Inference of gene regulation from expression datasets

A Thesis

Submitted to the Faculty

of

Drexel University

by

Yiqian Zhou

in partial fulfillment of the

requirements for the degree

of

Doctor of Philosophy

June 2015

Dedications

To my Mom and Dad

Acknowledgments

I would like to thank my advisor, Dr. Ahmet Sacan, who has always been giving me tireless support, taught me how to tackle difficult problems efficiently, and helped me develop my skills as a researcher.

I would like to thank my committee members, Dr. Seena Ajit, Dr. Lin Han, Dr. Uri Hershberg and Dr. Andres Kriete, for their helpful review and constructive criticism.

I would like to thank my lovely lab mates, Francis Bell, Rehman Qureshi, Heng Yang and Chunyu Zhao.

Thank you all.

Table of Contents

List of Tables

List of Figures

Abstract

Inference of gene regulation from expression datasets
Yiqian Zhou

The development of high throughput techniques and the accumulation of large scale gene expression data provide researchers great opportunities to more efficiently solve important but complex biological problems, such as reconstruction of gene regulatory networks and identification of miRNA-target interactions. In the past decade, many algorithms have been developed to address these problems. However, prediction and simulation of gene expression data have not yet received as much attention. In this study, we present a model based on stepwise multiple linear regression (SMLR) that can be applied for prediction and simulation of gene expression, as well as reconstruction of gene regulatory networks by analysis of time-series gene expression data, and we present its application in analysis of paired miRNA-mRNA expression data.

Chapter 1. Introduction

1.1. Gene regulatory network

1.1.1. The Importance of Biochemical Interaction Networks

Life can be regarded as a complex system in which genes, gene products, and other metabolites interact with each other. Uncovering and understanding these biochemical interactions are essential in system biology. With the accumulation of known interactions, one can organize them into a biochemical regulatory network. By reconstruction of regulatory network, which may contain a large number of components, scientists obtain a wider view of the biological system and a better understanding of its dynamic nature. The availability of regulatory networks helps scientists obtain answers for questions like: how does a specific biological system respond to external stimulus or treatment; what is the stable state of a cellular process under certain conditions; and how a biological process will behave if some portion of the system were abnormal? With the insights gained from biochemical regulatory network, scientists have the ability to understand, control and optimize biological systems, which leads to many practical applications in biotechnology and medicine.

1.1.2. Gene regulatory network (GRN)

In a typical biochemical gene regulatory network (Figure 1.1a), different kinds of biochemical interactions may take place at different levels, such as DNA level, transcript level and protein level. Reconstruction of a real regulatory network describing these biochemical interactions requires different kinds of knowledge and experimental data. The integration of different information is a complex process. Feist et al. [1] described a detailed framework in which various experimental data are integrated to reconstruct biochemical networks in microorganisms.

A gene regulatory network (GRN) is an abstract representation of the complex biochemical gene regulatory system (Figure 1.1b). Instead of describing the different types of real biochemical interactions across different levels, GRN is a simplified model describing gene-to-gene interactions [2], which are indirect relationships skipping intermediate proteins, non-coding RNAs, and other metabolites. GRN can be represented as a graph, in which nodes represent genes and edges represent relationship between genes (Figure 1.1b).

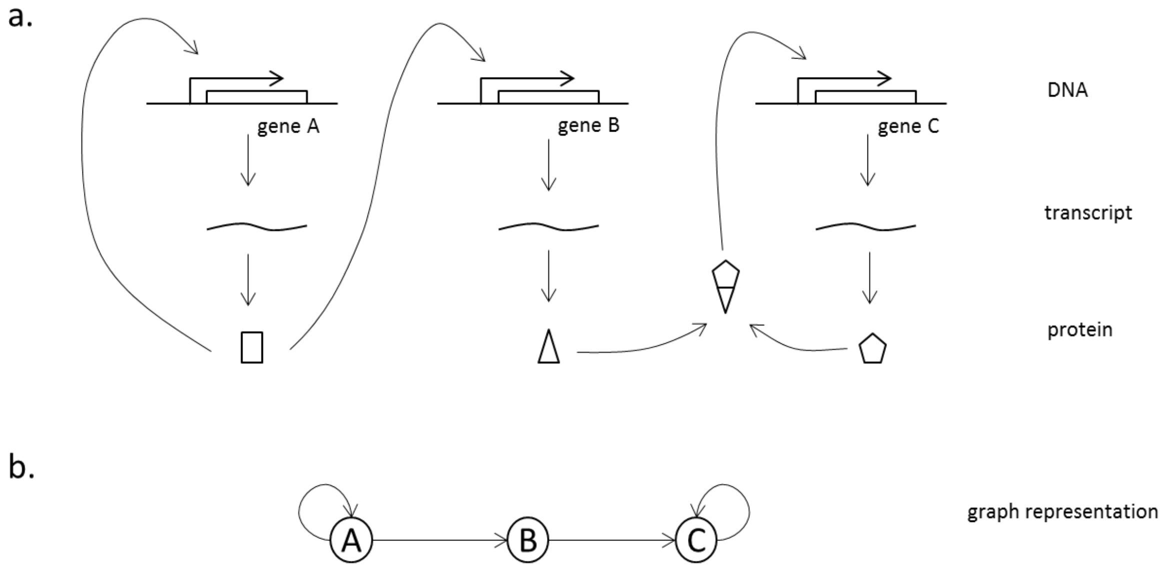


Figure 1.1. A simple regulatory gene network. Gene regulatory network (GRN) can be regarded as an abstract representation of a real biochemical regulatory network. (a) The regulation of gene expression may happen at different levels, and there are different types of interactions. (b) A gene regulatory network is an abstract representation describing how genes influence each other.

1.1.3. Reconstruction of GRN

By using the expression data of thousands of genes obtained from high throughput technologies, such as microarrays and RNA-Seq [3], it is possible to study how expression of a gene relates to expressions of other genes, even in the absence of direct data on the concentrations of protein products or other metabolites. The expression of genes can indirectly affect the expression of other genes: a gene can inhibit or stimulate expression of another gene. These activation and inhibition relationships can be represented as a directed graph with nodes representing genes and edges representing the effect of one gene on another. There are several methods of reverse-engineering for modeling gene regulatory networks from gene expression experiments. These methods include Boolean networks [2, 4-6], correlation networks [7-10], differential equation models [11-16], Bayesian network models [2, 7, 17]. We will give a brief introduction of each method in chapter 2.1.

1.2. Microarray

1.2.1. Introduction of Microarray

The advent of microarray technologies has enabled a high throughput evaluation of gene expression, providing a large scale snapshot of the cellular activity at the molecular level. In microarray experiments, gene expression is quantified by determining the relative amounts of mRNA transcripts. Firstly mRNA is harvested from a sample and then reverse-transcribed into cDNA. This cDNA is labeled with a fluorescent molecule and then allowed to bind to DNA probes that are attached to the surface of a microarray chips. cDNA is captured specifically by its complementary probe by hydrogen bonds formed between them. After washing-off non-specific cDNA sequence, the chips are scanned and the fluorescence values are measured to quantify relative amounts of specific cDNA present, thus determines the relative gene expression. Microarray has allowed researchers to explore the behavior of entire transcriptome under different experimental conditions, in a search for mechanistic basis of various cellular behaviors. Analysis of these microarray experimental results has led to new breakthroughs in the understanding, diagnosis, prognosis, and treatment of disease, as well as insights into the functioning of the basic biology of various organisms.

1.2.2. Time series gene expression data

Time series microarray experiments involve harvesting mRNA from samples at regular time-intervals. This experimental design evaluates gene expression over time in a high-throughput manner. Time-series expression data has the potential to provide more comprehensive information about the underlying behavior and inter-relationships of genes than the traditional time-invariant experiments. Furthermore, it allows interpretations of dynamic behaviors of complex biological systems [18]. Time-series microarray data has many applications, including the analysis of circadian rhythms, disease progression, drug response, and the study of the cell cycle [18-20]. In chapter 2.1, we review methods for reconstruction of GRN using time series data.

In addition to the inference of GRN structure, time series data can also be used for learning how regulation takes place at different time points of the experiment. Ernst et al. developed a method called Dynamic Regulatory Events Miner (DREM) to study how cells respond to stimuli, by combining times series gene expression data and TF-DNA binding data [21] (http://www.sb.cs.cmu.edu/drem/). In their study, a hidden Markov model is used with each hidden state associated with one time point. During training, a tree structure where each state is allowed to have two children is enforced to model bifurcation events in the time course. After training, genes are assigned to their most likely paths in the tree structure based on their expression profiles. Then TFs are associated to different paths of the splits by enrichment score of the genes that they target. This approach is used to infer the activation times of TFs and examine how they regulate the response to stimuli.

Unlike DREM, which groups genes based on expression profiles over the entire time course, Zaslavsky et al. developed an approach named Time-Dependent Activity Linker (TIDAL) [22] (http://tsb.mssm.edu/primeportal/?q=tidal_prog) that identifies genes with common patterns based on the initial up-regulation time and ignores down-regulation based on the observation that timing of changes for down-regulated genes is not correlated across experimental replicates. TIDAL associates TFs to different phases of the time series, in a similar way as DREM, by statistical enrichment of TF target genes. Because of the scarcity of identified TF targets, TF targets were identified computationally from the presence of TF-target binding sites. TFs are then connected into a transcriptional cascade in a temporal manner, such that a TF is placed in the time slice in which its mRNA is first up-regulated in the time series.

1.2.3. Data-driven modeling

The development of microarray technologies has enabled a high throughput evaluation of gene expression, providing large scale quantitative data of cellular activity at the molecular level. The availability of large scale gene expression data has changed the starting point of the knowledge generation cycle of individual biological regulatory networks and spawned the development of data-driven modeling of biological systems [23]. Traditionally, a hypothesis was constructed from background knowledge and tested by experiments, and then the hypothesis was either verified or modified for further experimental testing. This cycle was repeated until the solution of the problem was satisfactory (Figure 1.2). However, when there are many possible testable hypotheses and their detailed experimental validation is not feasible, it is more practical and resource effective to generate and prioritize candidate hypotheses from the prior data using computer based inductive reasoning. Inference of biological regulatory networks is a complex task and the search space of possible interactions is far too large. There is, however, a great amount of gene expression data accumulated from microarray studies, thus how to utilize those data to generate sound hypotheses is of great interest. Data-driven inference of a network, which generates a reasonable hypothesis, is a key step in understanding the biological systems.

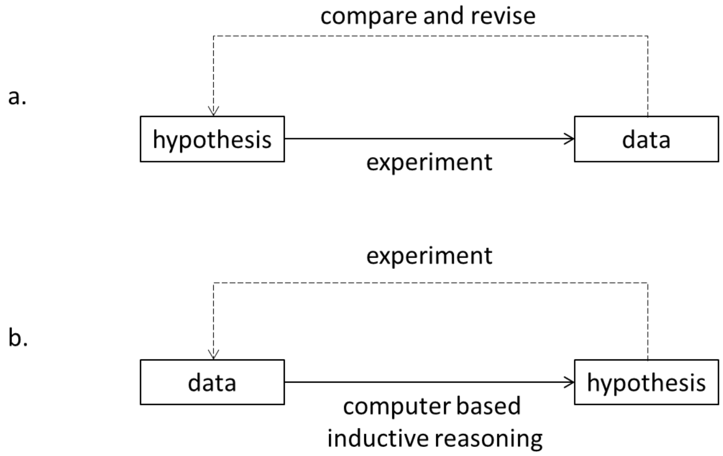


Figure 1.2. Cycle of knowledge. The availability of large scale gene expression data changed the start point of the cycle of knowledge generation. (a) Traditionally, a hypothesis is generated based on background knowledge and then tested by experiments. (b) With the large amount of available gene expression data and little suitable background knowledge, the hypothesis is generated based on the analysis of data.

1.3. Micro RNA

1.3.1. microRNA as fine tuner

MicroRNAs (miRNAs) are small (~22 nucleotide) non-coding endogenous RNA that play important roles in gene-regulatory events in both animals and plants, by pairing the messenger RNA (mRNA) of protein-coding genes [24]. miRNAs participate in a wide range of biological process [25] and it is predicted that miRNAs affect the expression of over 60% of mammalian gene [26], and they are regarded as fine-tuners that adjust the expression of protein-coding genes to optimize their expression patterns [27, 28]. Over the past decade, it has become clear that miRNAs contribute to almost all known physiological and pathological processes, in which cancer is of particular interests. Since the dysregulation of miRNA gene expression controls/ are controlled by the dysregulation of multiple oncogenes or tumour suppressors, studying the biological process of miRNA provide important opportunities for the development of miRNA-based diagnosis and treatment of cancer [29, 30].

1.3.2. Experimental identification of miRNA targets

To understand the functions of miRNAs, a central goal and major challenge is to determine their target mRNAs. There are many experimental techniques being used for target identification of miRNAs. These techniques apply different strategies that focusing on different components in the miRNA-directed regulation, as shown in Figure 1.3. These techniques can be categorized them as (1) transcriptome analysis, (2) proteome analysis, (3) RLM-RACE, (4) translation profiling, (5) miRNA manipulation, and (6) immunoprecipitation, which manipulate protein, mRNA, cleaved mRNA, ribosome, miRNA and RNA-induced silencing complex (RISC) respectively.

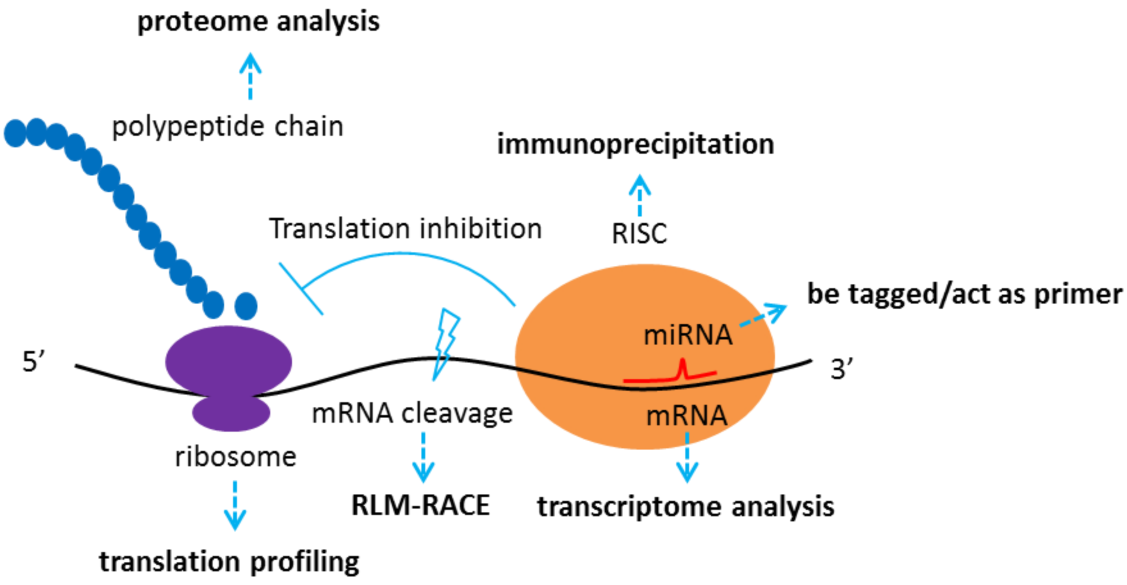


Figure 1.3. Categories of experimental techniques for miRNA target identification. These techniques are categorized based on which component in miRNA-directed regulation they focus on. Modified from [31]

1.3.2.1. Transcriptome analysis

To screen for miRNA targets, the most common strategy is to differentially express a single miRNA. On one hand, over-expression of miRNA can be performed by transfection with in vitro synthesized ds-RNA precursor, which mimic the endogenous miRNAs. After that, transcriptome analysis (or proteome analysis, discussed below in 1.3.2.2) is used to identify the mRNAs (or proteins, discussed below in 1.3.2.2), whose expression levels are affected as a consequence of miRNA over-expression. To measure the expression of mRNA, there are essentially three techniques, real-time quantitative PCR (qPCR), microarrays and next generation sequencing (NGS) (RNA-seq for example). Take cost and scale into consideration, currently microarray is mostly widely used for routine studies. For example, Lim et al. transfected miRNAs into human cells and examined changes in mRNAs using microarray, and found that miRNA reduce the level of large number of target mRNA transcripts [32]. Alternatively to microarray, Xu et al. used NGS to analyze transcriptome changes induced by the human miR-155 [33].

On the other hand, silencing of miRNA functions can be performed by gene knockout or expression of antisense that bind mature miRNA. For example, Sekine et al. disrupted Dicer, an enzyme essential for miRNA processing, to study the consequences of loss of miRNAs in conditional knockout mouse livers [34]. Krutzfeldt et al. designed chemically modified, cholesterol-conjugated single-stranded RNA analogues termed ‘antagomirs’ that are complementary to miRNAs for in vivo silencing, and showed that intravenous administration of antagomirs resulted in a marked decrease of corresponding miRNA levels [35]. Elmén et al. reported antagonism of miR-122 using unconjugated LNA (locked nucleic acid)-antimiR oligonucleotide based on the stable heteroduplexes between the LNA-antimiR and miR-122 [36]. Ebert et al. developed miRNA inhibitors termed ‘microRNA sponges’ that can be expressed in cells [37]. When expressed, these competitive inhibitors containing multiple, tandem binding sites to a miRNA of interests, and those binding sites are designed such that they can block an entire miRNA seed family. They showed microRNA sponges repress miRNA targets no less strongly compared to chemically modified antisense oligonucleotides. Haraguchi et al. developed RNA decoys, TuD RNAs (tough decoy RNAs), to achieve long-term suppression of miRNA [38]. The inhibitory RNAs are designed to be expressed in lentiviral vectors and to be transported into cytoplasm. They demonstrated that TuD RNA induce efficient suppression of specific miRNAs for more than one month in mammalian cells.

1.3.2.2. Proteome analysis

Besides measuring mRNA expression, there are also proteomic approaches for studying miRNA target regulation. Proteomic approaches can be employed to measure the ‘final effect’ of miRNAs since miRNAs regulate gene expression by both mRNA cleavage and translational repression.

Stable isotope labeling with amino acids in cell culture (SILAC) is a technique based on mass spectrometry that measure relative protein abundance among samples labeled with stable isotopes. Proteins are labeled by growing cells in medium containing amino acids labeled with heavy isotopes. Difference in protein abundance can be determined by the ratio of peak intensities in the mass spectrum. Vinther et al. applied SILAC to investigate the effect of miRNA-1 on HeLa cell proteome and found 12 out of 504 detected proteins were repressed by miRNA-1 transfection [39]. Yang et al. employed SILAC to identify targets of miR-143 and found 93 out of over 1200 identified proteins down-regulated more than 2-fold in miR-143 mimic transfected MiaPaCa2 pancreatic cancer cells as compared to controls [40]. Baek et al. [28] and Selbach et al. [41] both measured the response of thousands of proteins after miRNA transfection or endogeneous miRNA knockdown, by applying SILAC and pSILAC (cells in two samples are pulse-labeled with two different heavy versions of amino acids so that newly synthesized proteins will be ‘heavy’ or ‘medium-heavy’) respectively. Based on those large-scale study, they both found that single miRNAs can repress the production of hundreds of proteins and that most repressions are modest, which suggest that miRNAs act as fine-tuner for protein synthesis.

Two-dimensional differentiation in-gel electrophoresis (2D-DIGE) technique was also applied to identify miRNA targets. Proteins from two samples are labeled with different fluorescent dyes and then separated in two dimensions on gel by isoelectric focusing and SDS-PAGE. After that, proteins are identified by mass spectrometry. Zhu et al. applied 2D-DIGE to analysis total proteins extracted from cells treated with and without inhibitor of miR-21 (antisense miR-21) and identified tumor suppressor tropomyosin 1 (TPM1) as a potential miR-21 target [42].

1.3.2.3. RLM-RACE

RNA ligase mediated-5′ rapid identification of cDNA ends (5’ RLM-RACE) is a technique that can be used to identify miRNA targeting where target mRNA is directly cleaved. In 5’ RLM-RACE procedure, by using T4 RNA ligase, RNA adaptor can covalently attach to uncapped 5’ end of mRNA produced from Ago2-directed cleavage. After this ligation, with a forward primer complementary to the adaptor and a gene specific reverse primer, the RNA can be reverse transcript, subsequently PCR amplified and identified. Applying RLM-RACE, Yekta et al. validated the miR-196 directed cleavage of HOXB8 in mouse embryos [43]. Franco-Zorrilla et al. used RLM-RACE combining microarray to identify small RNA targets [44]. In their study, miRNA/siRNA-mediated-cleaved transcripts were isolated by RLM-RACE and then subjected amplifications and microarray hybridizations. An approach named Parallel analysis of RNA ends (PARE) is developed to identify products from 5’ RLM-RACE in large scale by high-throughput sequencing. Bracken et al. performed PARE to detect potential miRNA-directed mRNA cleavages in mouse embryo and adult tissues and found that numerous mRNA are potentially cleaved at low level by endogenous miRNAs [45].

1.3.2.4. Translation profiling

The analysis of mRNA associate with polysome profiling provides researchers information about the targets of miRNA and mechanism of miRNA-mediated translational repression. Nakamoto et al. proposed a method for identification of target mRNA by detecting the shifts in mRNA abundance in polysome profiles after miRNA knockdown [46]. They assumed that the position of mRNA in polysome profiles partially reflects the degree of its translation. Combining miRNA knockdown and microarray analysis, mRNA moving toward heavier polysome reflects the relief of miRNA-mediated translational repression. Guo et al. used ribosome profiling to measure the effect of miRNA on translational efficiency [47]. Together with the measure of change of mRNA level from microarray analysis, they found mRNA found that lowered mRNA level account for most of the decreased protein output.

1.3.2.5. miRNA manipulation

There are several approaches focusing on direct manipulation of miRNA.

Biotin-tagged

Ørom et al. presented an approach for experiential identification of miRNA targets based on affinity purification of tagged miRNAs [48]. In this method, they transfected cells with biotinylated miRNA duplexes and captured miRNA-RISC-mRNA complexes using streptavidin-sepharose beads. After that, RNAs were isolated and ready for downstream analysis by qRT-PCR or microarray. By employing this direct affinity-based procedure, Ørom et al. showed miR-10a interacts with the 5'-UTR region of ribosomal proteins encoding mRNAs to enhance their translation [49]. This method can be applied to examine the direct interaction of miRNA with its targets [50, 51] .With this technique, researchers may pull down targets of a particular miRNA of interest. But the potential limitations include that it is unknown how biotin tag affect the miRNA binding and whether it capture miRNA targets comprehensively.

Digoxigenin-labeled

Hsu et al. developed a simple approach to find potential targets of specific miRNA in vitro, which they termed as Labeled miRNA pull-down (LAMP) assay [52]. In their procedure, precursor miRNA (pre-miRNA) are synthesized and labeled with digoxigenin (DIG) and mixed with cell extracts, in which endogenous Dicer process pre-miRNAs and generates mature miRNAs in vitro. After DIG-labeled miRNA attaching to its target mRNA by the endogenous RISC, DIG-labeled miRNA-mRNA complex are pulled down by anti-DIG antiserum. The isolated RNA is then ready for further process such as RT-PCR, microarray. The LAMP assay is relatively simple and cost-effective, but the DIG label may influence Dicer processing and there are potentially non-specific binding between target mRNA and RISC.

As primer for RT

Since miRNA is (partially) complementary to the 3’-end of target mRNA, it can act as primer for target cDNA by reverse transcription (RT). Based on this principle, Vatolin et al. proposed a novel two-step reverse transcription method to detect miRNA-mRNA interaction in eukaryotic cells [53]. They firstly synthesize cDNA on an mRNA template using mRNA as endogenous cytoplasmic primer. This step extends miRNA and overcome the problem of low complementary miRNA-mRNA binding. In the second of RT, the purified hybrid 3’-cDNA-miRNA-5’ molecules are used to anneal target mRNA specifically. Andachi described a method based on the same idea focus on identification of targets of miRNA of interest [54]. They applied this method to *C. elegans* miRNA lin-4 and successfully detected interactions between miRNA lin-4 and lin-14.

1.3.2.6. Immunoprecipitation

Unlike transcriptome and proteome approaches, which cannot distinguish direct and indirect miRNA effects, biochemical approaches are generally more reliable, since it reveals the direct miRNA-mRNA interactions. miRNA-mRNA pairs can be purified by the immunoprecipitation of the RNA-induced silencing complex (RISC) components. The direct target mRNAs that are co-immunoprecipitated with

RISC can be identified by microarray or deep sequencing. Karginov et al. combined RISC immunopurification with microarray analysis of associated mRNAs for miRNA target discovery [55]. Similarly, Hendrickson et al. transfected HEK293T cells with epitope-tagged Ago2, immunopurified Ago2 associated with miRNAs and mRNAs, and determined the RNAs by microarray [56]. Easow et al. found enrichment of mRNAs containing miRNA seed complementary sites in 3’ by using immunoprecipitation of hemagglutinin (HA)-tagged Ago1 protein in Drosophila Melanogaster Schneider SL2 (S2) cells [57]. Beitzinger et al. used highly specific monoclonal antibodies against members of the Ago protein family to co-immunoprecipitate Ago-bound mRNAs in HEK 293 cells [58]. Tan et al. applied approach called Ribonucleoprotein ImmunoPrecipitation-gene Chip (RIP-Chip) [59]. In their study, wild-type human Ago2 protein is directly immunoprecipitated from untreated cells using antibodies and Ago2-associated mRNA transcripts are analyzed by microarray to identify miRNA-targetome.

There are limitations of these immunoprecipitation approaches. It is possible that the associations of RNA-binding protein and its target mRNA may result from reassociation of molecules subsequence to cell lysis, thus the immunoprecipitattion approaches does not always reflect the in vivo interactions [60]. In addition, while these approaches require interactions stable enough to survive immunoprecipitation process, potential targets may be missed during the process. There are several novel approaches developed that handle these limitations.

Crosslinking Immunoprecipitation

Cross-linking and immunoprecipitation assays (CLIP) is a technique that combines UV cross-linking with immunoprecipitation to study the protein-RNA binding sites. Upon exposure to UV light, covalent bonds are formed between proximal proteins and RNA. The cells are then lysed and the proteins of interest are isolated by immunoprecipitation. Based on CLIP, high-throughput sequencing of RNA isolated by crosslinking immunoprecipitation (HITS-CLIP), also known CLIP-seq, employ high-throughput sequencing techniques to identify RNA segments. By using HITS-CLIP approach, Chi et al. isolated argonaute protein-RNA complexes in mouse brain [61]. They identify interaction sites between miRNAs and target mRNA by analyzing data of Ago-miRNA binding sites and Ago-mRNA binding sites. Zisoulis et al. isolated endogenous mRNA target sequences bound by Argonaute protein ALG-1 in *C. elegans* and identified ALG-1 interactions with both 3′-UTR and coding exon sequences [62].

HITS-CLIP method is limited by the low efficiency of UV 254 nm RNA–protein cross-linking and the difficulty in precisely locating binding sites in sequenced fragments. To address the problem, a variant CLIP method, named photoactivatable-ribonucleoside-enhanced crosslinking and immunoprecipitation (PAR-CLIP) is developed [63]. PAR-CLIP enhances the UV cross-linking efficiency by incubating cells with a photoactivatable nucleoside. And more importantly, PAR-CLIP is capable of identifying cross-link site because of higher frequency of UV-induced mutations in cross-linked than non-cross-linked sites. By using PAR-CLIP, Hafner et al. determined the binding sites for several intensely studied RNA-binding proteins (RBPs) and miRNA-containing ribonucleoprotein complexes (miRNPs) [64].

As for HITS-CLIP, recent advances in data analysis refine the resolution of RNA-binding map [65]. In the HITS-CLIP procedure, the remaining cross-linked amino acids attached RNA impose an obstacle for reverse transcription, thus mutations may induced. By analyzing these cross-link-induced mutation sites (CIMS), RNA-protein interactions at single-nucleotide resolution can be obtained.

1.3.3. Validation of miRNA-mRNA interaction

miRNA-mRNA interactions identified by large scale screening methods described above should be further examined for their authenticity. For individual miRNA-mRNA interaction, researchers may turn to more reliable but also more labor intensive methods such as luciferase reporter assays. In this assay, 3’-UTR of the potential target gene is cloned immediately downstream of the luciferase open reading frame in the reporter plasmid. Altered luciferase activity is measured as a result of manipulation of a targeting miRNA, which demonstrate a direct miRNA effect. To ensure authenticity of a functional miRNA-mRNA pair, Kuhn et al. reviewed several miRNA targets validation methods and proposed multiple criteria for confirmation of miRNA validated targeting [66]. Those criteria include validated miRNA-mRNA interaction, co-expression of both miRNAs and mRNAs, effects of miRNA on amount of target protein, effects of miRNA on biological functions.

There are several databases house collection of experimentally identified miRNA-mRNA interactions, such as TarBase [67], miRTarBase [68], miRecords [69]. Since interactions identified by immunoprecipitation methods generally are more reliable than these identified by transcriptome and proteome analysis, it's reasonable to mention experimental evidence strong or not. In miRTarBase, miRNA-mRNA interactions are deemed as having strong support when they are validated by western blot, qPCR, or reporter assays, and having weak support with pSILAC and microarray experiments [70].

1.3.4. Sequence-based computational methods

So far thousands of miRNAs have been identified in animals and plants. There are 1872 homo sapiens miRNA sequence annotated in mirBase (v20, June 2013) [71]. However, only a small fraction of miRNA targets have been validated experimentally with confidence because of the relatively low efficiency and high cost of experimental procedures. Sequence-based computational methods have been developed to fill the gap by generating putative lists of miRNA-mRNA pairs, which have greatly reduce the number of interactions researchers need to validate.

Most sequence-based methods are based on experimentally determined rules of miRNA-mRNA interactions, including sequence complementarity between the nucleotides 2-7 of miRNA (called 'seed region') and 3'-UTR of the mRNAs [72], energetically favorable hybridization [73], evolutionary conservation of the binding sites among different species [74], RNA secondary structure accessibility [75] and multiple target sites [76]. Based on these rules, many approaches have been developed. Widely used methods includes TargetScan [72], miRanda [77], PicTar [78], TargetScanS [79], PITA [75], DIANA-microT [80]. Comprehensive review and discussion of these methods are available [81, 82].

