16910(4), At5g17240(4), At5g23870(4), At5g52920(4), At5g54180(4), At5g58510(4), At5g59490(4), At5g61010(4), At5g64310(4)*
miR159	MYB family transcription factors	*At2g26950/AtMYB104(3), At2g32460/AtMYB101(2), At3g11440/AtMYB65(3)*[a]*, At3g60460(3), At5g06100/AtMYB33(3)*[a]*, At4g26930/AtMYB97(4), At2g26960/AtMYB81(4), At5g55020/AtMYB120(4)*
	Expressed protein	*At1g29010(3), At3g53570(4), At4g37770(4)*[a]*, At5g04020(4)*
miR160	Auxin Response Factors	*At2g28350/ARF10(2)*[a]*, At4g30080/ARF16(3), At1g77850/ARF(1)*[a]
miR161	PPR repeat containing proteins	*At1g63080(3), At1g63400(3), At5g41170(3), At1g63150(3), At5g16640(3), At1g62720(3), At1g64580(3), At1g06580(3), At1g62670(3), At1g08610(4), At1g10270(4), At1g12700(4), At1g62590(4), At1g62860(4), At1g62910(4), At1g62930(4), At1g63070(4), At1g63130(4), At1g63230(4), At1g63330(4), At1g63630(4), At2g16880(4), At3g18020(4), At4g26800(4), At5g65560(4)*
	Expressed protein	*At4g17910(4), At5g27400(4)*
miR162	Expressed protein	*At1g03080(4), At1g48430(4), At3g20260(4), At3g21140(4), At3g50530(4), At3g54010(4), At5g55330(4)*
miR163	F-box family protein	*At1g64840(4)*
miR164	NAC domain proteins	*At1g56010/NAC1(2)*[a]*, At3g15170/CUC1(3)*[a]*, At5g07680(2)*[a]*, At5g53950/CUC2(3)*[a]*, At5g61430(2)*[a]*, At5g39610(4)*[a]
	Expressed protein	*At1g10530(4), At1g77770(4), At4g01210(4), At4g27520(4)*
miR165	HD-Zip transcription factors	*At1g30490/PHV(3)*[a]*, At2g34710/ATHB-14(3)*[a]*, At4g32880/ATHB-8(3), At5g60690/REC(3)*[a]*, At1g52150(4)*
	Expressed protein	*At1g32750(4), At5g24150(4)*
miR166	HD-Zip transcription factor	*At1g52150/ATHB-8(3), At1g30490/HB-9(4)*[a]*, At2g34710/HB-14(4)*[a]*, At4g32880/HB-8(4), At5g60690(4)*[a]
	Expressed protein	*At1g76590(4), At2g39415(4), At4g22620(4), At5g24150(4), At5g49250(4), At5g54390(4)*
miR167	Auxin Response Factor	*At5g37020/ARF8(3)*[a]*, At1g30330/ARF6(4)*
	Expressed proteins	*At1g67080(4), At2g38920(4), At3g06060(4), At5g64830(4)*
miR168	ARGONAUTE protein	*At1g48410/AG01(3)*[a]

	RNA helicase / RNAseIII	*At3g20420(4)*
	Expressed protein	*At3g58030(4), At2g25070(4), At4g15420(4)*
miR169	Expressed protein	*At1g50840(4), At2g17930(4)*
miR170	Scarecrow transcription factor family protein	*At2g45160(2), At3g60630/SCL6(2)* [a], *At4g00150/SCL6(2)* [a]
	F-box family protein	*At1g59675(4), At2g44700(4)*
	Expressed protein	*At1g27710(4), At3g27470(4), At3g55410(4)*
miR171	Scarecrow transcription factor	*At2g45160(0), At3g60630/SCL6(0)* [a], *At4g00150/SCL6(0)* [a]
	Expressed protein	*At2g22030(4), At4g04850(4)*
miR172	AP2 domain containing transcription factors	*At4g36920(2)* [a], *At5g12900(3), At5g60120(1)* [a], *At5g67180(3)* [a], *At2g28550(3)* [a] , *At2g35130(4)*
	MYB family transcription factor	*At1g09710(4), At5g65790(4)*
	Expressed protein	*At1g15960(4), At1g05805(4), At1g17030(4),At1g21060(4),At1g30920(4), At1g32340(4), At1g72980(4), At2g37590(4), At2g39110(4), At2g45720(4), At3g11570(4), At3g47360(4), At3g47670(4), At4g04650(4), At4g23950(4), At5g19560(4), At5g60310(4)*
miR173	PPR repeat-containing protein	*At1g12300(4), At1g12770(4), At3g16710(4)*
miR319	MYB family transcription factor	*At2g26950(2), At3g11440/MYB65(2)* [a], *At5g06100/MYB33(2)* [a], *At1g52000(4), At2g26960(4), At3g60460(4), At5g55020(4)*
	TCP family transcription factor	*At1g30210(4)* [a], *At1g53230(4)* [a], *At2g31070(4)* [a], *At3g15030(4)* [a], *At4g18390(4)* [a]
	Expressed protein	*At1g14200(3), At3g06450(3), At1g27800(4), At1g34720(4), At1g48090(4), At1g50610(4), At1g74200(4), At2g07787(4), At2g16290(4), At3g10980(4), At3g21140(4), At3g24370(4), At3g25720(4), At3g66658(4), At4g19860(4), At4g39850(4), At5g18100(4), At5g19260(4), At5g19790(4), At5g36840(4), At5g67090(4)*

[a] Experimentally validated targets.

Most of the miRNAs are found to be complementary to more than one mRNA and some of the potential targets are members from the same family. Many of the predicted instance, there were 16 SBP (Squamosa promoter binding protein) reported in *Arabidopsis*, 11 of them had miR156 complementary sites. They were *At1g27360 (SPL11), At1g27370 (SPL10), At1g69170 (SPL6), At2g42200 (SPL9), At3g57920 (SPL9), At5g43270 (SPL2), At5g50570 (SPL9), At5g50670 (SPL9), At2g33810 (SPL3), At1g53160 (SPL4)*, and *At3g15270 (SPL5)*. Other transcription factor families such as MYB, HD-Zip and TCP also had members to be predicted as potential miRNA targets.

Most predicted targets for these 19 miRNAs had function annotation involved in the plant development. For example, miR160 targeted auxin response factors *At2g28350 (ARF10), At4g30080 (ARF16)* and *At1g77850 (ARF1)*. Auxin response factors (ARFs) are family of transcription factors that bound to TGTCTC auxin response elements in promoters of early auxin response genes and mediated gene expression in response to auxin [167,168]. Genes known to encode F-box family proteins, such as *At1g22000*, *At1g32140*, *At3g28860*, *At1g59675*, and *At2g44700*, were also predicted to be potentially targeted by miRNAs. F-box proteins regulate diverse cellular processes, including cell cycle transition, transcriptional regulation and signal transduction [169]. Lu et al. [170] identified that miR774 targets the mRNA of at least on F-box protein. Till now, seven F-box mRNAs have been identified to be targeted by miRNAs, suggesting that the protein degradation machinery is subject to considering miRNA regulation.

miR164 was predicted to target NO APICAL MERISTEM (NAM) family proteins, i.e. *At1g56010 (NAC1), At3g15170 (CUC1)* and *At5g53950 (CUC2),* which are required for apical meristem formation [122,171]. HD-Zip transcription factor *At1g30490* encodes PHAVOLUTA (PHV), which regulates axillary meristem initiation and leaf development [122,172]. Moreover, one of the experimentally validated targets, *At5g06100,* encoding AtMYB33, was reported to bind to the promoter of the floral meristem-identity gene *LEAFY* [122,173]. Some target gene families have the function in DNA/RNA binding, such as PPR repeat containing proteins. PPR proteins were reported to be sequence-specific RNA- or DNA-binding proteins and play constitutive roles in mitochondria and chloroplasts, probably via binding to organellar transcripts [174]. TCP transcription factors were also reported to implicate in processes related to cell proliferation, which are recruited during evolution to control cell division and growth in various developmental processes [175-178]. APETALA2 (AP2) domain containing transcription factors play important role in the control of *Arabidopsis* flower and seed development [179]. miR156 and miR157 both target SBP which regulates the *Antirrhinum* floral meristem-identity gene SQUAMOSA [122,180]. The mRNA complementary sites for miR161, miR165, miR170 and miR171 are within the conserved domain among family members. And there is evidence that miRNAs regulate homologous mRNAs in basal plants with various reproductive structures and leaf morphology, which leads to the speculation that miRNAs are parts of ancient, conserved regulatory modules underlying developmental outcomes [150,181]. These two facts suggested that some miRNAs might mediate the formation of the conserved domain which defined particular protein families. For other miRNAs, namely miR156, miR160, miR164 and miR169, the complementary sites were not within the conserved domain among family members. For example, two of the 11 SBP proteins that were found to be targeted by miR156, namely *At2g33810* and *At1g53160*, were located on the 3' untranslated regions (3' UTR), whereas the other 9 were located on coding regions.

3.3.2 Simulation study using randomly shuffled sequences

The simulation study was applied to test whether an algorithm (method) could distinguish a miRNA from its shuffled version during the detecting process. If the high level of complementarity was due to the miRNA functional requirement, the complementary sites obtained for a randomly shuffled sequence should be substantially less than the miRNA target motifs obtained with the same parameters, for the pattern (miRNA−target) recognized to induce downstream function (e.g. posttranscriptional gene silencing) would be destroyed by the randomization. In contrast, if the number of complementary sites was only due to the base composition of miRNAs, there would be little difference in the results for miRNAs and their shuffled sequences. Herein four kinds of randomly shuffled sequences were generated, i.e. *monoshuffled*, *zeroshuffled*, *firstshuffled* and *dishuffled* sequences.

The *monoshuffled* method generated a truly permuted random sequence while the *dishuffled* further made the count of each dinucleotide the same as that of miRNAs. Although the later kept more information, it was less general, since the intention to maintain the dinucleotide count would result in relatively fewer kinds of permutations allowed. The other two randomization methods, namely *zeroshuffled* and *firstshuffled*, maintained the mononucleotide and dinucleotide distribution, respectively, instead of the exact counts, while the exact count of each mononucleotide and dinucleotide would fluctuate around that in miRNA [156].

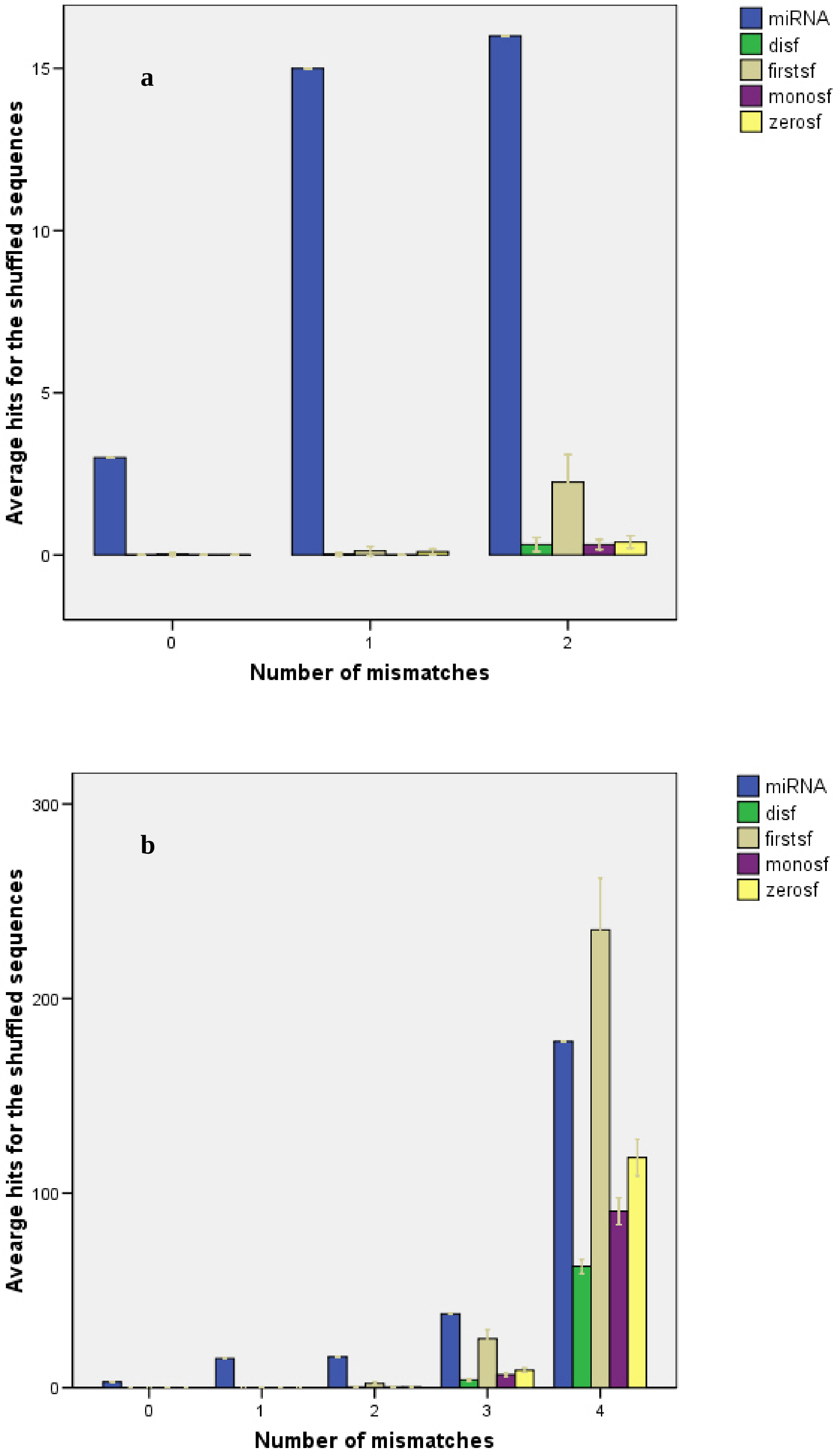

a
Average hits for the shuffled sequences
0
5
10
15
0
1
2
Number of mismatches
miRNA
disf
firstsf
monosf
zerosf
b
Avearge hits for the shuffled sequences
0
100
200
300
0
1
2
3
4
Number of mismatches
miRNA
disf
firstsf
monosf
zerosf

Figure 3.3.1 Signal to noise ratio I

The number of potential targets detected by the direct search approach for the miRNAs compared to that for 50 cohorts of randomly shuffled sequences. The blue bars represented the number of complementary sites for 19 miRNAs. The average number of complementary sites from *dishuffled*, *firstshuffled*, *monoshuffled* and *zeroshuffled* methods was represented by the green, khaki, purple and yellow bars, respectively. The error bars represented 2 standard error (SE).

miRNAs had significantly more complementary sites than that of randomly shuffled sequences (Figure 3.3.1a) when relatively fewer mismatches were allowed, i.e. $N \le 2$. If no mismatches were allowed, the complementary sites for the randomly shuffled sequences from *dishuffled*, *firstshuffled*, *monoshuffled* and *zeroshuffled* methods were 0, 1, 0, and 0, respectively. So for the 50 cohorts, complementary site had only been found once for one of the *firstshuffled* sequence, whereas three perfectly complementary sites had been detected for miR171. However, the high signal to noise ratio could only be observed when two or fewer mismatches were allowed, namely $150:1\ (N=1)$ and $8:1\ (N=2)$. In view of the low probability that so many complementary sites occurred by chance, we suggested that the complementarity between miRNA-mRNA reflected a functional requirement, thus these protein coding genes were quite likely to be regulatory targets of miRNAs. When more mismatches were allowed $(N=3,4)$, the signal to noise ratio decreased (Figure 3.3.1b). Since three (that from *dishuffled*, *monoshuffled* and *zeroshuffled* randomization methods) of the four signal to noise ratios remained substantially high when N was 4, we used 4 as the maximal allowed number of mismatches in our direct search approach and we chose the 215 genes (listed in Table 3-1) as the training set for the inhomogeneous HHM. The differences among the four different shuffled methods will be further discussed below (see Discussion).

When we compared the genes complementary either to miRNAs or to randomly shuffled sequences, we found that the functional annotations for those genes complementary to randomly shuffled sequences were lack of enrichment. In contrast, target candidates for miRNAs had enriched functional annotations as members of some protein families, most of which belonged to transcription factor families (Table 3-1). For example, the three genes, i.e. *At2g45160*, *At3g60630* and *At4g00150*, that perfectly complementary to miR171 all belonged to the scarecrow transcription factor family. When N was set to 4, the fact that miRNAs tended to target transcription factors and other regulatory genes belonged to the same protein family was still clear, whereas the function annotation of genes having complementary sites with randomly shuffled sequences became quite diverse.

3.3.3 Detecting miRNA target motifs using inhomogeneous HMM

Various algorithms have been developed to predict plant miRNA targets based on the miRNA-mRNA complementarity and most of the algorithms predicted targets through detecting mRNA sequences that had up to 4 mismatches to miRNA sequences [124,152]. However, there do exist natural miRNA targets with more than 4 mismatches [153], such that they were not able to be found by these algorithms. Moreover, we suggested that sequences with the same number of mismatches might not have the same possibility to be targeted cleavage by miRNAs owing to the mechanism of RISC. In several cases, particular miRNA-target mismatches are conserved through the evolutionary distance that separated *Arabidopsis* and rice, suggesting that certain mismatches might be under positive selective pressure rather than merely be tolerated [122,182]. Furthermore, properly placed mismatches might improve the enzyme turnover rate [75,182]. Schwab suggested that the presence of $G:U$ plays only a minor role in plant miRNA-target interaction compared to other mismatches, so we regarded the $G:U$ pair as mismatch [153].

We chose inhomogeneous HMM because of its capability of capturing the position specific information about particular matches/mismatches. In spite of the diverse miRNA sequences, the complementarity between miRNA-target duplex might follow some rules according to the RISC mechanisms, and we believed that the inhomogeneous HMM could find these hidden rules by learning from a training set of potential miRNA targets obtained for only 19 mature miRNAs contained in miRBase 3.0, a three years old release, and in this way we also assess the ability of our method to extrapolate from a limited prior knowledge [183]. To obtain the training set, we set the maximum number of mismatches tolerated at 4, and the direct search approach detected 215 genes whose mRNAs had the complementary site with at least one of the 19 miRNAs. The inhomogeneous HMM was trained using the Baum-Welch algorithm based on EM (expectation maximization). The optimal state chain of each miRNA-mRNA pair was computed using Viterbi algorithm, which represented one possible miRNA-target duplex that could be recognized by RISC and cleaved by its Argonaute component.