Table 1-1. Sequence-based methods for miRNA target indentification

Database	Reference	Link
TargetScan	[72]	http://www.targetscan.org/
miRanda	[77]	http://www.microrna.org
PicTar	[78]	http://pictar.mdc-berlin.de/
TargetScanS	[79]	http://genes.mit.edu/tscan/targetscanS.html
PITA	[75]	http://genie.weizmann.ac.il/pubs/mir07/
DIANA-microT	[80]	http://diana.cslab.ece.ntua.gr/microT/

1.3.5. Paired miRNA-mRNA expression data

As technology advances and it becomes clear that miRNAs are key regulator in regulatory network, miRNA expression profiling become popular and a multitude of miRNA profiling platform have become available [83]. Currently, there are three major well-established approaches: quantitative reverse transcription PCR(qRT-PCR), hybridization-based methods (for example, DNA microarrays) and high-throughput sequencing (RNA-seq) [83]. Each approach has its own limitations and advantages [84-86]. They together provide plenty of miRNA profiling data, which enable researchers to pinpoint important miRNAs and their roles in a particular biological process. Moreover, more and more paired miRNA-mRNA expression profiles have been achieved to investigate miRNA's role, especially in cancer (Table 1-2).

Table 1-2. Paired miRNA-mRNA expression data for cancer study.

GEO ID	miRNA platform	mRNA platform	sample number	sample type	reference
GSE35602	GPL8227	GPL6480	59	colorectal cancer	[87]
GSE32688	GPL7723	GPL570, GPL6801	96	pancreatic Cancer	[88]
GSE40355	GPL8227	GPL13497	48	Bladder cancer	[89]
GSE22220	GPL8178	GPL6098	426	breast cancer	[90]
GSE19536	GPL8227	GPL6480	215	breast cancer	[91]
GSE19783	GPL8227	GPL6480	216	breast cancer	[92]
GSE28544	GPL10850	GPL6244	56	breast cancer	
GSE19350	GPL8227	GPL570, GPL8887	41	central nervous system germ cell tumors	[93]
GSE35982	GPL14767	GPL4133	32	colorectal cancer	[94]
GSE21687	GPL8227	GPL570, GPL1261	339	Ependymoma	[95, 96]
GSE37372	GPL8178	GPL96, GPL570	78	Chordomas	[97]
GSE17227	GPL8178	GPL6370	20	Glioblastoma	[98]
GSE25632	GPL8179	GPL6884	108	Glioblastoma	[99]
GSE21849	GPL9081	GPL1708	65	Lymphoma	[100]

GSE17498	GPL8227	GPL2005, GPL9021	102	Myeloma	[101]
GSE17306	GPL9081	GPL570	106	myeloma	[102]
GSE28425	GPL8227	GPL13376	46	osteosarcoma	[103]
GSE20161	GPL8178	GPL6102	215	prostate cancer	[104]
GSE21032	GPL8227	GPL4091, GPL5188, GPL10264	743	prostate cancer	[105]
GSE25692	GPL9081	GPL7363	43	prostate cancer	[106]

1.3.6. Integrative approaches

Currently reliable prediction of miRNA-mRNA interactions remains to be a challenge. Prediction solely based on sequence information has high false positive rates [107]. To improve the performance, novel integrative approaches that combine sequence based predictions and experimental data are needed. With the accumulation of high throughput expression data, especially paired miRNA-mRNA experiments, several methods that incorporate these high throughput data have been developed. These methods can be roughly categorized as correlation based, multiple linear regression based and Bayesian based. We will give a brief introduction of each method in chapter 2.2.

1.3.7. Functional annotation of miRNAs

One central goal of miRNA study is to infer its biological functions. The common strategy for the functional annotation of miRNAs is to perform gene sets enrichment analysis with their targets mRNAs. This strategy is based on the accumulation of large amount of knowledge of genes and the assumption that miRNAs have similar functions to their targets. It starts with a list of miRNAs of interest, usually differentially expressed ones between different subgroups of samples. The target genes of these miRNAs are identified by sequence-based target prediction algorithms, and/or searching in validated target databases or literatures. When a set of targets is available, by using bioinformatics enrichment tools [108, 109], relevant biological functions can be assigned. Using this strategy, several tools have been developed for functional annotation, such as miRGator [110], miRDB [111], and FAME [112].

Chapter 2. Model

In this chapter, we will introduce methods that use gene expression data for reconstruction of gene regulatory network and methods that use paired miRNA-mRNA expression profiles for inference of miRNA-target interactions.

2.1. Reconstruction of gene regulatory network (GRN)

There are many models and approaches that have been proposed to infer the GRN based on gene expression data from microarrays. They can be categorized as Boolean networks, Bayesian networks, co-expression networks and differential equation models (Figure 2.1). We describe each of these categories in more detail below.

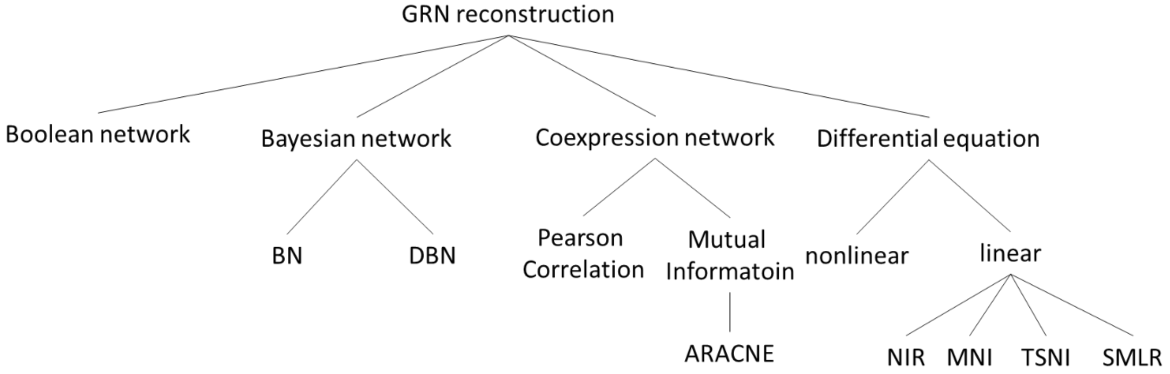


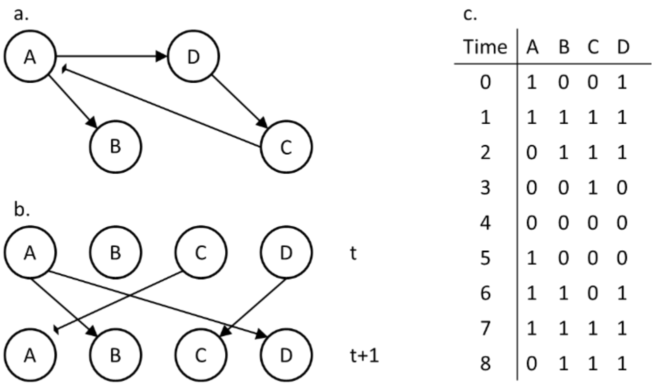
Figure 2.1. Gene regulatory network reconstruction methods described in this chapter.

2.1.1. Boolean networks

Boolean Network model was first proposed by Kauffman in 1969 [4]. The model represents gene at discrete time steps as a binary (on or off) variable. Boolean function is used to determine the gene state at the next discrete time step. Let there be n genes in the network, a Boolean network $G(V, F)$ is a set of genes $V = \{v_1, v_2 \dots , v_n \}$ and a list of Boolean functions $= \{f_1, f_2 \dots , f_n \}$, where a Boolean function

$f_i(v_{i1}, v_{i2}, \dots, v_{ik})$ is assigned for each node v_i, with a maximum of k inputs. In order to reduce the computational complexity and generate stable results, k is usually kept small. Networks are defined by the n and k values used to construct them, for example, a $n = 1000, k = 3$ network.

There are different ways for visualizing a Boolean network. Directed graphs and wiring diagrams are the most common ones. In a directed graph, nodes represent genes and edges indicate Boolean functions. In a wiring diagram, there are two levels of nodes; each level represents a discrete time step and the connections between two levels indicate Boolean functions. It can also be helpful to show trajectory tables for Boolean networks. A trajectory is a sequence of states of the whole network following an initial state. Trajectory tables provide the output of the network given various inputs. Figure 2.2 illustrates a simple $n = 4, k = 1$ Boolean network. Figure 2.2a shows the network as a directed graph. Figure 2.2b displays the network as a wiring diagram showing state transitions from a discrete time step to another. Figure 2.2c is a table of gene states at discrete time steps.



Time	A	B	C	D
0	1	0	0	1
1	1	1	1	1
2	0	1	1	1
3	0	0	1	0
4	0	0	0	0
5	1	0	0	0
6	1	1	0	1
7	1	1	1	1
8	0	1	1	1

Figure 2.2. A simple $\boldsymbol{n = 4, k = 1}$ Boolean network. The Boolean network has 4 nodes. Node $\boldsymbol{A}$ is on unless node $\boldsymbol{C}$ was on in the previous time. Nodes $\boldsymbol{B}$ and $\boldsymbol{D}$ are only on if node $\boldsymbol{A}$ was on in the previous step. (a) Directed graph representation. Edges with arrows imply stimulation. Edges with diamonds imply suppression. (b) Wiring diagram. The expression at time $\boldsymbol{t}$ determines the expression at time $\boldsymbol{t+1}$. (c)

State Table. When nodes $\boldsymbol{A}$ and $\boldsymbol{D}$ are expressed and nodes $\boldsymbol{B}$ and $\boldsymbol{C}$ are repressed, the network engages in a 6 state dynamic attractor. The values for times 7 through 13 will be the same as times 1 through 7.

Most Boolean network models are synchronous, that is, all states are updated simultaneously. Future states of the network are determined from previous states. For a network containing n genes, there are at most 2^n different states, thus a trajectory must reach a previously visited state within 2^n time steps and cycle eventually emerges. These cycles are called attractors, which represent steady states of a model. Point attractors contain 1 state for steady state conditions, and dynamic attractors contain multiple states. For example, the phase changing of the cell-division cycle is a dynamic cycle. Attractors represent stable phenotypic structures [113], and provide information about the system being modeled.

To create Boolean network for gene regulation, the expression data must be discretized (on or off). With discretized data, a network that explains the data can be generated. The REVerse Engineering ALgorithm (REVEAL) is an early algorithm that accomplishes this task [114]. REVEAL calculate mutual information between sets of genes via Shannon entropy, and extract the wiring relationships that accurately explain the state of an output gene. Based on REVEAL, Akutsu et al. provided a proof that $O(\log n)$ state transition pairs are enough to identify Boolean network with high probability when the number of input k for Boolean function is bounded by a constant [115]. They later presented a Monte-Carlo algorithm with improved time-complexity [116]. Since real gene expression data may be quite noisy, a learning paradigm called Best-Fit extension was developed with goal to learn a network with as few misclassifications as possible [117]. Lähdesmäki et al. introduced a method that learn GRN under Best-Fit extension with better efficiency [118]. Several studies constructed Boolean networks from time series expression data, including cell cycle in *S. pombe* [119], and IL-2 stimulated T cell responses in *M. musculus* [120].

One limitation of simple Boolean networks is that they do not represent the gene regulatory mechanisms exactly as they appear in living cells. In reality, gene expression is often varied on a continuous scale, not switched on or off as a Boolean toggle. Szallasi and Liang introduced ‘realistic Boolean genetic networks’ that address the issue of biological relevance of Boolean networks [121]. The model incorporate regulatory biochemical and physiological parameters but require additional computational time and space requirements. Another limitation of simple Boolean networks is its synchronous nature, that all genes in the system are subjected to changes simultaneously, which is not necessarily true in cells. Updating the network asynchronously may be a way to solve this problem. Some work with asynchronous Boolean networks under stochastic update has been performed [122, 123].

Another limitation of simple Boolean networks for modeling GRN based on gene expression data is their inability to compensate for noise and a lack of prior knowledge, which are exacerbated by noise inherent in high throughput gene expression data and the incomplete knowledge of genetic interactions. A model that account for noise associated with gene expression named Probabilistic Boolean Network (PBN) was proposed [124]. The basic idea of PBN is that it has more than one possible Boolean function for each node: each node is assigned a set of Boolean functions, which are called predictors. A PBN $G(V, F)$ is defined as a set of nodes $V = \{v_1, v_2 \dots, v_n\}$ and a list of sets of predictors $F = \{F_1, F_2 \dots, F_n\}$, where inside each set of predictors $F_i = \{f_i^1, f_i^2 \dots, f_i^{l(i)}\}$, f_i^j is a Boolean function and $l(i)$ is the number of predictors for node v_i. Notice that $N = \prod_{i=1}^{n} l(i)$ is the total number of possible PBN realizations (if $f_1, f_2 \dots, f_n$are independent). When only one function is used for each node, that is, $l(i) = 1$ for all i, then $N = 1$, then the model reduce to a standard simple Boolean network. Each predictor in a predictor set is associated with a probability given the current state of the network. In accordance to the probabilities, a predictor is randomly selected for each node at each time step. To determine the probability of a predictor, it involves a complicated procedure when PBNs utilize dependent predictors [124]. To reduce the computational complexity, most models avoid dependent predictors.

2.1.2. Bayesian networks

Bayesian networks (BN) models treat each gene as a random variable governed by a probability distribution, whose function is determined by a product of conditional probabilities. BNs are ideal for describing processes where value of each component is dependent upon values of a small number of other components [125]. The networks produced by BNs are directed acyclic graphs (DAG). Each gene in the network is dependent on a set of other genes, which are called its parents. In a BN, the probability distribution function for each gene is a product of the conditional probabilities of all of the ‘ancestor’ genes in the network. One of the major issues for BN models is the large number of possible networks that can be constructed from a particular set of genes. A particular directed acyclic graph that best fits the data must be determined.

Consider a finite set of random variables, $= \{X_1, X_2, X_3, \dots, X_n\}$. Two components make up a Bayesian network: the directed acyclic graph G, in which each vertex corresponds to a variable in the set S and the conditional probability distribution of each variable based on its parents in G. The Markov assumption, which is visualized by the directed acyclic graph, states that given a set of parent vertices, each vertex in G is conditionally independent of vertices that are not its descendants. Using this conditional independence, the joint probability distribution for the entire network can be expressed as

$$P(X_1, X_2, X_3, \dots, X_n) = \prod_{i=1}^{n} P(X_i | Pt(X_i)) \tag{2.1}$$

where $Pt(X_i)$ represents the set of parents of X_i. The conditional probability distributions can arise from discrete or continuous variables. Suppose the parents of a variable are expressed as follows

$$Pt(X_i) = \{X_i^1, X_i^2 \dots, X_i^{k(i)}\} \tag{2.2}$$

where $k(i)$ is the number of parents of X_i. If each parent possesses a discrete value from a set of finite size, then the probability of X_i given its parents can be represented as a table that specifies the value of X_i for each value set of its parents. If there are m possible values for each discrete variable then the table will specify $m^{k(i)}$ possible conditions. When continuous real valued variables, such as those present in gene expression datasets, are used then there are infinite possible distributions. In such cases, linear Gaussian conditional probability densities can be used. Each variable X_i is assumed to follow a normal distribution with a mean that is dependent on values of its parents. However, the variance of this distribution will be independent of the parents. The resulting joint distribution for the network will be a multivariate Gaussian distribution [125].

The large search space of possible directed acyclic graphs for a given set of nodes makes it difficult to identify a network that is ideal for a particular dataset. The most common solution to this problem is the usage of a scoring function to evaluate potential graphs. The two most common scoring functions are the Bayesian Information Criteria (BIC) and the Bayesian Dirichlet equivalence (BDe). These scoring methods incorporate penalties to prevent over-fitting the dataset [126]. However, finding the graph with the maximum score out of all possible graphs is known to be an NP-hard problem [127]. A priori biological information can also be utilized in order to restrict the number of possible graphs, for example, Ong et al. utilized the fact, that certain *E. coli* genes are co-transcribed and thus co-regulated, to identify edges between these genes in their networks [128]. Due to the exponential size of the search space for possible networks, heuristic search methods, such as greedy-hill climbing, Markov Chain Monte Carlo, and simulated annealing, are employed. Furthermore, model averaging, or bootstrapping techniques can be utilized to select the ideal network from several highly scoring networks identified by the selected search heuristics, and can also be employed to determine confidence intervals for the interactions [126]. Information theory-based scoring functions such as mutual information can also be employed.

Bayesian networks can handle incomplete or noisy data, combine heterogeneous data types, and can avoid over-fitting. However, BNs are unable to model feedback loops, since they don't allow graphs with cycles. Dynamic Bayesian networks (DBNs) were created to overcome this limitation. Unlike the regular Bayesian networks, dynamic Bayesian networks require time-series gene expression data instead of static gene expression data. Dynamic Bayesian networks model the gene regulatory network as a graph with two layers of nodes. The first layer of nodes represents the expression of the genes at time $t-1$, and the second layer represents the gene expression at time t. The expression of each gene is dependent on the expression of its parents at previous time point. This arrangement of nodes allows for the representation of a network containing cycles with an acyclic graph. See Figure 2.3 for an illustration of how a gene regulatory network can be represented by a dynamic Bayesian network.

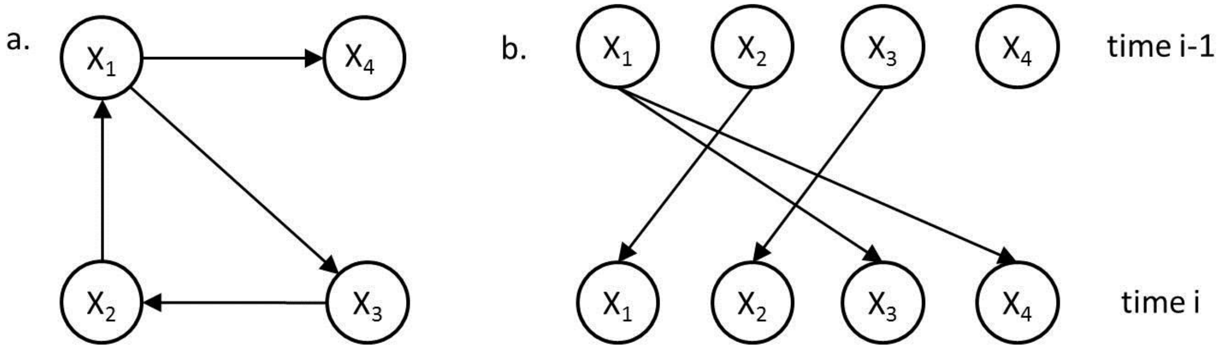


Figure 2.3. Representation of a network as Dynamic Bayesian Network**.** On the left is an example network. On the right is how the network would be represented by a dynamic Bayesian network. The dynamic Bayesian framework is able to overcome the limitation of Bayesian networks and can model the cyclic behavior of the gene regulatory network using two layers of nodes. Each row of nodes in the dynamic Bayesian network represents a time point.

A time-series microarray dataset with n samples, each corresponding to a unique time point, and p genes, can be expressed as a $n \times p$ matrix, where each row represents the gene expression at a discrete time point, and each column corresponds to a particular gene. We can express the joint probability as

$$P\left(x_{1,1}, x_{1,2}\ , \dots, x_{1,n}\ \ \dots, x_{p,n}\right) = P(X_1)P(X_2|X_1)\ \dots\ P(X_n|X_{n-1}) \qquad (2.3)$$

where $x_{i,t}$ on the left hand side represents the expression of gene i at time t, and X_t on the right hand side represents the column vector of the expression of all genes at time t. It is assumed that the structure of the network does not change between time points. (There are DBN studies about time-varying network structure that we will introduce later.) The conditional probability of one time point based on the previous time point can be expressed as the product of the conditional probabilities of each gene given its set of parent genes

$$P(X_t|X_{t-1}) = P(x_{1,t}|Parent_{1,t-1})P(x_{2,t}|Parent_{2,t-1}) \dots\ P(x_{p,t}|Parent_{p,t-1}) \qquad (2.4)$$

where $Parent_{i,t-1}$ is the vector of parents of the gene i at time $t-1$ [129, 130]. This equation is analogous to the joint probability distribution of the standard Bayesian network. However it differs that the expression of a gene is dependent on its parents' expressions at the previous time point. Like standard Bayesian networks a scoring function is needed to evaluate the possible network topologies. Similar scoring functions and search algorithms and techniques can be applied to dynamic Bayesian networks. Bayesian and dynamic Bayesian networks have been widely applied to the gene regulatory network reconstruction problem [131-134].

Recently there are studies of DBN on time-varying GRN structure. In those studies, the topology of GRN were no longer assumed to be static, but is varying during the time course. Song et al. proposed a formalism in which $P(X_t|X_{t-1})$ is time dependent [135]. They decompose the problem by finding the neighbor of each gene separately. To learn the neighborhood, they assume the network is sparse and vary smoothly, and transform the problem to L_1-regularized square linear regression problem. Lèbre et al. proposed a method called Auto Regressive TIme VArying models (ARTIVA) to learn time varying GRN from time series expression data [136]. For each gene, regression models are learned for district phases separated by change point.

2.1.3. Co-expression networks

Like clustering analysis [137], the construction of a co-expression network is based on the measure of similarity between gene expression profiles. The rationale behind co-expression networks is straightforward: if two genes have similar gene expression profiles, they are likely to interact with each other. Thus, if a metric can be established to evaluate the similarity, a GRN can be constructed by connecting genes that have similarity over certain cut-off threshold.

Pearson correlation is one of the simplest metrics for similarity, and it is suitable for large scale networks because of its computational cost efficiency. In 2003, Stuart et al. [10] utilized the Pearson correlation in their network reconstruction study of 3182 DNA microarrays from 4 different organisms (humans, flies, worms, yeasts). They first constructed a set of ‘metagenes’ across multiple organisms and then utilized the Pearson correlation to identify pairs of genes that had significantly correlated expression values. They obtained a correlation rank of all pairs of genes and calculated the probability of observing a particular configuration of ranks by chance. Finally, they obtained a co-expression network that contained 3416 genes and 22,163 interactions.

Despite its low data requirement and simplicity, Pearson correlation cannot handle non-linear similarity. As the number of available microarray datasets has steadily been increasing, researchers began to use other metrics that utilize larger sample sizes. Mutual information, based on information theory, is a popular alternative which is capable of capturing the general similarity between two variables. The definition of mutual information is based on Shannon entropy [138]. For a discrete random variable X with n outcomes $\{x_1, x_2, \dots, x_n\}$, the entropy $H(X)$, is defined as [138]

$$H(X) = -\sum_{i=1}^{n} p(x_i) logp(x_i) \tag{2.5}$$

where $p(x_i)$ is the corresponding probability of outcome x_i.

Entropy $H(X)$ measures the uncertainty of variable X. The conditional entropy of two variables X and Y taking values $\{x_i\}$ and $\{y_i\}$ is defined as

$$H(X|Y) = -\sum_{i,j=1}^{n} p(x_i, y_i) log \frac{p(x_i, y_i)}{p(y_i)} \tag{2.6}$$

where $p(x_i, y_i)$ is the probability that $X = x_i$ and $Y = y_i$.

Conditional entropy $H(X|Y)$ can be regarded as the uncertainty in the random variable X given Y. Thus, mutual information can be expressed as

$$I(X; Y) = H(X) - H(X|Y) \tag{2.7}$$

which measures the average amount of information Y conveys for X, or the reduction of uncertainty about X if Y is given. For gene expressions, which are continuous random variables, mutual information can be estimated by approaches based on discretization or approaches based on kernel density estimation [139]. Mutual information provides a metric for the general similarity between variables.

Butte et al. first proposed a methodology, termed Relevance Networks (RelNet), that computes pairwise mutual information for all genes [140]. In the study, they used 79 microarrays containing 2,467 genes in yeast, and calculated pairwise mutual information 3,041,811 times in total. Pairs with mutual information higher than the threshold were kept.

Later Margolin et al. developed the model called ARACNE (Algorithm for the Reconstruction of Accurate Cellular Networks) [7], which is also based on pairwise mutual information. ARACNE aimed to improve the inference performance by eliminating the majority of indirect interactions. The property called data processing inequality in information theory is applied. The idea is that if gene $g1$ and $g3$ interact through a third gene, $g2$, then the data processing inequality (DPI)

$$I\ (g1; g3) \leq min\{\ I(g1; g2), I(g2; g3)\} \qquad (2.8)$$

must be satisfied. That is, if three genes are connected, the edge with the smallest mutual information would be removed. The ARACNE model was successfully applied to study of microarray expression profiles of human B cells [8]. While ARACNE infer GRN from steady-state data, Zoppoli et al. [141] proposed a method called TimeDelay-ARACNE that works with time-series data. TimeDelay-ARACNE tries to extract dependencies between genes at different time delay with a stationary Markov Random Field.

Because of its relatively low computational cost and low data requirement, co-expression networks are usually used to infer global properties of large-scale of regulatory networks. However, the drawback of using a co-expression network is that since the similarity metric only accounts for a pairwise relationship, the models do not consider interactions including multiple genes.

2.1.4. Differential equation models

Differential equations are widely used for modeling dynamic systems in engineering. As for gene regulatory networks, a system of ordinary differential equations (ODE) can be used. The change rate of gene expression is described as a function of expressions of other genes and external perturbation

$$\frac{dx_i}{dt} = f_i(\vec{x}, \vec{u}, \vec{\theta}) \qquad (i = 1,2, \dots, n) \qquad (2.9)$$

Variable x_i represents the expression level of gene i, vector $\vec{x} = (x_1, x_2, \dots, x_n)$ represents the expressions of all genes in the system and $\vec{u}$ represents the external perturbations, like gene knockouts or chemical treatments. And vector $\vec{\theta}$ is the set of parameters. By estimation of the function $f_i, (i = 1,2, \dots, n)$, the structure of a GRN can be established.

Unlike co-expression models, differential equation models are capable of describing the dynamic behavior of GRN quantitatively, thus they can be used not only to study the topology of GRN but also to simulate GRN dynamics. Depending on the type of function f, the differential equation models for GRN can be non-linear or linear. Generally non-linear models are more complex and require more data to estimate the parameters. In addition, non-linear models usually require prior knowledge about the system to choose proper form of function. On the other hand, linear models require less data and usually no prior knowledge for parameters learning. Given the fact that microarray data are noisy and under-sampled, linear models are more suitable for GRN modeling. As described in the following section, there are several well developed methods based on linear differential equation models.

2.1.4.1. Nonlinear differential equation models

In general, the actual biochemical regulatory systems are complex and nonlinear [142] and numerous nonlinear ODE models have been proposed to describe the regulatory networks. However, due to the complexity of the models, the estimation of the parameters requires a large amount of data. Thus nonlinear ODE models are often only suitable for small-scale networks. For example, Sakamoto et al. used genetic programming to infer the right hand side of Equation (2.9) [16], which can be of arbitrary form. They limited the number of genes to around three. For larger networks, further assumptions and preprocessing are needed.

One of the most well-studied nonlinear ODE models is S system, which can be regarded as the canonical form of general non-linear differential equations [143]. The model for a network containing n genes is a system of nonlinear differential equations

$$\frac{dx_i}{dt} = \alpha_i \prod_{j=1}^{n} x_j^{g_{i,j}} + \beta_i \prod_{j=1}^{n} x_j^{h_{i,j}} \qquad (i = 1,2,\dots,n) \qquad (2.10)$$

where x_i is expression level of gene i. Parameters α_i and β_i are called rate constants, and $g_{i,j}$ and $h_{i,j}$ are called kinetic orders. Since the number of parameters in the model is proportional to the square of number of genes, it is a challenge to reconstruct large-scale differential equation networks. Kimura et al. proposed a method for inferring the large-scale network with the S system model [144]. The method is based on a problem decomposition strategy of dividing a problem into sub-problems.

Another strategy for implementing a nonlinear model is to restrict the nonlinear function f to specific types. However, picking a suitable function require prior knowledge and experience. As a result, many data-driven methods based on linear model have been proposed. For more information on nonlinear genetic regulatory system modeling, please refer to section 6 of the review by de Jong [145] .

2.1.4.2. Linear models

Linear models generally do not require extensive prior knowledge about the regulatory network, and they are suitable for larger scale networks because of their relative simplicity. As mentioned above, the behavior of gene expression can be modeled by nonlinear differential equations. Within the small neighborhood of a particular point of interest, this nonlinear system can approximated to the first order by a system of linear equations. Consider a system containing n genes and p perturbations, then for each gene i, the rate of change of gene expression is described as a function of expression of other genes and external perturbation

$$\frac{dx_i}{dt} = \sum_{j=1}^{n} a_{i,j} x_j + \sum_{l=1}^{p} b_{i,l} u_l \qquad (i = 1,2, \dots, n) \tag{2.11}$$

where x_i is the mRNA concentration of gene i and $a_{i,j}$ can be regarded as the influence of gene j on gene i and u_l is the lth external perturbation and $b_{i,l}$ can be regarded as the influence of perturbation l on gene i in this experiment.

For compactness, (2.11) can be expressed in matrix form

$$\frac{d\vec{x}}{dt} = A\vec{x} + B\vec{u} \tag{2.12}$$

where $\vec{x} = (x_1, x_2, \dots, x_n)^T$, $\vec{u} = (u_1, u_2, \dots, x_p)^T$ and A is a $n \times n$ matrix containing coefficients $a_{i,j}$, B is a $n \times p$ matrix containing coefficients $b_{i,l}$.

If there are m experiments, then there are m equations like (2.12), combining them together in matrix form results in the matrix form of the linear equation system

$$\frac{dX}{dt} = AX + BU \tag{2.13}$$

where $X = [\vec{x}_1 | \vec{x}_2 | \dots | \vec{x}_m]$, in which $\vec{x}_k$ is column vector containing gene expression level of n genes in the kth experiment. Thus X is a $n \times m$ matrix representing the expression level of all n genes in m experiments, and similarly $U = [\vec{u}_1 | \vec{u}_2 | \dots | \vec{u}_m]$ is a $p \times m$ matrix representing the external perturbation in m experiments.

Notice that matrix A describes how genes interact with each other, thus by estimating matrix A we can infer the GRN. However, in most cases, $m \ll n$ so that matrix A cannot be calculated directly. Furthermore, due to limitation of measuring techniques, the expression data are noisy which makes it more challenging. In recent years, several methods have been proposed to solve this problem.