In total 103 non–redundant optimal state chains were produced by using Viterbi algorithm. Most of the resulted 103 optimal chains are not limited to 4 mismatches, which is consistent with our aim of developing the inhomogeneous HMM. Most predicting algorithms for miRNA targets are limited to potential targets with up to 4 mismatches, but there do exist targets with more than 5 mismatches [153].

After scanning the genome, we found about 160,000 potential miRNA target motifs. This result covered almost all the experimentally validated miRNA targets (90/91) in *Arabidopsis* [124,150,152,153,159-166]. The majority of the 91 experimentally validated miRNA targets (58/91) were the targets for those miRNAs that were not included in the training set, namely miR393 to miR870 [124,153,160-163]. Our results substantiated the notion that the inhomogeneous HMM could enhance the power of capturing the information about position-specific mismatches/matches between miRNA-target duplex.

3.4 Discussion

3.4.1 Four methods used to generate randomly shuffled sequences

We used four different ways to generate randomly shuffled sequences, namely *monoshuffled*, *zeroshuffled*, *dishuffled* and *firstshuffled* methods [68,111,156,184]. As aforementioned, these methods preserved four different kinds of sequence specific characteristics of each miRNA, say mononucleotide count (*monoshuffled*), dinucleotide count (*dishuffled*), mononucleotide distribution (*zeroshuffled*) and dinucleotide distribution (*firstshuffled*). And the results from these were not identical. Among the four methods, the shuffled sequences generated by the *firstshuffled* always had more complementary sites than any of the other shuffled methods. Why were there substantially more complementary sites for *firstshuffled* sequences, even more than that for miRNAs, when N was 4? It reminded us that Workman and Krogh reported that when the folding free energies of mRNA and tRNA were compared to that of randomly shuffled sequences, there was no significant difference between them if the dinucleotide distribution of RNA was preserved. But this was not true for the randomly generated sequences with the same mononucleotide distribution, which suggested that dinucleotide composition of RNA sequences played a more important role in the RNA stability [156,184]. From Figure 3.3.1, we could conclude that the specific mononucleotide distribution of miRNAs would not result in many false positives. Another finding was that the *firstshuffled* method was more sensitive than the other three methods in capturing the sequence requirement for both RNA stability and recognition. In our study, the randomly shuffled sequences generated by *firstshuffled* always had more complementary sites, and Bonnet et al. reported that the folding free energy of sequences obtained by this method was always smaller than that of sequences from other methods [184]. Bonnet suggested the former might be due to the fluctuations in the energies of the Markov sequences [156,184]. Although the *dishuffled* method maintained even more information than *firstshuffled* method, namely both dinucleotide distributions as well as dinucleotide count, the complementary sites for this kind of shuffled sequences were less than that from *firstshuffled* method. Based on the method itself, the shuffled sequences might contain similar pattern, such as AAAAUUUU, which would not change any more in the further shuffling process.

The aforementioned methods were also applied to test if there were substantially more complementary sites for miRNAs than that for randomly shuffled sequences, based on the 103 optimal state chains obtained from HMM.

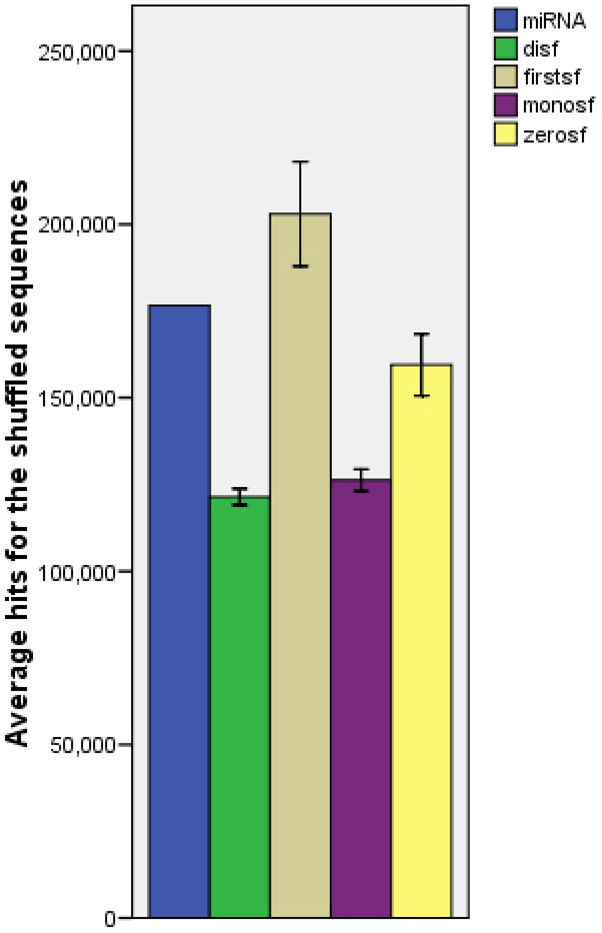


Figure 3.4.1 Signal to noise ratio II.

The number of potential targets detected by inhomogeneous HMM compared to that of 10 cohorts of randomly shuffled sequences. The blue bar represented the number of complementary sites for 123 miRNAs. The average number of complementary sites from *dishuffled*, *firstshuffled*, *monoshuffled* and *zeroshuffled* methods was represented by the green, khaki, purple and yellow bars, respectively. The error bars represented 2 standard error (SE).

The signal to noise ratio observed for inhomogeneous HMM was similar to that of the direct search approach $(N = 4)$, i.e. the signal was greater than the noise from *monoshuffled*, *zeroshuffled*, and *dishuffled* methods, whereas less than the noise from *firstshuffled* method (Figure 3.4.1).

By using the inhomogeneous HMM, we found that the average number of potential targets for each miRNA increased by two orders of magnitude, and the complementary sites for miRNAs were still significantly more than that for three kinds of shuffled sequences. The results suggested that the inhomogeneous HMM could dramatically increase the number of miRNA target motifs while retained the comparable specificity as that of the direct search approach ($N = 4$).

3.4.2 miRNA targets prediction without conservation requirement

Most miRNA target prediction methods applied an additional conservation constraint to each target candidate in order to increase the specificity, at the cost of discarding species-specific candidates [124,164,185,186]. We did not include this constraint in our miRNA target prediction algorithm so that our study was not merely on conserved miRNA targets. The number of predicted target sites for miRNAs increased dramatically when the conservation constraint was not used. However, there is increasing evidence that many of these nonconserved target sites may be indeed functional [187,188]. For instance, 30% to 50% nonconserved miRNA binding sites in the human genome might be functional when the miRNA and mRNA are expressed in the same tissue [189,190].

3.4.3 miRNAs are biased toward target TFs and other regulatory genes

Noncoding RNAs can specifically recognize another nucleic acid [191]. One of the main advantage of RNA as regulatory molecule is its compact size and sequence specificity. Moreover, RNA duplex can allow for stable mismatches and form particular structures [192]. Many miRNAs, such as miR156, miR159, miR160, miR164, miR166, miR172 and miR319, are reported to target TFs and play important functional roles in cell differentiation [150]. TFs work as “on” or “off” switches [193]. As cell differentiates, it needs to switch from one set of TFs to another. If it has to merely rely on the promoter binding proteins to inhibit DNA transcription of the earlier TFs, those TFs would not be turned off until all the TF transcripts are degenerated; on the contrary, a miRNA may function during plant differentiation to clear regulatory gene transcripts thereby facilitating more rapid and robust transitions to new expression programs [111,122,150]. Furthermore, TFs and miRNAs can act cooperatively on their targets in a largely combinatorial manner, that is, many different TFs or miRNAs control a particular gene [193].

3.4.4 Posttranscriptional regulation of gene expression by miRNA

Although there are substantial differences between the animal and plant kingdoms in regards to the mechanisms and scopes of miRNA-mediated gene regulation, the new discovery from recent studies on animal miRNAs can still help to shed light on the future plant miRNA studies. Lall et al. predicted that miRNAs regulate at least 10% of nematode genes [119]. The algorithms based on the near perfect matching to the 5’ miRNA sequence to predict targets found that more than one third of human genes might be under miRNA regulation [62,63,119,121]. Furthermore, the difference in the miRNA-mediated gene regulation between animal and plant miRNAs, namely mRNA cleavage by plant miRNA and translational repression by animal miRNA, may not always be compelling. For example, Lim et al. showed that animal miRNAs reduced the transcript abundance of a large number of genes with limited sequence complementarity [149]. And there was increasing evidence that miRNA could directly induce mostly weak but significant negative effects on the steady-state mRNA levels of their targets [194]. Sood et al. found that the mRNA levels of nucleus 3’ UTRs were significantly lower in the tissue of cognate miRNA expression compared with a background set simply comprising all genes [194]. A quite interesting finding was that even the mere presence of the central recognition motifs, referred to as “nucleus” or “seed” sequence [61,62,121,195] for each miRNA in human 3’ UTRs which was typically a few thousand, without any cross-species analysis, was sufficient for observing expression changes in mRNAs [194]. If this is also true for plant miRNA targets, that a limited complementarity between miRNA-mRNA duplex is enough to result in the observed reduction in the steady-state mRNA level, then current prediction algorithm relied on the near perfect complementarity and cross-species comparison almost certainly underestimated the number of genes under plant miRNA regulation, which was suggested to comprise less than 1% of protein-coding genes in *Arabidopsis* [150].

Chapter 4: Reconstructing Regulatory Networks

We conducted a systematic study on the transcriptional and posttranscriptional regulatory role at sequence level in *Arabidopsis*. Both miRNA target motifs (miRNA-mediated posttranscriptional regulatory sites) and TFBS (transcription factor binding motifs) were incorporated with microarray time-course gene expression profiles to determine their probabilistic dependences. We could correctly predict expression patterns for more than 50% of 1,132 genes, which was statistically significant, based solely on the sequence motifs adopted in the network model.

4.1 Introduction

Owing to the complete sequencing of a large number of genomes and the growing amount of high-throughput gene expression data, a comprehensive understanding of the regulatory mechanisms of gene expression becomes the next important issue of genomics [138,196]. It has been generally assumed that genes responding to a common environmental challenge should be co-regulated and show similar patterns of expression [79]. General components of regulatory networks are the genes involved in a specific system and the transcription factors (TFs) that regulate the system. DNA microarrays provide rapid and parallel surveys of gene-expression profiles for hundreds or thousands of genes in a single assay.

When expression data are not enough to accurately reconstruct these networks, a possible intermediate solution is to construct networks from gene modules – sets of genes that are assumed to share a common function or be involved in the same pathway [97], and these co-regulated genes are believed to be mediated by short DNA elements called regulatory sequences, which include TF binding sites [103]. A large number of experimental and computational studies have been done on locating transcriptional regulator binding DNA sequences and understanding their functions [9-11]. TFs regulate gene expression by binding selectively to sequence sites in promoters of genes, and genes regulated by the same TFs have been assumed to share the common binding sites in their promoter regions and exhibit similar expression profiles [12]. These binding motifs can be used as building blocks of large networks and several approaches were developed to identify how the set of *cis*-regulatory elements in a gene's promoter region governed its behavior and explained the observed expression profiles [6-8,10,197]. Kim et al. reported a Z-score based method that combined gene expression data analysis with promoter region sequence analysis to infer transcription regulatory elements of human genes [196]. Using Adaboost algorithm, Kundaje et al. learned a decision rule for predicting whether a gene was up- or down-regulated in a particular microarray experiment based on the presence of specific motifs and the expression levels of TFs [198]. Using different approaches, Segal et al. and Beer and Tavazoie both showed that a substantial fraction of yeast gene expression profiles could be explained in terms of the combination of *cis*-regulatory elements [6,7]. The combinatorial code underlying gene expression is composed of logic gates (OR, AND, NOT) and spatial configurations [7,8,197]. However, these approaches have not been broadly applied in multicellular organisms in spite of the reported success in model organisms. A key limitation of such approaches is that many regulators are regulated posttranscriptionally [8]. While progresses have been made in mapping transcriptional regulatory networks, posttranscriptional regulatory roles just begin to be uncovered. Posttranscriptional regulatory mechanism had been reported to occur through the binding of miRNAs to their targets [76,145,146]. However, the role of *Arabidopsis* miRNA in network topology and dynamics remains unexplored [199]. To address this need, we developed a combinatorial approach to determine the transcriptional and posttranscriptional regulatory elements based on gene expression profiles. We applied this approach to a *CONSTITUTIVE PHOTOMORPHOGENIC1* (*COP1*) mutant time course microarray dataset kindly provided by Dr. Deng Xingwang's lab in Yale Department of Biology to detect sequence elements that selectively bind to TFs and miRNAs in the process. Inspired by Beer and Tavazoie [7], we used Bayesian network -- a probabilistic model that integrated both the gene expression data and transcription factor binding sites (TFBS) as well as miRNA target motifs (discussed in Chapter 3) to deduce the combination of sequence elements that modulate gene expression, and we tried to explain the observed gene expression patterns in terms of the transcriptional and posttranscriptional regulatory motifs [128-130,133]. Firstly, genes in the *cop1* mutant time course microarray dataset were clustered into 12 expression patterns and overrepresented sequence elements in the upstream of the genes belonged to the same cluster were detected using AlignACE [110]. Secondly, Bayesian network strategy was applied to selecting these motifs in both upstream sequences and transcript sequences that were most related to the gene expression patterns. Lastly, we measured the degree to which gene expression patterns could be determined merely by these adopted regulatory motifs. Figure 4.1.1 illustrated the flow diagram of the approach.

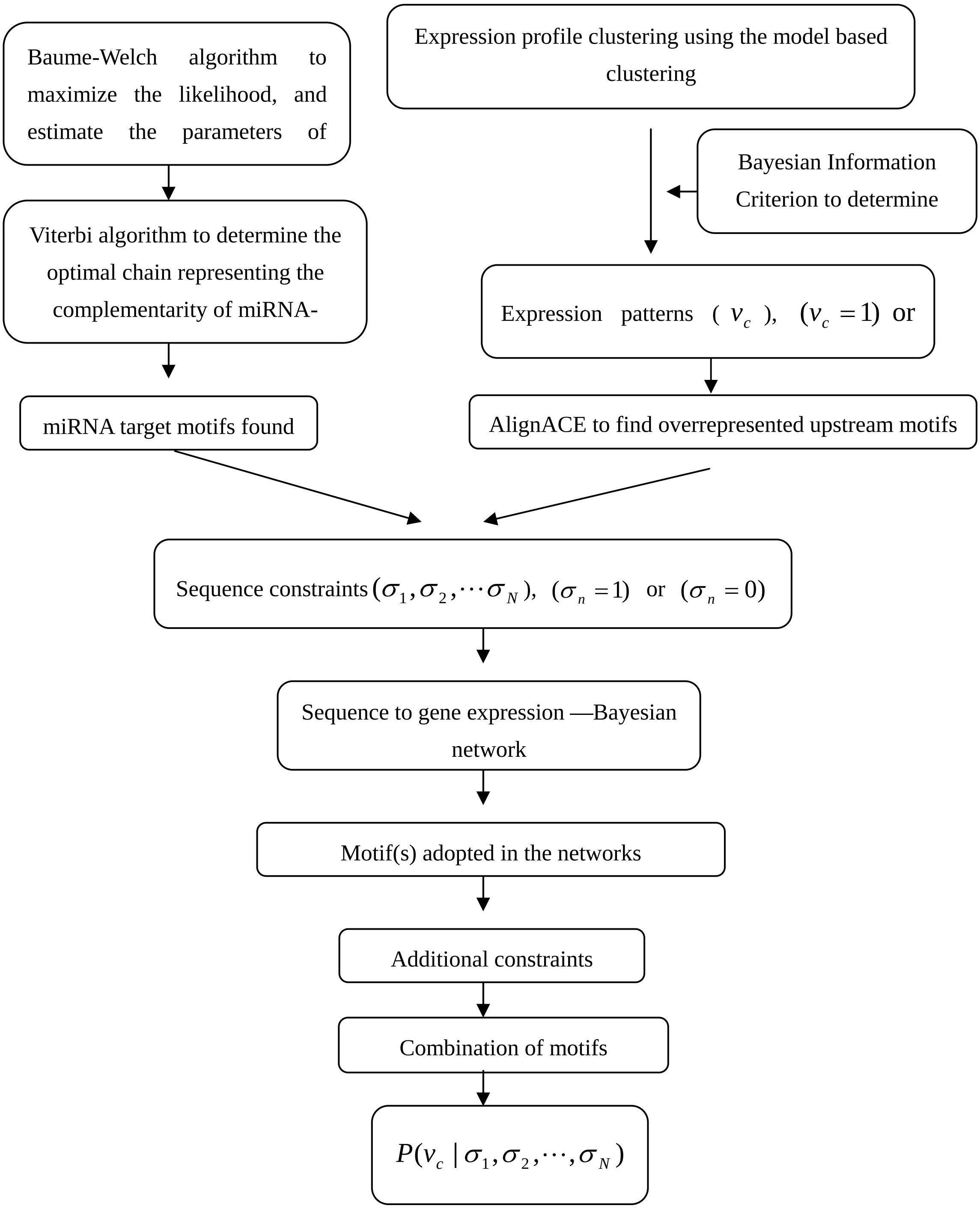

Baume-Welch algorithm to maximize the likelihood, and estimate the parameters of
Viterbi algorithm to determine the optimal chain representing the complementarity of miRNA-
miRNA target motifs found
Expression profile clustering using the model based clustering
Bayesian Information Criterion to determine
Expression patterns (v_c), $(v_c = 1)$ or
AlignACE to find overrepresented upstream motifs
Sequence constraints $(\sigma_1, \sigma_2, \cdots \sigma_N)$, $(\sigma_n = 1)$ or $(\sigma_n = 0)$
Sequence to gene expression —Bayesian network
Motif(s) adopted in the networks
Additional constraints
Combination of motifs
$P(v_c \mid \sigma_1, \sigma_2, \cdots, \sigma_N)$

Figure 4.1.1 Flowchart of the combinatorial approach.

4.2 Materials and methods

4.2.1 Materials

4.2.1.1 Upstream sequences of *Arabidopsis* genes

The entire intergenic region or 3000 bp, whichever was shorter, in the upstream of the transcriptional start site (TSS) for each *Arabidopsis* gene was retrieved from the TAIR (ftp://ftp.arabidopsis.org/seq_analysis_updates/) released in Mar 2006.

4.2.1.2 GO annotation of *Arabidopsis* genes

GOSLIM annotation file of *Arabidopsis* genes was downloaded from the TAIR (ftp://ftp.arabidopsis.org/home/tair/Ontologies/Gene_Ontology/) released in April, 2007.

4.2.1.3 Microarray time-course dataset of *Arabidopsis cop1* mutant

We used an *Arabidopsis cop1* mutant time course microarray dataset kindly provided by Prof Deng Xingwang's lab in Yale Department of Biology. Light is an important environmental signal that governs plant growth and development. One important light-signaling component involved in plant light responses is COP1 (CONSTITUTIVE PHOTOMORPHOGENIC1), which regulates not only photomorphogenesis but also other developmental processes [200]. Both wild type (reference sample) and *cop1* mutant (test sample) were grown at 30 degree for a series of time periods (0, 12, 24 hr …) before transferred to 22 degree.

This experiment was designed to determine COP1 regulated genes by modulating the endogenous activity of COP1. The protocols for hybridization to the *Arabidopsis* microarray, microarray slide washing, and scanning were as described previously in Ma et al. [201]. Microarray spot intensity signals were acquired by using Axon GenePix Pro 3.0 software package (Axon Instruments Inc). All the ratios were the expression intensities of *cop1* mutant divided by that of wild type seedling, respectively. Average normalized log-transformed expression ratios of 5,689 genes were subjected to clustering analysis.

4.2.2 Methods

4.2.2.1 Clustering and motif finding

To take into account the temporal relationship between time-points, a HMM based clustering approach was chosen [90,202,203]. The related software was downloaded from: http://ghmm.org/gql. BIC (Bayesian Information Criterion) was used to determine the 'optimal' number of clusters for the dataset and the 5,689 genes were divided into 12 clusters. AlignACE was then used to detect overrepresented sequence motifs (TFBS candidates) in the 3000 bp upstream of the genes in the same cluster [84,109,110]. The upstream sequences of all the genes were scanned using ScanACE for the motifs found by AlignACE [7,109].

4.2.2.2 Building Bayesian network

We followed the approach established by Beer and Tavazoie [7] and considered two-layer networks with parent nodes representing sequence motifs (TFBSs or miRNA target motifs) and descendent nodes representing gene expression profiles. Edges were directed and connected only sequence elements to expression profiles. The network structure could be described with a 0-1 matrix, with M rows, as many as genes under consideration, and N columns, where N was the number of nodes [138]. The descendent nodes were gene expression pattern v_c, where $c = 1,2,...C$, and C was the total number of clusters (expression patterns). The parent nodes were sequence motifs with specific constraints. The constraint of a sequence element was its presence in the 5’ upstream region of a gene as a TFBS or its presence in the transcript as a miRNA target motif, its orientation, its distance to TSS, and the presence or absence of other TFBSs or miRNA target motifs. If two or more TFBSs or miRNA target motifs were present, the interactive constraints were the distance between them, and/or their order relative to TSS, respectively. Let $\omega = (\sigma_1, \sigma_2, ..., \sigma_N)$ be the sequence constraints. If a constraint n was satisfied for a particular gene, then we have $\sigma_n = 1$, otherwise $\sigma_n = 0$. The final network encoded the distribution of $P(v_c \mid \sigma_1, \sigma_2 ..., \sigma_N)$, the probability of the gene being ($v_c = 1$) or not being ($v_c = 0$) a member of cluster c, given the states of the sequence constraints ω. About 80% of the total genes were used as training set and the rest 20% genes were used as test set [7].

From Bayes' theorem, we had:

$$p(S \mid D) = p(S)p(D \mid S) / p(D),$$

where D was the data and S was the network structure. The probability $p(D)$ did not depend on the structure, and $p(D \mid S)$ was the marginal likelihood. Assuming unrestricted multinomial distribution, parameter independence, Dirichlet priors and complete data, the $p(D \mid S)$ was given by:

$$p(D \mid S) = \prod_{i=1}^{d} \prod_{j=1}^{q_i} \frac{\Gamma(a_{ij})}{\Gamma(a_{ij} + N_{ij})} \cdot \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})},$$

where d, r_i and q_i were the number of descent nodes, the number of unique instantiations for descent node i and the number of the parent nodes of node i, respectively. In our case, $d = 1$ and $r_i = 2$. We used N_{ijk} to denote the number of cases in D in which variable v_c had the value k and its parent was instantiated as j, and $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$. We assumed uniform priors, such that $a_{ijk} = 1$ and $\alpha_{ij} = \sum_{k=1}^{r_i} \alpha_{ijk}$. Parents were added progressively to a node until no additional parent could increase the structure probability [130,143].

4.2.2.3 Predicting gene expression patterns using the Bayesian network model

A model with the highest log marginal likelihood (or the highest posterior probability, assuming equal priors on structure) is the best sequential predictor of the data D. For any given gene, the probability that this gene exhibits the expression pattern c could be calculated by [144]:

$$p(v_c = 1 | D, S_c) = \prod_{i=1}^{d} \frac{\alpha_{ijk} + N_{ijk}}{\alpha_{ij} + N_{ij}}.$$

The algorithms for Bayesian network building and gene expression pattern prediction were implemented as C++ programs and the whole process took about 1 hour on a desktop PC with 1GB memory.

4.2.2.4 Enrichment of functional annotation terms from Gene Ontology

The number of genes in a specific group with the same annotation term from Gene Ontology (GO) was compared to the total number of genes having this GO annotation term in the *Arabidopsis* genome. P-Value which indicated the significance of enrichment was calculated from the hypergeometric distribution:

$$P = \frac{\binom{C}{c}\binom{G-C}{g-c}}{\binom{G}{g}},$$

where C was the number of genes with a particular GO annotation term in the *Arabidopsis* genome, G was the total number of genes in *Arabidopsis* which was 25,676, c was the number of genes in a specific group with the particular GO annotation term and g was the total number of genes in that group.

4.3 Results

4.3.1 Gene expression dynamics in *cop1* mutant time course experiment

In the *cop1* mutant time course experiment, there were in total 10 time points, i.e. 0th hour (0 h), 12th hour (12 h) , 24th hour (24 h), 36th hour (36 h), 48th hour (48 h), 60th hour (60 h), 72nd hour (72 h), 4th day (4 d), 5th day (5 d) and 6th day(6 d). The log expression ratio reflected the difference between the expression level of *cop1* mutant and that of wild-type for each gene, and rapid and transient changes were observed in the log expression ratios of many genes. At each time point, log expression ratios are very diverse (Table 4-1), for example, 1,218 genes and 1,180 genes had the highest and lowest log expression ratios at time point 48 h, respectively. There might be many explanations for the observation, two of which were given as below. One was that transcription factors may switch on/off at a particular time point, which resulted in the up-regulation or down-regulation of their targets. Another possibility was that a certain transcription factor might be turned on at a particular time point, which activated the transcription of certain protein-coding genes as well as miRNA genes through binding to their 5' proximal promoters. The up-regulation of miRNAs would cause the down-regulation of their targets, and this down-regulation could further cascade to the targets of the targets of miRNAs [204].

Table 4-1 Changes in the log expression ratios

Time point	The number of genes with the highest log expression ratios among the 10 time points	The number of genes with the lowest log expression ratios among the 10 time points
0 h	686	649
12 h	585	474
24 h	447	459
36 h	134	403
48 h	1218	1180
60 h	690	1016
72 h	120	88
4 d	335	268
5 d	639	460
6 d	835	692

Maximal log likelihood value obtained by BIC showed that the optimal number of clusters was 12, so we divided the 5,689 genes into 12 clusters using GQLCluster [90,202]. Each cluster contained 755, 157, 400, 509, 275, 638, 725, 374, 658, 422, 186 and 590 genes, respectively.

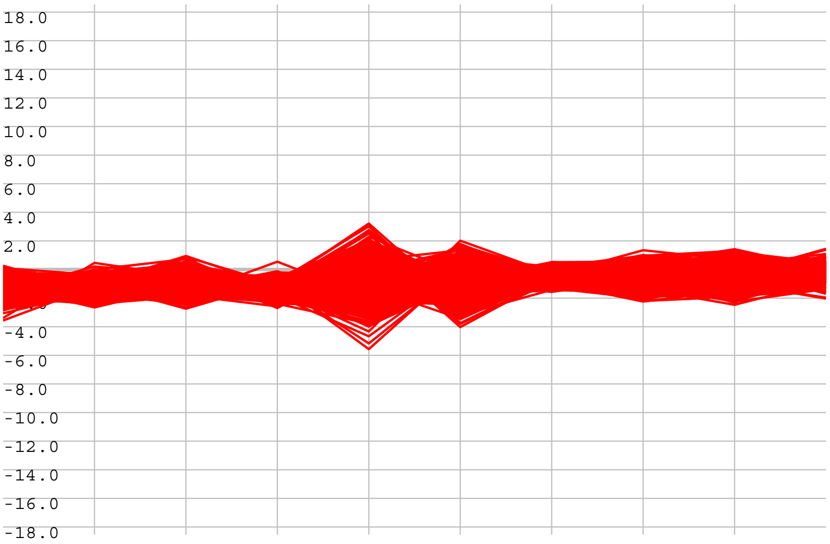


Figure 4.3.1 The expression patterns of the genes in cluster 1 (output of GQLCluster software).

Cluster 1 comprised 755 genes, and there was no enriched functional annotation found in this cluster.

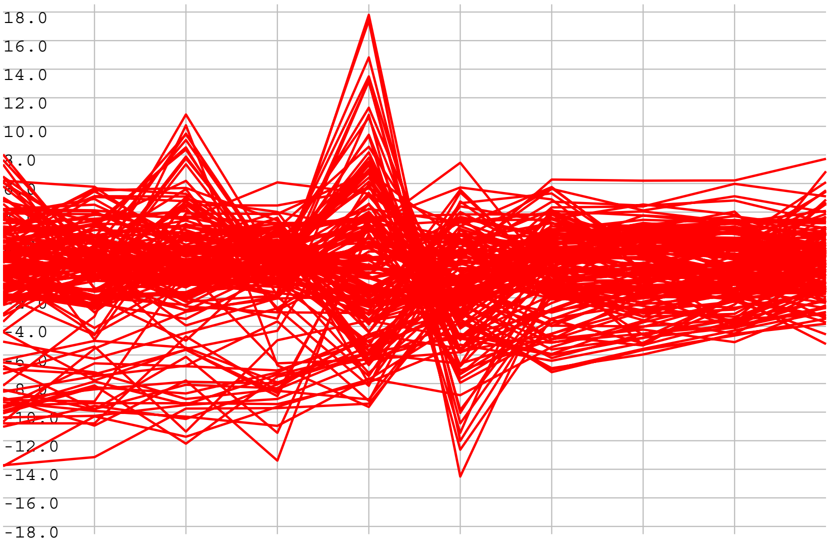


Figure 4.3.2 The expression patterns of the genes in cluster 2 (output of GQLCluster software).

Cluster 2 comprised 157 genes. Many genes in cluster 2 had annotations related to the stress response of plants, such as “response to water” or “response to light stimulus".

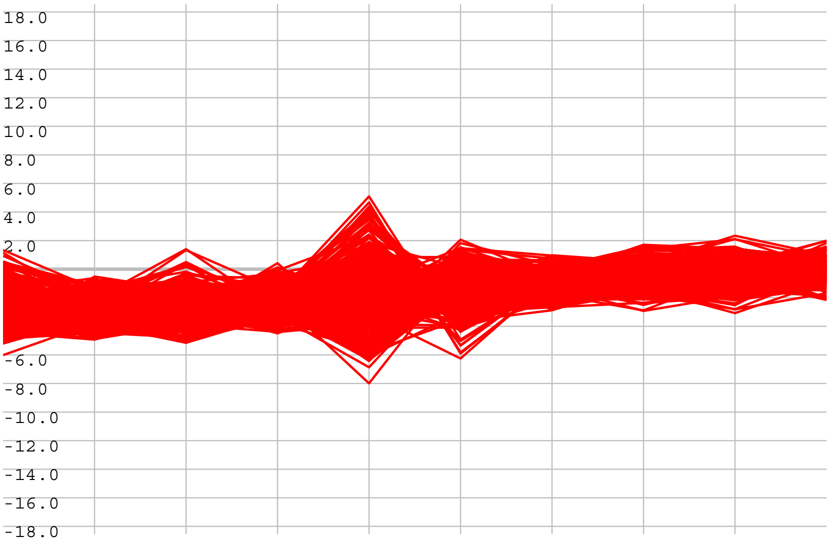


Figure 4.3.3 The expression patterns of the genes in cluster 3 (output of GQLCluster software).

Cluster 3 comprised 400 genes. The average log expression ratio kept increasing after the time point 72 h. There was no enriched functional annotation found in this cluster.

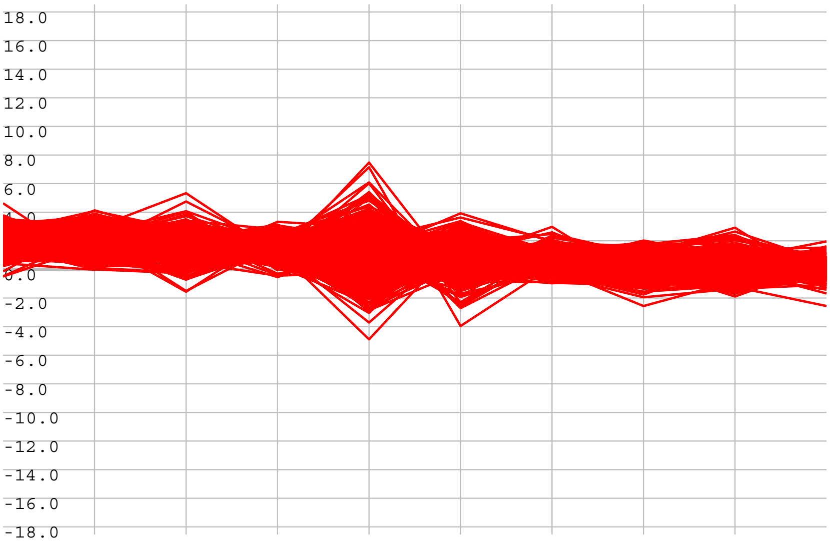


Figure 4.3.4 The expression patterns of the genes in cluster 4 (output of GQLCluster software).

Cluster 4 comprised 509 genes. Different from the expression patterns of the genes in cluster 3, the average log expression ratio kept decreasing after the time point 72 h. Many genes in cluster 4 had GO annotations related to the ribosome function (Table 4-2).

Table 4-2 The number of genes in cluster 4 which had GOSLIM annotation related to ribosome

GOSLIM annotation	The number of genes
RNA binding	9
Structural constituent of ribosome	47
Translation initiation factor activity	10
Ribosome	30
Cytosolic small ribosomal subunit	9
Eukaryotic translation initiation factor 3 complex	2
Eukaryotic translation elongation factor 1 complex	2
Translation	43
Translational initiation	6
Translational elongation	6
Translational termination	1

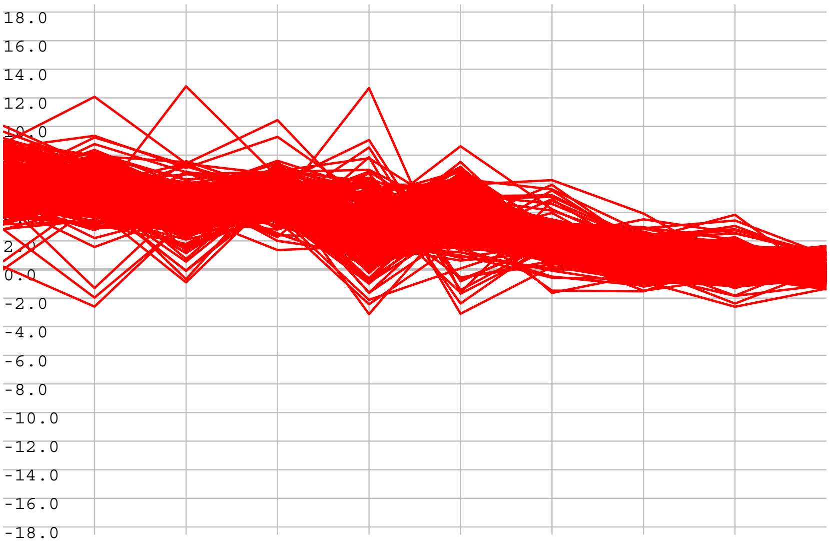


Figure 4.3.5 The expression patterns of the genes in cluster 5 (output of GQLCluster

software).

Cluster 5 comprised 275 genes and the average log expression ratio monotonically decreased along the time course (in total 6 days). Similar to that in cluster 2, some genes in cluster 5 also had annotations related to the stress response of plants, such as “response to jasmonic acid stimulus” or “response to auxin stimulus”. The gene *At5g63110* had the GO annotation term “posttranscriptional gene silencing”, and a dozen of genes had GO annotation terms related to photosynthesis.

Table 4-3 Genes in cluster 5 which had GOSLIM annotation related to photosynthesis

GOSLIM annotation	**Gene name**
Response to light stimulus	*At1g58290, At1g77760, At3g26650, At3g62410*
Photosystem I	*At1g55670*
Photosystem I reaction center	*At2g20260, At3g16140*
Photosynthesis, light harvesting	*At3g08940*
Photosynthetic electron transport in photosystem I	*At1g55670, At2g46820*
Photosynthetic NADP+ reduction	*At1g55670*
Photorespiration	*At2g35370*
Response to UV-B	*At5g63860*
Phytochrome binding	*At2g20180*
Photosynthesis	*At1g55670, At3g16140, At4g05180*
Chloroplast photosystem I	*At1g55670, At2g46820*
Chloroplast photosystem II	*At4g05180*
Photosystem I stabilization	*At1g55670*
Photosystem II stabilization	*At5g01920*

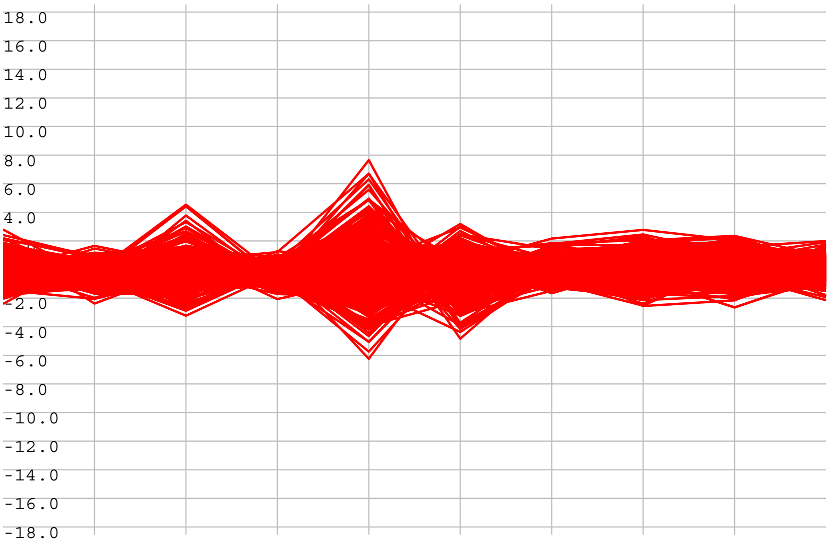


Figure 4.3.6 The expression patterns of the genes in cluster 6 (output of GQLCluster software).