2.1.4.2.1. GRN Inference using Singular Value Decomposition

Yeung et al. proposed an approach to reconstruct GRN based on Singular Value Decomposition (SVD) and robust regression [146]. The goal is to deduce matrix A in equation (2.13) by using the measured data of X, BU, and $\frac{dX}{dt}$. The approach is a two-step procedure. First they use singular value decomposition to solve the system of ODE, equation (2.13), and get a family of matrix A. The family of solutions represents networks that are consistent with the measured data. Next the best candidate can be chosen according to the prior knowledge of the biological network. If no prior knowledge is available, inspired by the observation that GRNs are sparse, they pick the network among candidates by maximizing the number of zero entries in matrix A using robust regression based on L_1 norm. More specifically, by using SVD to decompose X^T

$$X^T = U\Sigma V^T \tag{2.14}$$

Where U and V are orthogonal, $UU^T = I$ and $VV^T = I$. Then plugging this into equation (2.13) and rearranging results in the following

$$AV\Sigma U^T = \frac{dX}{dt} - BU \tag{2.15}$$

and one particular solution

$$A_0 = \left(\frac{dX}{dt} - BU\right) U\Sigma^{-1} V^T \tag{2.16}$$

Thus, the family of solutions consistent with the measurement is

$$A = A_0 + CV^T \tag{2.17}$$

where C is an arbitrary scalar coefficient matrix.

The next step is to choose C such that A has maximum number of zero entries. Their idea was to set $A = 0$ in equation (2.17) and obtain an over-determined equation $CV^T = -A_0$, and then find the exact fit plane passing through as many points as possible. In order to do this, they use L_1 norm regression that minimizes the sum of absolute values of errors to calculate C. This SVD based approach was tested and validated in numerous experiments on model gene networks in their study [146].

2.1.4.2.2. Network identification by multiple regression

One potential drawback of the approach above is that it requires data to estimate the time derivative $\frac{dX}{dt}$. Gardner et al. proposed a method called network identification by multiple regression (NIR) [12], which used only steady-state expression measurements such that $\frac{dX}{dt} = 0$.

If only one gene is perturbed in each experiment, equations (2.11) and (2.12) become

$$\frac{dx_i}{dt} = \sum_{j=1}^{n} a_{i,j}x_j + u_i \qquad (i = 1,2,\ldots,n) \tag{2.18}$$

$$\frac{d\vec{x}}{dt} = A\vec{x} + \vec{u} \tag{2.19}$$

If there are m such perturbation experiments, similar to equation (2.13), we have

$$\frac{dX}{dt} = AX + U \tag{2.20}$$

Since the concentration of n genes are at steady state, $\frac{d\vec{x}}{dt} = 0$, equation (2.20) becomes

$$AX = -U \tag{2.21}$$

By retrieving the connectivity matrix A, we can describe the network. Thus the remaining problem is to solve the linear system. Since sample size is usually limited, $m \ll n$, plus that measurement of gene expression is noisy, it is preferred to have an over-determined system (more equations than unknowns) and use statistic regression. NIR solves this problem by assuming that for each gene, there is a maximum number of regulators, k. Then the solutions of all possible combinations are calculated by least-square regression. Then the best solution is chosen based on the significance of the regression based on the *F-test*.

NIR was applied to reconstruct regulatory networks containing nine genes in the SOS pathway in *E. coli*. Each perturbation was accomplished by overexpressing one of the nine genes with arabinose-controlled episomal expression plasmid.

2.1.4.2.3. Mode-of-action by network identification

Since NIR required a well-designed perturbation experiment, it is not applicable to many datasets. Bernard et al. proposed a method called mode-of-action by network identification (MNI) to find the solution of the system without information about permutation u [13]. Thus, MNI is suitable for the analysis of a wider range of microarray data. The idea of MNI is based on the assumption that any external stimuli acts on only a small number of genes, thus most coefficients, $a_{i,j}$, in equation (2.18) will be zero.

For each gene i, there are m steady-state experiments,

$$\sum_{j=1}^{n} a_{i,j} x_{j,l} = -u_{i,l} \qquad (l = 1,2, \dots, m) \tag{2.22}$$

Extract the experiments in which $-u_{i,l'} = 0$, and obtain

$$\sum_{j=1}^{n} a_{i,j} x_{j,l'} = 0 \qquad (l' = 1,2, \dots, m') \tag{2.23}$$

When implementing MNI, only experiments in which the perturbation, $-u_{i,l}$, is zero are considered. Determining these experiments is not trivial. Di Bernardo et al. proposed a recursive method starting with an initial estimate of $\hat{a}_{i,j}$[13]. The external influence, $\hat{u}_{i,l}$, is calculated using equation (2.22). Any experiment with an external influence greater than a pre-determined threshold is removed for further calculations. Equation (2.23) is used to obtain a new estimate of $\hat{a}_{i,j}$. The method is repeated using the new estimate and continues until $\hat{a}_{i,j}$ and $\hat{u}_{i,l}$ converge.

As in NRI, equation (2.23) is underdetermined. Thus additional constraints are needed to find a reliable solution. Unlike NRI, which uses subset regression to identify a small set of non-zero coefficients, MNI uses the fact that expression profiles of many genes are similar. Thus genes expression can be represented by a reduced set of 'characteristic' or 'meta' genes by using SVD. The original space of gene expression with dimension n is first mapped into the space of metagenes with reduced dimension. The recursive procedure described above is then used to identify network for the metagenes**.** After all the work, the estimated perturbation is mapped back into n gene space.

MNI was proposed for the application of finding target genes of a particular treatment. Di Bernardo et al. applied MNI to the analysis 515 whole-genome yeast gene expression datasets resulting from different perturbation experiments and correctly enriched the target gene and pathway for most compounds.

2.1.4.2.4. Time Series Network Identification

Bansal et al. proposed a method based on time series expression experiments, called Time Series Network Identification (TSNI) [14]. Similar to NIR, at a particular time point, the rate of synthesis of a transcript is represented as a function of the expression of the other genes and the external perturbation, and the differential equation in (2.13) is converted to its corresponding discrete form, a difference equation

$$X_{t+1} = A_d X_t + B_d U_t \qquad (2.24)$$

where X_{t+1} is the gene expression at time point $t+1$ and X_t at time point t. This equation states that the gene expression level at one time point depends on expression profile at previous time point and external perturbation.

Rewriting this in a more compact form yields

$$X = H * Y \tag{2.25}$$

where $X = X_{t+1}$, $H = [A_d \quad B_d]$ and $Y = \begin{bmatrix} X_t \\ U_t \end{bmatrix}$.

Similar to MNI, TSNI applies SVD to decompose matrix Y, and solve the equation (2.25) with reduced dimensions and then maps the obtained solution back into the original space to obtain $A_d$and B_d. TSNI is suitable for the reconstruction of GRN containing genes of interest by analysis of time series gene expression data resulting from a particular perturbation. Bansal et al. applied TSNI to recover a nine gene subnetwork,part of the DNA-damage response pathway in *E. coli* using experimental data obtained by treatment of Norfloxacin [14].

2.2. Inference of miRNA-mRNA interaction

Regulatory relationships between miRNA and mRNA can be inferred from the expression data of paired miRNA-mRNA samples. And expression-based results can be combined with results from sequence-based prediction and experimental results. The general idea of integrative methods is represented in Figure 2.4. For analysis of miRNA-mRNA expression data, methods can roughly be categorized as 1) correlation based, 2) multiple linear regression based and 3) Bayesian based.

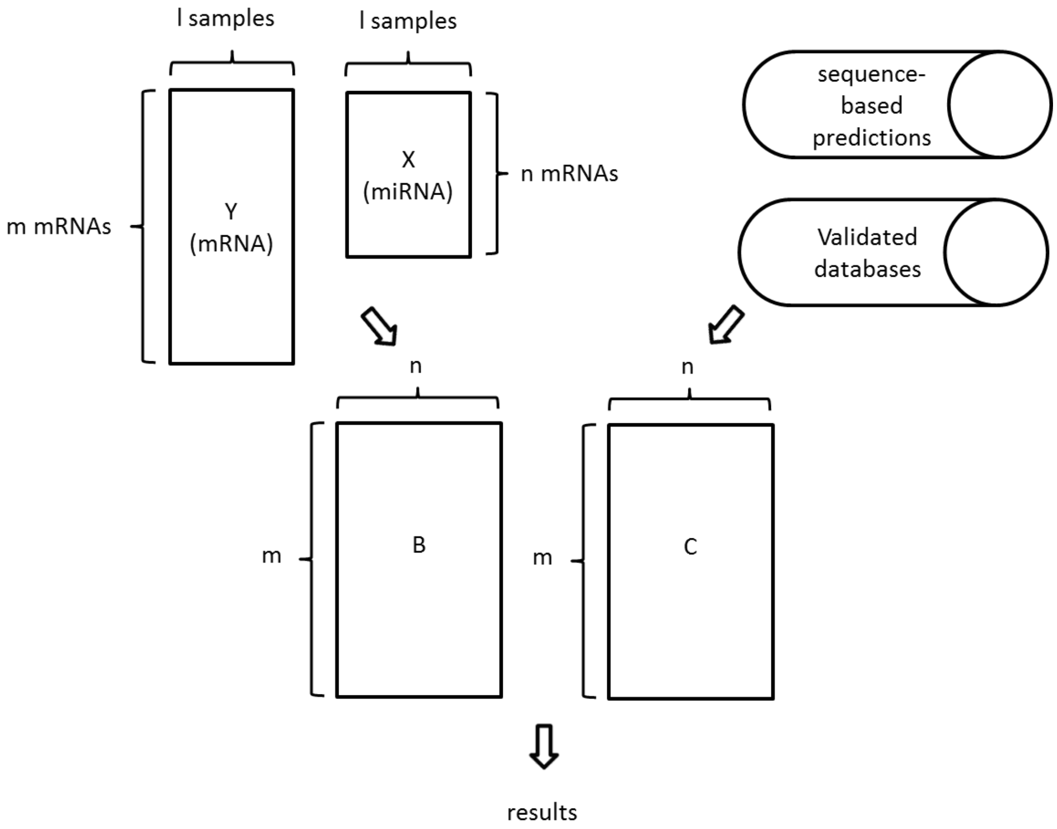


Figure 2.4. Integrative analysis of miRNA-mRNA interaction. Regulatory relationships between miRNA and mRNA can be inferred by analysis of paired miRNA-mRNA expression data or analysis of complementary sequences. The inferred interactions can also be combined to experimental validated interactions.

2.2.1. Correlation based/ mutual information

There are several web tools available for integrative analysis of miRNA-mRNA expression data and sequence-based miRNA target predictions. The most straightforward way to identify miRNA-mRNA regulatory pair using expression data is calculating their pairwise correlations or mutual information.

Nam et al. developed a database **miRGator** (http://genome.ewha.ac.kr/miRGator/) [147] that integrates target prediction, function analysis and genome annotation. In this database, functions of miRNA are inferred from the list of target genes that are predicted by miRanda, PicTar and TargetScans programs. Statistical enrichment test of target genes in gene ontology, pathway and disease annotations is also available. The database integrates public expression data and the correlation between miRNA and mRNA can be calculated and compared. In their most recent update version, miRGator v3.0 [148] (http://mirgator.kobic.re.kr/), public available deep sequencing miRNA data are compiled and several utilities for study of iso-miRs, miRNA editing and modifications are included. For miRNA target analysis, 3 databases of validated targets, 6 databases of predicted targets as well as the results of inverse correlation analysis of matched miRNA-mRNA gene expression (based on miRNA-seq and RNA-seq data from the same sample) are integrated.

Nam et al. later developed a web tool called **MMIA** (http://epigenomics.snu.ac.kr/MMIA/) [149], which incorporate commonly used miRNA target prediction algorithms and miRNA-mRNA expression data. For a miRNA of interest, on one hand, its predicted target mRNAs are selected from TargetScan, PictTar and PITA, on the other hand, its significantly inversely correlated mRNAs are identified from miRNA-mRNA expression data. Intersection of these two set of mRNAs is then used for gene set analysis (GSA) to discover miRNA-associated phenotypes and biological functions.

Peng et al. proposed an integrative approach to identify the miRNA-mRNA regulatory modules in Hepatitis C virus (HCV) infection [150]. They calculate the miRNA-mRNA correlation matrix based on standard Pearson correlation of paired microarray expression data. The correlation matrix is converted into a binary miRNA-mRNA correlation network by estimating false positive rate and choosing a proper cut-off value. In parallel, a binary miRNA-target matrix is created by computational target prediction based on seed match. At last, the miRNA-mRNA regulatory network and regulatory modules are extracted from the combination of miRNA-mRNA correlation matrix and the corresponding miRNA-target matrix.

Ritchie et al. developed an online resource named **mimiRNA** (http://mimirna.centenary.org.au) [151], which integrates expression data from samples across different tissues and cell types. A sample classification algorithm named ExParser is used to group together identical miRNA or mRNA experiments from different sources. Based on these expression data, mimiRNA provide visualization of relationship between miRNA and mRNA expression by calculating Pearson correlation coefficient of miRNA-mRNA pairs. Results from target prediction algorithm (TargetScan, miRBase, RNA22, PicTar) are included to assist assessments of potential targets.

Sales et al. developed a web tool named **MAGIA** (http://gencomp.bio.unipd.it/magia) [152] that integrates sequence-based target prediction and analysis of expression data. MAGIA extract target prediction from Pita, miRanda, TargetScan and user can take the intersection or union of those predictions. MAGIA then refines target predictions using miRNA-mRNA gene expression data. User can choose different metrics, Spearman's correlation, Pearson correlation, mutual information, GenMir++ (see below) and meta-analysis (only for non-matched biological samples) to compute the interaction measures.

Huang, G.T. developed web interface called **mirConnX** (http://www.benoslab.pitt.edu/mirconnx) [110] for inferring miRNA-mRNA regulatory network by integrating sequence information and gene expression data. At first a prior network is built based on computationally predicted transcription factor (TF)-gene associations, miRNA target prediction and literature. In parallel, an association network is inferred from expression data. The two networks are then combined using user-specified weight functions. mirConnX provide choice of different correlation measure (Spearman ρ correlation, Kendall τ rank correlation) besides Pearson correlation, and a simple weight function for integration of network.

The approaches mentioned above only consider pairwise miRNA-mRNA correlations. However, one mRNA may be targeted by several miRNAs and its expression profile may be affected by multiple miRNAs the same time.

2.2.2. Multiple linear regressions

Regards to the many-to-many miRNA-mRNA relationships, that is, one miRNA may targets multiple mRNAs and one mRNA may be affected by multiple miRNAs, an ordinary multiple linear model is quite natural for modeling miRNA-mRNA regulation. Suppose there are N potential miRNA regulators and M target mRNAs, a multiple linear equation is as below

$$y_i = \beta_0 + \sum_{j=1}^{N} \beta_{ij} x_j + \varepsilon_i \tag{2.26}$$

where y_i and x_j are variables representing the expression of mRNA i and miRNA j respectively with $i = 1,2,\dots,M$ and $j = 1,2,\dots,N$; β_0 is the constant term, and β_{ij} characterize the regulatory effect of miRNA j on mRNA i. Suppose there are L samples, if we denote the expression data across samples as column vector, $\boldsymbol{y}_i = [y_{1i}, y_{2i}, \dots, y_{Li}]^T$, $\boldsymbol{x}_j = \left[x_{1j}, x_{2j}, \dots, x_{Lj}\right]^T$ and $\boldsymbol{\beta}_i = [\beta_0, \beta_{i1}, \beta_{i2}, \dots, \beta_{iN}]^T$, then put column vectors into a matrix, $\boldsymbol{X} = [\boldsymbol{1}|\boldsymbol{x}_1|\boldsymbol{x}_2| \dots |\boldsymbol{x}_N]$, equation (2.26) for all i and j can be written in matrix form

$$\boldsymbol{y}_i = \boldsymbol{X}\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_{\mathrm{i}} \tag{2.27}$$

Given the expression data, the goal is to calculate optimal coefficients $\widehat{\boldsymbol{\beta}}$ to that satisfying some criteria, in most cases, minimizing the sum of square of error terms. This least square solution is given by normal equation

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{2.28}$$

Wang et al. modified the multiple linear regression model to find high-confidence targets among potential targets from sequence-based algorithm [153]. Follow the notation in equation (2.27) and (2.28), the estimated mRNA expression can be calculated by $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}.$ With total sum of square is defined as

$SS_{total} := (\boldsymbol{y} - \bar{y})^T(\boldsymbol{y} - \bar{y})$, explained sum of square is defined as $SS_{reg} := (\hat{\boldsymbol{y}} - \bar{y})^T(\hat{\boldsymbol{y}} - \bar{y})$, the coefficient of determination $R^2 := SS_{reg}/SS_{total}$ is a statistic that measure the goodness of fitting. Based on the concept of R^2, they define relative R^2 as R_k^2/R_N^2, where R_k^2 is the R^2 for a partial model contain k regulators and R_N^2 is the R^2 for a full model contain all N potential regulators. If R_k^2/R_N^2 is larger than some cut-off value, the k miRNAs in the partial model are regarded to high-confidence. In their approach, the coefficient $\hat{\boldsymbol{\beta}}$ is calculated by normal equation given in equation (2.28).

A problem rise when the data is collinear or the number of samples is less than the number of regulators ($M < N$), the matrix $\boldsymbol{X}^T\boldsymbol{X}$ becomes singular and there is no stable solution. In this case, the system is called underdetermined, which is the problem for most microarray expression data. For an underdetermined system, alternative methods or extra constrains are needed.

Regularized least square

To solve underdetermined linear systems, an alternative is regulation, which adds extra penalties besides the least-square requirement. That is, in addition to minimize the sum of square, or equivalently, the L_2 norm $||\boldsymbol{y} - \boldsymbol{X\beta}||_2^2$, extra penalty term $P(\boldsymbol{\beta})$ is added and then the model can be formulated as an optimization problem

$$\min_{\boldsymbol{\beta}}\{||\boldsymbol{y} - \boldsymbol{X\beta}||_2^2 + \lambda P(\boldsymbol{\beta})\} \tag{2.29}$$

Where is λ is the tuning factor. The most common penalty terms are L_1 norm, i.e., $P(\boldsymbol{\beta}) = ||\boldsymbol{\beta}||_1$ and L_2 norm, i.e., $P(\boldsymbol{\beta}) = ||\boldsymbol{\beta}||_2$. It is called LASSO regression when L_1 norm is used, and Ridge regression when L_2 norm is used. A combination of these two, i.e., $P(\boldsymbol{\beta}) = \lambda_1||\boldsymbol{\beta}||_1 + \lambda_2||\boldsymbol{\beta}||_2$ is called elastic-net penalized regression. By adding a penalty term, the coefficients, $\boldsymbol{\beta}$, of the solution can be forced to be ‘small’. For highly correlated predictors, Ridge regression shrink the coefficients to each other [154], which is proper for many predictors with non-zero coefficients. On the other hand, LASSO tends to pick one predictor and ignore the its highly correlated counterparts. LASSO is ideal when we expect many coefficients to be zero, thus LASSO is more suitable for modeling miRNA-mRNA regulatory interactions.

Kim et al. [155] implemented the multiple linear regression model to study miRNA-mRNA targeting in the regulation of colorectal cancer by analyzing expression data. To find the optimal solution with L_1 norm objective function, they employ an iterative algorithm called Broyden–Fletcher–Goldfarb–Shanno (BFGS), in which Hessian matrix is approximated and updated to obtain the direction of iteration.

Lu et al. proposed a **LASSO** regression model for miRNA-mRNA relationship inferences [156]. Since binding to RNA-induced silencing complex (RISC) is essential for miRNA functioning, they include the concentration of Argonaute (Ago) proteins into the model. Based on the ordinary model in equation (2.26), they added variable that characterize the concentration of Ago1-4 protein

$$y_i = \beta_0 + \sum_j \beta_{2,ij} Ago_{2,j} x_j + \sum_j \beta_{134,ij} Ago_{134,j} x_j + \varepsilon_i \tag{2.30}$$

In the above equation, mRNA expression levels of Ago genes are used to represent the Ago protein level. Ago1, 3, 4 proteins are merged into one term since they are all regarded as the competitor of Ago2 in binding with RISC. For regression model in equation (2.30), the penalty term being used is $\sum_j(|\beta_{134,ij}| + |\beta_{2,ij}|)$ corresponding to L_1 norm. In this way, LASSO enforces the sparseness requirement on solutions.

Based on LASSO model, Muniategui et al. propose a method named **TaLasso** (miRNA-Target LASSO) that add non-positivity constraints, $\beta_j \leq 0$, to ensure that the solution includes only negative miRNA-mRNA relationship [157], since miRNA usually repress the target mRNA. The web-tool for human miRNA is available at http://talasso.cnb.csic.es/. Beck et al. used **elastic-net** penalty, which combined L_1

norm and L_2 norm of the coefficients to handle correlated and sparse problems [158]. The weights of penalty can be determined by cross validations.

2.2.3. Bayesian inference

In addition approaches based on regression, there are several approaches been developed using Bayesian techniques.

Huang et al. developed a widely used approach called **GenmiR++** (Generative model for miRNA regulation) [159]. Like the multiple linear regression methods, GenmiR++ model mRNA expression as a multiple liner function of regulator miRNAs. Specifically, for the expression of mRNA i in tissue t, y_{it}, can be formulated as

$$y_{it} = \mu_t - \gamma_t \sum_j \lambda_j s_{ij} x_{jt} + \varepsilon_{it} \tag{2.31}$$

where j is the index of miRNA, x_{jt} is the expression of miRNA j in tissue t, μ_t is a background expression parameter of tissue t, $\gamma_t$is a positive scaling factor in tissue t accounting for differences in hybridization conditions and normalization between miRNA and mRNA , λ_j represents the down regulatory effects of miRNA j, and s_{ij} is an unobserved random variable for each candidate miRNA-mRNA pair such that $s_{gk} = 1$ if miRNA j indeed targets mRNA i. A set of candidate miRNA-mRNA pairs can be extracted from sequence-based algorithm and represented in form of binary matrix C, in which $c_{ij} = 1$ if miRNA j is predicted to target mRNA i and $c_{ij} = 0$ otherwise.

Now, the problem of finding real miRNA-mRNA is formulated as calculating the posterior probabilities of $s_{ij} = 1$ given $c_{ij} = 1$ for all (i, j) pairs. Based on the relationship in equation (2.31), a probabilistic graphical model and Bayesian modeling framework are built. Since exact inference of posterior probability distribution of s_{ij} is difficult, GenmiR++ implements a variational Bayesian algorithm which assumes a factorized distribution (s_{ij}, λ_j, γ_t are independent of each other) and searches for optimal result iteratively via expectation-maximization (EM) algorithm. Later, Huang et al. developed this model to GenmiR3 [160], with a new prior distribution that integrates sequence feature of miRNA-mRNA binding, such as energy of hybridization and conservation of target sites.

Based on GenMir++, Su et al. developed an algorithm called **HCTarget** (High Confident targets) [161] that also formulate a linear model. Unlike GenMir++ where miRNA regulatory effects are regarded as constant among all tissues, HCTarget re-defines the parameters of miRNA effect, equation (2.31) then becomes

$$y_{it} \; = \; \mu_t - \sum_j \beta_{jt} s_{ij} x_{jt} + \varepsilon_{it} \qquad (2.32)$$

Where β_{jt} represents the regulatory effect of miRNA j in tissue t, and this term can be regarded a combined factor of as $\gamma_t$and $\lambda_j$in equation (2.31). Next, unlike GenMir++ using variational Bayesian algorithm, HCTarget use Markov Chain Monto Carlo algorithm to infer posterior distribution by iterative sampling directly.

Stingo et al. proposed a **Bayesian graphical modeling** approach to infer miRNA-mRNA regulatory relationship [162]. Similar to GenMir++, they represent the regulation in a linear model. But unlike GenMir++ and HCTarget assuming constant regulatory effect of miRNA on all of its targets, they consider distinct regulatory effect of miRNA on different mRNAs, which is reasonable since miRNA has different sequence complementarity with its different targets. Equation (2.32) become

$$y_i \; = \; \sum_j \beta_{ij} s_{ij} x_j + \varepsilon_i \qquad (2.33)$$

where the variables and subscripts have meanings as those in equation (2.31) and (2.32). Since difference of tissue is not considered in this model, the subscript t is dropped. The parameter β_{ij} represents the distinct regulatory effect of miRNA j on mRNA i. Like GenMiR3, this model also considers the reliability of sequence-based prediction and combines them all into prior distribution. Next they use Metropolis–Hastings algorithm for posterior inference of s_{ij}. Since there are many repressors, they assume that most of mRNA are regulated by a small number of miRNAs and use Stochastic Search Variable Selection (SSVS) method to explore the huge posterior space.

Liu et al. presented a method to capture miRNA-mRNA relationship using **Bayesian network** structure learning [163]. Given a set of miRNA, X, and a set of mRNA, Y, their regulatory relationship can be represented in a graph $\{X, Y, E\}$, in which directed edges in E indicate dependencies between nodes X and Y. The aim of Bayesian learning is to identify a graph that is best supported by the expression data. Since searching space of possible graphs is hyper-exponentially with the number of nodes, they reduce the searching space with information from sequence-based prediction algorithm and then search for the optimal solution exhaustively. To avoid over-fitting, they adopt an averaging procedure on several candidate graph with highest scores.

The Bayesian based approaches have potential limitation that they are not suitable for large scale regulatory networks and require prior knowledge to refine the searching space, such as searching interactions in the set of sequence-based prediction.

2.2.4. Other

In addition to the databases and approaches described above, there are some approaches developed with different prospective of view.

Gennarino et al. proposed an approach called **HOCTAR** (host gene oppositely correlated targets) to predict miRNA targets by expression of miRNA host gene [164]. Based on the observation that expression of intragenic miRNAs and expression of their host genes are similar and thus it may be possible to use expression data of host genes to infer the expression of corresponding embedded miRNA, they hypothesized that expression of miRNA host gene is inversely correlated to the targets of the embedded miRNA. Thus, candidate miRNA targets can be ranked based on their inverse correlation to their prospective miRNA host genes. The main advantage of HOCTAR is that it can provide a way to predict miRNA target by analyzing the huge amount of microarray experiments that monitor the expression of both miRNAs (through their host genes) and candidate targets. Results of HOCTAR that contain ranked list of predicted targets of annotated human intragenic miRNAs are available at http://hoctar.tigem.it/.

Bandyopadhyay et al. proposed a method called **TargetMinner** (http://www.isical.ac.in/~bioinfo_miu/) [165] to incorporate miRNA-mRNA expression data to improve the performance of miRNA targets prediction. Their focus is to find a better training miRNA-mRNA pairs for target-predicting machine learning. Besides the experimentally verified positive miRNA-mRNA targeting pairs, negative ones are also included. To identify negative miRNA-mRNA pairs, they first select miRNA-mRNA pairs from sequence-based predictions (miRanda, TargetScanS, PicTar, DIANA-mircoT). Among those potential pairs, they proposed a four stage filtering procedure, in which they identified tissue specific miRNA and mRNAs using miRNA-mRNA expression data and then regard miRNA-mRNA pairs as negative if they both overexpressed in the same tissue. These candidates negative pairs are then filtered by testing with independent expression datasets, considering thermodynamic stability and seed-site conservation. After gathering training datasets containing positive and negative, they used a support vector machine (SVM) based classifier for miRNA target prediction and shown improved performance. The SVM model use miRNA-targeting site context-specific features, which do not include expression pattern.

Li et al. proposed a Bayesian inference model for miRNA target prediction with miRNA, mRNA and protein expression data [166]. In their study, two Bayesian models are built. First model combines miRNA expression and protein abundance to identify a set of confident miRNA-protein pairs. For those top miRNA-protein pairs, a second model includes mRNA expression data for calculation of miRNA-mRNA expression correlation. Thus two regulatory mechanisms (mRNA degradation and translational repression) can be distinguished. For protein expression data, they use negative binomial model to characterize the peptide count. For the mRNA expression data, they discrete them to binary value 1 or 0 representing high or low expression. The Bayesian modeling frameworks are similar to GenMir++ and the Gibbs sampling is implemented for inference of posterior distribution.

Pihur et al. introduced an approach based on **partial least square** (PLS) regression for reconstruction of genetic association networks from microarray data [167]. In PLS, unlike ordinary multiple regression in equation (2.26), a number $R < N$ of orthogonal latent factors $t_i, i = 1,2 \dots R$, are sequentially constructed as linear combinations of $x_j, j = 1,2, \dots N$ firstly. Next a linear model

$$y_i = \sum_{k=1}^{R} c_{ik} t_k + \varepsilon_i \tag{2.34}$$

is constructed using the latent factors. By combining $c_{ik}, k = 1,2 \dots R$, and the coefficients used for latent factors construction, the association by the regulator x_j and response variable y_i can be computed. The latent factors are constructed trying to explain variability in both regulators and response variable as much as possible. Li et al. applied PLS regression approach to analysis the association between miRNAs and mRNAs [168]. Firstly differentially expressed miRNAs and mRNAs are identified by *t*-test. Then the PLS model and statistical tests based on bootstrapping were performed to find significant miRNA-mRNA pairs.

2.3. Stepwise Multiple Linear Regression (SMLR)

Correlation or mutual information based methods are fast and straightforward for discovering gene regulatory interactions. However, unlike multiple linear models, they only capture pairwise interactions and ignore that one gene may be regulated by several regulators.

Bayesian models model gene expression by treating each gene as a random variable governed by a probability function determined by the product of conditional probabilities. Bayesian model can handle incomplete or noisy data, combine heterogeneous data types, and avoid over-fitting. Prior knowledge can be included in the model naturally. Due to the exponential size of the search space for possible networks, heuristic search methods are utilized to identify the network. Bayesian network are more suitable for inference of smaller network.

Multiple linear models are natural choice for modeling of many-to-one regulatory relationships. The learnt models can be used for expression prediction. In addition, unlike the Bayesian models, which introduce latent variable s_{ij} and use posterior $P(s_{ij} = 1)$ to access the confidence of prediction, regression based model can use hypothesis test to evaluate the confidence of miRNA j regulate mRNA i. However, due to small sample size, a linear system is usually underdetermined and optimal solution is unattainable. To address the problem, researchers use dimension reduction techniques, such as SVD, or introduce extra penalty term in the model.

In our study, we use stepwise regression to solve the underdetermined system. The expression level of each gene is modeled as a linear function of expression levels of its regulator genes

$$y = a^0 + \sum_{i=1}^{N} a^i x^i \qquad (2.35)$$

The coefficients a_i are identified using stepwise multiple linear regression (SMLR) with a forward selection strategy [169, 170]. The predictors for a given gene are identified starting with the inclusion of the constant term a^0. In each forward selection step, individual predictor variables are considered for addition based on their statistical significance in the regression fitting. The p-value of an F-statistic for each variable is calculated to test the model including and excluding that variable using the null hypothesis that its weight coefficient is zero, using the following equation:

$$F = \frac{SSE^* - SSE}{SSE/(n-p-1)} \tag{2.36}$$

where SSE is the sum of squared error according to the expanded model using $p+1$ predictor variables, and SSE^* is the sum of squared error according to the reduced model using only p predictor variables as follows

$$\begin{gathered} SSE = \sum (y_i - \hat{y}_i)^2 \\ SSE^* = \sum (y_i - \hat{y}_i^*)^2 \end{gathered} \tag{2.37}$$

where $\hat{y}_i$ and $\hat{y}_i^*$ are the values predicted by the expanded and reduced models, respectively.