Cluster 6 comprised 638 genes, and 15 of them had the GO annotation term "RNA binding". Furthermore, one gene, namely *At1g01040*, had the GO annotation term "miRNA-mediated gene silencing, mRNA cleavage".

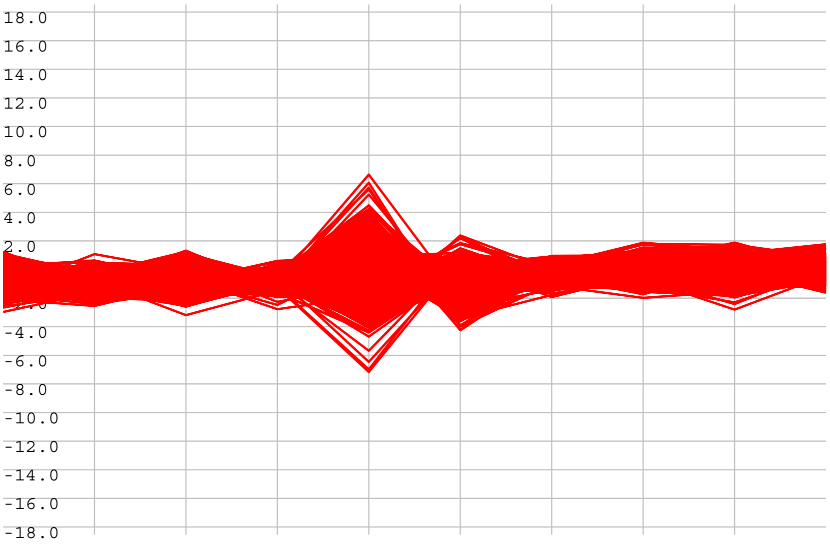


Figure 4.3.7 The expression patterns of the genes in cluster 7 (output of GQLCluster software).

Cluster 7 comprised 725 genes, and 13 of them had the GO annotation term "oxidoreductase activity" and 8 genes had the GO annotation term "oxygen binding". It has been well known that oxidoreductase are involved in response to many stresses, including light stress [204-207].

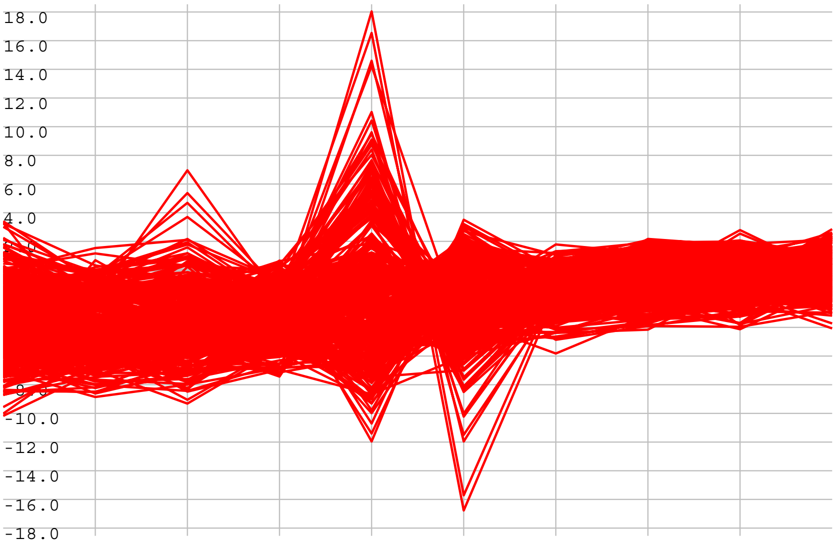


Figure 4.3.8 The expression patterns of the genes in cluster 8 (output of GQLCluster software).

Cluster 8 comprised 374 genes. Similar to that in cluster 2 and cluster 5, a large portion of genes in cluster 8 had annotations related to the stress response of plants.

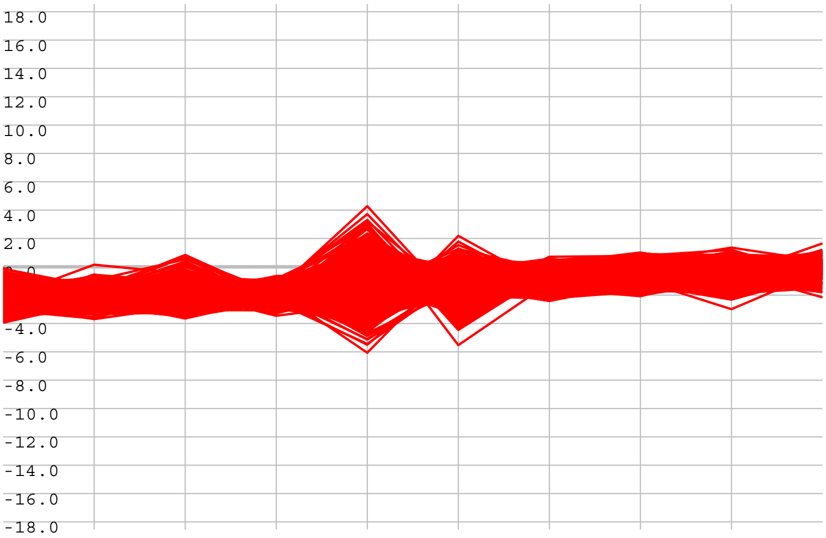


Figure 4.3.9 The expression patterns of the genes in cluster 9 (output of GQLCluster software).

Cluster 9 comprised 658 genes and we did not find obvious functional annotations enriched in this cluster.

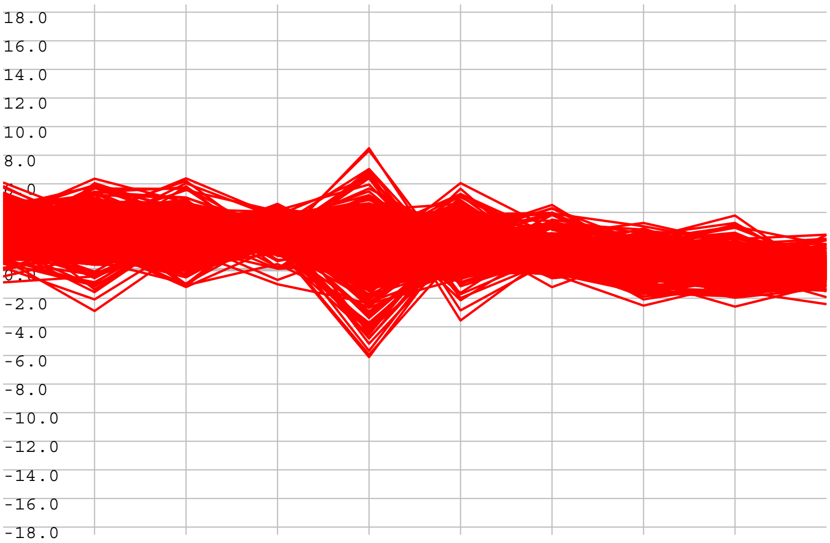


Figure 4.3.10 The expression patterns of the genes in cluster 10 (output of GQLCluster software).

Cluster 10 comprised 422 genes and 13 genes had the GO annotation term “RNA binding”.

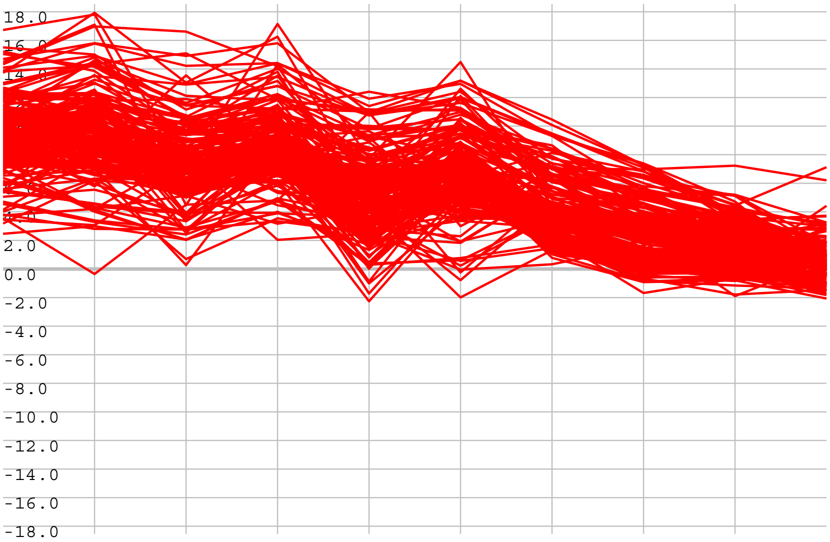


Figure 4.3.11 The expression patterns of the genes in cluster 11 (output of GQLCluster

software).

Cluster 11 comprised 186 genes and the average log expression ratio decreased along the time course. Among the 186 genes, 15 genes and 10 genes had GO annotation terms “plastoglobule” and “chlorophyll binding”, respectively. Furthermore, many genes had GO annotation terms related to photosynthesis (Table 4-4).

Table 4-4 Genes in cluster 11 with GOSLIM annotations related to photosynthesis

GOSLIM annotation	Gene name
Photosystem I	*At1g30380, At5g64040*
Photosystem II	*At1g67740, At1g79040*
Photosystem I reaction center	*At1g03130, At1g31330, At4g12800, At4g28750*
Phototropism	*At2g30520*
Photosynthetic electron transport	*At1g60950*
Photosynthesis, light harvesting in photosystem I	*At3g54890, A3g61470*
Photosynthesis, light harvesting in photosystem II	*At2g34420, At2g34430*
Photosynthetic electron transport in photosystem I	*At5g64040*
Photosystem I antenna complex	*At3g61470*
Photosystem II antenna complex	*At1g15820, At4g10340*
Photorespiration	*At1g23310, At1g63750, At2g13360, At5g04140*
Photoinhibition	*At3g15850, At5g50820, At5g66570*
Photosystem II assembly	*At3g50820, At5g66570*
Response to UV-B	*At1g51400, At3g55120*
Photosystem II oxygen evolving complex assembly	*At1g79040*
Response to light stimulus	*At1g60950, At5g04140*
Photosynthesis	*At1g03130, At1g15820, At1g29930, At1g30380, At1g31330, At1g67740, At1t79040, At2g05070, At2g34420, At2g34430, At3g47470, At3g54890, At3g61470, At4g10340,*

	At5g01530, At5g54270
Light-harvesting complex	*At1g29930, At2g05070, At2g34420, At2g34430, At3g47470, At3g54890, At3g61470, At4g10340, At5g01530, At5g54270*
Chloroplast photosystem I	*At5g64040*
Chloroplast photosystem II	*At1g67740, At3g50820, At5g66570*
Photosystem II stabilization	*At3g50820, At5g66570*

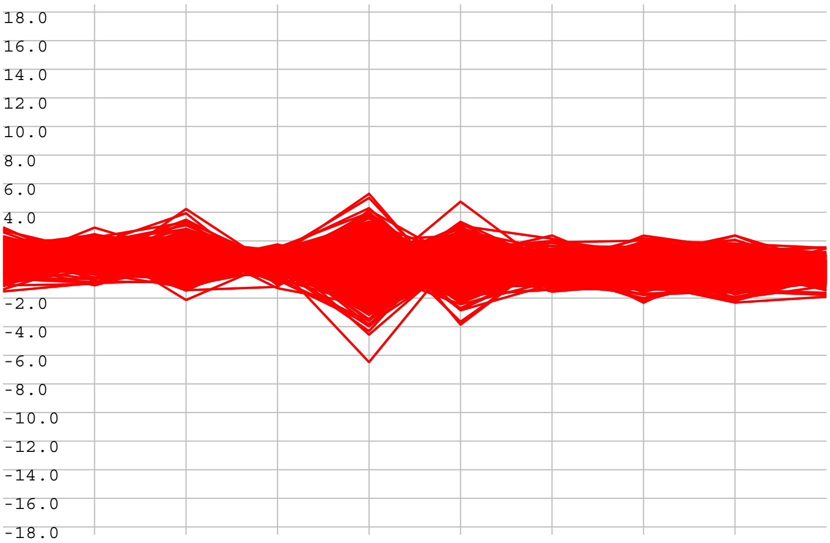


Figure 4.3.12 The expression patterns of the genes in cluster 12 (output of GQLCluster software).

Cluster 12 comprised 590 genes. We found that some genes in this cluster had function related to translation regulation, for example, 29 genes had the GO annotation term “structural constituent of ribosome”. There were 30 genes in cluster 12 had the GO annotation term “translation”, among which 4, 6 and 1 genes had the annotation terms “translation initiation”, “translation elongation” and “translation termination”, respectively.

Sequences 3000 bp upstream of TSS were retrieved for each gene and potential transcription factor binding motifs (TFBS) were detected using AlignACE for the genes belonged to the same cluster. There were 33, 21, 27, 24 31, 25, 27, 20, 30, 23, 27, 33 TFBSs found for each cluster, respectively. We also added 15 known hexamer motifs (Table 4-5) described in Gao et al. to the TFBS dataset [208].

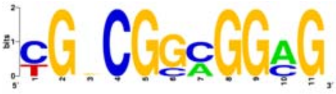
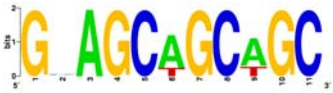
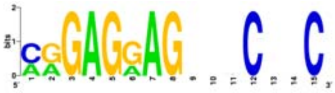
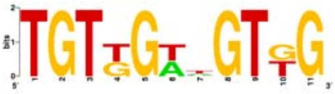
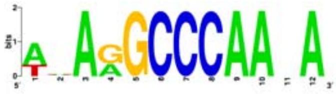
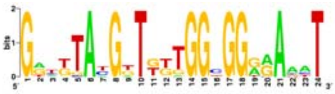
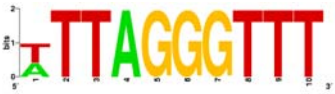
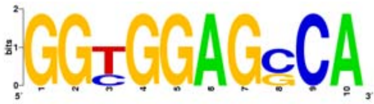
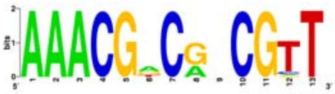
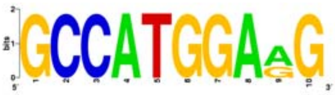
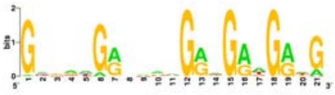
Table 4-5 Hexamer motifs described in PlantCARE database

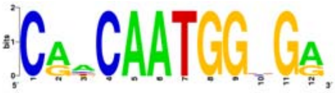
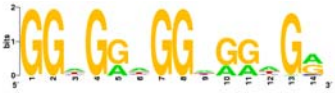
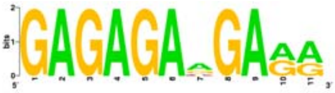
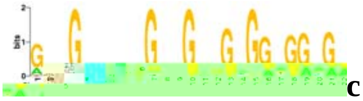
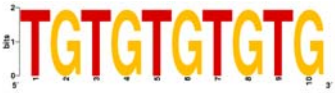
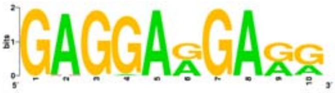
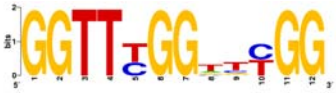
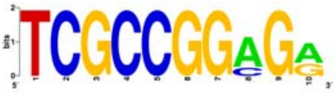
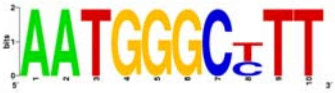
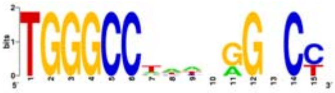
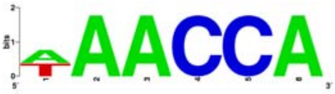
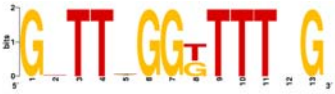
Motif name	Consensus sequences	Annotation
G-box	GACGTG	Light response element
HexamerAtH4	CCGTCG	Hexamer motif of histone H4 promoter
MYCAtERD1	CATGTG	MYC recognition motif
TBoxAtGAPB	ACTTTG	Tbox found in GAPB gene promoter
GCC-core	GCCGCC	Core of GCC-box
My-CAtRD22	CACATG	Binding site for MYC (rd22BP) in dehydration-responsive gene, rd22
CAT-box	GCCACT	*Cis*-acting regulatory element related to meristem expression
CCGTCC-box	CCGTCC	*Cis*-acting regulatory element related to meristem specific activation
GT1-motif	GGTTAA	Light responsive element
MBS	TAACTG	MYB binding site involve in drought-inducibility
TCT-motif	TCTTAC	Part of a light responsive element
Wbox	TTGACC	Wounding and pathogen response
Cbox	TGACGT	Light responsive element
I-box	GATAA[T/G]	Part of a light responsive element
MYB1At	[A/T]AACCA	MYB recognition site found in the promoters of the dehydration-responsive genes

4.3.2 Discovery of transcriptional and posttranscriptional regulatory motifs using *cop1* mutant time-course microarray data

The TFBSs and miRNA target motifs were fed to the Bayesian network model and the model weighted sequence motifs according to their contribution to the expression profiles. There had been no evidence that the TF binding to a gene's upstream region could also posttranscriptionally affect its cleavage by miRNAs and vice versa, therefore the TFBSs and miRNA target motifs were treated independently in the network construction. No interactions were allowed between two motifs of different kinds, whereas for motifs of either kind, their distances to TSS, their orientations, their copy numbers and the interaction between any two adopted motifs were all taken into account. The *cop1* mutant microarray time course experiment was not specially designed to test miRNA targets expression, so we gave upstream motifs the priority in the network construction. Therefore, a network might only have upstream motif nodes without any miRNA target nodes, but could not only have miRNA target nodes instead. About 80% of the genes (4,557) were used to train the Bayesian network model and the rest 20% genes (1,132) were used to estimate the proportion of the genes whose expression patterns could be correctly predicted by merely the adopted transcriptional and posttranscriptional regulatory motifs in the networks.