If the F-statistic is significant, the null hypothesis is rejected, and that particular predictor variable is included in the model. Our forward selection procedure considers the full set of predictor variables, returning a p-value for each one. If any predictor variable had a p-value less than an entrance tolerance, it was added to the model. This ensures that variables with marginal contributions (with a coefficient close to zero) are omitted from the model.

Chapter 3. Regulatory network reconstruction and simulation

3.1. Introduction

The advent of microarray technologies has enabled a high throughput evaluation of gene expression, providing a large scale snapshot of the cellular activity at the molecular level. The availability of these tools has allowed researchers to explore the behavior of entire genomes under different experimental conditions, in a search for mechanistic basis to various cellular behaviors. The analysis of these microarray experimental results has led to new breakthroughs in the understanding, diagnosis, prognosis, and treatment of disease, as well as insights into the functioning of the basic biology of various organisms [171-174].

Gene expression can often be quantified by determining the relative amounts of mRNA transcripts. In microarray experiment, mRNA is harvested from a sample and then reverse-transcribed into cDNA. The cDNA is labeled with fluorescent molecule and then allowed to bind to DNA probes attached to the surface of the microarray chip. The process of complementary binding between the cDNA and the DNA probes on the chip is known as hybridization. The fluorescence values that are measured from the chip enable the quantification of the relative amounts of cDNA present in each sample, which determines the relative gene expression [175]**.**

The techniques for analyzing steady state microarray data are well-characterized [176-179]. However, these techniques are ill-suited to the analysis of time-series microarray data. Time series microarray experiments involve harvesting mRNA from samples at regular time-intervals. This experimental design leads to multiple data points for each gene that can be used to evaluate gene expression over time in a high-throughput manner. Time-series expression data has the potential to provide more comprehensive information about the underlying behavior and inter-relationships of genes than the traditional time-invariant experiments. Furthermore, it can allow for the interpretation of dynamic behaviors in complex biological systems [18]. Time-series microarray data has many applications including the analysis of circadian rhythms, disease progression, drug response, and the study of the cell cycle [18-20].

Each gene's expression can be modified or controlled by various biochemical processes. Transcription factors can directly regulate the synthesis of mRNA, but the expression of genes can indirectly affect the expression of other genes. A gene can inhibit the expression of another gene or it can stimulate the expression of another gene. These activation and inhibition relationships can be represented as a directed graph with nodes representing genes and edges representing the effect of one gene on another. There are several methods of reverse-engineering or modeling gene regulatory networks from two-condition differential expression experiments and time series experiments. These methods include Boolean networks [2, 4-6], correlation networks [7-10], differential equation models [11-16], Bayesian network models [2, 7], and dynamic Bayesian network models [17]. For more detail, please refer to chapter 2.1.

While a great deal of focus has been placed on the network reconstruction problem, the prediction and simulation of gene expression values have not received as much attention. We note that methods developed to infer the presence or absence of regulatory interactions are not directly applicable to the prediction problem. On the other hand, the methods that focus on the prediction problem may not lend themselves to the interpretation of their model for inference of interactions. In this study, we use a linear model to represent gene interaction networks and simultaneously solve the network reconstruction and gene expression prediction problems. The neural network approach of Maraziotis et al [180]. (referred here as FuzzyNet) is closest in its goals to the problems being investigated in this study. In FuzzyNet, a recurrent neural fuzzy network is trained for time series data. While neural networks are generally not amenable to interpretation, the rules generated by FuzzyNet allow identification of regulatory interactions. However, unlike the approach described herein, Fuzzynet does not predict the strength of the predicted interactions and also does not provide a confidence measure for its predictions.

In this study, we present a linear model for time series data and use stepwise multiple linear regression (SMLR) to learn the model parameters from the training dataset(s). To the best of our knowledge, this is the first time a linear model of interaction has been reported to solve the prediction, simulation, and reconstruction problems. The rest of this report is organized as follows. In Section 3.2, we formally define the computational problems and describe our linear model and the process of fitting it to data by using stepwise multiple linear regression. In Section 3.3, we describe the datasets used in the experiments, and present empirical justification for the choice of parameters, including the number of interactions, the statistical significance threshold for interactions, and the number of time points considered in the input. We then present results for the next time step prediction of expression values, the simulation of the entire time course data, and finally, the inference of the regulatory network. Results are compared with similar studies where applicable. We conclude with a summary of our contributions, contrasting with existing solutions.

3.2. Methods

Time series microarray data can be described as a $N \times T$ data matrix, representing the mRNA levels of N genes over T consecutive time points. In this study, we focus on three related computational problems, as illustrated in Figure 3.1. In the single time point prediction problem (Figure 3.1b), one attempts to learn a function that can generate the expression levels in time t from the expression levels at the preceding time point(s). Each pair of time points in Figure 3.1a provides a training instance for learning such a function. In the time series data simulation problem (Figure 3.1c), the entire time series data is generated from only the initial conditions given at the first time point. In this study, we model the simulation problem simply as iterations of the single time point prediction problem, leaving more complex approaches accounting deficiencies of this straightforward extension, such as error accumulation, as future work. In the network reconstruction problem (Figure 3.1d), one attempts to discover the underlying gene regulatory network from the microarray data. While network reconstruction problem is often solved independently [9, 181], we perform network reconstruction via post-processing of the single time step prediction function.

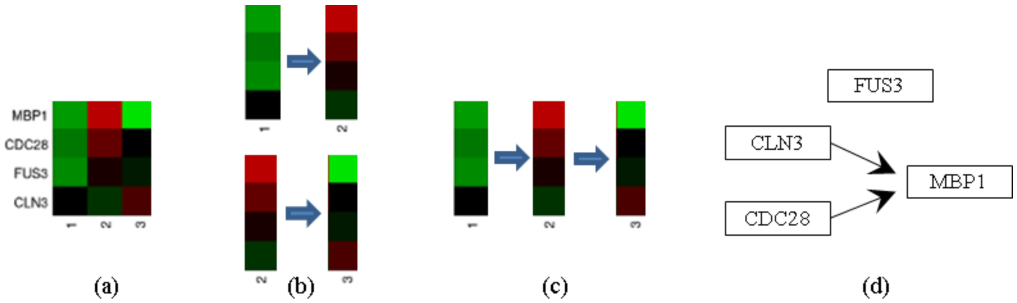


Figure 3.1. Demonstration of microarray time series data and the computational prediction problems investigated in this study. (a) Sample time series microarray data with 4 genes and 3 time points. Red, green, and black colors denote high, low, and medium expression levels, respectively. (b) Single time point prediction problem showing prediction of expression levels at time $\boldsymbol{t}$ from time $\boldsymbol{t-1}$. (c) Simulation of entire time series data from the initial expression levels at time $\boldsymbol{t=1}$. (d) An example reconstructed network involving the four genes, where arrows indicate transcriptional regulation.

We model the expression level of each gene as a linear function of the expression levels of the genes in the preceding time step (this model is generalized to consider multiple previous time points as shown later)

$$g_t^j = w^0 + \sum_{i=1..N} w^i g_{t-1}^i \tag{3.1}$$

where $g_t^j (j = 1,2, \dots, N)$ is the expression level of a response gene g^j at time t, g_{t-1}^i is the expression level of the candidate predictor gene at the preceding time step, N is the number of genes being studied, and w^0 is a constant bias term. F-statistic is calculated as in equation (2.36) in chapter 2.3. If the F-statistic is significant, the null hypothesis is rejected, and that particular predictor variable is included in the model. The forward selection procedure considers the full set of predictor variables, returning a p-

value for each one. If any predictor variable had a p-value less than an entrance tolerance, it was added to the model. This ensures that variables with marginal contributions (with a coefficient close to zero) are omitted from the model.

Since the data were already normalized the constant term w^0 can be set to zero, and without loss of generality, the expression levels of all the genes at time t can be written as

$$G_t = G_{t-1} \times M \tag{3.2}$$

where G is an $N \times 1$ vector of gene expression values and M is an $N \times N$ matrix of weight coefficients. The coefficient matrix M can be converted into a sparse matrix, replacing insignificant interactions with zeros.

The model described above utilizes only the most preceding time point as input. This single time point provides only a static snapshot of the changing gene expression levels. It is not, for instance, directly possible to infer whether the expression level of a gene was going up or down during the preceding time point. We therefore consider the more general case of utilizing prior τ time points, where the expression level of a gene g^j is now modeled as a linear function of all the genes from the preceding τ time points

$$g_t^j = w^0 + \sum_{q=t-\tau \dots t-1} \sum_{i=1..N} w_q^i g_q^i \tag{3.3}$$

Correspondingly, the expression levels of all the genes at time t can now be written as:

$$G_t = [G_{t-\tau}; G_{t-\tau+1}; \dots G_{t-2}; G_{t-1}] \times M_\tau \tag{3.4}$$

where G_t is again an $N \times 1$ vector of predicted gene expression values at time t; the expression levels of the genes at all previous τ time points are concatenated into a single τN vector, and M_τ is a coefficient matrix of size $\tau N \times N$, containing the coefficients from all genes at the previous τ points. The value of τ

can be determined empirically from the mean squared error on the training data, as described in the experiments below. Starting from the first τ time points of a given experiment, the learned coefficient matrix is used to incrementally simulate the rest of the time points.

The weight matrix M (and M_τ) describes the influence of each predictor gene on the response genes. The magnitude of these weights indicates the strength of the interaction and their sign indicates whether the interactions are activating or inhibitory. Each weight is also associated with a p-value, indicating the statistical significance of the corresponding interaction. We rank the interactions by their p-values and use the top k most significant interactions in the network reconstruction, where k can be pre-defined from the average number of interactions observed in real networks or discovered empirically, as presented below to minimize the training error. The accuracy of the reconstructed network is evaluated with respect to a reference network, such as the pathways available in the KEGG compendium [182], using the following measures

$$\begin{aligned} precision &\stackrel{\text{def}}{=} \frac{\#\ of\ correctly\ predicted\ edges}{\#\ of\ predicted\ edges} \\[1em] recall &\stackrel{\text{def}}{=} \frac{\#\ of\ correctly\ predicted\ edges}{\#\ of\ edges\ in\ the\ known\ network} \qquad (3.5) \\[1em] F\!-measure &\stackrel{\text{def}}{=} 2 * \frac{precision * recall}{precision + recall} \end{aligned}$$

Precision, recall, and F-measure each take values in the range between 0 and 1, with 1 being the best score. The ability to rank the interactions by their significance allows us to control the precision-recall trade-off, which is presented as precision-recall plots below. Note that existing approaches produce or report a single precision and recall result; we use the same number of predicted edges in the network for comparison with earlier studies. In comparisons, we denote our approach as SMLR (stepwise multiple linear regression).

3.3. Experiments and Results

3.3.1. Datasets

The time-series datasets modeled in this study are from Spellman et al. [19]. These datasets were generated using four different methods to synchronize Saccharomyces Cerevisiae cell cultures to the same phase of the cell-cycle [19, 20]. The experiments utilized multiple strains of yeast and mRNA was harvested from cells extracted from the cultures at predetermined time intervals. The usage of different methods of synchronizing the cultures resulted in four unique datasets, each named after the synchronization method. Each of the datasets consisted of yeast cells whose cell-cycles had been arrested at a different phase. This results in the different datasets beginning at different phases of the cell cycle.

One dataset (ALPHA) utilized the alpha factor to arrest the cell-cycle and consisted of 18 time points separated by intervals of 7 minutes. A second dataset separated cells by elutriation (ELU dataset). By separating cells of different sizes the investigators were able to extract cells of similar size that were likely to be in the same phase of the cell cycle. They collected daughter cells that were not budding into new cells. This dataset consisted of 14 time points separated by intervals of 30 minutes. These first two datasets were collected by Spellman et al. [19, 175]. Spellman et al. included two further datasets from Cho et al. [20] in their analysis. The third dataset used CDC15 strain of yeast cells, where the cell cycle was arrested by raising the temperature of the culture. This dataset had 24 time points separated by 10 or 20 minute time intervals. We excluded the time points that were separated by 20 minute intervals from our analysis. The fourth dataset consisted of the strain of yeast possessing CDC28 and also was synchronized by temperature change. This dataset had 17 time points separated by 10 minute intervals. All the expression data was normalized so that the mean log2 ratio of the data was 0 [20].

3.3.2. Identification of parameter values

The performance of the linear model was first investigated for the next time step prediction problem. To do so, all ‘predictor-responder’ pairs (i.e., all input-output pairs in Figure 3.1b) were extracted from the four datasets and combined into a single set. In a four-fold cross-validation scheme, three fourths of these pairs were randomly selected for training and the remaining pairs were used for testing. The performance was evaluated in terms of the mean squared error (MSE) of the predicted testing data compared with the real data.

Since we used the *p*-values calculated from the multiple linear regression to determine which genes would be used as predictors of the response gene under consideration, finding a proper cut-off *p*-value was important and prevented us from over-fitting our model to the training data by excluding many insignificant predictors. As demonstrated in Figure 3.2a, the average number of predictors per response gene was directly related to the cut-off *p*-value, but there was no clear plateau for the number of predictors with respect to the *p*-value cut-off. By examining the MSE versus the average number of predictors (Figure 3.2b), we were able to identify an average number of predictors giving a minimum MSE value. Optimum MSE values on the test dataset are obtained for the average number of predictors ranging from 2 to 3, which is in line with the number of interactions observed or estimated by others [23, 121, 183-185]. Using fewer than 2 predictors was insufficient to capture the expression pattern, while using more than 3 predictors resulted in over-fitting. Our experiments using multiple preceding time points also showed similar behavior. Thus, in subsequent experiments, we chose a *p*-value cut-off of 0.025, which provided 3 predictors for each gene on average.

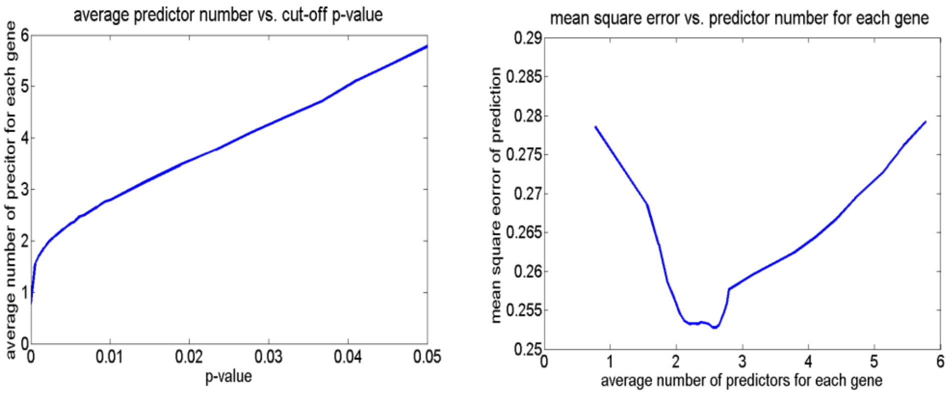


Figure 3.2. (a) The average number of predictors versus the cut-off *p*-value calculated including only the most preceding time step. (b) The mean square error versus the average number of predictors. Similar results were obtained when multiple preceding time steps were considered.

We followed a similar approach for determining the optimal number of preceding time points, τ, to consider in the model. Figure 3.3 shows the MSE for various number of time points used in prediction. The 4-fold cross-validation experiment was repeated 1000 times and the error bars indicate the standard error of the mean for the average MSE in these 1000 runs. The MSE obtained when 2 preceding time points were used was significantly better than the MSE for other values of τ (*p*-value of two sample *t*-test between the MSE for $\tau = 2$ and $\tau = \{1,3,4,5\}$ are 0.0008, 0.0003, 0.0002, 0.0002, respectively). When 2 time points were used, 53% and 47% of the predictors were from the first and second preceding time points, respectively.

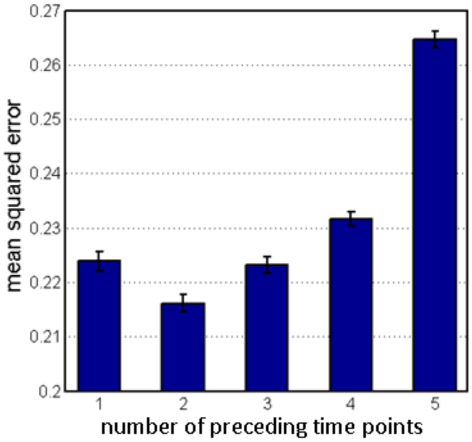


Figure 3.3. The mean squared error versus the number of preceding time points used for prediction. Bars show the average MSE from of 1000 4-fold cross-validation experiments. Error bars show the standard error of the mean.

In order to compare our results with those reported in FuzzyNet [180], we also used CDC15 for training and CDC28 and ALPHA datasets for testing. The result of this comparison is shown in Table 3-1. We note that the training error obtained by our approach is significantly better than those from FuzzyNet for all but two genes. Overall, SMLR is able to provide lower error rates for 50% of the predictions. SMLR incurs very high error rates in the testing datasets for two of the genes, namely the transcription factor MBP1 and the S-phase entry cyclin-6 gene CLB6. We attribute the high error rate in the testing datasets for MBP1 to the fact that the gene expression pattern for MBP1 in the training dataset does not show a clear cyclic expression pattern like the other genes do (Figure 3.5), whereas in the testing datasets, such an expression pattern is observed (Figure 3.6 and Figure 3.7). This may be due to MBP1 being under different regulatory pressures for different cell cycle synchronization methods.

CLB6 has three regulators in the KEGG pathway for the genes used in this study, and we suspect that the nonlinear interaction between these regulators is not sufficiently captured by our linear model. Although we anticipated such cases, we leave inclusion of higher order terms into our model under limited data availability conditions as future work. Despite the high MSE values, SMLR is able to detect the gene expression pattern for CLB6 (Figure 3.6 and Figure 3.7).

Table 3-1. Mean squared error comparison with FuzzyNet for the next time step prediction problem. For both FuzzyNet and SMLR, CDC15 dataset was used for training and CDC28 and ALPHA were used for testing. The average MSE calculated for each gene were compared. Superior MSE values for each dataset and gene are shown in bold.

Dataset:	**CDC15**		**CDC28**		**ALPHA**		**Average**	
Gene	FuzzyNet	SMLR	FuzzyNet	SMLR	FuzzyNet	SMLR	FuzzyNet	SMLR
CLB5	0.17	**0.06**	0.18	**0.08**	0.45	**0.05**	0.27	**0.06**
SWI4	0.36	**0.09**	0.49	**0.08**	0.12	**0.05**	0.32	**0.07**
SIC1	0.45	**0.08**	0.41	**0.31**	0.74	**0.24**	0.53	**0.21**
CDC20	0.55	**0.48**	0.37	**0.07**	0.62	**0.09**	0.51	**0.21**
SW16	0.28	**0.07**	**0.33**	0.36	0.50	**0.13**	0.37	**0.19**
CLN2	0.56	**0.08**	0.58	**0.26**	0.73	**0.23**	0.62	**0.19**
CLN3	0.25	**0.06**	**0.25**	0.27	**0.15**	0.46	**0.22**	0.26
CDC28	**0.13**	0.40	**0.07**	0.41	**0.06**	0.47	**0.09**	0.43
CLN1	**0.19**	0.30	**0.36**	0.61	**0.67**	0.89	**0.41**	0.60
CDC6	0.37	**0.05**	**0.34**	0.97	**0.42**	0.98	**0.38**	0.67
MBP1	0.27	**0.10**	**0.43**	1.91	**0.70**	2.13	**0.47**	1.38
CLB6	0.40	**0.07**	**0.36**	2.86	**0.25**	1.71	**0.34**	1.55

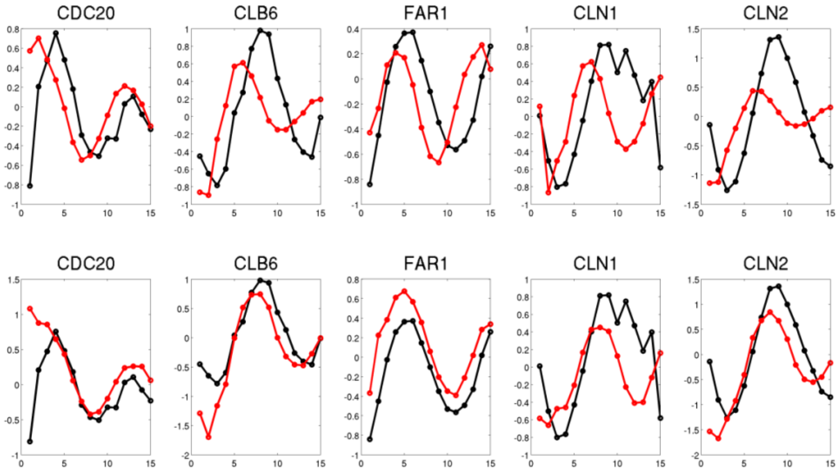


Figure 3.4. Including ELU dataset in training causes error in predicted periodicity. The models were tested on the CDC15 dataset. Upper: training with the ELU, CDC28, and ALPHA datasets. Lower: training with the CDC28 and ALPHA datasets. Real CDC15 data is shown in black, simulated expression levels are shown in red. Expression patterns for only 5 of the genes that best illustrate the error in the periodicity are shown.

3.3.3. Time series data simulation

Taking the next time step prediction function, we iterated the prediction over the entire time course. Only the first τ time points were given as input and the predicted expression levels are fed into the next iteration of the simulation. In each simulation experiment, one of the datasets was left out for testing and the model parameters were trained on the remaining datasets.

We have observed that the simulated expression patterns match that of the real data (Figure 3.4, top row), but with an increase or decrease in the frequency of the expression patterns. We attribute this change in the periodicity to the fact that the datasets were generated with different time intervals, causing the trained function to output an expression level that is not in-sync with the testing dataset. Specifically, the ELU dataset had a time interval of 30 minutes, which is larger than the others (7 or 10 minutes). Testing a model trained for the other three datasets on the ELU dataset would give predictions with an increased period compared to the real data. Conversely, including ELU in training data would give predictions that are beyond the time interval of other datasets, effectively giving accelerated cell cycle for the predicted test dataset. We confirm this by repeating the training with the exclusion of the ELU dataset. As expected, this exclusion corrects the phase-shift in the predictions (Figure 3.4, bottom row).

Excluding the ELU dataset, we performed three additional experiments, taking each of the remaining datasets for testing (Figure 3.5: CDC15, Figure 3.6: ALPHA, and Figure 3.7: CDC28). The simulations covered the gene expressions of 83 genes, which were known to be participating in the yeast cell cycle. We present the simulated results of only 14 genes that are later used for the regulatory network reconstruction. For each simulation, we show the predictions for models with $\tau = 1$ (red) and $\tau = 2$ (green). We observe that the overall expression patterns of the predictions are very well matched with the real data. However, the predictions tend to be conservative in their amplitude compared to the real data (especially see CDC6 and CLB5 in CDC15 dataset; SWI4, FAR1, CDC6, SIC1, and CLN2 in ALPHA dataset; SWI4, CDC20, and CLB6 in CDC28 dataset).

In general, the simulated expression levels follow a smoother trend compared to the real data. This is expected, considering that the real microarray measurements contain fluctuations due to biological variations or noise from the data collection technology. The predictions for $\tau = 1$ and $\tau = 2$ have a high degree of overlap. Using two preceding time points as input results in slightly better predictions (see for instance FAR1, CDC6, SIC1, and CLN2 in Figure 3.6).

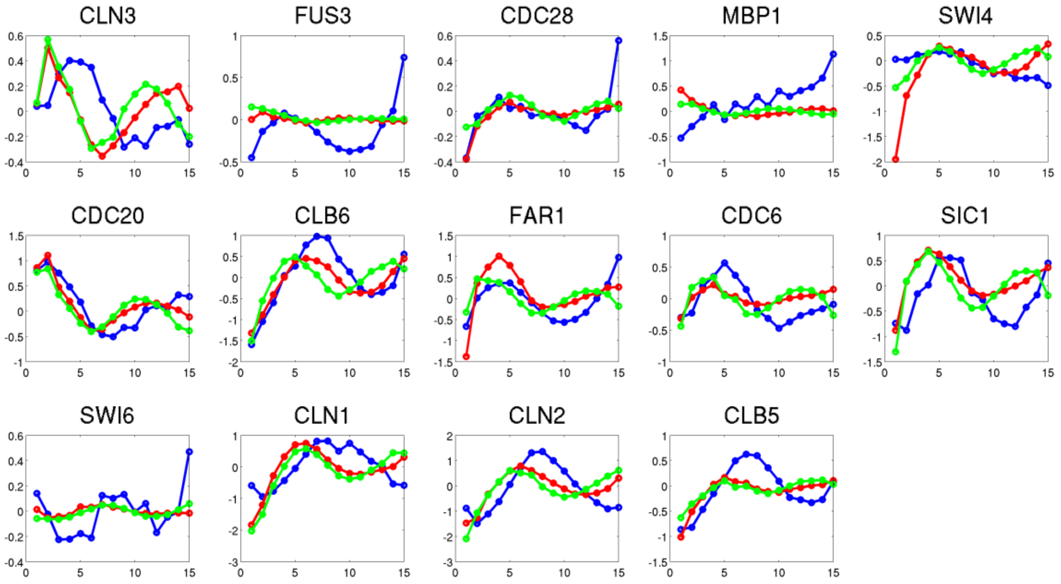


Figure 3.5. Simulated data of CDC15 from models trained from ALPHA and CDC28 datasets using one previous time point (red) or two previous time points (green). The real data is shown in blue.

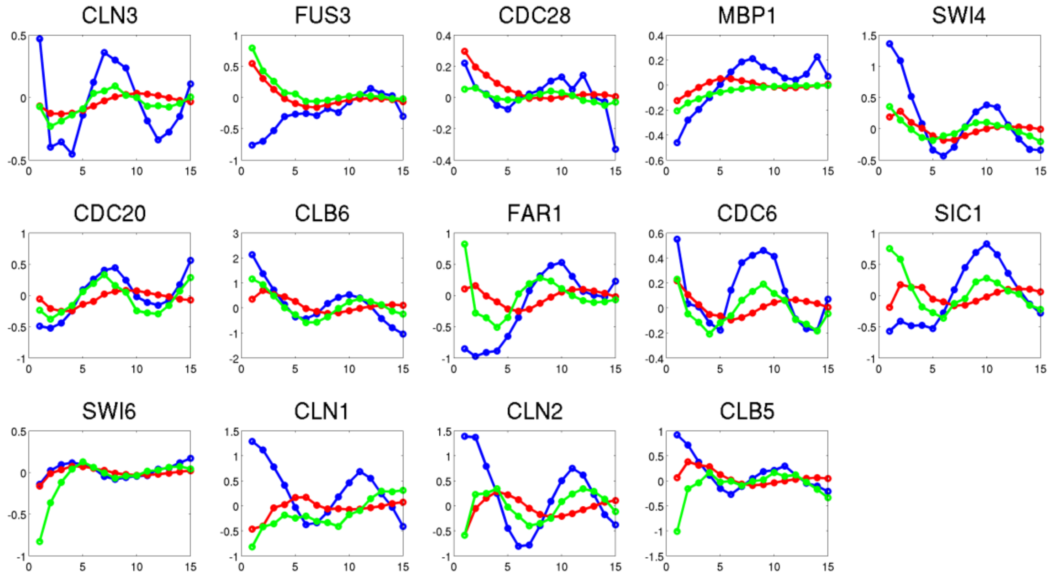


Figure 3.6. Simulated data of ALPHA from models trained from CDC15 and CDC28 datasets using one previous time point (red) or two previous time points (green). The real data is shown in blue.

In order to examine the large scale behavior of the gene expressions, we generated heat-maps for the real and simulated data (Figure 3.8). Two clusters of expression patterns have emerged from the heat-map for the real data. The simulated data using both 1 and 2 time preceding time points are able to preserve these expression clusters. A cluster of genes show up-regulation from 2nd to 7th time points and begin to be up-regulated in the next cycle starting from 14th time point. A second cluster of genes show up-regulation between the 5th and 10th time points. The genes CDC20, MBP1, SWI6 show expression patterns different from the other genes. The highly fluctuating behavior of MBP1 explains the high mean squared error reported for MBP1 in Table 3-1.

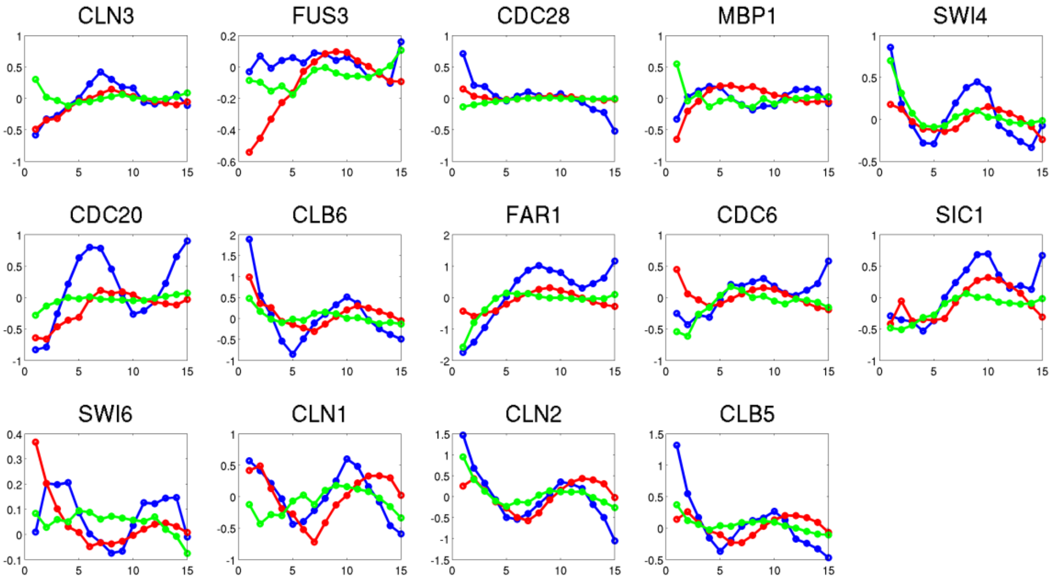


Figure 3.7. Simulated data of CDC28 from models trained from CDC15 and ALPHA datasets using one previous time point (red) or two previous time points (green). The real data is shown in blue.

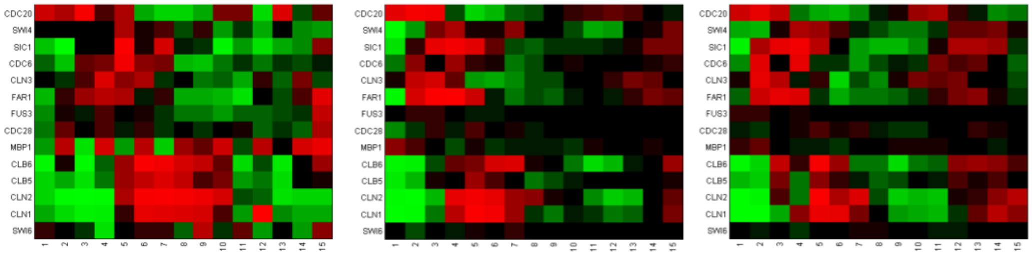


Figure 3.8. The heat-maps show the periodic behavior of the genes over time steps. Left: real data. Middle: simulated data using one time points. Right: simulated data using two time points.

3.3.4. Network reconstruction

Having created a model of the expression of each gene as a linear function of the expression levels of the genes at preceding time points, we were able to directly apply this model to the gene regulatory network reconstruction problem. A central intuitive assumption in this application is that the coefficients of the predictor genes directly reflect the strength of their influence on the respective target response genes in the gene interaction network. The predictors of all genes were compiled into a single list of predicted regulatory interactions, ranked by their *p*-values. These *p*-values were corrected for false discovery rate using the Benjamini-Hochberg method [186]. For a given *p*-value cut-off, the interactions with greater statistical significance were used to reconstruct the regulatory network.

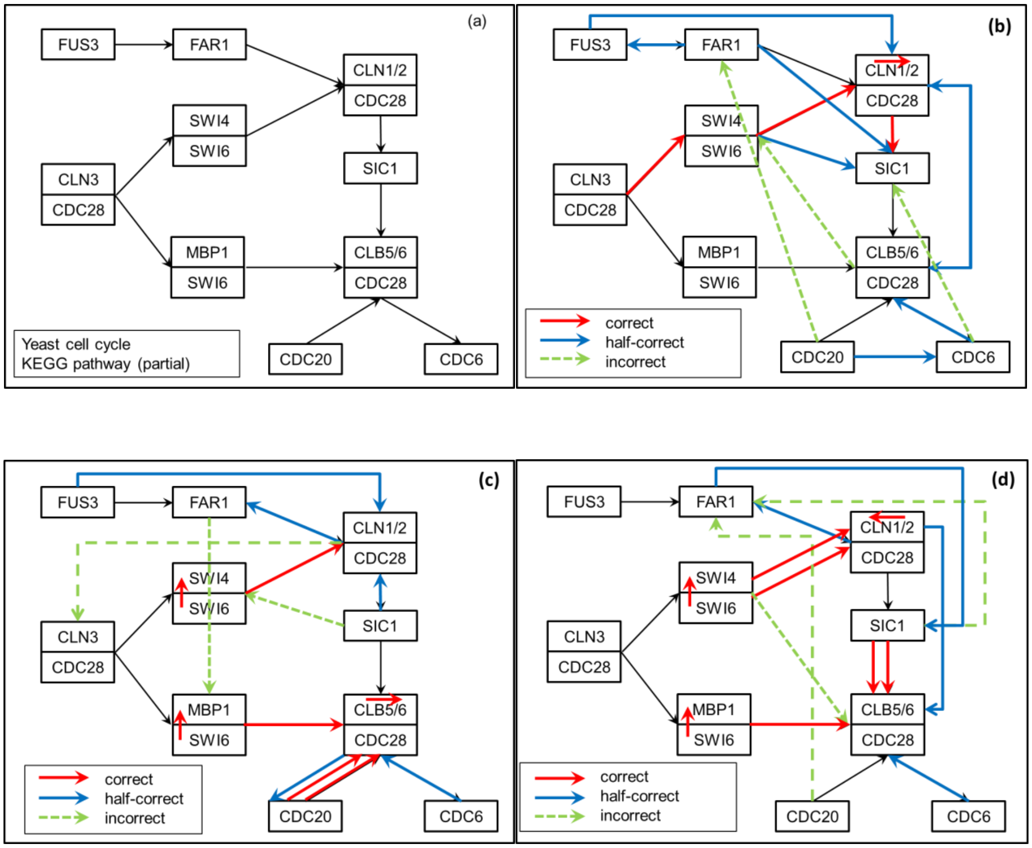


Figure 3.9. Regulatory network reconstruction. (a) Sub-network extracted from Yeast cell-cycle pathway obtained from KEGG. The KEGG pathway contains 51 edges in total; multiple edges between covarying modules are not displayed here. (b) Regulatory interactions predicted by the DBN model [129, 187]. (c) Interactions predicted by a model trained on the CDC28 dataset. (d) Integration of predictions from the four data sets.

The regulatory network was reconstructed by connecting the selected predictors with their response gene using a directed edge. The magnitude of the weights in the model represents the strength of the regulatory interaction, and their sign determines whether it is an activating or inhibitory regulation. For comparison with existing methods that only determine the presence or absence of the interactions, we constructed the regulatory network as an unweighted, directed graph. We compared the gene regulatory networks reconstructed from our model to the networks reconstructed using DBN [187] and FuzzyNet [180] methods. The target network contained 14 genes, as shown in Figure 9a. Using all of the datasets for training, the DBN model predicted 15 edges consisting of 4 correct, 8 half-correct, and 3 incorrect edges (Figure 3.9b), where the correct and incorrect edges are the edges present or absent, respectively, in the KEGG pathway and half-correct edges are those that either capture indirect effects or the reverse direction of interaction. For the same number of edges predicted from the CDC28 dataset alone, our model is able to predict 7 correct, 5 half-correct, and 3 incorrect edges (Figure 3.9c).

Since each edge in our model is associated with a p-value, a straightforward method of integrating the results from all of the datasets is to pool the predicted edges from different datasets and re-rank them by their p-values. Integrating the interactions predicted from each of the datasets in this fashion increases the number of correctly predicted edges to 8 and decreases the number of half-correct predictions to 4 (Figure 3.9d). Each dataset provided support for a different but overlapping set of interactions, where three of the interactions ($\mathrm{SWI6} \rightarrow \mathrm{SWI4}$, $\mathrm{SWI6} \rightarrow \mathrm{MBP1}$, and $\mathrm{SWI4} \rightarrow \mathrm{CLN2}$) were determined highly significant across all datasets. The performance of our method when trained on individual datasets and when trained on all four datasets is summarized in Table 3-2. Excluding ELU and training on the remaining three datasets did not affect the network reconstruction performance.

Table 3-2. Prediction performance of our method for each of the four different datasets separately and for integrating the results from all of the training datasets. Evaluations in this table are based on the top 15 most significant edges predicted from each dataset.

Dataset(s)	**Precision %**	**Recall %**	**F-measure %**
CDC28	46.7	13.7	21.2
CDC15	33.3	9.8	15.1
ALPHA	33.3	9.8	15.1
ELU	26.7	7.8	12.1

Integrated	53.3	15.6	29.9

Next we compare the performance of our method (SMLR) to those of other methods. DN, DBN, and FuzzyNet have reported 14, 15, and 36 predicted interactions, respectively. For direct comparison, we generated three networks by varying the cut-off *p*-value in SMLR, such that the same numbers of edges are obtained. SMLR achieves better precision, recall, and F-measure values when compared with these methods (Figure 3.10). Particularly, the predictions made by SMLR are at least twice more precise and complete when compared with the same number of predictions made by BN and DBN. FuzzyNet makes a larger number of predictions than BN and DBN and performs slightly worse than SMLR for the same number of predictions.

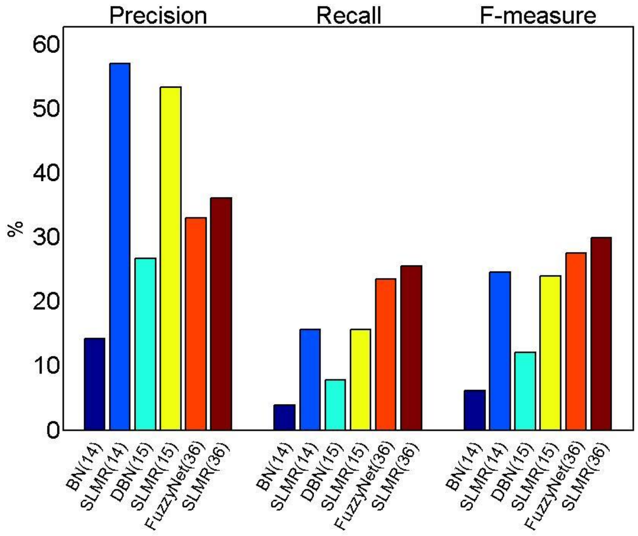


Figure 3.10. Comparison of network reconstruction performance for SMLR and other methods. The number of estimated interactions reported by each method is indicated in the parentheses. The *p*-value threshold of SMLR was adjusted to generate three networks, such that the same number of edges is reported with the method it is being compared to.

Note that our method is additionally able to rank the predicted interactions using their associated statistical significance values, such that any desired number of interactions can be generated. The precision-recall curves of the predictions made by our method for varying *p*-values are shown in Figure 3.11. Integrated predictions outperform predictions from individual datasets in precision, up to a recall of 20%. We attribute this partially to our integration strategy, which focuses on collecting predictions with high statistical significance from individual datasets, biasing the improvement to the top predictions. The performance of our method is slightly better than that of FuzzyNet for comparable precision and recall values.

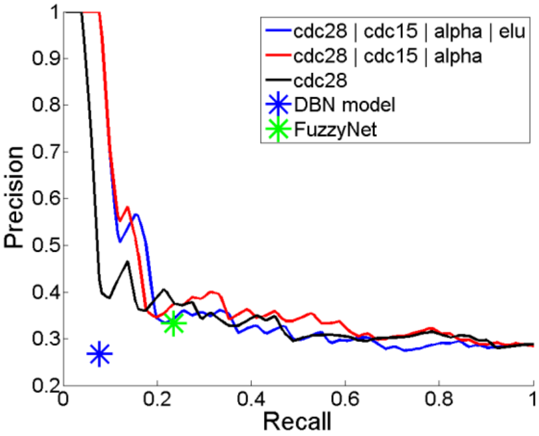


Figure 3.11. Comparing the precision-recall curves for our method with that of others. Results from our method on integrating all datasets, excluding ELU, and using only CDC28 are shown; other individual datasets are omitted for clarity.

In addition to the comparison of SMLR to the methods that are suitable for time-series data, we also compared SMLR to the methods that use steady state microarray data, including ARACNE [7, 17], which is a state-of-art method based on mutual information (MI) calculation. Here ARACNE was used to reconstruct the regulatory network using the same four datasets, where the microarray samples at each time point in the time series were regarded as different steady state samples. The edges predicted by

ARACNE were sorted by their associated MI scores. Besides ARACNE, we also attempted to reconstruct the network by calculating the gene expression correlation between each pair of genes. The edges representing the gene pairs were sorted by the *p*-value of the correlation. Since the results of ARACNE and correlation calculation lack the edge directionality, for the purpose of comparison we consider the presence of an edge as correct if the edge is observed in the known network, without regarding its direction. Figure 3.12 demonstrates that the performance of SMLR using time series data is superior to that of both ARACNE and correlation-based reconstruction. This indicates that utilizing the time series data as a dynamic and dependent set of measurements instead of static independent samples results in a more reliable reconstructed network.

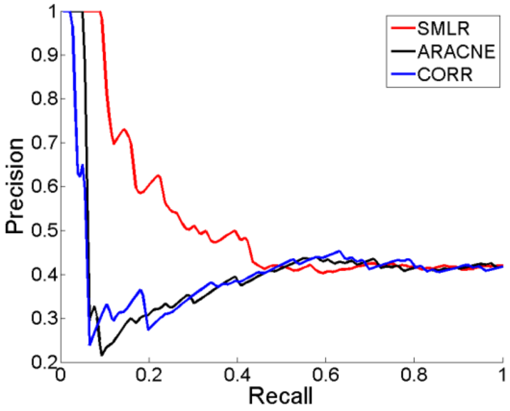


Figure 3.12. Comparison of our method (SMLR) to ARACNE and the correlation-based reconstruction (CORR). Note that unlike the results reported in Figure 3.11, the direction of the edges is disregarded and the interactions predicted by SMLR in either direction were considered as correct. ARACNE and CORR only report un-directed interactions.

In order to further evaluate how well the predicted network is statistically supported from the data, we performed random permutations of the time points and analyzed the resulting predicted interactions (Figure 3.13). The integrated predictions perform consistently better than randomly permuted data sets, at two standard deviations better precision than randomized data sets. This shows that the predictions made by our method are not simply due to spurious expression patterns in the data set due to noise or systematic errors. On the other hand, predictions from individual data sets degrade quickly, and one can be confident in their accuracy only for the top few best predictions. Figure 3.13 also demonstrates the effectiveness of using the *p*-value for ranking the predictions, as concluded from the general trend of the overall monotonicity in the reduction of the precision as more edges are predicted. Surprisingly, the performance of the DBN model is close to the results obtainable by our method for randomized data indicating that the results of DBN may not be statistically supported from the datasets.

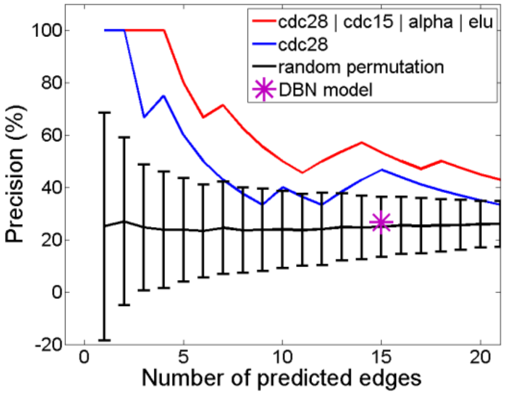


Figure 3.13. Comparison of predictions of our method to its predictions from randomized data. Error bars for the recall of the randomly permuted datasets show its standard deviation in the 100 random trials.

Table 3-3. Coefficients and *p*-values of the predicted interactions from integrated 4 datasets. The sign of the coefficients is compared against the interactions available in the KEGG Yeast cell cycle pathway. Incorrect predictions naturally do not have corresponding information in KEGG. For co-regulated genes, we considered an activating relationship to be correct.

Source Gene	Target Gene	Accuracy	*p*-value (log10)	Coefficient	Sign correct
SWI4	CLN2	Correct	-7.20	1.36	yes
SIC1	CLB6	Correct	-5.79	0.92	no

SWI4	CLN1	Correct	-5.73	1.02	yes
SWI6	SWI4	Correct	-5.12	-1.83	no, co-regulated
FAR1	SIC1	Half-correct	-5.04	0.77	yes (indirect)
SWI4	CLB6	Incorrect	-4.95	2.24	---
SWI6	MBP1	Correct	-4.70	0.94	yes, co-regulated
CDC6	CDC28	Half-correct	-4.29	0.40	yes
CLN2	FAR1	Half-correct	-4.24	-0.65	yes
SIC1	FAR1	Incorrect	-4.20	0.69	---
CDC20	FAR1	Incorrect	-4.06	0.58	---
SIC1	CLB5	Correct	-4.03	0.46	no
CLN2	CLN1	Correct	-4.03	0.64	yes, co-regulated
SWI6	CLB5	Correct	-3.99	-2.29	no
CLN1	CLB6	Half-correct	-3.94	-1.55	no

Another important advantage of our approach over existing methods is the interpretability of the inferred coefficients as the strength of the interactions. We have listed the coefficients for the top 15 predicted interactions in Table 3-3. There are currently no quantitatively annotated datasets for regulatory networks, so we are not able to validate the magnitude of these coefficients directly. On the other hand, the KEGG pathway contains information regarding whether an interaction activates or inhibits the target gene. We observe that the signs of the correctly predicted coefficients match for some of the top predictions. The positive sign of the half-correct interaction $\mathrm{FAR1} \rightarrow \mathrm{SIC1}$ maps to two consecutive inhibitory interactions in the KEGG pathway ($\mathrm{FAR1} \rightarrow \mathrm{CLN1/2}, \mathrm{CDC28} \rightarrow \mathrm{SIC1}$), which effectively makes it an activating interaction.

3.4. Discussion

In this paper, we have employed a multiple linear regression model to predict and simulate time-series microarray data and also to reconstruct gene regulatory networks from this model. Linear models provide a compelling alternative to other existing approaches due to their simplicity, robustness against noise, and low computational requirements. Our approach introduces two additional parameters, in addition to the coefficients estimated in the linear model. Specifically, we have shown that the number of prior time points used to train the model and the *p*-value cut-off of genes to include in the gene expression prediction function can be determined empirically from the training data. We have demonstrated that the proposed model is able to make correct predictions for the yeast cell cycle pathway, and simulate the expressions of the genes involved. The predicted gene expressions showed similar cyclic behavior and similar clustering, when compared with the real data. The linear model presented here is able to model the presence, directionality, and the strength and sign of the interactions in a reconstructed regulatory network. This is an important advantage over most of the existing methods that at best predict the directionality of the interactions.

The statistical significance associated with each predicted interaction provides a convenient way of assessing the reliability of the prediction. Given that most computational prediction approaches to biological problems aim to produce new hypotheses that can be validated with further biological experiments, the prioritization of the predictions becomes an invaluable feature for these time and labor intensive and low-throughput downstream experiments. The statistical significance also provides a straightforward means of integrating multiple time-series data sets, collected under different experimental conditions and time scales. Whereas very short time intervals mean that consecutive time points may not reveal regulatory interactions, longer time points risk missing the regulatory window of action. While each regulatory interaction is likely to operate at different time scales, the integration of the datasets with varying time intervals would be able to collect such interactions into a single predicted network.

Although the network reconstruction was robust to the heterogeneity of the training datasets, the simulation of the time course data was sensitive to the time intervals of these datasets. Of the four datasets used in this study, the elutriation dataset (ELU) was collected at a thirty minute time interval, which was three times longer than any of the other datasets. Inclusion of this data did not prevent the model from capturing the cyclic behavior of genes; however our simulation contained a phase shift compared to the real data. When the elutriation dataset was included in the training (or testing) set, our model predicted changes in the gene expression to occur at earlier (or later) times than they actually occurred in the real data. We conclude that the model should be trained with data collected at similar time intervals to the testing data in order to achieve better performance. Approaches to interpolate the expression levels and thus artificially generate new datasets with the same time interval may be pursued as a potential solution when dataset exclusion is not desirable. In particular, the datasets can be re-sampled from a continuous representation using linear interpolation[18] or spline interpolation [188, 189]. These continuous representations additionally allow re-alignment of datasets to minimize the effects of varying phase and periodicity of the datasets. Such dataset integration methods will be especially useful pre-processing steps when the method introduced in this paper is applied to large scale, heterogeneous datasets.

In this study, the gene regulatory network is reconstructed solely based on time-series expression datasets. By incorporating other types of data and additional a priori knowledge, the performance of GRN reconstruction is expected to be improved. For instance, candidate predictors for a particular response gene can be filtered by focusing on only subset of genes that share common Gene Ontology annotations to the response gene, or by analysis of similarity of protein domains and binding sites of TF targets. Previous known regulatory interactions and transcription factor (TF)-gene relationship identified from ChIP-chip and ChIP-seq data can be included in the structure of network and their effects can be considered during model learning.

In order to identify the predictor genes and fit the model parameters to the data, we have used a stepwise multiple linear regression with a forward selection strategy. This greedy stepwise optimization strategy may not discover a globally optimal solution. Using more comprehensive sampling approaches such as Monte Carlo methods [190], or utilizing related model fitting methods, such as ridge regression [191, 192] and partial least squares regression [193] may improve the model fitting and consequently increase the accuracy of the reconstructed regulatory network, at the cost of increased training time. Known regulatory interactions can also be incorporated as constraints in the search and sampling of predictors during the model fitting stage. Incorporation of known transcription factors improves network reconstruction [194]; consequently, the predictors in our model fitting can be limited to the set of known transcription factors to improve the reconstruction accuracy.

It may be argued that using a linear model for representing regulatory interactions is incorrect or limited. While in this study we do not claim that a linear model should represent the kinetics of regulatory interactions, we have shown that in the context of expression prediction, time-course simulation, and network reconstruction problems, the linear model provides a sufficient approximation to the otherwise complex regulatory interactions. Furthermore, using more complex functional forms would incur a larger number of parameters that need to be estimated from the data, bringing the sufficiency of the available data into question.

In evaluating the accuracy of different methods, we used the interactions available in the KEGG pathways as the ground truth. We acknowledge that future discoveries may change the known interactions in the cell cycle pathway investigated in this study, and alter the evaluations presented in this paper. We also expect that the discrepancies between our predictions and currently known interactions may guide such new discoveries. Furthermore, the view that interactions between pairs of genes should be an either always or never phenomena is limiting, since gene regulation is dynamic and certain interactions may be present only under certain temporal and experimental conditions. The investigation of interactions as emerging or disappearing relationships and the predictions of these dynamic behaviors have attracted recent attention [181].

To conclude, we demonstrated our approach on a relatively small dataset and compared its results to those from Bayesian Network, dynamic Bayesian Network [187] and Fuzzy Neural Network [180] models. Our method generally produced a lower mean squared error for the simulated data than the neural fuzzy network method. We also achieved better accuracy than these methods in reconstructing the Yeast cell cycle pathway. These early comparisons are promising; however a large scale evaluation using a more comprehensive set of synthetic and real datasets and different types of reconstruction methods as well as handling differences in sampling rates is left for future work. Finally, we note that it may be possible to develop a meta-method that combines the predictions of various methods into a single improved regulatory network.

Chapter 4. Inference of miRNA-mRNA interaction

4.1. Introduction

As introduced in the chapter 1.3, so far only a small fraction of miRNA targets have been experimentally validated with confidence. To reduce the number of interactions researchers need to validate, sequence-based computational methods have being used to generate putative lists of miRNA-mRNA pairs. However, predictions solely based on sequence information have high false positive rates [107]. To improve the performance, novel integrative approaches that combine sequence based predictions and experimental data are needed. With the accumulation of high throughput expression data, especially paired miRNA-mRNA datasets, several methods that incorporate these high throughput data have been developed. As introduced in chapter 2.2, there are correlation based model, multiple linear regression model and Bayesian model for analysis of miRNA-mRNA expression.

Correlation-based methods are suitable for fast identification of miRNA-target pairs of interest in large datasets while multiple linear models have advantages that they consider the many-to-one biological interactions between miRNAs and mRNA. In this chapter, we propose an algorithm that combines correlation-based method and multiple stepwise regression models for identification of miRNA-mRNA interaction. By analysis of paired miRNA-mRNA expression data, the algorithm generates a putative list of miRNA-target pairs, which are sorted by confidence. The performances of proposed method are evaluated using miRNA-target pairs found in MirTarBase [195] as true positive.

The proposed method was then used for a case study of hsa-miR-939. Previous study has shown that the expression of miR-939 was significantly altered in samples from patients with complex regional pain syndrome (CRPS) versus control samples [196]. CRPS is a chronic disorder that often triggered by trauma or injury. Inflammation is known to play an important role in CRPS. Inference of hsa-miR-939 targets using paired miRNA-mRNA expression data help the investigation of its role in mediating inflammation and pain.

4.2. Methods

For identification of miRNA-mRNA interaction using paired expression data, we propose an algorithm that hybridizes correlation method and stepwise regression, named as forward stepwise correlation (forwardCorr). As stepwise regression using forward selection, described in section 2.3, the algorithm starts with no predictors included in the multiple linear model. The correlation and the p-values are calculated, and the most correlated predictor (the one with smallest p-value) will be selected first. Next the effect of the selected predictor on remaining predictors and response variable will be removed, and the correlation and p-values are calculated again for the remaining potential predictors.

To remove the effect of selected predictors, suppose that there are p predictor genes selected in the model, then the training matrix $\boldsymbol{X}$ will be $L \times p$ with L samples and p predictors. Instead of using normal equation (2.28) directly, QR matrix factorization is used to solve the least square problem because of its better numerical property. In linear algebra, the QR factorization of the matrix $\boldsymbol{X}$ is written as

$$\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R} \tag{4.1}$$

where $\boldsymbol{Q}$ is a $L \times L$ orthogonal matrix and $\boldsymbol{R}$ is a $L \times p$ upper triangular matrix. There are several methods for computing QR factorization, such as Householder transformations and Gram-Schmidt orthogonalization process. Notice that $\boldsymbol{R}$ has zeros below the main diagonal and $L \geq p$ (the number of selected predictors, p, is small in the beginning of forward selection process), then equation (4.1) can be rewritten as

$$\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R} = [\boldsymbol{Q}_1 \quad \boldsymbol{Q}_2]\begin{bmatrix}\boldsymbol{R}_1 \\ \boldsymbol{0}\end{bmatrix} = \boldsymbol{Q}_1\boldsymbol{R}_1 \tag{4.2}$$

where $\boldsymbol{Q}_1$ contains the first p columns of matrix $\boldsymbol{Q}$ and $\boldsymbol{Q}_2$ contains the remaining $L-p$ columns; $\boldsymbol{R}_1$ is a $p \times p$ upper triangular matrix. After substitute equation (4.2) into the normal equation (2.28), we get

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = (\boldsymbol{R}_1^T\boldsymbol{R}_1)^{-1}\boldsymbol{R}_1^T\boldsymbol{Q}_1^T\boldsymbol{y} = \boldsymbol{R}_1^{-1}(\boldsymbol{R}_1^T)^{-1}\boldsymbol{R}_1^T\boldsymbol{Q}_1^T\boldsymbol{y} = \boldsymbol{R}_1^{-1}\boldsymbol{Q}_1^T\boldsymbol{y} \tag{4.3}$$

Thus the fitted value $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{Q}_1\boldsymbol{R}_1\boldsymbol{R}_1^{-1}\boldsymbol{Q}_1^T\boldsymbol{y} = \boldsymbol{Q}_1\boldsymbol{Q}_1^T\boldsymbol{y}$ and the residue become

$$\boldsymbol{y}_r = \boldsymbol{y} - \widehat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{Q}_1\boldsymbol{Q}_1^T\boldsymbol{y} \tag{4.4}$$

which can be regarded as response variable with the effect of those p predictors removed. Similarly, the effect on the remaining predictors can also be removed. Let $\boldsymbol{X}^{out}$ be the training matrix of remaining potential predictors, it has size $L \times (N-p)$ where N is the number of total predictors and p is the number of predictors already selected. $\boldsymbol{X}_r^{out}$ below will be the new training matrix for selecting next predictor

$$\boldsymbol{X}_r^{out} = \boldsymbol{X}^{out} - \boldsymbol{Q}_1\boldsymbol{Q}_1^T\boldsymbol{X}^{out} \tag{4.5}$$

Now with $\boldsymbol{y}_r$ and $\boldsymbol{X}_r^{out}$ available, one more predictor is selected based on newly calculated correlation and *p*-values. Then $\boldsymbol{X}$ will have one more column and become $L \times (p+1)$; $\boldsymbol{X}^{out}$ have one less column and become $L \times (N-p-1)$. The procedures are repeated such that each step one predictor is selected into the model. Across each step, the *p*-values are recalculated and recorded. In the end, the predictors are ranked according to their 'best' coefficient and *p*-value across all steps.

As an example for demonstration, *p*-values for six predictor miRNAs across first five steps are shown in Table 4-1. The final *p*-value for predictor hsa-miR-138 is 0.44, which is obtained in step 3 while its *p*-value in first step is 0.45. It means that even though the *p*-value for simple correlation between hsa-miR-138 and the response gene is 0.45, hsa-miR-138 can still be considered as a better predictor at step 3, when the multiple linear model contains two predictors already.

Table 4-1. Example of forwardCorr algorithm. *p*-values are calculated at each step and the best *p*-value is used for ranking.

Predictor	Step 1	Step 2	Step 3	Step 4	Step 5	…
hsa-miR-137	**0.02**	0.03	0.16	0.21	0.35	
hsa-miR-376c-3p	**0.39**	0.73	0.89	0.88	0.66	
hsa-miR-585-3p	0.45	0.58	**0.44**	0.68	0.8	
hsa-miR-302c-3p	**0.59**	0.99	0.83	0.9	0.76	
hsa-miR-487a-3p	0.59	**0.39**	0.51	0.67	0.89	
hsa-miR-202-3p	0.77	0.92	0.8	0.95	**0.58**	

One advantage of forwardCorr when compare it to simple correlation methods is that it considers the effect from multiple predictors. Consider an example in Figure 4.1, response gene A is regulated by predictor B, C and D, and there is another potential predictor E that is indirectly correlated to A. By using simple correlation methods, the pair A-E may have smaller *p*-value than the real regulatory pair A-D do, because the effect of D on A is shadowed by the stronger effects from regulator B and C. By removing the effects from B and C, and recalculate the correlation and *p*-value, there is a 'second chance' for predictor D to obtain a proper *p*-value that smaller than the indirect predictor E.

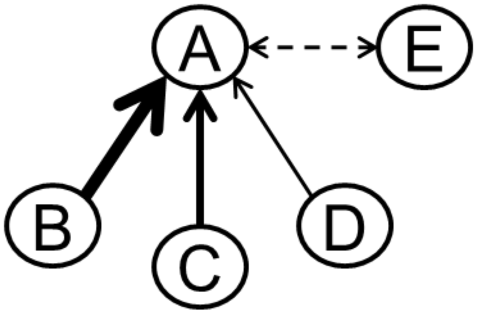


Figure 4.1. ForwardCorr is suitable for many-to-one regulation. Response gene A is regulated by multiple regulators, B, C and D. The simple correlation methods may fail to capture D as predictor of

gene A, since its regulatory effect to A may be mixed with and shadowed by those from its stronger co-regulators B and C.

We can control the number of steps by setting a cut-off *p*-value, pEnter or setting the number of max steps, maxStep. The algorithm stops when there is no significant predictor with *p*-value smaller than pEnter, or the step number exceeds maxStep, whichever comes first.

4.3. Result

4.3.1. Forward stepwise correlation

Forward stepwise correlation (fowardCorr) algorithm is a hybrid of simple correlation and stepwise regression. The difference between simple correlation method and fowardCorr can be seen in Figure 4.2. It shows the results of forwardCorr using two breast cancer datasets, with pEnter = 0.01 and maxStep = infinite. The horizontal axis is the step number and the vertical axis indicates the percentage of final *p*-values come from each step. In the left, for dataset GSE19536, there are about 39%, 22% and14% predictors obtained their best *p*-value at step 1, step 2 and step 3 respectively. Notice that 39% final *p*-value coming from step one means that the 39% *p*-value will be the same when using simple correlation.