Table 4-6 TFBS nodes adopted in the networks

Network	Motif nodes adopted	Motif nodes constraints
1	 a	Reverse strand Distance to TSS up to 2060 bp
	 b	Distance to TSS up to 2200 bp
		Motifs **a** and **b** with a distance up to 1020 bp
2	 a	Distance to TSS up to 380 bp
	 b	Distance to TSS up to 460 bp
	 c	Distance to TSS up to 1040 bp
		Motifs **a** and **b** with a distance up to 80 bp
3	 a	
	 b	Distance to TSS up to 20 bp
	 c	Distance to TSS up to 1160 bp
		Motif **b** not present together with motif **a**
4	 a	Distance to TSS up to 340 bp
	 b	Distance to TSS up to 900 bp
5	 a	Distance to TSS up to 1420 bp

	 b	Distance to TSS up to 340 bp
		Motifs **a** and **b** with a distance up to 340 bp
6	 a	
	 b	
	 c	Distance to TSS up to 420 bp
	 d	Reverse strand
	 e	
	 f	
	 g	
7	 a	Distance to TSS up to 2040 bp
8	 a	Distance to TSS up to 80 bp
	 b	More than one copy
		Motifs **a** and **b** with a distance up to 1020 bp
9	 a	Distance to TSS up to 60 bp

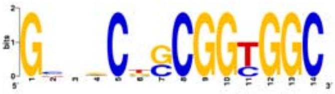
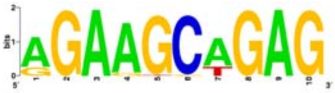
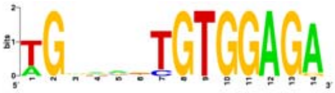
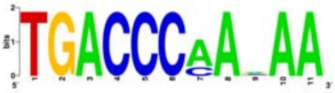
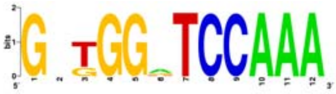
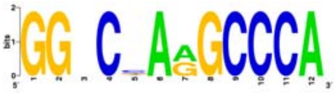
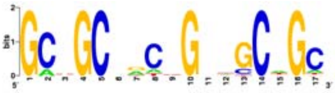
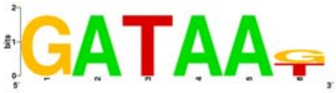
	 b	
10	 a	Distance to TSS up to 120 bp
11	 a	Distance to TSS up to 2520 bp
	 b	Forward strand Distance to TSS up to 160 bp
	 c	Distance to TSS up to 300 bp
		Motifs **a** and **b** present together
		Motifs **a** and **c** present together
		Motifs **a** and **c** with a distance up to 1340 bp
12	 a	Distance to TSS up to 2960 bp
	 b	Distance to TSS up to 980 bp
	 c	More than one copy
		Motif **b** not present together with motif **a**
		Motifs **c** and **b** with a distance up to 2480 bp

Table 4-7 miRNA target nodes adopted in the networks

Network	miRNA nodes adopted	miRNA nodes constraints
1	miR405	Distance to TSS up to 700 bp
	miR835-3p	Distance to TSS up to100 bp
2	miR835-5p	

3	miR414	
	miR401	Distance to TSS up to 100 bp
4	miR852	
	miR836	
		miR852 and miR836 targets with a distance up to 300 bp
5	miR823	
6	miR390	
	miR406	
	miR842	
7	miR413	Distance to TSS up to 2100 bp
	miR833-5p	Distance to TSS up to 1200 bp
	miR826	Distance to TSS up to 800 bp
		miR413 and miR833-5p targets with a distance up to 1900 bp
8	miR163	Distance to TSS up to 200 bp
9	miR165	
	miR166	Distance to TSS up to 600 bp
	miR823	
	miR773	Distance to TSS up to 400 bp
		miR165 targets more distant to TSS than miR166 targets
		miR823 and miR773 targets with a distance up to 1700 bp
10	N/A	
11	miR396	
12	miR845	Distance to TSS up to 500 bp

The average number of nodes was 5 for the 12 networks, and in average 3 were upstream motif nodes and 2 were miRNA target nodes (listed in Tables 4-6 and

4-7). Most networks had at least one miRNA target node except network 10, which only had a single upstream motif node. The most frequent constraint added to each node (both upstream motif node and miRNA target node) was its distances to TSS. Two known upstream motif nodes had been added, respectively, to two networks, namely MYB1At to network 8 and I-box to network 12 (Table 4-8). We compared the adopted upstream motifs with the known motifs stored in PlantCARE database [209,210], and found that ten of our adopted upstream motifs contained functional sites of the known motifs. Moreover, four of them contained motifs that had been annotated to function as light responsive elements (Table 4-8).

Table 4-8 Known TF binding motifs adopted in the networks

Network	Known motif adopted	Function description
1	TGACG-motif ^{c}	*Cis*-acting regulatory element involved in the MeJA-responsiveness (binding site of *Arabidopsis* bZIP protein TGA1a)
2	CAAT-box ^{c}	Common *cis*-acting element in promoter and enhancer regions
4	TGA-1 ^{c}	Auxin-responsive element
5	GAG-motif ^{c}	Part of a light responsive element; part of the rbcA conserved DNA module array (rbcA-CMA1) involved in light responsiveness
	GT1-motif ^{c}	Light responsive element binding site for GT1 nuclear protein factor
	GATA-motif ^{c}	Part of a light responsive element; part of the fed conserved DNA module array (fed-CMA1) involved in light responsiveness
	GA-motif ^{c}	Glycine max; part of a light responsive element; part of the LRE rbcS-(I-G) unit in a rbcS gene

	CAAT-box [c]	Common *cis*-acting element in promoter and enhancer regions
6	TGACG-motif [c]	*Cis*-acting regulatory element involved in the MeJA-responsiveness (binding site of *Arabidopsis* bZIP protein TGA1a)
	GAG-motif [c]	Part of a light responsive element; part of the rbcA conserved DNA module array (rbcA-CMA1) involved in light responsiveness
8	CAAT-box [c]	Common *cis*-acting element in promoter and enhancer regions
	MYB1At [s]	MYB recognition site found in the promoters of the dehydration-responsive gene rd22 and many other genes in *Arabidopsis*
11	CAAT-box [c]	Common *cis*-acting element in promoter and enhancer regions
12	I-box [s]	Part of light responsive element

[c] The known motif was contained within a adopted motif

[s] The known motif was adopted in the networks

In total 20 miRNA target nodes were adopted in the 12 networks, which were miR163, miR165, miR166, miR390, miR396, miR401, miR405, miR406, miR413, miR414, miR773, miR823, miR826, miR833-5p, miR835-3p, miR835-5p, miR836, miR842, miR845 and miR852 (Table 4-9).

Table 4-9 miRNAs whose target motifs were adopted in the networks and the protein classes of these potential targets

Network	miRNA nodes adopted	Protein classes of potential targets
1	miR405	DTW domain-containing protein

	miR835-3p	Zinc finger family protein F-box family protein
2	miR835-5p	Exostosin family protein MYB transcription factors
3	miR414	F-box family protein PHD finger family protein PWWP domain-containing protein RNA recognition motif (RRM)-containing protein WD-40 repeat family protein
	miR401	Exostosin family protein
4	miR852	WRKY family transcription factor
	miR836	Heat shock family protein Exostosin family protein
5	miR823	Protease-associated (PA) domain-containing protein Exostosin family protein
6	miR390	Leucine-rich repeat family protein One of 3 loci encoding tasiR-ARF (a small interfering RNA that regulates the accumulation of ARF2, 3 and 4)
	miR406	Pentatricopeptide (PPR) repeat-containing protein
	miR842	Jacalin lectin family protein Glycoside hydrolase starch-binding domain-containing protein WRKY family transcription factor
7	miR413	Basic helix-loop-helix (bHLH) family protein
	miR833-5p	Long-chain-fatty-acid--CoA ligase ABC transporter family protein
	miR826	2-oxoglutarate-dependent dioxygenase (AOP2)
8	miR163	WRKY family transcription factor F-box family protein
9	miR165	Homeobox-leucine zipper transcription factor
	miR166	Homeobox-leucine zipper transcription factor
	miR823	Protease-associated (PA) domain-containing protein Exostosin family protein
	miR773	ATPase, plasma membrane-type PHD finger transcription factor
10	N/A	

11	miR396	Growth regulation factor transcription factors Elongation factor Tu family protein
12	miR845	Jacalin lectin family protein

Only one of the 20 adopted miRNA nodes, miR823, was present in more than one network, namely in both network 5 and network 9. About half of the miRNA target nodes adopted in our networks were newly identified [185]. We found that for the 9 adopted miRNA target nodes which were newly identified, at least 6 had been reported to express in seedlings. Furthermore, 3 of them, namely miR823, miR842 and miR845 were preferentially sequenced in seedlings [185].

Mallory and Vaucheret suggested that most miRNAs are involved in overlapped regulatory networks rather than working independently, pointing to a coordinating role in fine-tuned adjustment of mRNA levels within these networks [211]. Our result supported the hypothesis that miRNAs might act together to regulate target mRNAs [112]. We found that some miRNA target motifs were present together, for example the distance between target motifs of miR836 and miR852 were always within 450 bps (Table 4-7).

4.3.3 Predicting gene expression patterns

We used the upstream motif nodes and miRNA target nodes adopted in the Bayesian network model to predict gene expression patterns. Each of the 1,132 genes was assigned to the network having the highest probability $p(v_c = 1 | D, S_c)$. Some expression patterns were quite similar, so we first calculated the correlation coefficient of the mean expression profiles between any two of the 12 clusters. If two expression patterns had a correlation coefficient greater than 0.9, the two clusters were regarded as overlapped expression patterns. And a gene assigned to any of the overlapped expression patterns would also be regarded as correctly assigned [7].

About 50% genes (572/1132) were correctly assigned. We did simulation study by randomly assigning the 1,132 genes to 12 clusters for 100,000 times. The number of correctly assigned genes was 329 in average, and the P-Value of correctly assigning 572 genes was less than 10^{-8}. Moreover, 552 out of the 572 genes could still be correctly assigned without miRNA nodes and the introducing of miRNA nodes could further correctly assign 20 genes. We retrieved the functional annotations of these 20 genes (Table 4-10) and found that *At5g12840*, a CCAAT-binding domain-containing protein, was one of the experimentally validated miRNA targets. Furthermore, *At5g54900* and *At5g57870* both had the term “RNA-binding” in GO and *At5g59780 (MYB59)* was one of the MYB transcription factors, which was reported to serve as LONG HYPOCOTYL5 (HY5) binding target and response to GA, JA, salt stress, ABA, ethylene, and auxin [212].

Table 4-10 The annotation of the 20 genes that could not be correctly assigned without involving miRNA target motifs in the networks

Gene name	Gene annotation
At5g19520	Mechanosensitive ion channel domain-containing protein
At5g44380	FAD-binding domain-containing protein

At5g12840	CCAAT-binding transcription factor (CBF-B/NF-YA) family protein
At5g14800	Pyrroline-5-carboxylate reductase
At5g54370	Late embryogenesis abundant protein-related
At5g57870	Eukaryotic translation initiation factor 4F, putative / eIF-4F, RNA binding
At5g63280	Zinc finger (C2H2 type) family protein
At5g18210	Short-chain dehydrogenase/reductase (SDR) family protein
At5g64500	Membrane protein-related
At5g54900	RNA-binding protein 45 (RBP45)
At5g47730	SEC14 cytosolic factor
At5g63790	No apical meristem (NAM) family protein
At5g35460	Expressed protein
At5g36290	Expressed protein
At5g57910	Expressed protein
At5g53850	Haloacid dehalogenase-like hydrolase family protein
At5g27150	Sodium proton exchanger
At5g59780	MYB family transcription factor (MYB59)
At5g24760	Alcohol dehydrogenase, putative, contains Pfam zinc-binding dehydrogenase domain
At5g26570	Contains InterPro domain glycoside hydrolase, starch-binding domain

4.4 Discussion

4.4.1 COP1 acting as a repressor in *Arabidopsis* photomorphogenic development

Various plant growth and development processes are critically influenced by light [213-215]. Wild type *Arabidopsis* seedling development follows two patterns, etiolation in darkness and photomorphogenesis in the light [216]. *COP/DET/FUS (CONSTITUTIVE PHOTOMORPHOGENIC/DE-ETIOLATED/FUSCA)* is a class of genes which were identified as downstream signaling components of all photoreceptors [200,217,218]. Mutation in *COP/DET/FUS* causes constitutive photomorphogenic development even in the dark [200,219].

One of the important light signaling components involved in plant light response is COP1. The constitutive photomorphogenic phenotype of *cop1* mutation indicated that COP1 acts as a negative regulator, or a light-inactivated repressor, of photomorphogenesis [216]. Shin et al. showed that COP1 is required to repress a flower-specific transcription factor, *AtMYB21*, in seedlings. They also suggested that the abundance of some genes in *cop1* mutant is at least partially due to the ectopic expression of *MYB* genes, including *AtMYB21* [200]. Shin et al. further suggested that *COP1* regulated not only photomorphogenesis, but also other developmental processes, that is, in addition to the constitutive photomorphogenic phenotypes, the mutants stop further development in the seedling stage and ectopically express genes that are not related to light signaling [215,220].

In the 12 clusters, 15 genes had the GO annotation term “photomorphogenesis” (Table 4-11). Among them, 5 were annotated to encode COP9 signalosome subunits (CSN).

Table 4-11 Genes which had GOSLIM annotation related to photomorphogenesis

GOSLIM annotation	Gene name	Cluster
Photomorphogenesis	*At2g46370, At3g28860, At3g61140/CSN8, At5g14250/CSN3*	Cluster 1
	At5g56280/CSN6A	Cluster 6
	At2g36910	Cluster 7
	At1g79810	Cluster 8

	At1g22920/CSN5B, At2g26990/CSN2, At3g55370, At4g08920	Cluster 9
Regulation of photomorphogenesis	*At4g02440*	Cluster 5
	At2g24790	Cluster 12
Negative regulation of photomorphogenesis	*A5g46210*	Cluster 10
	At4g05420	Cluster 12

Most *COP/DET/FUS* genes, such as *COP9(AtCSN8)*, encode components of the COP9 signalosome which resemble the components of the 26S proteosome [200,221-224]. We retrieved the expression profiles of the 15 genes (Figure 4.4.1) and found that the log expression ratios of these genes, especially the 5 genes that were annotated to encode COP9 signalosome subunits, had not changed much during the time course (Figure 4.4.2).

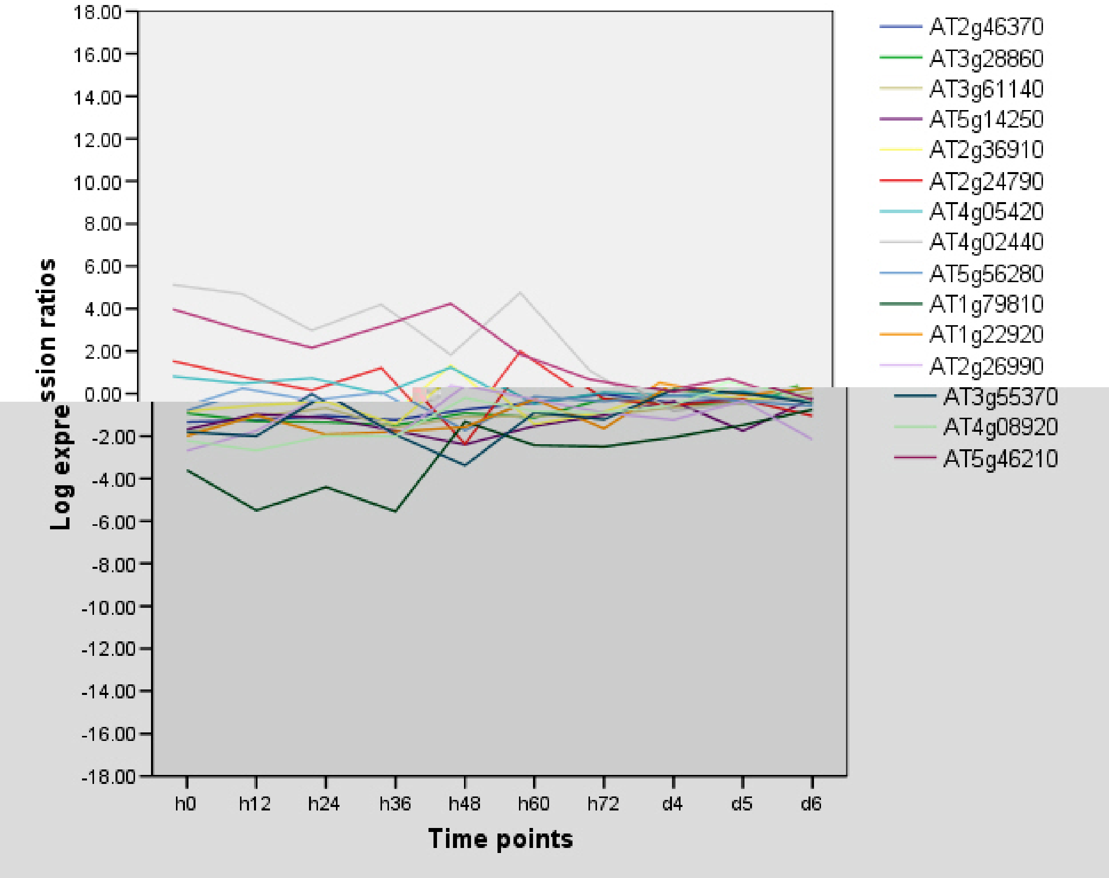


Figure 4.4.1 The expression profiles of the 15 genes that had the GO annotation term “photomorphogenesis”.

If *cop1* mutant is grown at 20 degree, it will show mutant phenotype; whereas it will show wild type phenotype if it is grown at higher temperature (e.g. 30 degree). The longer it has been grown at 30 degree, the more similar the phenotype of wild type and that of *cop1* mutant will be. This was consistent with the observation that despite the up-regulation or down-regulation of the 15 genes at the earlier time points (before 72h), the expression differences (*cop*1 mutant versus wild type) diminished gradually at later time points.

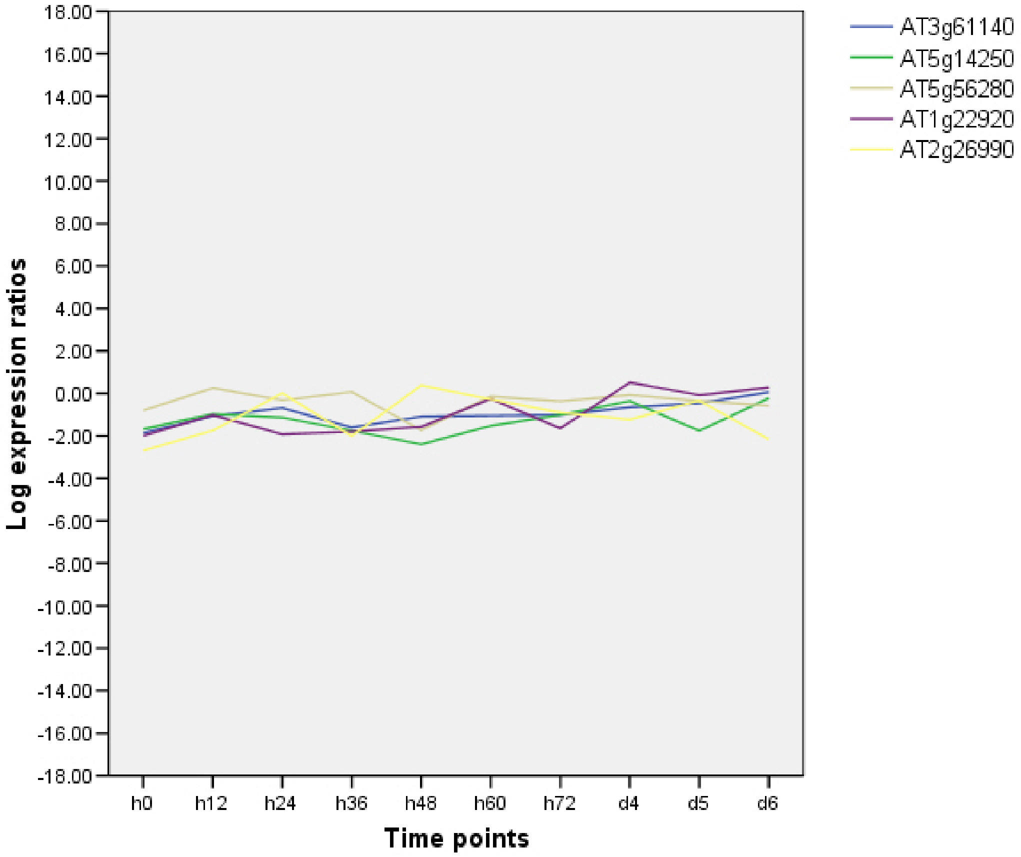


Figure 4.4.2 The expression profiles of the 5 genes encoding COP9 signalosome subunits.

The log expression ratios of the 5 genes encoding COP9 signalosome had not changed significantly across the 10 time points, which suggested that though COP9 signalosome regulated the nucleocytoplasmic portioning of COP1 [216], there was no observed feedback regulation under the present experimental condition.

4.4.2 Transcriptional and posttranscriptional regulatory networks

There seems to be more computationally predicted motifs without a known matching transcription factor, than transcription factors without a known binding sequence [225]. Since motifs are usually short, they may randomly present in the upstream regions of many genes without a functional role [6].

Bayesian network models weight sequence motifs according to their contribution and are greatly helpful for selecting motifs that are functional under specific experimental conditions [226,227]. Among the many possible machine-learning methods that could be applied to predicting interactions, Bayesian network possesses two advantages: (a) it allows features of more than one data type to be represented together and converted into a common probabilistic framework without unnecessary simplification [5,227]; (b) it is capable of dealing with incomplete information and encoding dependencies between variables [7]. In this study, Bayesian network was used to infer the probabilistic dependence between sequence elements and expression patterns [7,129,133].

The steady-state of an mRNA results from the balance between transcription and decay [228]. TFs and miRNAs are two groups of well-known factors involved in the regulation of gene transcription as well as decay. Most genomic studies of gene expression regulation focused on transcription rather than on mRNA decays. Based on a model in which upstream motifs contribute additively to the log-expression level of a gene, Bussemaker presented a computational method for discovering *cis*-regulatory elements that circumvented the need to cluster genes based on their profiles [127]. Beer and Tavazoie correctly predicted 70% of the gene expression patterns by use of Bayesian network based only on upstream motifs [7]. Li et al. developed a promoter classification method using a Relevance Vector Machine (RVM) and Bayesian statistical principles to identify discriminatory features in the promoter sequences of genes that could classify transcriptional responses and they correctly predicted 70% genes as being up- or down-regulated, based on a small set of discriminative promoter motifs [229]. In the meanwhile, Foat et al. identified functional 3' UTR motifs (including miRNA target sites) that best correlated with the observed changes in mRNA levels [112,228]. Sood et al. used computational methods to explore the effects of endogenous miRNA expression on endogenous steady-state mRNA levels. In their model, changes in mRNA levels of a given gene (measured by the microarray experiment) are written as a sum over contributions from all sequence motifs in the 3' UTR of that gene, which could explain changes in mRNA levels for 50% genes [194]. Although Beer and Tavazoie as well as Rajewsky both in their work suggested the integration of posttranscriptional and transcriptional motifs in the future study of gene regulatory networks, respectively [7,112], none of the groups had correlated both transcriptional and posttranscriptional regulatory elements together with the mRNA steady-state level.

4.4.3 Microarray data used in constructing the regulatory networks

Gene expression is thought to be a temporal process. Under different conditions, different proteins are required for different functions; even under stable conditions, due to protein degradation mRNA is transcribed continuously and new proteins are generated. A TF gene also takes time to express its protein product and then affect (directly of indirectly) the transcript level of its target gene [230]. In static experiments, the expression of genes in different samples is measured, whereas in time series experiments, the expression of genes during a temporal process is measured. Static data are assumed to be independent and identically distributed, whereas time series data exhibit a strong autocorrelation between successive time points [97].

Beer and Tavazoie used large data set obtained under many experimental conditions and identified, by reverse engineering, the regulatory elements dictating their expression patterns [3,7]. Foat et al. used 750 microarray datasets in 750 conditions, respectively, in his study [228]. However, the algorithm that is able to predict expression in a single condition solely based on promoter sequences has not been described [3]. To address this need, we used a single condition time course data in order to gauge the capability of our method to capture the underlying mechanism of gene expression in this process at sequence level.

4.4.4 Contribution of miRNAs in gene regulation networks

Lee et al. investigated the regulation of 61 TFs expressed in a tissue-enriched manner in *Arabidopsis* roots. They suggested that the sequence within 3 kb upstream of a TF was sufficient for driving the endogenous mRNA expression pattern in 80% (35/44) of the cases. 25% of the TFs underwent posttranscriptional regulation via miRNA (2/24) or via intercellular protein movement (6/24). Lee et al. suggested that promoter region had major contribution to the steady-state of *Arabidopsis* TF transcripts, while posttranscriptional regulation of gene expression was also frequently observed [146].

In our study, 3.4% of the 572 genes could only be correctly assigned after introducing miRNA target nodes, which might suggest that the consequence of miRNA-mediated posttranscriptional regulation was marginal in our time course expression profiles although miRNA was considered as one of the most important posttranscriptional gene regulators. This might result from a possible bias in the predictive power of TFBS since the motif finding was done for each fixed cluster. In view of this, we did a reference test using only the aforementioned 15 known hexamer motifs (Table 4-5) and miRNA target motifs. Using the 15 known hexamer motifs, we could only correctly assign 296 genes, which was even less than that from random assignment (P-Value = 0.98) and this suggested that the observed expression profiles could not be explained solely by the combination of the15 known motifs. After adding miRNA target nodes, we could correctly assign

500 genes (P-Value $<10^{-8}$). On the other hand, without TFBSs adopted, we could only correctly assign 386 genes solely on the adopted miRNA target motifs. The result suggested that miRNAs might confer additional layers of robustness on gene regulation networks. Exploration of miRNA regulatory mechanism together with known transcriptional regulatory interactions and other functional genomics data might help to further elucidate the function of miRNAs at a system-wide level [112].

Plants have evolved sophisticated gene regulatory networks that mediate developmental changes in response to light [231]. In *Arabidopsis*, COP1 interacts with specific TFs to repress their activities in the dark and HY5 is one of such TFs [232-234]. The first nuclear target identified for COP1 was the bZIP- type transcription factor HY5, characterized as a positive regulator of photomorphogenic development and as a suppressor of *cop1* mutation [216,235,236]. HY5 inhibits hypocotyl elongation in the light and is a key TF in seedling photomorphogenesis [231]. A recent Chip-on-chip approach systematically identified direct targets of HY5 [212,231]. Five of the aforementioned 204 genes, that could not be correctly assigned solely by the 15 known motifs but could be correctly assigned once miRNA nodes were adopted, were reported to be HY5 binding targets [212]. The five genes were *At5g23010*, *At5g05410*, *At5g49280*, *At5g59780* and *At5g59820*, which showed functional enrichment (P-Value =0.01). This result suggested that miRNA might also involve in the light responsive network. Furthermore, we suggested that many of the 204 genes might have function related to miRNA regulation mechanism. Hence we grouped the 204 genes according to their GO annotations. Some genes had significantly enriched GO annotation terms. It was not surprising to find that both functional annotation terms "DNA or RNA binding" and "transcription factor activity" were enriched as it was well-known that plant miRNAs were biased toward to target TFs and other regulatory genes [150]. Functional annotations "response to abiotic or biotic stimulus" and "response to stress" were also enriched with high significance (both P-Value $<10^{-8}$), which was consistent with the fact that miRNA played important roles in plant responses to environmental stresses as well as in development and genome maintenance [159].

Table 4-12 Functional enrichment of the 204 genes from GOSLIM annotations

GOSLIM annotation	Within group (204 genes)	Not within group (25,472 genes)	All (25,676 genes)	P-Value
DNA or RNA binding	22	1806	1828	0.020
Transcription factor activity	21	1667	1688	0.016
Transferase activity	27	1625	1652	$<10^{-8}$
Kinase activity	21	1677	1698	0.017
Nucleotide binding	20	1227	1247	0.001
Nucleus	28	2409	2437	0.019
Mitochondria	24	990	1114	$<10^{-8}$
Transport	29	1259	1288	$<10^{-8}$
Response to abiotic or biotic stimulus	34	1276	1310	$<10^{-8}$
Response to stress	18	76	94	$<10^{-8}$
Chloroplast	38	2457	2495	$<10^{-8}$

Chapter 5: Conclusion

It comes the era of systems biology, which has two key elements, namely large-scale molecular measurements and computational modeling [237]. The molecular constituents of a system and their variations across a series of dynamic phenotypic changes can be measured and the measurements are collectively referred to as "omics", such as genomics, transcriptomics, proteomics, metabolomics, pharmacogenomics, physiomics etc. They are quantitative study of sequences, expression, metabolites and so on. The computational systems biology uses computational approaches to understand biological systems in system-level, which integrates various types of data in multiple levels and phases; one of such efforts is to reconstruct transcriptional regulatory networks using various types of gene expression data and regulatory sequence motifs.

To date, a lot of studies have been done to explain the gene expression profiles in terms of the combination of transcriptional factor binding motifs or *cis*-regulatory elements [6,7]. However, these approaches have not been broadly applied in multi-cellular organisms in spite of the reported success in model organisms. A key limitation of such approaches is that many regulators are regulated posttranscriptionally [8]. While progress is made in mapping transcriptional regulatory networks, posttranscriptional regulatory networks just begin to be uncovered. Aiming at integrating both transcription factor binding motifs and posttranscriptional regulatory motifs toward a better quantitative modeling of changes in mRNA level, we proposed a probabilistic approach to determine the context-dependent role of genomic TF binding motifs together with miRNA binding motifs in transcriptional and posttranscriptional regulation. Regardless the simple strategy employed, our method may provide an incomplete or coarse-grained portrait of the underlying transcriptional and posttranscriptional regulatory network, which can correctly predict the expression patterns for more than 50% genes in our test dataset. Consequently, our method facilitated the incorporation of diverse sources with limited prior knowledge. The relationship between sequence motifs and gene expression patterns could be investigated more precisely from datasets that observe expression profiles of miRNAs, mRNAs and proteins from the same samples simultaneously. Furthermore, other gene regulatory mechanism besides that regulated by TFs and miRNAs should also be taken into consideration in the future network model, such as : cell signaling, miRNA splicing, polyadenylation and localization; chromatin modifications; and mechanisms of protein localization, modification and degradation [189].

To facilitate our transcriptional regulatory network study, we proposed a novel Hidden Markov Model based method for prediction plant miRNA target motifs. Comparing to the exiting methods of plant miRNA prediction, our method does not have conservation constraints and is capable of capturing the position specific information about particular matches/mismatches. We assume here that the complementarity between miRNA-target duplex might follow some rules according to the RISC mechanisms, and the HMM method could find these hidden rules by learning from a training set of known plant miRNA targets. Despite the substantial advances in the past few years, molecular network biology was still in its infancy [2]. Future progress may be expected in two directions, namely the development of new theoretical methods in order to improve the capability of charactering the network topology and the development of highly sensitive tools to enhance our data collecting abilities. Once the concentrations, fluxes and interactions of various types of molecules at high resolution both in time and space can be identified and quantified, the global signaling networks can be comprehensively studied.

Reference

1. D'Haeseleer P, Liang S, Somogyi R (2000) Genetic network inference: from co-expression clustering to reverse engineering. Bioinformatics 16: 707-726.

2. Barabási A-L, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nature Reviews Genetics 5: 101-113.

3. Blais A, Dynlacht BD (2005) Constructing transcriptional regulatory networks. Genes Dev 19: 1499-1511.

4. Szallasi Z. Genetic Network Analysis in Light of Massively Parallel Biological Data Acquisition; 1999. pp. 5-16.

5. Butte A (2002) The use and analysis of microarray data. Nature Reviews Drug Discovery 1: 951-960.

6. Segal E, Yelensky R, Koller D (2003) Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. Bioinformatics 19: i273-282.

7. Beer MA, Tavazoie S (2004) Predicting Gene Expression from Sequence. Cell 117: 185-198.

8. Segal E, Friedman N, Kaminski N, Regev A, Koller D (2005) From signatures to models: understanding cancer using microarrays. Nature Genetics 37: 38-45.

9. Elemento O, Tavazoie S (2005) Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach. Genome Biology 6: R18.

10. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, et al. (2002) Transcriptional Regulatory Networks in Saccharomyces cerevisiae. Science 298: 799-804.

11. Stormo GD (2000) DNA binding sites: representation and discovery. Bioinformatics 16: 16-23.

12. Hvidsten TR, Wilczynski B, Kryshtafovych A, Tiuryn J, Komorowski J, et al. (2005) Discovering regulatory binding-site modules using rule-based learning. Genome Res 15: 856-866.

13. Riechmann JL, Heard J, Martin G, Reuber L, Z C, et al. (2000) Arabidopsis Transcription Factors: Genome-Wide Comparative Analysis Among Eukaryotes. Science 290: 2105-2110.

14. Du T, Zamore PD (2005) microPrimer: the biogenesis and function of microRNA. Development 132: 4645-4652.

15. Cai X, Hagedorn CH, Cullen BR (2004) Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. RNA 10: 1957-1966.

16. Lee Y, Kim M, Han J, Yeom K-H, Lee S, et al. (2004) MicroRNA genes are transcribed by RNA polymerase II The EMBO Journal 23: 4051-4060.

17. Parizotto EA, Dunoyer P, Rahm N, Himber C, Voinnet O (2004) In vivo investigation of the transcription, processing, endonucleolytic activity, and functional relevance of the spatial distribution of a plant miRNA. Genes Dev 18: 2237-2242.

18. Lee MLT (2004) Analysis of Microarray Gene Expression Data. Norwell, Massachusetts: Kluwer Academic Publishers.

19. Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T (2001) Identification of Novel Genes Coding for Small Expressed RNAs. Science 294: 853-858.

20. Lau NC, Lim LP, Weinstein EG, Bartel DP (2001) An Abundant Class of Tiny RNAs with Probable Regulatory Roles in Caenorhabditis elegans. Science 294: 858-862.

21. Lee Y, Jeon K, Lee J-T, Kim S, Kim VN (2002) MicroRNA maturation: stepwise processing and subcellular localization The EMBO Journal 21: 4663-4670.

22. Rodriguez A, Griffiths-Jones S, Ashurst JL, Bradley A (2004) Identification of Mammalian microRNA Host Genes and Transcription Units. 14: 1902-1910.

23. Rodriguez A, Griffiths-Jones S, Ashurst JL, Bradley A (2004) Identification of

Mammalian microRNA Host Genes and Transcription Units. Genome Res 14: 1902-1910.

24. Lee Y, Ahn C, Han J, Choi H, Kim J, et al. (2003) The nuclear RNase III Drosha initiates microRNA processing. Nature 425: 415-419.

25. Denli AM, Tops BBJ, Plasterk RHA, Ketting RF, Hannon GJ (2004) Processing of primary microRNAs by the Microprocessor complex. Nature 432: 231-235.

26. Gregory RI, Yan K-p, Amuthan G, Chendrimada T, Doratotaj B, et al. (2004) The Microprocessor complex mediates the genesis of microRNAs. Nature 432: 235-240.

27. Han J, Lee Y, Yeom K-H, Kim Y-K, Jin H, et al. (2004) The Drosha-DGCR8 complex in primary microRNA processing. Genes Dev 18: 3016-3027.

28. Landthaler M, Yalcin A, Tuschl T (2004) The Human DiGeorge Syndrome Critical Region Gene 8 and Its D. melanogaster Homolog Are Required for miRNA Biogenesis. Current Biology 14: 2162-2167.

29. Yi R, Qin Y, Macara IG, Cullen BR (2003) Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. Genes Dev 17: 3011-3016.