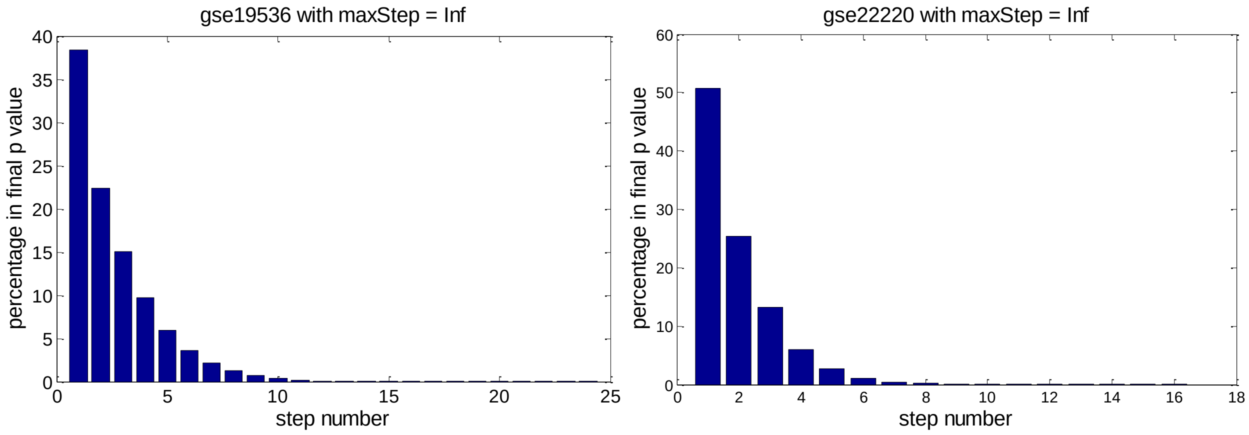


Figure 4.2. Percentage of final *p*-values found at each step.

When the step number becomes large, the residues became so small that we should ignore, otherwise it cause over-fitting and reduce the importance of calculations from beginning steps. In the following experiments, we were using *p*-value cut-off pEnter = 0.01 and limiting maximum step number maxStep = 5.

4.3.2. Comparison to correlation method

ForwardCorr algorithm was run with two breast cancer datasets (GSE22220 and GSE19536) and two prostate cancer datasets (GSE20161 and GSE21032) as we listed in Table 1-2. Using miRNA-mRNA pairs found in MirTarBase [195] as true positive, the precision-recall plot of predicted ranking for each dataset is shown in Figure 4.3. It shows that ForwardCorr have similar area under curve with simple correlation methods in dataset GSE22220, and outperform simple correlation in three other datasets regarding the area under curve, especially in dataset GSE19536. The value of precision is low, in the magnitude of 10^{-3}. One reason is that the validated miRNA-mRNA pairs is only a very small portion of all miRNA-mRNA pairs (MirTarBase/all: 23110/9367820, 32752/16770255, 12781/13802635, 32338/16195287 for four datasets respectively).

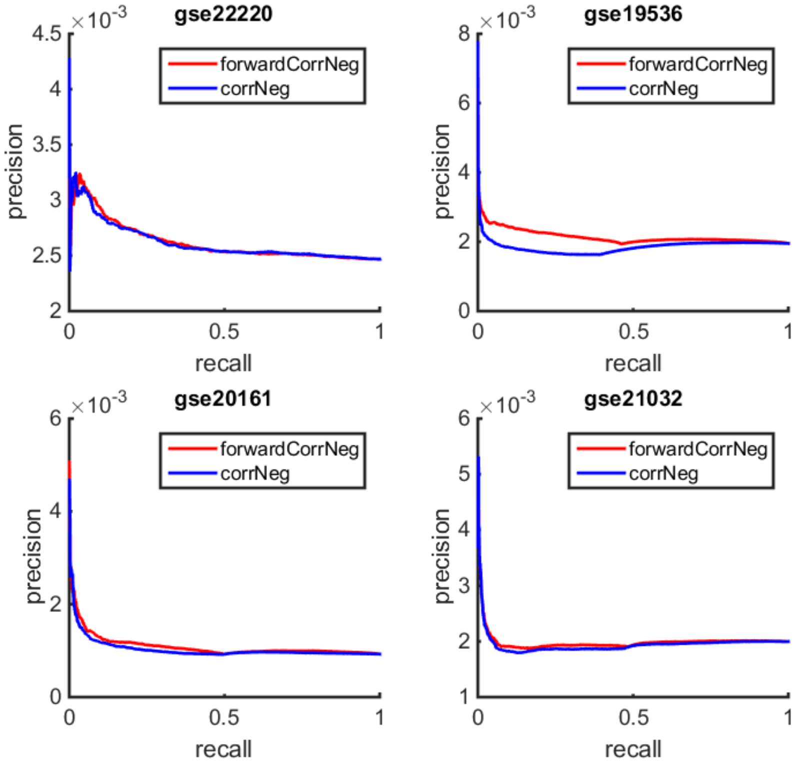


Figure 4.3. Precision-recall. Comparison between ForwardCorr and simple correlation.

4.3.3. Integration results from different cancer

Given that the miRNA-mRNA regulatory pattern depends on cell type and context, a model built from breast cancer datasets may not be suitable for other type of cancer, prostate cancer for instance. The accumulation of paired miRNA-mRNA datasets for various types of cancer (see Table 1-2, as well as The Cancer Genome Atlas (TCGA) project [197, 198]) together provide us the opportunities to infer cancer-specific miRNA-mRNA interaction and integrate the results from each dataset.

In total 16 datasets were used for this experiment. After learning, each dataset generates a list of miRNA-mRNA pairs, ranked by the *p*-value from small to large. Each potential miRNA-mRNA pair has different rank in different dataset. To combine the results, interaction pairs are sorted by the product of their ranks across all datasets.

The precision-recall plot for the combined list of miRNA-mRNA pairs are shown in Figure 4.4. It shows three different subset of all miRNA-mRNA pairs: those predicted by TargetScan (left), those not predicted by TargetScan (middle) and all of them. For each subset of miRNA-mRNA pairs, forwardCorr (red curve) is better than simple correlation method (blue curve). Notice that the subset of TargetScan predicted miRNA-mRNA pairs have higher precision, which may be due to the fact that researchers usually rely on sequence-based methods to filter potential pairs and design validation experiments, thus the density of validated targets is higher within TargetScan.

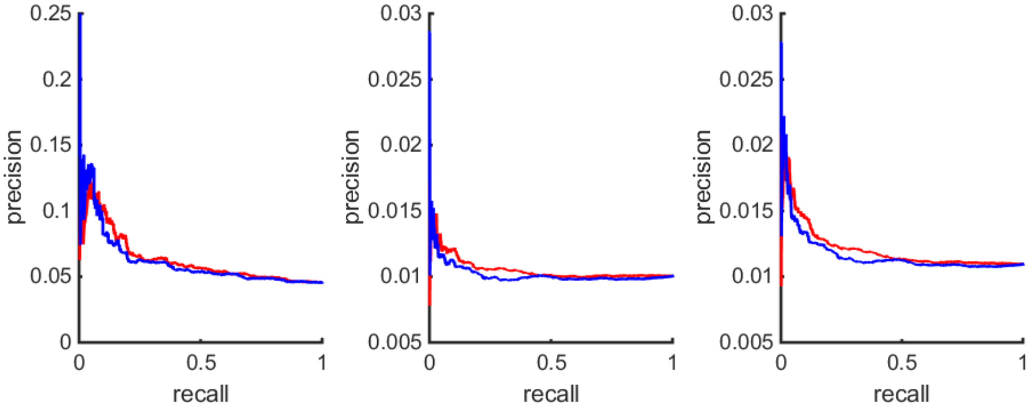


Figure 4.4. Precision-recall curve of combined results from different datasets with forwardCorr and simple correlation.

The area-under-curves (AUCs) in Figure 4.4 are shown in Table 4-2. ForwardCorr have 3%-4% improvement than simple correlation. For this experiment, more than 60% (189/313) well validated pairs (by western blot, qPCR, or reporter assays) in MirTarBase are not predicted by TargetScan. Expression-based methods thus serves as a complementary to sequence-based methods to discover true miRNA-target interactions. For those potential pairs not predicted by TargetScan, forwardCorr have 3.5% better AUC than simple correlation method.

Table 4-2. Area under curve for forwardCorr and simple correlation.

Precision-recall	Within TargetScan	Outside TargetScan	All pairs
Negative correlation	0.0610	0.0102	0.0117
Forward selection	0.0628	0.0105	0.0121
Delta	+2.9%	+3.5%	+3.5%

4.3.4. Integration result from sequence-based methods

In addition to discovering miRNA-target interaction that not identified by sequence-based methods, expression-based results by forwardCorr can be used to improve the performance of sequence-based prediction. The miRNA-mRNA pairs covered by TargetScan are ranked by context+ score [199] and its precision-recall plot is shown in black in Figure 4.5. For the same set of miRNA-mRNA pairs, the result of forwardCorr algorithm is shown in green. By calculating the rank product of results from TargetScan and forwardCorr, the list of miRNA-mRNA is re-ranked and have better AUC than either TargetScan or forward by itself. The numerical values for AUC are listed in Table 4-3.

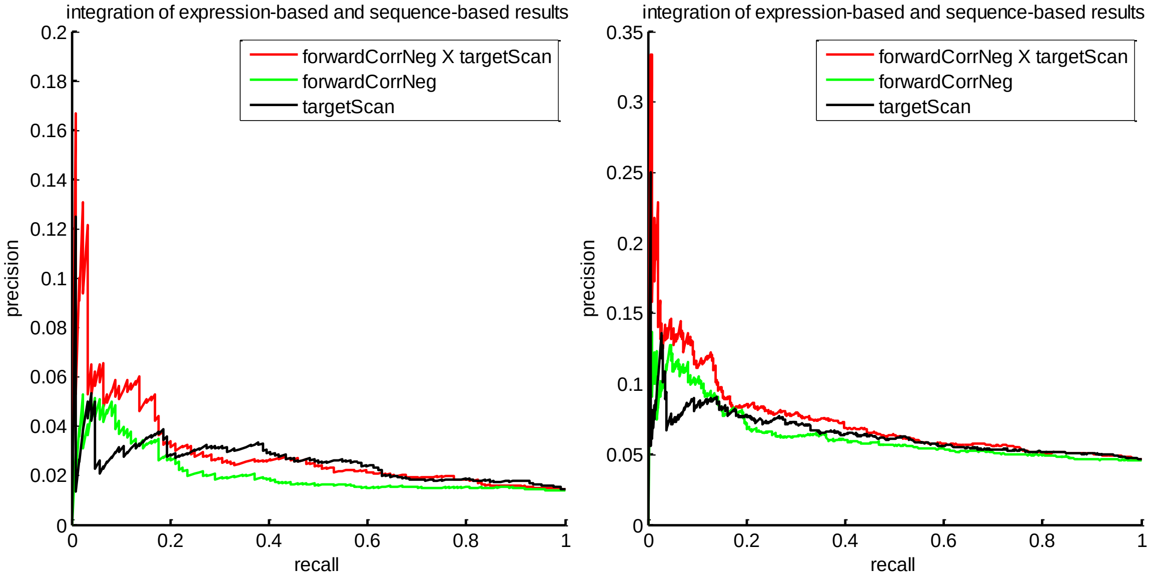


Figure 4.5. Precision-recall curve of combined results of sequence-based and expression-based methods. Integrated results from expression-based methods and sequence-based methods have better area under curve. Green, results from ForwardCorr using gene expression data. Algorithm; Black, results from sequence-based algorithm TargetScan. Red, rank product of both results.

Table 4-3. Area under curve for integrative analysis of miRNA-target interaction

Area under curve	**MTB strong evidence**	**MTB**
targetScan	0.0253	0.0641
forwardCorrNeg	0.0206	0.0628
forwardCorrNeg × targetScan	**0.0291**	**0.0738**
corrNeg	0.0201	0.0610
corrNeg × targetScan	0.0283	0.0724

4.3.5. Case study

To investigate how hsa-miR-939 participate in inflammatory gene regulatory network and its role in mediating inflammation and pain, we first generate gene regulatory network using a seed list of gene, which include known hsa-miR-939 targets from previous study, literature [196] and MirTarBase database (TNF, IL6, TNFAIP1, NOS2, VEGFA, NFKB1, IL1RN, CCL2, SH3BP2 and AMPD2) . In addition, regulatory neighbors of NFKB [200] were also included (PPP2CA, TRAF6, IRAK1, NKRF, HDAC9 , AKT1 , PDCD4, CYLD, PTEN, CHUK and IKBKB). This gene list was uploaded to geneMANIA web server [201], and a network contains 41 genes were generated, as in Figure 4.6.

For target prediction based on expression, 5 paired miRNA-mRNA expression datasets that measures hsa-miR-939 are included. After learning, *p*-values were adjusted [202] and false discovery rate (FDR) 0.05 was used as cut-off, which generate 1002 mRNAs as targets of hsa-miR-939. Seven out of 1002 predicted targets were found in the network, whose interactions with miR-939 are shown as red edges in Figure 4.6. Also there were 178 targets for hsa-miR-939 predicted by TargetScan, and 3 of them are also found in the network, shown as blue edges.

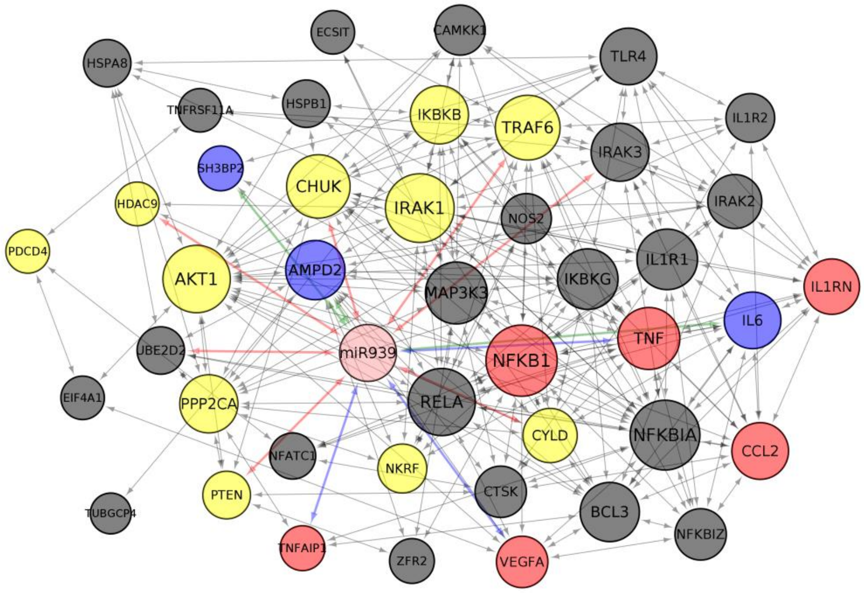


Figure 4.6. Network for hsa-miR-939. Nodes: known hsa-miR-939 targets from previous study (red), from MirTarBase (blue), known neighbors of NF-κB (yellow) and nodes returned from geneMANIA (black). Edges: expression-based prediction using forwardCorr (red), found MarTarBase (green), predicted by sequence-based method targetScan (blue) and retrieved by from geneMANIA (black)

Chapter 5. Functional annotation of miRNA

5.1. Introduction

MicroRNAs (miRNAs) are small (~22 nucleotides) non-coding endogenous RNAs that play important roles in gene regulation by targeting the messenger RNA (mRNA) of protein-coding genes [24]. In most cases, though not always [203], miRNAs act to repress the expression of their target gene [204, 205]. miRNAs guide the repression by either degrading the mRNA molecules, decreasing the translational efficiency, or both. When a miRNA and its target mRNA are highly complementary, the pairing is extensive and the miRNA directs the cleavage of the mRNA, which is the predominant mode of miRNA-guided repression in plants. In animals, extensive miRNA-mRNA complementary pairing and the consequent cleavage of mRNA is less prevalent. Nevertheless, recent studies indicate that target mRNA degradation provides a major contribution to translational repression in animals [47, 206].

miRNAs participate in a wide range of biological processes, affecting the expression of over 60% of mammalian genes [26]. Over the past decade, it has become clear that miRNAs contribute to almost all known physiological and pathological processes, cancer being of particular interest. Since dysregulation of miRNAs is closely linked with dysregulation of oncogenes and tumor suppressors, studying the biological processes of miRNAs provides unique opportunities for the development of miRNA-based diagnostics and treatment of cancer [29, 30].

To understand the functions of miRNAs, a central goal and major challenge is to determine their target mRNAs. There are many experimental techniques for target identification of miRNAs of interest (see chapter 1.3.2). These experimentally identified miRNA-mRNA interactions are collected in several repositories, such as TarBase [67] and miRTarBase [195]. So far thousands of miRNAs have been identified in animals and plants, but only a small fraction of targets for these miRNAs have been validated experimentally, because of the low efficiency and high cost of experimental validation. Sequence-based computational methods have been developed to fill this gap by generating putative lists of miRNA-mRNA pairs, which have greatly reduced the number of interactions researchers need to validate experimentally. Widely used miRNA target prediction methods include TargetScan [26], miRanda [207], PicTar [208], TargetScanS [79], and DIANA-microT [209].

Currently, reliable prediction of miRNA-mRNA interactions remains a challenge. Predictions based solely on sequence information have high false positive rates [107]. In order to improve the performance, novel integrative approaches that combine sequence based predictions and miRNA experimental data are needed. Genome-wide mRNA expression measurement has become an indispensable tool in molecular biology. Similarly, technological advances have spawned a multitude of miRNA profiling platforms [83]. They together provide paired miRNA-mRNA expression profiles that enable researchers to pinpoint important miRNAs and their roles in particular biological processes.

Several methods that incorporate these high throughput data have been developed to find miRNA-mRNA regulatory pairs (see chapter 2.2), including those based on correlation [110, 147, 150, 151] or mutual information [152]. The findings from gene-expression analysis can be integrated with those from sequence-based methods by intersection [149] or weighted sum [110]. These simple approaches are efficient in extracting potential interactions from big datasets but they only consider independent pairwise miRNA-mRNA associations. Since a mRNA can be targeted by several miRNAs and its expression profile is affected by multiple miRNAs at the same time, multiple linear regression models have been proposed [153, 155]. When the data is co-linear or the number of samples is less than the number of regulators, the linear model is underdetermined and optimal solution is unattainable. This can be circumvented by introducing penalty terms to the system, such as L_1 norm, L_2 norm , or combination of both, of the coefficients of regulators [158]. In addition to regression-based approaches, several Bayesian models have been developed, inferring the posterior probability of real miRNA-mRNA interactions based on the expression data, such as implemented in GenmiR++ [159] and its variations [160-162]. Bayesian network structure learning has also been proposed [163], in which regulatory relationships are represented as a graph and the graph that is best supported by the expression data is sought after.

The approaches proposed so far have focused on inference and validation of the “structure” of the miRNA-mRNA regulatory networks from the paired miRNA-mRNA expression data. Although knowing which genes are targeted by which miRNAs is of great value, it is not sufficient for determining whether a gene would be differentially expressed in a particular cellular context.

We have previously shown that a simple linear model is able to quantitatively predict and simulate gene expression levels in time-series data [210]. In this study, we investigate the application of a similar linear model for quantitative estimation of mRNA expression levels from miRNA data. The present study is unique in its focus on explicit quantitative modeling of gene expression levels, rather than just identifying miRNA targets.

5.2. Methods

We infer miRNA-mRNA regulatory interactions by analyzing paired miRNA-mRNA expression data using stepwise multiple linear regression (SMLR) [210]. Suppose there are M mRNAs and N miRNAs of interest, the expression level of each mRNA is modeled as a linear function of the expression levels of the miRNAs

$$y_i = \beta_{i0} + \sum_{j=1}^{N} \beta_{ij} x_j + \varepsilon_i \tag{5.1}$$

where y_i and x_j are variables representing the expression of mRNA i and miRNA j respectively, with $i = 1,2,\ldots,M$ and $j = 1,2,\ldots,N$; ε_i is the error term and β_{i0} is a constant term representing the baseline mRNA expression. The β_{ij} term characterizes the regulatory effect of miRNA j on mRNA i. We identify the coefficient weights β_{ij} using stepwise multiple linear regression with a forward selection strategy, as described in chapter 2.3. Briefly, the predictors for a given gene y_i are identified starting with the inclusion of the constant term. In each forward selection step, individual predictor variables are considered for addition based on their statistical significance in the regression fitting. The *p*-value of an *F*-statistic for each variable is calculated to determine whether to include or exclude that variable in the model, using the null hypothesis that its weight coefficient is zero.

Suppose there are L samples, we can denote the expression of mRNA i and miRNA j across samples as row vectors: $\boldsymbol{y}_i = [y_{i1}, y_{i2}, \dots, y_{iL}]$ and $\boldsymbol{x}_j = \left[x_{j1}, x_{j2}, \dots, x_{jL}\right]$. More compactly, let $\boldsymbol{X} = [\boldsymbol{1}; \boldsymbol{x}_1; \boldsymbol{x}_2; \dots; \boldsymbol{x}_N]$ and $\boldsymbol{Y} = [\boldsymbol{y}_1; \boldsymbol{y}_2; \dots; \boldsymbol{y}_M]$, with semicolon separating row vector and each row representing a mRNA or miRNA and each column representing a sample. If the data is already normalized, the constant term $\boldsymbol{1}$ in $\boldsymbol{X}$ can be dropped, leaving $\boldsymbol{X}$ and $\boldsymbol{Y}$ with dimensions of $M \times L$ and $N \times L$, respectively, and representing the experimental data of M miRNAs and N mRNAs across L samples. Let $\boldsymbol{\beta}_i = [\beta_{i0}, \beta_{i1}, \beta_{i2}, \dots, \beta_{iN}]$ and $\boldsymbol{B} = [\boldsymbol{\beta}_1; \boldsymbol{\beta}_2; \boldsymbol{\beta}_3, \dots, \boldsymbol{\beta}_M]$. Then the SMLR model can be written in a simple matrix form

$$\boldsymbol{Y} = \boldsymbol{B} * \boldsymbol{X} \tag{5.2}$$

The coefficient matrix $\boldsymbol{B}$ is $M \times N$, which represents miRNA-mRNA regulatory interactions from M miRNAs and N mRNAs. Note that the coefficient matrix $\boldsymbol{B}$ is sparse, since the coefficients of insignificant interactions are set to zero.

Before estimating the interaction coefficients from training data and predicting gene expression levels, we need to perform necessary data pre-processing. Since we want to have a general model that works for expression datasets from different platforms and given the fact that most expression data available on Gene Expression Omnibus (GEO) database have already been normalized based on different assumptions regarding the specific platform, we avoid extra normalization across each sample unless necessary. First, we remove probes (genes) that have more than 3 missing data points and impute the missing value using the k-nearest-neighbor method with $k = 3$. Next, we center and scale the expression of each probe (gene) to have a mean value of zero and a standard deviation of one. This transformation does not alter the correlation between genes or the results of t-test for samples from different subgroups. Data preprocessing ensures that expression levels from different samples are on the same scale and that our predicted values can be directly compared with those from the real data. After preprocessing, we estimate the interaction coefficients $\boldsymbol{B}$ using stepwise multiple linear regression [210].

We evaluate the accuracy of the model predictions on both the training and independent testing datasets. In particular, we focus on how well the predictions preserve the differential expression profiles, as the list of differentially expressed genes is one of the most important outcomes from microarray studies. For both the real and predicted data, we perform Student's t-test to identify the genes that are significantly differentially expressed between experimental groups and analyze the overlap between the lists of genes generated from the real and predicted data.

Another type of downstream analysis from miRNA and mRNA profiling experiments is to identify functionally enriched biological annotations and pathways. Enrichment from miRNA profiles relies on first determining the target genes of a list of miRNAs of interest (usually those differentially expressed in a particular experimental subgroup) using sequence-based target prediction algorithms or using experimentally validated targets. The gene list thus obtained is then analyzed for common biological annotations using gene set enrichment methods [108, 109]. Using this strategy, several tools have been developed for functional annotation of miRNAs, including miRGator [110] and FAME [112]. However, these approaches suffer from the limitations of the available miRNA-mRNA interaction data: experimentally validated datasets are far from complete and computational methods produce many false positives. Furthermore, miRNA-mRNA interactions are highly tissue and development-specific [147]. Dependency of individual miRNA-mRNA interactions on a particular cellular context is highlighted by the recent discovery of a network of competing genes and pseudogenes that act as miRNA 'sponges' [211].

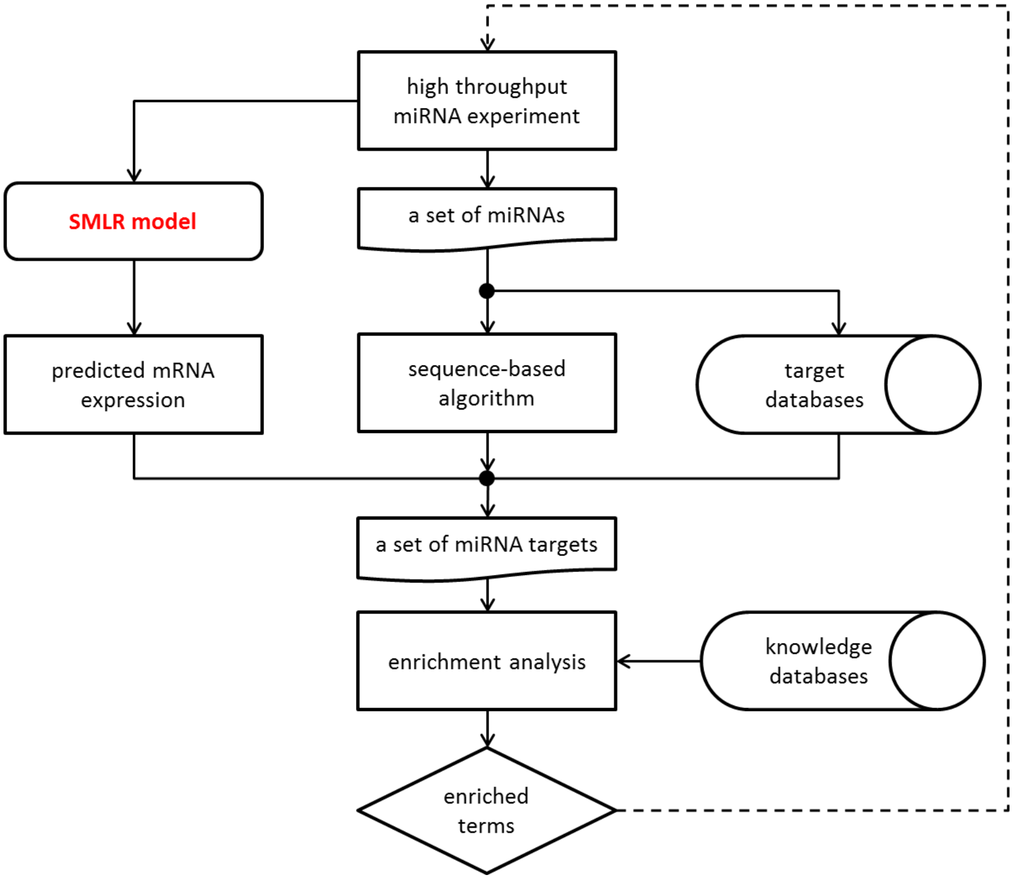


Figure 5.1. Flowchart of miRNA functional annotation for a specific biological process. The conventional strategy is to find a set of miRNAs of interest and perform gene set enrichment analysis with their target mRNAs. Here we propose a different strategy that, instead of starting with selected miRNAs, considers all miRNAs in the experiment for prediction of mRNA expressions based on SMLR model and then identifies differentially expressed mRNAs from estimated expression levels, which are then used in gene set enrichment analysis. Modified from [212].

Here, we propose to use the mRNA levels estimated from our SMLR model for downstream functional annotation tasks (See Figure 5.1). Considering any negative coefficient in the matrix $\boldsymbol{B}$ to indicate a targeting interaction, we evaluate the ability of our approach to discover mRNA targets and compare its performance to the TargetScan target prediction method [79] and to a negative correlation method where negatively correlated miRNA-mRNA are assumed to be targeting interactions (Pearson $p < 0.01$). Note that our method does not distinguish direct interactions from transitive ones or from those arising from co-regulation. Regardless of the source of the coefficients, our approach generates estimates of mRNA expression values, just as if they were obtained from a microarray gene expression experiment study. Once we obtain these estimated gene expression levels, we calculate a predicted list of differentially expressed genes and then perform gene set enrichment analysis using the DAVID web service [213]. Functional annotation is performed against OMIM, GO terms, BBID pathway, and KEGG pathway databases. We evaluate the performance on the functional enrichment task by comparing the resulting functional categories with those obtained from the real mRNA data and those obtained using target prediction methods.

In the following section, we first illustrate the application of SMLR to predict gene expression levels and functional categories, using a breast cancer expression profiling dataset. We then evaluate the ability of the model coefficients estimated from one dataset to generalize to another dataset generated from different experimental platforms. We compare the gene lists and functional categories predicted from miRNA data to those obtained from the real data and from TargetScan.

5.3. Results

5.3.1. Leave-one-out-cross-validation

In order to evaluate the ability of the SMLR model to predict gene expression levels from miRNA data, we first used public available dataset from a paired miRNA-mRNA study [91], in which miRNA and mRNA profiles were obtained from the same primary breast cancer carcinomas (see Table 5-1), where the

TP53 mutational status and estrogen receptor (ER) status of each sample are also available. These samples are part of a larger cohort from the Oslo region [214].

Table 5-1. Breast cancer datasets used in this study.

GEO ID:	**GSE19536**	**GSE22220**
Reference:	[91]	[90]
# samples:	101	207
miRNA platform:	Agilent-019118 Human miRNA Microarray 2.0 G4470B (miRNA ID version) (GPL8227)	Illumina Human v1 MicroRNA expression beadchip (GPL8178)
mRNA platform:	Agilent-014850 Whole Human Genome Microarray 4x44K G4112F (Probe Name version) (GPL6480)	Illumina humanRef-8 v1.0 expression beadchip (GPL6098)
# miRNAs:	489	735
# mRNAs:	40,989	24,332

After data preprocessing, we obtained normalized expression profiles for 489 miRNAs and 40996 genes. We then performed leave-one-out-cross-validation (LOOCV) to evaluate the model, where we set aside one of the samples as the test sample and calculated the interaction coefficients from the remaining 100 training samples. The resulting model is then applied to the miRNA profiles from the test sample separately. This procedure is repeated with each sample in the dataset used as a test sample.

Hierarchical clustering of the 1000 most differentially expressed mRNAs in the real data is shown in Figure 5.2 left. For comparison, a heatmap of the predicted expression levels is shown side-by-side in

Figure 5.2 right with the same row and column arrangements. The predicted data displays surprisingly similar expression patterns, supporting the idea that the miRNA expression alone provides a good summary of the gene expression state of the cell.

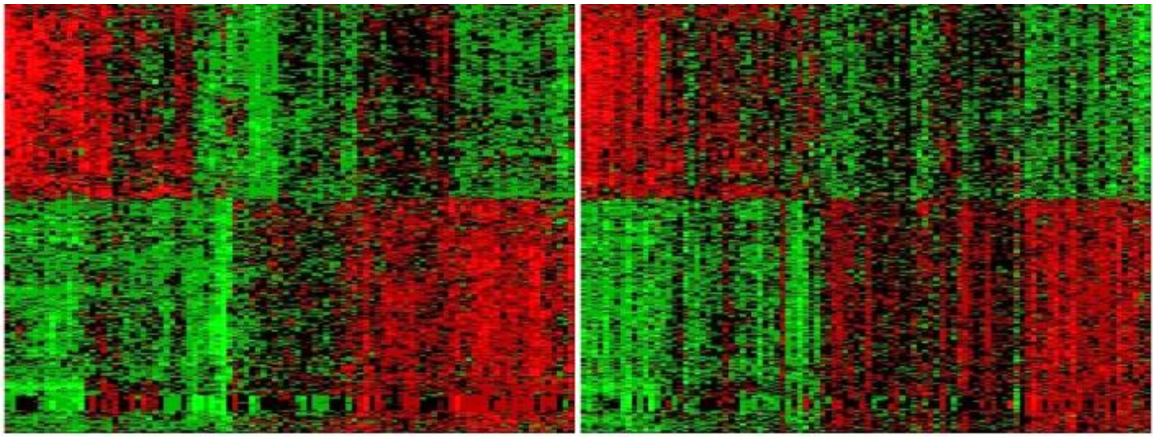


Figure 5.2. Hierarchical clustering of mRNA expression. Left: Hierarchical clustering of the 1000 most differentially expressed mRNAs from the GSE19536 dataset. Right: expression levels of the same mRNAs predicted from the paired miRNA expression data, using SMLR with leave-one-out-cross-validation strategy. Rows are mRNA probes and columns are samples. Predicted data is shown with the same row and column arrangement as the real data. Root mean squared error (RMSE) of all predicted values was 1.11.

In order to further evaluate the reliability and usefulness of the gene expressions predicted from miRNA data, we examined whether the predicted values can identify a similar set of differentially expressed mRNAs. A two-sampled *t*-test on predicted gene expression data was performed between the ER-positive and ER-negative subgroups of samples. The *p*-values of the *t*-test are compared to those obtained from the original gene expression data (See Figure 5.3). These two set of *p*-values are highly correlated (r = 0.77). The mRNAs that are differentially expressed in the real data were likely to be found differentially expressed in the predicted data as well.

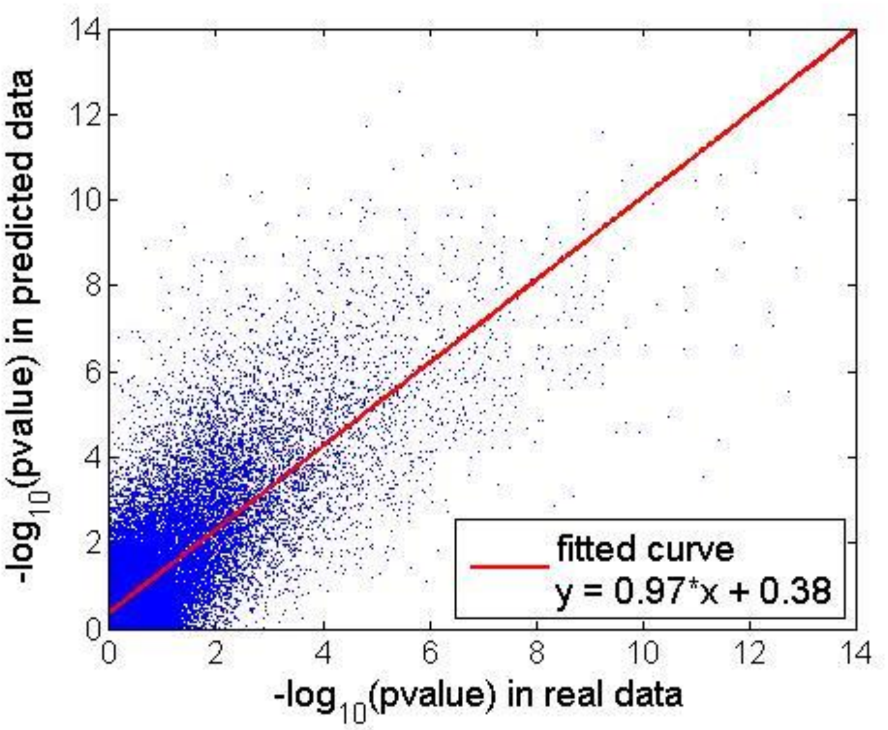


Figure 5.3. Comparison of differentially expressed mRNAs identified from the real and predicted expression data. Each point represents a mRNA, where the x and y axes show the -log$_{10}$ transformed *p*-values obtained from an unpaired *t*-test in real and predicted data, respectively, comparing ER-positive and ER-negative breast cancer samples. The least-square fitted line is shown in red.

Genome-wide microarray analysis is often used to prioritize a set of genes for follow-up wet-lab experimentation, such as reporter assays to confirm transcription, measurement of protein levels by northern blots, or knock-out experiments to evaluate phenotypic outcomes resulting from the absence of a gene. As such, it is important that our predictions preserve the ranking of the differentially expressed genes. Figure 5.4 shows the overlap between the top-k most differentially expressed gene sets obtained from the real and predicted data. The figure also shows the amount of overlap for gene sets obtained with the commonly used *p*-value thresholds of 0.01 and 0.05. At different top-k or *p*-value cut-offs, about half of the genes from the predicted gene set are in common with the result from real gene set.

Considering the noisy nature of gene expression data and the biological complexity of the rules governing translation of mRNAs to different protein isoforms, differential expression detected in microarray experiments is not conclusive for similar expression of the encoded proteins or for regulation of a particular phenotype the genes are involved in. Gene set enrichment is commonly utilized to find biological functions affected by the concerted changes in a set of genes.

Based on the prediction of mRNA expression, we propose a strategy for functional annotation in miRNA studies, as illustrated in Figure 5.1. For a miRNA study, the functional annotations of miRNAs of interest can be obtained by enrichment analysis with a set of their target mRNAs. Traditionally, the set of miRNAs of interest are selected according to their differential expression patterns and their targets are selected from sequence-based target prediction algorithms or from experimentally validated targets. All targets of differentially expressed miRNAs are then (falsely) assumed to also be differentially regulated, even though these target genes are also targeted by other non-differentially expressed miRNAs. This is an unrealistic assumption that results in thousands of genes, limiting the statistical power of the enrichment analysis. This is demonstrated in Figure 5.4, where we compare the accuracy of the genes assumed to be differentially regulated from negative correlation and TargetScan predictions (17% and 22%, respectively) with those obtained from our method (63% and 67% for the same number of genes). Compared to context-agnostic target-prediction methods, we more effectively utilize the cellular context available from the state of all miRNAs in determining whether a gene is differentially expressed.

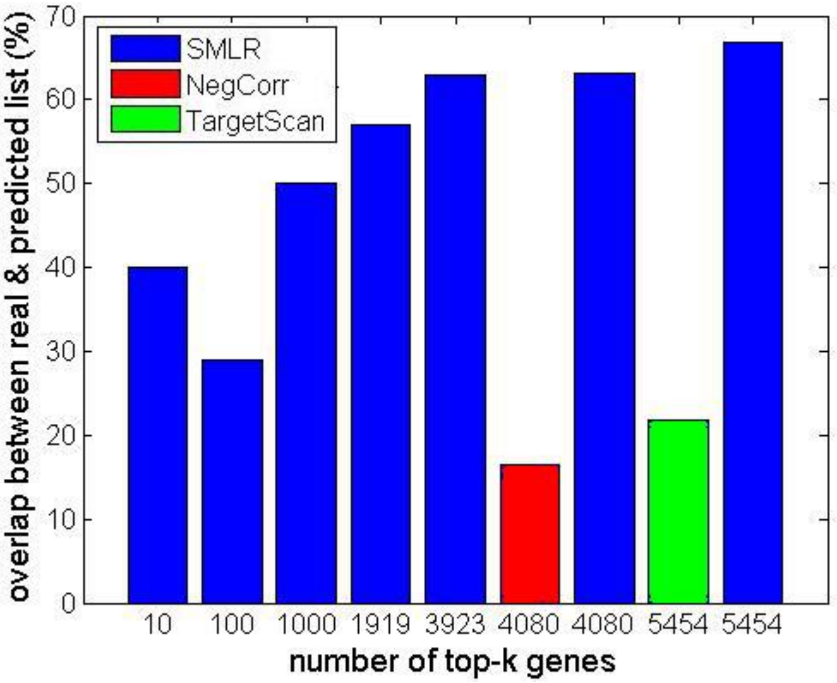


Figure 5.4. Amount of overlap between the lists of differentially expressed genes in real and predicted data. Percentage overlap between the most differentially expressed gene sets obtained from real and predicted data is shown. Each bar shows gene sets obtained with either a top-k or *p*-value criteria. After false discovery rate (FDR) correction, there were 1923 and 3942 mRNAs with *p*-value < 0.01 and 0.05, respectively.

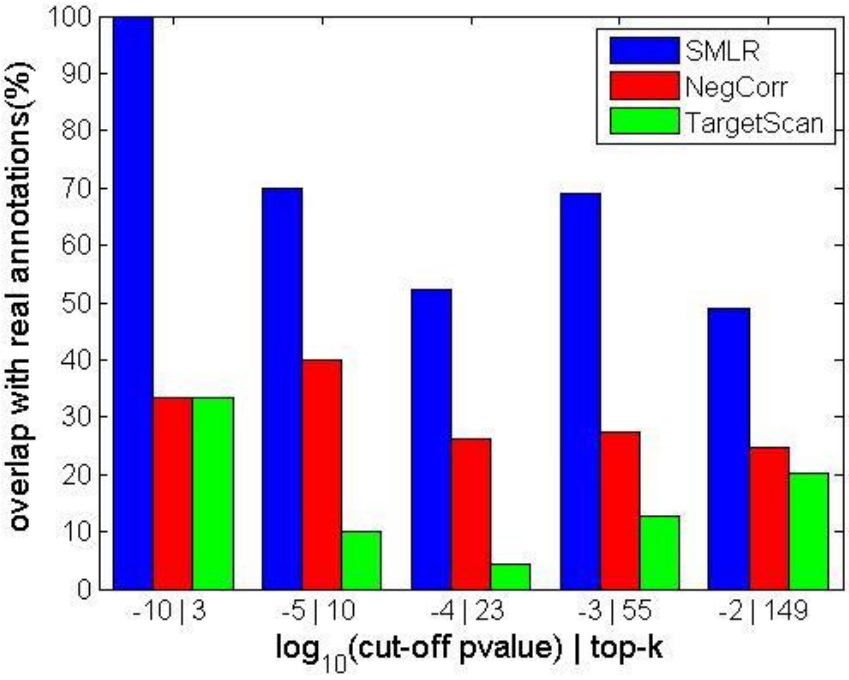
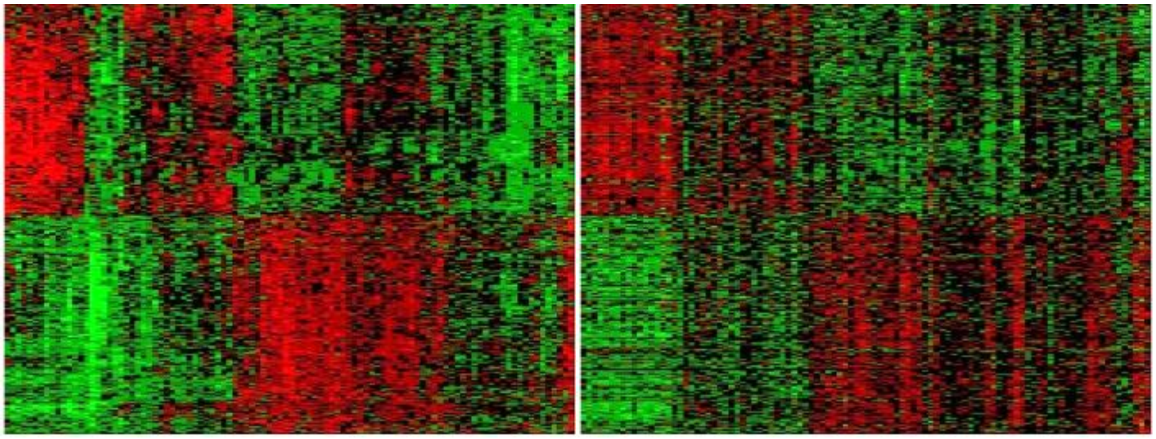


Figure 5.5. Functional enrichment from different methods. Percent overlap of functional annotations obtained from different methods with those obtained from real data are shown. At each *p*-value modified Fisher Exact P-Value from DAVID, the same number of top-k annotations from each method are compared.

In order to compare the functional annotations obtained from different methods, we used the DAVID web service to perform gene set enrichment [213]. The annotation obtained from real expression data was used as the ground truth. To generate differentially expressed gene lists, cut-off *p*-value 0.01 was used for *t*-test with real and predicted expression. For TargetScan and negative correlation methods, the gene lists were formed by combining all of the targets of differentially expressed miRNAs ($p<0.01$). Overlap of the functional annotation terms obtained from different methods with those generated from the real data is shown in Figure 5.5. Top-3 functional categories enriched from the real data were: Phosphoprotein, Alternative Splicing, and Splice Variant. SMLR was able to generate the same three terms in its top-3, whereas TargetScan and negative correlation only ranked only one of them in their top-3 lists. For the top-10 functional annotations obtained from each method, 70% were in common between results from real data and SMLR prediction, sharing similar rankings in statistical significance; while 40% and 10% were in common for negative correlation and TargetScan methods. These results support the claim that gene expression values predicted from miRNAs alone can capture the affected biological processes and that the functional annotations from estimated mRNA values are more accurate than those from collection of predicted targets.

5.3.2. Cross-dataset prediction

The results in the previous section were obtained by leave-one-out cross-validation within a single experimental study, where each miRNA to mRNA mapping in a test sample was done using a model trained on the rest of the samples. In this section, we evaluate the cross-database performance of SMLR by applying the model trained from one study to another dataset from an independent experimental study. Specifically, we train a model on GSE22220 dataset of human primary breast cancer samples from Oslo region [91] and test its prediction performance on GSE19536 dataset from early breast cancer patients in Oxford [90]. Since miRNA-mRNA interactions are highly tissue-specific and development-specific, we focus on datasets from the same cancer type here. Although both datasets were from breast cancer samples, they used different microarray platforms for mRNA and miRNA profiling (See Table 5-1).



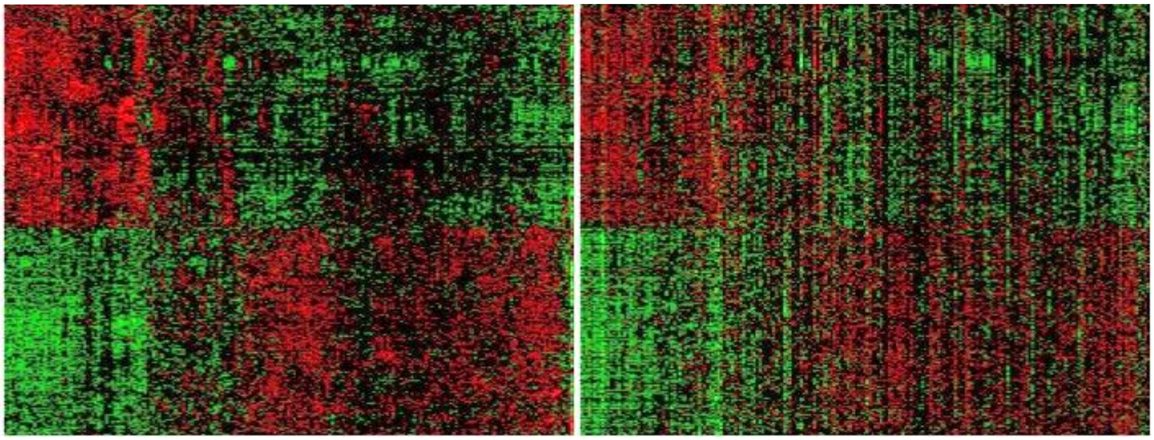


Figure 5.6. Hierarchical clustering of true (left) and cross-database predicted (right) mRNA expression. Top: SMLR is trained with GSE22220 dataset and tested on GSE19536 (RMSE=1.02). Bottom: SMLR is trained with GSE19536 dataset and tested on GSE22220 (RMSE=1.26). Top 1000 most differentially expressed mRNAs with respect to ER-status are shown. Hierarchical clustering is only done on the real data (left); and the same row-column ordering is used to display the predicted data (right).

In order to perform a cross-database application of the model, we first find the mRNAs and miRNAs that are in common between the two studies. Since the studies use different microarray platforms with different probe IDs, we convert the mRNA probe IDs to their GeneBank accession numbers and the miRNA probe IDs to their miRBase IDs. This results in 14873 mRNAs and 232 miRNAs that are in common between the two studies.

The comparison of the heat maps generated from real and predicted data illustrates that SMLR is able to predict the overall expression profiles that reflect the ER status of the samples (See Figure 5.6, top row). We observe the same behavior when the training and test datasets were switched (Figure 5.6, bottom row). Taking the differentially expressed mRNAs from the predicted GSE22220 data (*p*-value<0.01) and performing gene set enrichment, again finds functional annotations that are in better agreement with those obtained from the real data, when compared to the agreement of the annotations resulting from the TargetScan or negative correlation methods (Figure 5.7).

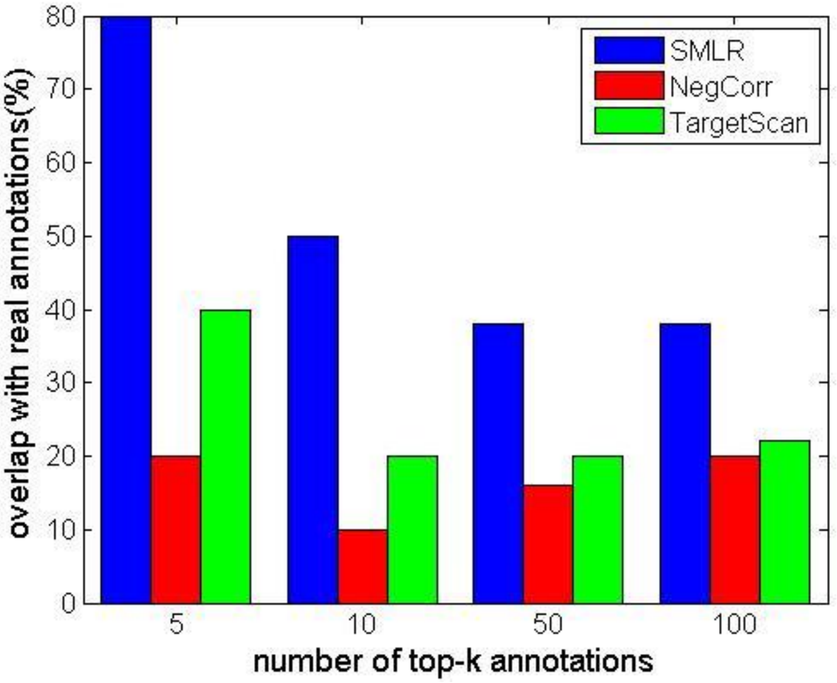


Figure 5.7. Comparison of functional enrichment in GSE19536 dataset. SMLR is trained using GSE22220 dataset and differentially expressed genes from the predicted GSE19536 data are used for gene set enrichment. Negative correlation and TargetScan methods use all the predicted targets of differentially expressed miRNAs in GSE19536.

5.3.3. miRNA-mRNA target prediction

Although our main focus in this study is quantitative prediction of mRNA expression levels, some of the underlying predictors discovered by SMLR model may be from direct miRNA-mRNA target interactions. Specifically, some of the coefficients β_{ij} in equation (5.1) (which make up the matrix $\boldsymbol{B}$ equation (5.2)) may represent direct miRNA-mRNA targeting interactions. We assess the extent in which SMLR can discover such targeting interactions by comparing these interactions with known miRNA targets in miRTarBase and predicted targets in TargetScan.

The SMLR model was trained on both GSE22220 and GSE19536 datasets combined and the miRNA-mRNA pairs in the model with negative coefficients, representing a potential targeting effect, were collected. Here, we consider only the 248 miRNAs for which there was at least one such targeting interaction. There were on the average 8 experimentally validated targets for each of these miRNAs, listed in miRTarBase. TargetScan had an average of 341 predicted targets per miRNA. Considering miRTarBase as the ground truth, the accuracy of miRNA-mRNA target pairs predicted by SMLR was 0.10% (41 correct out of 40,633 predictions), whereas TargetScan had an accuracy of 1.12% (944 out of 84,489 predictions) and the negative correlation method had an accuracy of 0.05% (222 out of 428,048 predictions).

Although SMLR had a lower accuracy than TargetScan, we must note that the coverage of miRTarBase is currently very limited. Consequently, these accuracy measures are sensitive to availability of further experimentally validated target data. Furthermore, whereas SMLR finds interactions specific to the datasets it is trained with, namely the breast cancer samples, miRTarBase dataset and TargetScan predictions do not provide any context-specific information for their target interactions. Regardless of these drawbacks in the analysis, combining the predictions from SMLR and TargetScan, by intersecting their miRNA-mRNA target pair lists, achieves an accuracy of 2.17% (23 correct out of 1,060 common predictions), which is better than application of either method alone.

5.4. Discussion

In this study, we took a radically different approach to miRNA-mRNA interactions and used a multiple linear regression model to directly estimate the mRNA expression levels from miRNA data. Whereas traditional methods try to determine targets of individual miRNAs and rely on these target lists for downstream functional analysis, we estimate mRNA levels from the cellular context captured by the collection of miRNAs. The benefits and opportunities provided by our approach are tremendous. For instance, our approach makes it possible to computationally predict mRNA levels for media, such as serum, where miRNAs are relatively stable and easy to extract and measure with current experimental techniques but mRNAs are less stable and more challenging to measure.

Traditionally, after identifying differentially regulated miRNAs, researchers would sift through hundreds or thousands of targets of these miRNAs and subjectively pick several targets of interest for further experimental validation, e.g., to test for binding of miRNA to mRNA or for differential regulation of the mRNA. Not only are these target lists non-specific to the tissue type, developmental stage, or environmental factors involved in an experimental study; they also ignore the fact that these genes are targeted by multiple miRNAs, some of which may not be differentially regulated or may be regulated in different directions. In our approach on the other hand, we build a model in a cell-type specific manner, connecting multiple miRNAs to each mRNA. We believe that a prioritization of the target genes based on estimated expression levels will result in a higher positive rate in validation experiments.

Our choice of the SMLR model for prediction of mRNA expression levels was based on its simplicity and interpretability. We believe that the linearity assumption used in SMLR provides an appropriate trade-off between the power and generality of the model and the number of parameters that can be correctly estimated from the currently available datasets. Furthermore, the interactions obtained from linear models were previously found to be better than those generated from Bayesian models and Neural Networks [210].

In this study, we mainly focused on breast cancer datasets and demonstrated that a model trained in one experimental platform can be successfully applied to miRNA data from an independent laboratory using different experimental platforms. Although it is possible to apply a model trained on one tissue type to miRNA data from another tissue type; the predicted gene expression values would not be as accurate as restricted the predictions to the same tissue and comparable experimental conditions. For example, applying the model trained on the breast cancer dataset GSE22220 to predict gene expression values from miRNA data in a prostate cancer study GSE20161 resulted in a mean squared error of 1.35, about 33%

higher than the error when it was applied to another breast cancer dataset GSE19536. In our future work, we will build a repository of models for different tissue types and experimental conditions of interest. The limiting factor for building such a repository will be the availability of high quality paired miRNA and mRNA data collected from the same samples.

Although our main focus was not identification of the direct miRNA-mRNA targeting interactions, we show that the interactions with negative coefficients in our model can be indicative of direct regulation. Note that the targets from our model were generated only from the two breast cancer studies. We expect that a large scale modeling from all publicly available paired miRNA-mRNA datasets will provide target predictions that are in better agreement with experimentally validated targets. Motivated by the observation that targeting interactions obtained from two breast cancer datasets can improve the accuracy of TargetScan predictions, we expect that our approach will provide a means of improving sequence-based target predictions in a context-specific manner.

Chapter 6. Summary

In our study, a stepwise multiple linear regression (SMLR) model is proposed to learn the gene-gene regulation and miRNA-mRNA interactions. With this model, we used time-series data for reconstruction of gene regulatory network (GRN) and paired miRNA-mRNA expression data for inference of miRNA-mRNA interactions.

SMLR model is suitable for prediction of gene expression. After learning from time-series data, the model is capable of predicting future gene expression and even simulating the whole time series. For miRNA-mRNA study, the expression profiles of mRNA can be predicted from miRNA expression using the learnt SMLR model.

With the predicted expression profile, one can perform further analysis that relies on expression while the real expression is not available. In this study, we performed the miRNA functional annotation by using the predicted mRNA expression.

Compared to other multiple linear models, SMLR is computationally cost-effective for solving the underdetermined system and its forward selection procedure is naturally suitable for sparse network such as GRN.

For future work, reliability of gene expression prediction via SMLR model should be further tested and improved by analysis of more expression profiles, especially from next generation sequencing. With the accumulation of expression profiles in cancer studies, specific SMLR model may be learnt for different tissue type and cancer type.

Bibliography

1. Feist, A.M., et al., *Reconstruction of biochemical networks in microorganisms.* Nat Rev Microbiol, 2009. **7**(2): p. 129-43.
2. Gardner, T.S. and J.J. Faith, *Reverse-engineering transcription control networks.* Physics of Life Reviews, 2005. **2**(1): p. 65-88.
3. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nat Rev Genet, 2009. **10**(1): p. 57-63.
4. Kauffman, S.A., *Metabolic stability and epigenesis in randomly constructed genetic nets.* J Theor Biol, 1969. **22**(3): p. 437-67.
5. Hecker, M., et al., *Gene regulatory network inference: data integration in dynamic models-a review.* Biosystems, 2009. **96**(1): p. 86-103.
6. Abul, O., R. Alhajj, and F. Polat, *Asymptotical Lower Limits on Required Number of Examples for Learning Boolean Networks Computer and Information Sciences – ISCIS 2006*, A. Levi, et al., Editors. 2006, Springer Berlin / Heidelberg. p. 154-164.
7. Margolin, A., et al., *ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context.* BMC Bioinformatics, 2006. **7**(Suppl 1): p. S7.
8. Basso, K., et al., *Reverse engineering of regulatory networks in human B cells.* Nat Genet, 2005. **37**(4): p. 382-90.
9. Faith, J.J., et al., *Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles.* PLoS Biol, 2007. **5**(1): p. e8.
10. Stuart, J.M., et al., *A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules.* Science, 2003. **302**(5643): p. 249-255.
11. van Someren, E.P., L.F. Wessels, and M.J. Reinders, *Linear modeling of genetic networks from experimental data.* Proc Int Conf Intell Syst Mol Biol, 2000. **8**: p. 355-66.
12. Gardner, T.S., et al., *Inferring Genetic Networks and Identifying Compound Mode of Action via Expression Profiling.* Science, 2003. **301**(5629): p. 102-105.
13. di Bernardo, D., et al., *Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks.* Nat Biotechnol, 2005. **23**(3): p. 377-83.
14. Bansal, M., G. Della Gatta, and D. di Bernardo, *Inference of gene regulatory networks and compound mode of action from time course gene expression profiles.* Bioinformatics, 2006. **22**(7): p. 815-22.
15. Chen, T., H.L. He, and G.M. Church, *Modeling gene expression with differential equations.* Pac Symp Biocomput, 1999: p. 29-40.
16. Sakamoto, E. and H. Iba. *Inferring a system of differential equations for a gene regulatory network by using genetic programming*. in *Evolutionary Computation, 2001. Proceedings of the 2001 Congress on*. 2001.
17. Margolin, A.A., et al., *Reverse engineering cellular networks.* Nat. Protocols, 2006. **1**(2): p. 662-671.
18. Aach, J. and G.M. Church, *Aligning gene expression time series with time warping algorithms.* Bioinformatics, 2001. **17**(6): p. 495-508.
19. Spellman, P.T., et al., *Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization.* Mol. Biol. Cell, 1998. **9**(12): p. 3273-3297.
20. Cho, R.J., et al., *A genome-wide transcriptional analysis of the mitotic cell cycle.* Mol Cell, 1998. **2**(1): p. 65-73.
21. Ernst, J., et al., *Reconstructing dynamic regulatory maps*. Vol. 3. 2007.