30. Bohnsack MT, Czaplinski K, Gorlich D (2004) Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs. RNA 10: 185-191.

31. Lund E, Guttinger S, Calado A, Dahlberg JE, Kutay U (2004) Nuclear Export of MicroRNA Precursors. Science 303: 95-98.

32. Zeng Y, Cullen BR (2004) Structural requirements for pre-microRNA binding and nuclear export by Exportin 5. Nucleic Acids Research 32: 4776-4785.

33. Bernstein E, Caudy AA, Hammond SM, Hannon GJ (2001) Role for a bidentate ribonuclease in the initiation step of RNA interference. Nature 409: 363-366.

34. Grishok A, Pasquinelli AE, Conte D, Li N, Parrish S, et al. (2001) Genes and Mechanisms Related to RNA Interference Regulate Expression of the Small Temporal RNAs that Control C. elegans Developmental Timing.

Cell 106: 23-34.

35. Hutvagner G, McLachlan J, Pasquinelli AE, Balint E, Tuschl T, et al. (2001) A Cellular Function for the RNA-Interference Enzyme Dicer in the Maturation of the let-7 Small Temporal RNA. Science 293: 834-838.

36. Ketting RF, Fischer SEJ, Bernstein E, Sijen T, Hannon GJ, et al. (2001) Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in C. elegans. Genes Dev 15: 2654-2659.

37. Chendrimada TP, Gregory RI, Kumaraswamy E, Norman J, Cooch N, et al. (2005) TRBP recruits the Dicer complex to Ago2 for microRNA processing and gene silencing. Nature 436: 740-744.

38. Förstemann K, Tomari Y, Du T, Vagin VV, Denli AM, et al. (2005) Normal microRNA Maturation and Germ-Line Stem Cell Maintenance Requires Loquacious, a Double-Stranded RNA-Binding Domain Protein. PLoS Biol 3: e236.

39. Jiang F, Ye X, Liu X, Fincher L, McKearin D, et al. (2005) Dicer-1 and R3D1-L catalyze microRNA maturation in Drosophila. Genes Dev 19: 1674-1679.

40. Saito K, Ishizuka A, Siomi H, Siomi MC (2005) Processing of pre-microRNAs by the dicer-1-loquacious complex in Drosophila cells. PLoS Biol 3: E235.

41. Khvorova A, Reynolds A, Jayasena SD (2003) Functional siRNAs and miRNAs Exhibit Strand Bias. Cell 115: 209-216.

42. Schwarz DS, Hutvagner G, Du T, Xu Z, Aronin N, et al. (2003) Asymmetry in the Assembly of the RNAi Enzyme Complex. Cell 115: 199-208.

43. Hammond SM, Bernstein E, Beach D, Hannon GJ (2000) An RNA-directed nuclease mediates post-transcriptional gene silencing in Drosophila cells. Nature 404: 293-296.

44. Hammond SM, Boettcher S, Caudy AA, Kobayashi R, Hannon GJ (2001) Argonaute2, a Link Between Genetic and Biochemical Analyses of RNAi. Science 293: 1146-1150.

45. Hutvagner G, Zamore PD (2002) A microRNA in a Multiple-Turnover RNAi

Enzyme Complex. Science 297: 2056-2060.

46. Zeng Y, Wagner EJ, Cullen BR (2002) Both Natural and Designed Micro RNAs Can Inhibit the Expression of Cognate mRNAs When Expressed in Human Cells. Molecular Cell 9: 1327-1333.

47. Doench JG, Petersen CP, Sharp PA (2003) siRNAs can function as miRNAs. Genes Dev 17: 438-442.

48. Song J-J, Liu J, Tolia NH, Schneiderman J, Smith SK, et al. (2003) The crystal structure of the Argonaute2 PAZ domain reveals an RNA binding motif in RNAi effector complexes. Nat Struct Mol Biol 10: 1026-1032.

49. Yan KS, Yan S, Farooq A, Han A, Zeng L, et al. (2003) Structure and conserved RNA binding of the PAZ domain. Nature 426: 469-474.

50. Lingel A, Simon B, Izaurralde E, Sattler M (2004) Nucleic acid 3[prime]-end recognition by the Argonaute2 PAZ domain. Nat Struct Mol Biol 11: 576-577.

51. Parker JS, Roe SM, Barford D (2004) Crystal structure of a PIWI protein suggests mechanisms for siRNA recognition and slicer activity The EMBO Journal 23: 4727-4737.

52. Ma J-B, Yuan Y-R, Meister G, Pei Y, Tuschl T, et al. (2005) Structural basis for 5[prime]-end-specific recognition of guide RNA by the A. fulgidus Piwi protein. Nature 434: 666-670.

53. Parker JS, Roe SM, Barford D (2005) Structural insights into mRNA recognition from a PIWI domain-siRNA guide complex. Nature 434: 663-666.

54. Liu J, Carmell MA, Rivas FV, Marsden CG, Thomson JM, et al. (2004) Argonaute2 Is the Catalytic Engine of Mammalian RNAi. Science 305: 1437-1441.

55. Rand TA, Ginalski K, Grishin NV, Wang X (2004) Biochemical identification of Argonaute 2 as the sole protein required for RNA-induced silencing complex activity. PNAS 101: 14385-14389.

56. Song J-J, Smith SK, Hannon GJ, Joshua-Tor L (2004) Crystal Structure of Argonaute and Its Implications for RISC Slicer Activity. Science 305:

1434-1437.

57. Baumberger N, Baulcombe DC (2005) Arabidopsis ARGONAUTE1 is an RNA Slicer that selectively recruits microRNAs and short interfering RNAs. PNAS 102: 11928-11933.

58. Qi Y, Denli AM, Hannon GJ (2005) Biochemical Specialization within Arabidopsis RNA Silencing Pathways. Molecular Cell 19: 421-428.

59. Rivas FV, Tolia NH, Song J-J, Aragon JP, Liu J, et al. (2005) Purified Argonaute2 and an siRNA form recombinant human RISC. Nat Struct Mol Biol 12: 340-349.

60. Lewis BP, Shih I-h, Jones-Rhoades MW, Bartel DP, Burge CB (2003) Prediction of Mammalian MicroRNA Targets. Cell 115: 787-798.

61. Brennecke J, Stark A, Russell RB, Cohen SM (2005) Principles of MicroRNA–Target Recognition. PLoS Biology 3: e85.

62. Lewis BP, Burge CB, Bartel DP (2005) Conserved Seed Pairing, Often Flanked by Adenosines, Indicates that Thousands of Human Genes are MicroRNA Targets. Cell 120: 15-20.

63. Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, et al. (2005) Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature 434: 338-345.

64. Olsen PH, Ambros V (1999) The lin-4 Regulatory RNA Controls Developmental Timing in Caenorhabditis elegans by Blocking LIN-14 Protein Synthesis after the Initiation of Translation. Developmental Biology 216: 671-680.

65. Park W, Li J, Song R, Messing J, Chen X (2002) CARPEL FACTORY, a Dicer Homolog, and HEN1, a Novel Protein, Act in microRNA Metabolism in Arabidopsis thaliana. Current Biology 12: 1484-1495.

66. Papp I, Mette MF, Aufsatz W, Daxinger L, Schauer SE, et al. (2003) Evidence for Nuclear Processing of Plant Micro RNA and Short Interfering RNA Precursors. Plant Physiology 132: 1382-1390.

67. Xie Z, Johansen LK, Gustafson AM, Kasschau KD, Lellis AD, et al. (2004) Genetic and Functional Diversification of Small RNA Pathways in Plants.

PLoS Biol 2: e104.

68. Reinhart BJ, Weinstein EG, Rhoades MW, Bartel B, Bartel DP (2002) MicroRNAs in plants. Genes Dev 16: 1616-1626.

69. Vazquez F, Gasciolli V, Crete P, Vaucheret H (2004) The Nuclear dsRNA Binding Protein HYL1 Is Required for MicroRNA Accumulation and Plant Development, but Not Posttranscriptional Transgene Silencing. Current Biology 14: 346-351.

70. Peragine A, Yoshikawa M, Wu G, Albrecht HL, Poethig RS (2004) SGS3 and SGS2/SDE1/RDR6 are required for juvenile development and the production of trans-acting siRNAs in Arabidopsis. Genes Dev 18: 2368-2379.

71. Park MY, Wu G, Gonzalez-Sulser A, Vaucheret H, Poethig RS (2005) Nuclear processing and export of microRNAs in Arabidopsis. PNAS 102: 3691-3696.

72. Boutet S, Vazquez F, Liu J, Beclin C, Fagard M, et al. (2003) Arabidopsis HEN1: A Genetic Link between Endogenous miRNA Controlling Development and siRNA Controlling Transgene Silencing and Virus Resistance. Current Biology 13: 843-848.

73. Yu B, Yang Z, Li J, Minakhina S, Yang M, et al. (2005) Methylation as a Crucial Step in Plant microRNA Biogenesis. Science 307: 932-935.

74. Llave C, Xie Z, Kasschau KD, Carrington JC (2002) Cleavage of Scarecrow-like mRNA Targets Directed by a Class of Arabidopsis miRNA. Science 297: 2053-2056.

75. Tang G, Reinhart BJ, Bartel DP, Zamore PD (2003) A biochemical framework for RNA silencing in plants. Genes Dev 17: 49-63.

76. Kidner CA, Martienssen RA (2005) The developmental role of microRNA in plants. Growth and development 8: 38-44.

77. Alvarez-Garcia I, Miska EA (2005) MicroRNA functions in animal development and human disease. Development 132: 4653-4662.

78. Qian J, Dolled-Filhart M, Lin J, Yu H, Gerstein M (2001) Beyond synexpression relationships: local clustering of time-shifted and inverted

gene expression profiles identifies new, biologically relevant interactions. Journal of Molecular Biology 314: 1053-1066.

79. Quackenbush J (2001) Computational Analysis of Microarray data. Nature Reviews Genetics 2: 418-427.

80. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. PNAS 95: 14863-14868.

81. West M, Blanchette C, Dressman H, Huang E, Ishida S, et al. (2001) Predicting the clinical status of human breast cancer by using gene expression profiles. PNAS 98: 11462-11467.

82. Simon RM, Korn EL, McShane LM, Radmacher MD, Wright GW, et al. (2003) Design and analysis of DNA microarray investigations. New York: Springer Press.

83. Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, et al. (2000) Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature 406: 536-540.

84. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM (1999) Systematic determination of genetic network architecture. Nature Genetics 22: 281-285.

85. MacQueen JB. Some Methods for classification and Analysis of Multivariate Observations; 1967; Berkeley. University of California Press. pp. 281-297.

86. Rahnenfuehrer J. Efficient clustering methods for tumor classification with microarrays; 2002; Mannheim, Germany.

87. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, et al. (1999) Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. PNAS 96: 2907-2912.

88. Gardner TS, di Bernardo D, Lorenz D, Collins JJ (2003) Inferring Genetic Networks and Identifying Compound Mode of Action via Expression Profiling. Science 301: 102-105.

89. Gardner TS, Faith JJ (2005) Reverse-engineering transcription control networks. Physics of Life Reviews 2: 65-88.

90. Schliep A, Schonhuth A, Steinhoff C (2003) Using hidden Markov models to analyze gene expression time course data. Bioinformatics 19: i255-263.

91. Ma P, Castillo-Davis CI, Zhong W, Liu JS (2006) A data-driven clustering method for time course gene expression data. Nucl Acids Res 34: 1261-1269.

92. Cheng C, Ma X, Yan X, Sun F, Li LM (2006) MARD: a new method to detect differential gene expression in treatment-control time courses. Bioinformatics 22: 2650-2657.

93. Bar-Joseph Z, Gerber G, Simon I, Gifford DK, Jaakkola TS (2003) Comparing the continuous representation of time-series expression profiles to identify differentially expressed genes. PNAS 100: 10146-10151.

94. Heard NA, Holmes CC, Stephens DA, Hand DJ, Dimopoulos G (2005) Bayesian coclustering of Anopheles gene expression time series: Study of immune defense response to multiple experimental challenges. PNAS 102: 16939-16944.

95. Qu Y, Xu S (2006) Quantitative Trait Associated Microarray Gene Expression Data Analysis. Mol Biol Evol 23: 1558-1573.

96. Luan Y, Li H (2003) Clustering of time-course gene expression data using a mixed-effects model with B-splines. Bioinformatics 19: 474-482.

97. Bar-Joseph Z (2004) Analyzing time series gene expression data. Bioinformatics 20: 2493-2503.

98. Wasserman WW, Sandelin A (2004) Applied Bioinformatics for the Identification of Regualtory Elements. Nature Reviews Genetics 5: 276-287.

99. Lee RC, Feinbaum RL, Ambros V (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell 75: 843-854.

100. Chan CS, Elemento O, Tavazoie S (2005) Revealing Posttranscriptional Regulatory Elements Through Network-Level Conservation. PLoS Computational Biology 1: e69.

101. DeRisi JL, Iyer VR, Brown PO (1997) Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale. Science 278: 680-686.

102. Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, et al. (2002) Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. PNAS 99: 757-762.

103. Prakash A, Tompa M (2005) Discovery of regulatory elements in vertebrates through comparative genomics. Nature Biotechnology 23: 1249-1256.

104. Lawrence CE, Reilly AA (1990) An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. Proteins 7: 41-51.

105. Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol 2: 28-36.

106. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, et al. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science 262: 208-214.

107. Liu JS, Neuwald AF, Lawrence CE (1995) Bayesian Models for Multiple Local Sequence Alignment and Gibbs Sampling Strategies. Journal of the American Statistical Association 90: 1156-1170.

108. Liu X, Brutlag DL, Liu JS (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. Pac Symp Biocomput: 127-138.

109. Hughes JD, Estep PW, Tavazoie S, Church GM (2000) Computational identification of Cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. Journal of Molecular Biology 296: 1205-1214.

110. Roth FP, Hughes JD, Estep PW, Church GM (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. Nature Biotechnology 16: 939 - 945.

111. Bartel DP (2004) MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. Cell 116: 281-297.

112. Rajewsky N (2006) microRNA target predictions in animals. Nature Genetics 38: S8 - S13.

113. Brennecke J, Hipfner DR, Stark A, Russell RB, Cohen SM (2003) bantam Encodes a Developmentally Regulated microRNA that Controls Cell Proliferation and Regulates the Proapoptotic Gene hid in Drosophila. Cell 113: 25-36.

114. Stark A, Brennecke J, Russell RB, Cohen SM (2003) Identification of Drosophila MicroRNA Targets. PLoS Biology 1: e60.

115. Enright A, John B, Gaul U, Tuschl T, Sander C, et al. (2003) MicroRNA targets in Drosophila. Genome Biology 5: R1.

116. John B, Enright AJ, Aravin A, Tuschl T, Sander C, et al. (2004) Human MicroRNA Targets. PLoS Biology 2: e363.

117. Rajewsky N, Socci ND (2004) Computational identification of microRNA targets. Developmental Biology 267: 529-535.

118. Kiriakidou M, Nelson PT, Kouranov A, Fitziev P, Bouyioukos C, et al. (2004) A combined computational-experimental approach predicts human microRNA targets. Genes Dev 18: 1165-1178.

119. Lall S, Grün D, Krek A, Chen K, Wang Y-L, et al. (2006) A Genome-Wide Map of Conserved MicroRNA Targets in C. elegans. Current Biology 16: 460-471.

120. Stark A, Brennecke J, Bushati N, Russell RB, Cohen SM (2005) Animal MicroRNAs Confer Robustness to Gene Expression and Have a Significant Impact on 3'UTR Evolution. Cell 123: 1133-1146.

121. Krek A, Grün D, Poy MN, Wolf R, Rosenberg L, et al. (2005) Combinatorial microRNA target predictions. Nature Genetics 37: 495-500.

122. Rhoades MW, Reinhart BJ, Lim LP, Burge CB, Bartel B, et al. (2002) Prediction of plant microRNA targets. Cell 110: 513-520.

123. Dsouza M, Larsen N, Overbeek R (1997) Searching for patterns in genomic data. Trends in Genetics 13: 497-498.

124. Jones-Rhoades MW, Bartel DP (2004) Computational Identification of Plant MicroRNAs and Their Targets, Including a Stress-Induced miRNA Molecular Cell 14: 787-799.

125. Akutsu T, Miyano S, Kuhara S (2000) Inferring qualitative relations in genetic networks and metabolic pathways. Bioinformatics 16: 727-734.

126. Yuh C-H, Bolouri H, Davidson EH (1998) Genomic Cis-Regulatory Logic: Experimental and Computational Analysis of a Sea Urchin Gene. Science 279: 1896-1902.

127. Bussemaker HJ, Li H, Siggia ED (2001) Regulatory element detection using correlation with expression. Nature Genetics 27: 167 - 174.

128. Beaumont MA, Rannala B (2004) The Bayesian Revolution in Genetics. Nature Reviews Genetics 5: 251-261.

129. Pearl J (1988) Probabilistic Reasoning in Intelligent Systems.

130. Cooper GF, Herskovits E (1992) A Bayesian Method for the Induction of Probabilistic Networks from Data. Machine Learning. pp. 309-347.

131. Breiman L (1996) Bagging predictors. Machine Learning 24.

132. Neapolitan RE (2004) Learning Bayesian networks (artificial intelligence). New York: Prentice-Hall.

133. Friedman N, Linial M, Nachman I, Pe'er D (2000) Using Bayesian Networks to Analyze Expression Data. Journal of Computational Biology 7: 601-620.

134. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, et al. (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nature Genetics 34: 166-176.

135. Nachman I, Regev A, Friedman N (2004) Inferring quantitative models of regulatory networks from expression data. Bioinformatics 20: i248-256.

136. Segal E, Taskar B, Gasch A, Friedman N, Koller D (2001) Rich probabilistic models for gene expression. Bioinformatics 17: S243-252.

137. Cheeseman P, Stutz J (1996) Bayesian Classification (AutoClass): Theory

and Results. Advances in Knowledge Discovery and Data Mining. Menlo Park, CA: AAAI Press. pp. 153-180.

138. Sabatti C, James GM (2006) Bayesian sparse hidden components analysis for transcription regulation networks. Bioinformatics 22: 739-746.

139. Durbin R, Eddy S, Krogh A, Mitchison G (1998) Biological sequence analysis -- Probabilistic models of proteins and nucleic acids. Cambridge, United Kingdom: Cambridge University Press. 63-65 p.

140. Baum LE (1972) An equality and associated maximization techique in statistical estimation for probabilistic functions of Markov processes. Inequalities 3: 1-8.

141. Forney GD, Jr. (1973) The viterbi algorithm. Proceedings of the IEEE 61: 268- 278.

142. Rodin AS, Boerwinkle E (2005) Mining genetic epidemiology data with Bayesian networks I: Bayesian networks and example application (plasma apoE levels). Bioinformatics 21: 3273-3278.

143. Madden MG. The Performance of Bayesian Network Classifiers Constructed using Different Techniques; 2003.

144. Heckerman D (1995) A Tutorial on Learning with Bayesian Networks. Redmond: Microsoft Research, Advanced Technology Division, Microsoft Corporation. MSR-TR-95-06 MSR-TR-95-06.

145. Wienholds E, Plasterk RHA (2005) MicroRNA function in animal development. FEBS Letters 579: 5911-5922.

146. Lee J-Y, Colinas J, Wang JY, Mace D, Ohler U, et al. (2006) Transcriptional and posttranscriptional regulation of transcription factor expression in Arabidopsis roots. PNAS 103: 6055-6060.

147. Yekta S, Shih Ih, Bartel DP (2004) MicroRNA-Directed Cleavage of HOXB8 mRNA. Science 304: 594-596.

148. Vaucheret H, Vazquez F, Crete P, Bartel DP (2004) The action of ARGONAUTE1 in the miRNA pathway and its regulation by the miRNA pathway are crucial for plant development. Genes Dev 18: 1187-1197.

149. Lim LP, Lau NC, Garrett-Engele P, Grimson A, Schelter JM, et al. (2005) Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. Nature 433: 769-773.

150. Jones-Rhoades MW, Bartel DP, Bartel B (2006) MicroRNAs and their regulatory roles in plants. Annual Review of Plant Biology 57: 19-53.

151. Bao N, Lye K-W, Barton MK (2004) MicroRNA Binding Sites in Arabidopsis Class III HD-ZIP mRNAs Are Required for Methylation of the Template Chromosome. Developmental Cell 7: 653-662.

152. Wang X-J, Reyes J, Chua N-H, Gaasterland T (2004) Prediction and identification of Arabidopsis thaliana microRNAs and their mRNA targets. Genome Biology 5: R65.

153. Schwab R, Palatnik JF, Riester M, Schommer C, Schmid M, et al. (2005) Specific Effects of MicroRNAs on the Plant Transcriptome. Developmental Cell 8: 517-527.

154. Griffiths-Jones S (2004) The microRNA Registry. Nucl Acids Res 32: D109-111.

155. Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ (2006) miRBase: microRNA sequences, targets and gene nomenclature. Nucl Acids Res 34: D140-144.

156. Workman C, Krogh A (1999) No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. Nucl Acids Res 27: 4816-4822.

157. Dugad R, Desai UB (1996) A Tutorial on Hidden Markov Models. Urbana, USA: Beckman Institute, ECE Department, University of Illinois. SPANN-96.1 SPANN-96.1.

158. Nam J-W, Shin K-R, Han J, Lee Y, Kim VN, et al. (2005) Human microRNA prediction through a probabilistic co-learning model of sequence and structure. Nucl Acids Res 33: 3570-3581.

159. Sunkar R, Zhu J-K (2004) Novel and Stress-Regulated MicroRNAs and Other Small RNAs from Arabidopsis. Plant Cell 16: 2001-2019.

160. Fujii H, Chiou T-J, Lin S-I, Aung K, Zhu J-K (2005) A miRNA Involved in

Phosphate-Starvation Response in Arabidopsis. Current Biology 15: 2038-2043.

161. Bari R, Datt Pant B, Stitt M, Scheible W-R (2006) PHO2, MicroRNA399, and PHR1 Define a Phosphate-Signaling Pathway in Plants. Plant Physiol 141: 988-999.

162. Allen E, Xie Z, Gustafson AM, Carrington JC (2005) microRNA-Directed Phasing during Trans-Acting siRNA Biogenesis in Plants. Cell 121: 207-221.

163. Lobbes D, Rallapalli G, Schmidt DD, Martin C, Clarke J (2006) SERRATE: a new player on the plant microRNA scene. EMBO Rep 7: 1052-1058.

164. Fahlgren N, Howell MD, Kasschau KD, Chapman EJ, Sullivan CM, et al. (2007) High-Throughput Sequencing of Arabidopsis microRNAs: Evidence for Frequent Birth and Death of MIRNA Genes. PloS ONE 2: e219.

165. Wu G, Poethig RS (2006) Temporal regulation of shoot development in Arabidopsis thaliana by miR156 and its target SPL3. Development 133: 3539-3547.

166. Sieber P, Wellmer F, Gheyselinck J, Riechmann JL, Meyerowitz EM (2007) Redundancy and specialization among plant microRNAs: role of the MIR164 family in developmental robustness. Development 134: 1051-1060.

167. Tiwari SB, Hagen G, Guilfoyle T (2003) The Roles of Auxin Response Factor Domains in Auxin-Responsive Transcription. Plant Cell 15: 533-543.

168. Schruff MC, Spielman M, Tiwari S, Adams S, Fenby N, et al. (2006) The AUXIN RESPONSE FACTOR 2 gene of Arabidopsis links auxin signalling, cell division, and the size of seeds and other organs. Development 133: 251-261.

169. Kuroda H, Takahashi N, Shimada H, Seki M, Shinozaki K, et al. (2002) Classification and Expression Analysis of Arabidopsis F-Box-Containing Protein Genes. Plant Cell Physiol 43: 1073-1085.

170. Lu C, Kulkarni K, Souret FF, MuthuValliappan R, Tej SS, et al. (2006)

MicroRNAs and other small RNAs enriched in the Arabidopsis RNA-dependent RNA polymerase-2 mutant. Genome Res 16: 1276-1288.

171. Aida M, Ishida T, Fukaki H, Fujisawa H, Tasaka M (1997) Genes Involved in Organ Separation in Arabidopsis: An Analysis of the cup-shaped cotyledon Mutant. Plant Cell 9: 841-857.

172. McConnell JR, Emery J, Eshed Y, Bao N, Bowman J, et al. (2001) Role of PHABULOSA and PHAVOLUTA in determining radial patterning in shoots. Nature 411: 709-713.

173. Gocal GFW, Sheldon CC, Gubler F, Moritz T, Bagnall DJ, et al. (2001) GAMYB-like Genes, Flowering, and Gibberellin Signaling in Arabidopsis. Plant Physiol 127: 1682-1693.

174. Lurin C, Andres C, Aubourg S, Bellaoui M, Bitton F, et al. (2004) Genome-Wide Analysis of Arabidopsis Pentatricopeptide Repeat Proteins Reveals Their Essential Role in Organelle Biogenesis. Plant Cell 16: 2089-2103.

175. Luo D, Carpenter R, Vincent C, Copsey L, Coen E (1996) Origin of floral asymmetry in Antirrhinum. Nature 383: 794 - 799.

176. Cubas P, Lauter N, Doebley J, Coen E (1999) The TCP domain: a motif found in proteins regulating plant growth and development. The Plant Journal 18: 215-222.

177. Doebley J, Stec A, Gustus C (1995) teosinte branched1 and the Origin of Maize: Evidence for Epistasis and the Evolution of Dominance. Genetics 141: 333-346.

178. Faivre-Rampant O, Bryan GJ, Roberts AG, Milbourne D, Viola R, et al. (2004) Regulated expression of a novel TCP domain transcription factor indicates an involvement in the control of meristem activation processes in Solanum tuberosum. J Exp Bot 55: 951-953.

179. Okamuro JK, Caster B, Villarroel R, Van Montagu M, Jofuku KD (1997) The AP2 domain of APETALA2 defines a large new family of DNA binding proteins in Arabidopsis. PNAS 94: 7076-7081.

180. Klein J, Saedler H, Huijser P (1996) A new family of DNA binding proteins includes putative transcriptional regulators of the Antirrhinum majus floral

meristem identity gene SQUAMOSA. Mol Gen Genet 250: 7-16.

181. Axtell MJ, Bartel DP (2005) Antiquity of MicroRNAs and Their Targets in Land Plants. Plant Cell 17: 1658-1673.

182. Bartel B, Bartel DP (2003) MicroRNAs: At the Root of Plant Development? Plant Physiol 132: 709-717.

183. Miranda KC, Huynh T, Tay Y, Ang Y-S, Tam W-L, et al. (2006) A Pattern-Based Method for the Identification of MicroRNA Binding Sites and Their Corresponding Heteroduplexes. Cell 126: 1203-1217.

184. Bonnet E, Wuyts J, Rouze P, Van de Peer Y (2004) Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. Bioinformatics 20: 2911-2917.

185. Rajagopalan R, Vaucheret H, Trejo J, Bartel DP (2006) A diverse and evolutionarily fluid set of microRNAs in Arabidopsis thaliana. Genes Dev 20: 3407-3425.

186. Adai A, Johnson C, Mlotshwa S, Archer-Evans S, Manocha V, et al. (2005) Computational prediction of miRNAs in Arabidopsis thaliana. Genome Res 15: 78-91.

187. Farh KK-H, Grimson A, Jan C, Lewis BP, Johnston WK, et al. (2005) The Widespread Impact of Mammalian MicroRNAs on mRNA Repression and Evolution. Science 310: 1817 - 1821.

188. Sethupathy P, Megraw M, Hatzigeorgiou AG (2006) A guide through present computational approaches for the identification of mammalian microRNA targets. Nature Methods 3: 881 - 886.

189. Chen K, Rajewsky N (2007) The evolution of gene regulation by transcription factors and microRNAs. Nat Rev Genet 8: 93-103.

190. Chen K, Rajewsky N (2006) Natural selection on human microRNA binding sites inferred from SNP data. Nat Genet 38: 1452-1456.

191. Eddy SR (2001) Non-coding RNA genes and the modern RNA world. Nat Rev Genet 2: 919-929.

192. Mattick JS (2004) RNA regulation: a new genetics? Nature Reviews

Genetics 5: 316-323.

193. Hobert O (2004) Common logic of transcription factor and microRNA action. Trends in Biochemical Sciences 29: 462-468.

194. Sood P, Krek A, Zavolan M, Macino G, Rajewsky N (2006) Cell-type-specific signatures of microRNAs on target mRNA expression. PNAS 103: 2746-2751.

195. Lai EC (2002) Micro RNAs are complementary to 3' UTR sequence motifs that mediate negative post-transcriptional regulation. Nature Genetics 30: 363 - 364.

196. Kim S-Y, Kim Y (2006) Genome-wide prediction of transcriptional regulatory elements of human promoters using gene expression and promoter analysis data. BMC Bioinformatics 7: 330.

197. Bar-Joseph Z, Gerber GK, Lee TI, Rinaldi NJ, Yoo JY, et al. (2003) Computational discovery of gene modules and regulatory networks. Nat Biotechnol 21: 1337-1342.

198. Kundaje A, Middendorf M, Shah M, Wiggins C, Freund Y, et al. (2006) A classification-based framework for predicting and analyzing gene regulatory response. BMC Bioinformatics 7: S5.

199. Barabasi AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5: 101-113.

200. Shin B, Choi G, Yi H, Yang S, Cho I, et al. (2002) AtMYB21, a gene encoding a flower-specific transcription factor, is regulated by COP1. The Plant Journal 30: 23-32.

201. Ma L, Li J, Qu L, Hager J, Chen Z, et al. (2001) Light Control of Arabidopsis Development Entails Coordinated Regulation of Genome Expression and Cellular Pathways. Plant Cell 13: 2589-2607.

202. Schliep A, Steinhoff C, Schonhuth A (2004) Robust inference of groups in gene expression time-courses using mixtures of HMMs. Bioinformatics 20: i283-289.

203. Costa IG, Schonhuth A, Schliep A (2005) The Graphical Query Language: a tool for analysis of gene expression time-courses. Bioinformatics 21:

2544-2545.

204. Zhou X, Wang G, Zhang W (2007) UV-B responsive microRNA genes in Arabidopsis thaliana. Mol Syst Biol 3: 103.

205. Kimura M, Yoshizumi T, Manabe K, Yamamoto YY, Matsui M (2001) Arabidopsis transcriptional regulation by light stress via hydrogen peroxide-dependent and -independent pathways. Genes to Cells 6: 607-617.

206. Kimura M, Manabe K, Abe T, Yoshida S, Matsui M, et al. (2003) Analysis of Hydrogen Peroxide–independent Expression of the High-light–inducible ELIP2 Gene with the Aid of the ELIP2 Promoter–Luciferase Fusion. Photochemistry and Photobiology 77: 668-674.

207. Apel K, Hirt H (2004) REACTIVE OXYGEN SPECIES: Metabolism, Oxidative Stress, and Signal Transduction. Annual Review of Plant Biology 55: 373-399.

208. Gao Y, Li J, Strickland E, Hua S, Zhao H, et al. (2004) An Arabidopsis Promoter Microarray and its Initial Usage in the Identification of HY5 Binding Targets in Vitro. Plant Molecular Biology 54: 683-699.

209. Lescot M, Dehais P, Thijs G, Marchal K, Moreau Y, et al. (2002) PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. Nucl Acids Res 30: 325-327.

210. Rombauts S, Dehais P, Van Montagu M, Rouze P (1999) PlantCARE, a plant cis-acting regulatory element database. Nucl Acids Res 27: 295-296.

211. Mallory AC, Vaucheret H (2006) Functions of microRNAs and related small RNAs in plants. Nature Genetics 38: S31 - S36.

212. Lee J, He K, Stolc V, Lee H, Figueroa P, et al. (2007) Analysis of Transcription Factor HY5 Genomic Binding Sites Revealed Its Hierarchical Role in Light Regulation of Development. Plant Cell 19: 731-749.

213. Arnim AGv, Deng X-W (1996) Light inactivation of arabidopsis photomorphogenic repressor COP1 involves a cell-specific regulation of its nucleocytoplasmic partitioning. Cell 79: 1035-1045.

214. Neff MM, Fankhauser C, Chory J (2000) Light: an indicator of time and place. Genes Dev 14: 257-271.

215. Shin B, Choi G, Yi H, Yang S, Cho I, et al. (2002) AtMYB21, a gene encoding a flower-specific transcription factor, is regulated by COP1. The Plant Journal 30: 23-32.

216. Osterlund MT, Ang L-H, Deng XW (1999) The role of COP1 in repression of Arabidopsis photomorphogenic development. Trends in Cell Biology 9: 113-118.

217. Miséra S, Müller AJ, Weiland-Heidecker U, Jürgens G (1994) The FUSCA genes of Arabidopsis: negative regulators of light responses. Molecular and General Genetics MGG 244: 242-252.

218. Kwok SF, Piekos B, Misera S, Deng XW (1996) A Complement of Ten Essential and Pleiotropic Arabidopsis COP/DET/FUS Genes Is Necessary for Repression of Photomorphogenesis in Darkness. Plant Physiol 110: 731-742.

219. Wei N, Deng XW (1996) The Role of the COP/DET/FUS Genes in Light Control of Arabidopsis Seedling Development. Plant Physiol 112: 871-878.

220. Mayer R, Raventos D, Chua NH (1996) det1, cop1, and cop9 Mutations Cause Inappropriate Expression of Several Gene Sets. Plant Cell 8: 1951-1959.

221. Karniol B, Malec P, Chamovitz DA (1999) Arabidopsis FUSCA5 Encodes a Novel Phosphoprotein That Is a Component of the COP9 Complex. Plant Cell 11: 839-848.

222. Serino G, Tsuge T, Kwok S, Matsui M, Wei N, et al. (1999) Arabidopsis cop8 and fus4 Mutations Define the Same Gene That Encodes Subunit 4 of the COP9 Signalosome. Plant Cell 11: 1967-1980.

223. Staub JM, Wei N, Deng XW (1996) Evidence for FUS6 as a Component of the Nuclear-Localized COP9 Complex in Arabidopsis. Plant Cell 8: 2047-2056.

224. Wei N, Deng X-W (1999) Making sense of the COP9 signalosome: a regulatory protein complex conserved from Arabidopsis to human. Trends

in Genetics 15: 98-103.

225. D'haeseleer P (2006) What are DNA sequence motifs? Nat Biotechnology 24: 423-425.

226. Zhong W, Sternberg PW (2006) Genome-Wide Prediction of C. elegans Genetic Interactions. Science 311: 1481-1484.

227. Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, et al. (2003) A Bayesian Networks Approach for Predicting Protein-Protein Interactions from Genomic Data. Science 302: 449-453.

228. Foat BC, Houshmandi SS, Olivas WM, Bussemaker HJ (2005) Profiling condition-specific, genome-wide regulation of mRNA stability in yeast. PNAS 102: 17675-17680.

229. Li Y, Lee KK, Walsh S, Smith C, Hadingham S, et al. (2006) Establishing glucose- and ABA-regulated transcription networks in Arabidopsis by microarray analysis and promoter classification using a Relevance Vector Machine. Genome Res 16: 414-427.

230. Zou M, Conzen SD (2005) A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. Bioinformatics 21: 71-79.

231. Jiao Y, Lau OS, Deng XW (2007) Light-regulated transcriptional networks in higher plants. Nature Reviews Genetics 8: 217-230.

232. Chattopadhyay S, Ang L-H, Puente P, Deng X-W, Wei N (1998) Arabidopsis bZIP Protein HY5 Directly Interacts with Light-Responsive Promoters in Mediating Light Control of Gene Expression. Plant Cell 10: 673-684.

233. Cao D, Lin Y, Cheng C-L (2000) Genetic Interactions between the Chlorate-Resistant Mutant cr 8 8 and the Photomorphogenic Mutants cop1 and hy5. Plant Cell 12: 199-210.

234. Ang L-H, Chattopadhyay S, Wei N, Oyama T, Okada K, et al. (1998) Molecular Interaction between COP1 and HY5 Defines a Regulatory Switch for Light Control of Arabidopsis Development. Molecular Cell 1: 213-222.

235. Ang LH, Deng XW (1994) Regulatory Hierarchy of Photomorphogenic Loci:

Allele-Specific and Light-Dependent Interaction between the HY5 and COP1 Loci. Plant Cell 6: 613-628.

236. Oyama T, Shimura Y, Okada K (1997) The Arabidopsis HY5 gene encodes a bZIP protein that regulates stimulus-induced development of root and hypocotyl Genes Dev 11: 2983-2995.

237. Hiesinger PR, Hassan BA (2005) Genetics in the Age of Systems Biology. Cell 123: 1173-1174.