22. Zaslavsky, E., et al., *Reconstruction of regulatory networks through temporal enrichment profiling and its application to H1N1 influenza viral infection.* BMC Bioinformatics, 2013. **14**(Suppl 6): p. S1.
23. Goodacre, R., et al., *Metabolomics by numbers: acquiring and understanding global metabolite data.* Trends Biotechnol, 2004. **22**(5): p. 245-52.
24. Bartel, D.P., *MicroRNAs: Genomics, Biogenesis, Mechanism, and Function.* Cell, 2004. **116**(2): p. 281-297.
25. He, L. and G.J. Hannon, *MicroRNAs: small RNAs with a big role in gene regulation.* Nat Rev Genet, 2004. **5**(7): p. 522-531.
26. Friedman, R.C., et al., *Most mammalian mRNAs are conserved targets of microRNAs.* Genome Research, 2009. **19**(1): p. 92-105.
27. Bartel, D.P., *MicroRNAs: Target Recognition and Regulatory Functions.* Cell, 2009. **136**(2): p. 215-233.
28. Baek, D., et al., *The impact of microRNAs on protein output.* Nature, 2008. **455**(7209): p. 64-71.
29. Croce, C.M., *Causes and consequences of microRNA dysregulation in cancer.* Nature Reviews Genetics, 2009. **10**(10): p. 704-714.
30. Lujambio, A. and S.W. Lowe, *The microcosmos of cancer.* Nature, 2012. **482**(7385): p. 347-355.
31. Ørom, U.A. and A.H. Lund, *Experimental identification of microRNA targets.* Gene, 2010. **451**(1–2): p. 1-5.
32. Lim, L.P., et al., *Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs.* Nature, 2005. **433**(7027): p. 769-773.
33. Xu, G., et al., *Transcriptome and targetome analysis in MIR155 expressing cells using RNA-seq.* RNA, 2010. **16**(8): p. 1610-1622.
34. Sekine, S., et al., *Disruption of Dicer1 induces dysregulated fetal gene expression and promotes hepatocarcinogenesis.* Gastroenterology, 2009. **136**(7): p. 2304-2315 e1-4.
35. Krutzfeldt, J., et al., *Silencing of microRNAs in vivo with 'antagomirs'.* Nature, 2005. **438**(7068): p. 685-9.
36. Elmén, J., et al., *Antagonism of microRNA-122 in mice by systemically administered LNA-antimiR leads to up-regulation of a large set of predicted target mRNAs in the liver.* Nucleic Acids Research, 2008. **36**(4): p. 1153-1162.
37. Ebert, M.S., J.R. Neilson, and P.A. Sharp, *MicroRNA sponges: competitive inhibitors of small RNAs in mammalian cells.* Nat Methods, 2007. **4**(9): p. 721-6.
38. Haraguchi, T., Y. Ozaki, and H. Iba, *Vectors expressing efficient RNA decoys achieve the long-term suppression of specific microRNA activity in mammalian cells.* Nucleic Acids Research, 2009. **37**(6): p. e43.
39. Vinther, J., et al., *Identification of miRNA targets with stable isotope labeling by amino acids in cell culture.* Nucleic Acids Research, 2006. **34**(16): p. e107.
40. Yang, Y., et al., *Identifying targets of miR-143 using a SILAC-based proteomic approach.* Molecular BioSystems, 2010. **6**(10): p. 1873-1882.
41. Selbach, M., et al., *Widespread changes in protein synthesis induced by microRNAs.* Nature, 2008. **455**(7209): p. 58-63.
42. Zhu, S., et al., *MicroRNA-21 Targets the Tumor Suppressor Gene Tropomyosin 1 (TPM1).* Journal of Biological Chemistry, 2007. **282**(19): p. 14328-14336.
43. Yekta, S., I.H. Shih, and D.P. Bartel, *MicroRNA-directed cleavage of HOXB8 mRNA.* Science, 2004. **304**(5670): p. 594-6.
44. Franco-Zorrilla, J.M., et al., *Genome-wide identification of small RNA targets based on target enrichment and microarray hybridizations.* The Plant Journal, 2009. **59**(5): p. 840-850.
45. Bracken, C.P., et al., *Global analysis of the mammalian RNA degradome reveals widespread miRNA-dependent and miRNA-independent endonucleolytic cleavage.* Nucleic Acids Research, 2011. **39**(13): p. 5658-5668.

46. Nakamoto, M., et al., *Physiological identification of human transcripts translationally regulated by a specific microRNA.* Human Molecular Genetics, 2005. **14**(24): p. 3813-3821.
47. Guo, H., et al., *Mammalian microRNAs predominantly act to decrease target mRNA levels.* Nature, 2010. **466**(7308): p. 835-840.
48. Ørom, U.A. and A.H. Lund, *Isolation of microRNA targets using biotinylated synthetic microRNAs.* Methods, 2007. **43**(2): p. 162-165.
49. Ørom, U.A., F.C. Nielsen, and A.H. Lund, *MicroRNA-10a Binds the 5'UTR of Ribosomal Protein mRNAs and Enhances Their Translation.* Molecular Cell, 2008. **30**(4): p. 460-471.
50. Christoffersen, N.R., et al., *p53-independent upregulation of miR-34a during oncogene-induced senescence represses MYC.* Cell Death Differ, 2010. **17**(2): p. 236-45.
51. Kedde, M., et al., *RNA-Binding Protein Dnd1 Inhibits MicroRNA Access to Target mRNA.* Cell, 2007. **131**(7): p. 1273-1286.
52. Hsu, R.-J., H.-J. Yang, and H.-J. Tsai, *Labeled microRNA pull-down assay system: an experimental approach for high-throughput identification of microRNA-target mRNAs.* Nucleic Acids Research, 2009. **37**(10): p. e77.
53. Vatolin, S., K. Navaratne, and R.J. Weil, *A Novel Method to Detect Functional MicroRNA Targets.* Journal of Molecular Biology, 2006. **358**(4): p. 983-996.
54. Andachi, Y., *A novel biochemical method to identify target genes of individual microRNAs: Identification of a new Caenorhabditis elegans let-7 target.* RNA, 2008. **14**(11): p. 2440-2451.
55. Karginov, F.V., et al., *A biochemical approach to identifying microRNA targets.* Proceedings of the National Academy of Sciences, 2007. **104**(49): p. 19291-19296.
56. Hendrickson, D.G., et al., *Systematic Identification of mRNAs Recruited to Argonaute 2 by Specific microRNAs and Corresponding Changes in Transcript Abundance.* PLoS ONE, 2008. **3**(5): p. e2126.
57. Easow, G., A.A. Teleman, and S.M. Cohen, *Isolation of microRNA targets by miRNP immunopurification.* RNA, 2007. **13**(8): p. 1198-1204.
58. Beitzinger, M., et al., *Identification of human microRNA targets from isolated argonaute protein complexes.* RNA Biol, 2007. **4**(2): p. 76-84.
59. Tan, L.P., et al., *A high throughput experimental approach to identify miRNA targets in human cells.* Nucleic Acids Research, 2009. **37**(20): p. e137.
60. MILI, S. and J.A. STEITZ, *Evidence for reassociation of RNA-binding proteins after cell lysis: Implications for the interpretation of immunoprecipitation analyses.* RNA, 2004. **10**(11): p. 1692-1694.
61. Chi, S.W., et al., *Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps.* Nature, 2009. **460**(7254): p. 479-86.
62. Zisoulis, D.G., et al., *Comprehensive discovery of endogenous Argonaute binding sites in Caenorhabditis elegans.* Nat Struct Mol Biol, 2010. **17**(2): p. 173-9.
63. Hafner, M., et al., *PAR-CliP--a method to identify transcriptome-wide the binding sites of RNA binding proteins.* J Vis Exp, 2010(41).
64. Hafner, M., et al., *Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP.* Cell, 2010. **141**(1): p. 129-41.
65. Zhang, C. and R.B. Darnell, *Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data.* Nat Biotechnol, 2011. **29**(7): p. 607-14.
66. Kuhn, D.E., et al., *Experimental validation of miRNA targets.* Methods, 2008. **44**(1): p. 47-54.
67. Vergoulis, T., et al., *TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support.* Nucleic Acids Research, 2011.
68. Hsu, S.D., et al., *miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions.* Nucleic Acids Res, 2014. **42**(Database issue): p. D78-85.
69. Xiao, F., et al., *miRecords: an integrated resource for microRNA-target interactions.* Nucleic Acids Res, 2009. **37**(Database issue): p. D105-10.

70. Hsu, S.D., et al., *miRTarBase: a database curates experimentally validated microRNA-target interactions.* Nucleic Acids Res, 2011. **39**(Database issue): p. D163-9.
71. Kozomara, A. and S. Griffiths-Jones, *miRBase: annotating high confidence microRNAs using deep sequencing data.* Nucleic Acids Research, 2014. **42**(D1): p. D68-D73.
72. Lewis, B.P., et al., *Prediction of Mammalian MicroRNA Targets.* Cell, 2003. **115**(7): p. 787-798.
73. REHMSMEIER, M., et al., *Fast and effective prediction of microRNA/target duplexes.* RNA, 2004. **10**(10): p. 1507-1517.
74. Friedman, R.C., et al., *Most mammalian mRNAs are conserved targets of microRNAs.* Genome Res, 2009. **19**(1): p. 92-105.
75. Kertesz, M., et al., *The role of site accessibility in microRNA target recognition.* Nat Genet, 2007. **39**(10): p. 1278-84.
76. Stark, A., et al., *Animal MicroRNAs Confer Robustness to Gene Expression and Have a Significant Impact on 3'UTR Evolution.* Cell, 2005. **123**(6): p. 1133-1146.
77. John, B., et al., *Human MicroRNA targets.* PLoS Biol, 2004. **2**(11): p. e363.
78. Krek, A., et al., *Combinatorial microRNA target predictions.* Nat Genet, 2005. **37**(5): p. 495-500.
79. Lewis, B.P., C.B. Burge, and D.P. Bartel, *Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets.* Cell, 2005. **120**(1): p. 15-20.
80. Maragkakis, M., et al., *DIANA-microT web server: elucidating microRNA functions through target prediction.* Nucleic Acids Res, 2009. **37**(Web Server issue): p. W273-6.
81. Saito, T. and P. Sætrom, *MicroRNAs – targeting and target prediction.* New Biotechnology, 2010. **27**(3): p. 243-249.
82. Watanabe, Y., M. Tomita, and A. Kanai, *Computational Methods for MicroRNA Target Prediction*, in *Methods in Enzymology*, J.R. John and J.H. Gregory, Editors. 2007, Academic Press. p. 65-86.
83. Pritchard, C.C., H.H. Cheng, and M. Tewari, *MicroRNA profiling: approaches and considerations.* Nature Reviews Genetics, 2012. **13**(5): p. 358-369.
84. Chen, Y., et al., *Reproducibility of quantitative RT-PCR array in miRNA expression profiling and comparison with microarray analysis.* BMC Genomics, 2009. **10**(1): p. 407.
85. Ach, R., H. Wang, and B. Curry, *Measuring microRNAs: Comparisons of microarray and quantitative PCR measurements, and of different total RNA prep methods.* BMC Biotechnology, 2008. **8**(1): p. 69.
86. Git, A., et al., *Systematic comparison of microarray profiling, real-time PCR, and next-generation sequencing technologies for measuring differential microRNA expression.* RNA, 2010. **16**(5): p. 991-1006.
87. Nishida, N., et al., *Microarray analysis of colorectal cancer stromal tissue reveals upregulation of two oncogenic miRNA clusters.* Clin Cancer Res, 2012. **18**(11): p. 3054-70.
88. Donahue, T.R., et al., *Integrative survival-based molecular profiling of human pancreatic cancer.* Clin Cancer Res, 2012. **18**(5): p. 1352-63.
89. Hecker, N., et al., *A New Algorithm for Integrated Analysis of miRNA-mRNA Interactions Based on Individual Classification Reveals Insights into Bladder Cancer.* PLoS ONE, 2013. **8**(5): p. e64543.
90. Buffa, F.M., et al., *microRNA-associated progression pathways and potential therapeutic targets identified by integrated mRNA and microRNA expression profiling in breast cancer.* Cancer Res, 2011. **71**(17): p. 5635-45.
91. Enerly, E., et al., *miRNA-mRNA Integrated Analysis Reveals Roles for miRNAs in Primary Breast Tumors.* PLoS ONE, 2011. **6**(2): p. e16915.
92. Enerly, E., et al., *miRNA-mRNA integrated analysis reveals roles for miRNAs in primary breast tumors.* PLoS One, 2011. **6**(2): p. e16915.

93. Wang, H.W., et al., *Pediatric primary central nervous system germ cell tumors of different prognosis groups show characteristic miRNome traits and chromosome copy number variations.* BMC Genomics, 2010. **11**: p. 132.

94. Fu, J., et al., *Identifying microRNA-mRNA regulatory network in colorectal cancer by a combination of expression profile and bioinformatics analysis.* BMC Syst Biol, 2012. **6**: p. 68.

95. Wani, K., et al., *A prognostic gene expression signature in infratentorial ependymoma.* Acta Neuropathol, 2012. **123**(5): p. 727-38.

96. Johnson, R.A., et al., *Cross-species genomics matches driver mutations and cell compartments to model ependymoma.* Nature, 2010. **466**(7306): p. 632-6.

97. Long, C., et al., *Integrated miRNA-mRNA Analysis Revealing the Potential Roles of miRNAs in Chordomas.* PLoS ONE, 2013. **8**(6): p. e66676.

98. Ernst, A., et al., *De-repression of CTGF via the miR-17-92 cluster upon differentiation of human glioblastoma spheroid cultures.* Oncogene, 2010. **29**(23): p. 3411-22.

99. Zhang, W., et al., *miR-181d: a predictive glioblastoma biomarker that downregulates MGMT expression.* Neuro Oncol, 2012. **14**(6): p. 712-9.

100. Montes-Moreno, S., et al., *miRNA expression in diffuse large B-cell lymphoma treated with chemoimmunotherapy.* Blood, 2011. **118**(4): p. 1034-40.

101. Lionetti, M., et al., *Identification of microRNA expression patterns and definition of a microRNA/mRNA regulatory network in distinct molecular groups of multiple myeloma.* Blood, 2009. **114**(25): p. e20-6.

102. Zhou, Y., et al., *High-risk myeloma is associated with global elevation of miRNAs and overexpression of EIF2C2/AGO2.* Proc Natl Acad Sci U S A, 2010. **107**(17): p. 7904-9.

103. Namlos, H.M., et al., *Modulation of the osteosarcoma expression phenotype by microRNAs.* PLoS One, 2012. **7**(10): p. e48086.

104. Wang, L., et al., *Gene networks and microRNAs implicated in aggressive prostate cancer.* Cancer Res, 2009. **69**(24): p. 9490-7.

105. Taylor, B.S., et al., *Integrative genomic profiling of human prostate cancer.* Cancer Cell, 2010. **18**(1): p. 11-22.

106. Rajasekhar, V.K., et al., *Tumour-initiating stem-like cells in human prostate cancer exhibit increased NF-kappaB signalling.* Nat Commun, 2011. **2**: p. 162.

107. Sethupathy, P., M. Megraw, and A.G. Hatzigeorgiou, *A guide through present computational approaches for the identification of mammalian microRNA targets.* Nat Methods, 2006. **3**(11): p. 881-6.

108. Khatri, P. and S. Drăghici, *Ontological analysis of gene expression data: current tools, limitations, and open problems.* Bioinformatics, 2005. **21**(18): p. 3587-3595.

109. Huang, D.W., B.T. Sherman, and R.A. Lempicki, *Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.* Nucleic Acids Research, 2009. **37**(1): p. 1-13.

110. Huang, G.T., C. Athanassiou, and P.V. Benos, *mirConnX: condition-specific mRNA-microRNA network integrator.* Nucleic Acids Research, 2011. **39**(suppl 2): p. W416-W423.

111. Wang, X., *miRDB: A microRNA target prediction and functional annotation database with a wiki interface.* RNA, 2008. **14**(6): p. 1012-1017.

112. Ulitsky, I., L.C. Laurent, and R. Shamir, *Towards computational prediction of microRNA function and activity.* Nucleic Acids Research, 2010. **38**(15): p. e160.

113. Liang, S., S. Fuhrman, and R. Somogyi. *REVEAL, a general reverse engineering algorithm for inference of genetic network architectures*. in *Pacific symposium on biocomputing*. 1998.

114. Liang, S., S. Fuhrman, and R. Somogyi, *Reveal, a general reverse engineering algorithm for inference of genetic network architectures.* Pac Symp Biocomput, 1998: p. 18-29.

115. Akutsu, T., S. Miyano, and S. Kuhara, *Identification of genetic networks from a small number of gene expression patterns under the Boolean network model.* Pac Symp Biocomput, 1999: p. 17-28.

116. Akutsu, T., S. Miyano, and S. Kuhara, *Algorithms for identifying Boolean networks and related biological networks based on matrix multiplication and fingerprint function.* J Comput Biol, 2000. **7**(3-4): p. 331-43.

117. Shmulevich, I., et al., *Inference of genetic regulatory networks via best-fit extensions*. 2002: Springer.

118. Lähdesmäki, H., I. Shmulevich, and O. Yli-Harja, *On learning gene regulatory networks under the Boolean network model.* Machine Learning, 2003. **52**(1-2): p. 147-167.

119. Davidich, M.I. and S. Bornholdt, *Boolean network model predicts cell cycle sequence of fission yeast.* PLoS One, 2008. **3**(2): p. e1672.

120. Martin, S., et al., *Boolean dynamics of genetic regulatory networks inferred from microarray time series data.* Bioinformatics, 2007. **23**(7): p. 866-874.

121. Szallasi, Z. and S. Liang. *Modeling the normal and neoplastic cell cycle with 'realistic Boolean genetic networks': Their application for understanding carcinogenesis and assessing therapeutic strategies*. in *Pacific Symposium on Biocomputing*. 1998.

122. Deng, X., H. Geng, and M.T. Matache, *Dynamics of asynchronous random Boolean networks with asynchrony generated by stochastic processes.* Bio Systems, 2007. **88**(1-2): p. 16.

123. Greil, F. and B. Drossel, *Dynamics of critical Kauffman networks under asynchronous stochastic update.* Physical review letters, 2005. **95**(4): p. 048701.

124. Shmulevich, I., et al., *Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks.* Bioinformatics, 2002. **18**(2): p. 261-274.

125. Friedman, N., et al., *Using Bayesian networks to analyze expression data.* J Comput Biol, 2000. **7**(3-4): p. 601-20.

126. Bansal, M., et al., *How to infer gene networks from expression profiles.* Molecular systems biology, 2007. **3**(1).

127. Imoto, S., et al. *Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network*. in *Bioinformatics Conference, 2002. Proceedings. IEEE Computer Society*. 2002.

128. Ong, I.M., J.D. Glasner, and D. Page, *Modelling regulatory pathways in E. coli from time series expression profiles.* Bioinformatics, 2002. **18 Suppl 1**: p. S241-8.

129. Kim, S.Y., S. Imoto, and S. Miyano, *Inferring gene networks from time series microarray data using dynamic Bayesian networks.* Briefings in Bioinformatics, 2003. **4**(3): p. 228-235.

130. Zou, M. and S.D. Conzen, *A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data.* Bioinformatics, 2005. **21**(1): p. 71-79.

131. Hartemink, A.J., et al., *Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks.* Pac Symp Biocomput, 2001: p. 422-33.

132. Segal, E., et al., *Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data.* Nat Genet, 2003. **34**(2): p. 166-76.

133. Nachman, I., A. Regev, and N. Friedman, *Inferring quantitative models of regulatory networks from expression data.* Bioinformatics, 2004. **20**(suppl 1): p. i248-i256.

134. Rangel, C., et al., *Modeling T-cell activation using gene expression profiling and state-space models.* Bioinformatics, 2004. **20**(9): p. 1361-1372.

135. Song, L., M. Kolar, and E.P. Xing. *Time-varying dynamic bayesian networks*. in *Advances in Neural Information Processing Systems*. 2009.

136. Lebre, S., et al., *Statistical inference of the time-varying structure of gene-regulation networks.* BMC Systems Biology, 2010. **4**(1): p. 130.

137. Eisen, M.B., et al., *Cluster analysis and display of genome-wide expression patterns.* Proceedings of the National Academy of Sciences, 1998. **95**(25): p. 14863-14868.

138. Shannon, C.E. and W. Weaver, *A mathematical theory of communication*, 1948, American Telephone and Telegraph Company.

139. Steuer, R., et al., *The mutual information: detecting and evaluating dependencies between variables.* Bioinformatics, 2002. **18**(suppl 2): p. S231-S240.

140. Butte, A.J. and I.S. Kohane. *Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements*. in *Pac Symp Biocomput*. 2000.

141. Zoppoli, P., S. Morganella, and M. Ceccarelli, *TimeDelay-ARACNE: Reverse engineering of gene networks from time-course data by an information theoretic approach.* BMC Bioinformatics, 2010. **11**(1): p. 154.

142. Klipp, E., et al., *Systems biology in practice: concepts, implementation and application*. 2008: Wiley-Blackwell.

143. Savageau, M.A. and E.O. Voit, *Recasting nonlinear differential equations as S-systems: a canonical nonlinear form.* Mathematical biosciences, 1987. **87**(1): p. 83-115.

144. Kimura, S., et al., *Inference of S-system models of genetic networks using a cooperative coevolutionary algorithm.* Bioinformatics, 2005. **21**(7): p. 1154-1163.

145. de Jong, H., *Modeling and simulation of genetic regulatory systems: a literature review.* J Comput Biol, 2002. **9**(1): p. 67-103.

146. Yeung, M.S., J. Tegnér, and J.J. Collins, *Reverse engineering gene networks using singular value decomposition and robust regression.* Proceedings of the National Academy of Sciences, 2002. **99**(9): p. 6163-6168.

147. Nam, S., et al., *miRGator: an integrated system for functional annotation of microRNAs.* Nucleic Acids Research, 2008. **36**(suppl 1): p. D159-D164.

148. Cho, S., et al., *MiRGator v3.0: a microRNA portal for deep sequencing, expression profiling and mRNA targeting.* Nucleic Acids Res, 2013. **41**(Database issue): p. D252-7.

149. Nam, S., et al., *MicroRNA and mRNA integrated analysis (MMIA): a web tool for examining biological functions of microRNA expression.* Nucleic Acids Research, 2009. **37**(suppl 2): p. W356-W362.

150. Peng, X., et al., *Computational identification of hepatitis C virus associated microRNA-mRNA regulatory modules in human livers.* BMC Genomics, 2009. **10**(1): p. 373.

151. Ritchie, W., S. Flamant, and J.E.J. Rasko, *mimiRNA: a microRNA expression profiler and classification resource designed to identify functional correlations between microRNAs and their targets.* Bioinformatics, 2010. **26**(2): p. 223-227.

152. Sales, G., et al., *MAGIA, a web-based tool for miRNA and Genes Integrated Analysis.* Nucleic Acids Research, 2010. **38**(suppl 2): p. W352-W359.

153. Wang, H. and W.H. Li, *Increasing MicroRNA target prediction confidence by the relative R(2) method.* Journal of Theoretical Biology, 2009. **259**(4): p. 793-8.

154. Friedman, J., T. Hastie, and R. Tibshirani, *Regularization Paths for Generalized Linear Models via Coordinate Descent.* J Stat Softw, 2010. **33**(1): p. 1-22.

155. Kim, S., M. Choi, and K.H. Cho, *Identifying the target mRNAs of microRNAs in colorectal cancer.* Computational Biology and Chemistry, 2009. **33**(1): p. 94-9.

156. Lu, Y., et al., *A Lasso regression model for the construction of microRNA-target regulatory networks.* Bioinformatics, 2011. **27**(17): p. 2406-2413.

157. Muniategui, A., et al., *Quantification of miRNA-mRNA Interactions.* PLoS ONE, 2012. **7**(2): p. e30766.

158. Beck, D., et al., *Integrative analysis of next generation sequencing for small non-coding RNAs and transcriptional regulation in Myelodysplastic Syndromes.* BMC Medical Genomics, 2011. **4**(1): p. 19.

159. Huang, J.C., Q.D. Morris, and B.J. Frey, *Bayesian inference of MicroRNA targets from sequence and expression data.* Journal of Computational Biology, 2007. **14**(5): p. 550-63.

160. Huang, J.C., B.J. Frey, and Q.D. Morris, *Comparing sequence and expression for predicting microRNA targets using GenMiR3.* Pacific Symposium on Biocomputing, 2008: p. 52-63.

161. Su, N., et al. *Predicting MicroRNA targets by integrating sequence and expression data in cancer*. in *Systems Biology (ISB), 2011 IEEE International Conference on*. 2011.

162. Stingo, F.C., et al., *A Bayesian graphical modeling approach to microRNA regulatory network inference.* Annals of Applied Statistics, 2010. **4**(4): p. 2024-2048.

163. Liu, B., et al., *Exploring complex miRNA-mRNA interactions with Bayesian networks by splitting-averaging strategy.* BMC Bioinformatics, 2009. **10**(1): p. 408.

164. Gennarino, V.A., et al., *MicroRNA target prediction by expression analysis of host genes.* Genome Research, 2008.

165. Bandyopadhyay, S. and R. Mitra, *TargetMiner: microRNA target prediction with systematic identification of tissue-specific negative examples.* Bioinformatics, 2009. **25**(20): p. 2625-2631.

166. Li, J., et al., *A probabilistic framework to improve microrna target prediction by incorporating proteomics data.* J Bioinform Comput Biol, 2009. **7**(6): p. 955-72.

167. Pihur, V., S. Datta, and S. Datta, *Reconstruction of genetic association networks from microarray data: a partial least squares approach.* Bioinformatics, 2008. **24**(4): p. 561-568.

168. Li, X., et al., *Modeling microRNA-mRNA Interactions Using PLS Regression in Human Colon Cancer.* BMC Medical Genomics, 2011. **4**(1): p. 44.

169. Hadi, S.C.A.S., *Regression Analysis by Example*. 4th ed. WILEY SERIES IN PROBABILITY AND STATISTICS, ed. N.A.C.C. David J. Balding, Nicholas I. Fisher,, et al. 2006: A JOHN WILEY & SONS, INC.

170. Draper, N. and H. Smith, *Applied Regression Analysis (Wiley Series in Probability and Statistics)*. 1998: Wiley-Interscience.

171. Golub, T.R. and D.K. Slonim, *Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression.* Science, 1999. **286**(5439): p. 531.

172. van de Vijver, M.J., et al., *A Gene-Expression Signature as a Predictor of Survival in Breast Cancer.* New England Journal of Medicine, 2002. **347**(25): p. 1999-2009.

173. Fan, X., et al., *DNA Microarrays Are Predictive of Cancer Prognosis: A Re-evaluation.* Clinical Cancer Research, 2010. **16**(2): p. 629-636.

174. Wong, D.J. and H.Y. Chang, *Learning More from Microarrays: Insights from Modules and Networks.* J Investig Dermatol, 2005. **125**(2): p. 175-182.

175. Zhu, G., et al., *Two yeast forkhead genes regulate the cell cycle and pseudohyphal growth.* Nature, 2000. **406**(6791): p. 90-94.

176. Quackenbush, J., *Microarray data normalization and transformation.* Nat Genet, 2002. **32 Suppl**: p. 496-501.

177. Mutch, D.M., et al., *The limit fold change model: a practical approach for selecting differentially expressed genes from microarray data.* BMC Bioinformatics, 2002. **3**: p. 17.

178. Tusher, V.G., R. Tibshirani, and G. Chu, *Significance analysis of microarrays applied to the ionizing radiation response.* Proc Natl Acad Sci U S A, 2001. **98**(9): p. 5116-21.

179. Jeffery, I.B., D.G. Higgins, and A.C. Culhane, *Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data.* BMC Bioinformatics, 2006. **7**: p. 359.

180. Maraziotis, I., A. Dragomir, and A. Bezerianos, *Gene networks inference from expression data using a recurrent neuro-fuzzy approach.* Conf Proc IEEE Eng Med Biol Soc, 2005. **5**: p. 4834-7.

181. Almansoori, W., et al., *Link prediction and classification in social networks and its application in healthcare and systems biology.* Network Modeling and Analysis in Health Informatics and Bioinformatics: p. 1-10.

182. Kanehisa, M. and S. Goto, *KEGG: Kyoto Encyclopedia of Genes and Genomes.* Nucleic Acids Research, 2000. **28**(1): p. 27-30.

183. Storey, J.D. and R. Tibshirani, *Statistical significance for genomewide studies.* Proceedings of the National Academy of Sciences, 2003. **100**(16): p. 9440-9445.

184. Thieffry, D., et al., *From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in Escherichia coli.* BioEssays, 1998. **20**(5): p. 433-440.

185. Wall, M., A. Rechtsteiner, and L. Rocha, *Singular Value Decomposition and Principal Component Analysis: A Practical Approach to Microarray Data Analysis*, D.P. Berrar, W. Dubitzky, and M. Granzow, Editors. 2003, Springer US. p. 91-109.

186. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society. Series B (Methodological), 1995. **57**(1): p. 289-300.

187. Kim, S., S. Imoto, and S. Miyano, *Dynamic Bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data.* Biosystems, 2004. **75**(1-3): p. 57-65.

188. Bar-Joseph, Z., et al., *Continuous representations of time-series gene expression data.* J Comput Biol, 2003. **10**(3-4): p. 341-56.

189. Bar-Joseph, Z., et al., *Comparing the continuous representation of time-series expression profiles to identify differentially expressed genes.* Proceedings of the National Academy of Sciences, 2003. **100**(18): p. 10146-10151.

190. Berg, B.A., *Markov Chain Monte Carlo Simulations and Their Statistical Analysis (With Web-Based Fortran Code)*. 2004, Hackensack, NJ: World Scientific.

191. Hoerl, A.E. and R.W. Kennard, *Ridge Regression: Biased Estimation for Nonorthogonal Problems.* Technometrics, 1970. **12**(1): p. 55-67.

192. Marquardt, D.W. and R.D. Snee, *Ridge Regression in Practice.* The American Statistician, 1975. **29**(1): p. 3-20.

193. Lindgren, F., P. Geladi, and S. Wold, *The kernel algorithm for PLS.* Journal of Chemometrics, 1993. **7**(1): p. 45-59.

194. Yao Fu, L.R.J., Julie A. Dickerson. *Gene Regulatory Network Reconstruction Based on Gene Expression and Transcription Factor Activities*. in *BIOCOMP*. 2010.

195. Hsu, S.D., et al., *miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions.* Nucleic Acids Research, 2014. **42**(Database issue): p. D78-85.

196. Orlova, I., et al., *MicroRNA modulation in complex regional pain syndrome.* Journal of Translational Medicine, 2011. **9**(1): p. 195.

197. Network, C.G.A.R., *Integrated genomic analyses of ovarian carcinoma.* Nature, 2011. **474**(7353): p. 609-615.

198. Jacobsen, A., et al., *Analysis of microRNA-target interactions across diverse cancer types.* Nature structural & molecular biology, 2013.

199. Garcia, D.M., et al., *Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs.* Nat Struct Mol Biol, 2011. **18**(10): p. 1139-46.

200. Ma, X., et al., *MicroRNAs in NF-κB signaling.* Journal of Molecular Cell Biology, 2011. **3**(3): p. 159-166.

201. Warde-Farley, D., et al., *The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function.* Nucleic Acids Research, 2010. **38**(suppl 2): p. W214-W220.

202. Benjamini, Y. and D. Yekutieli, *The control of the false discovery rate in multiple testing under dependency.* Annals of statistics, 2001: p. 1165-1188.

203. Vasudevan, S., Y. Tong, and J.A. Steitz, *Switching from Repression to Activation: MicroRNAs Can Up-Regulate Translation.* Science, 2007. **318**(5858): p. 1931-1934.

204. Hobert, O., *Gene Regulation by Transcription Factors and MicroRNAs.* Science, 2008. **319**(5871): p. 1785-1786.

205. Fabian, M.R., N. Sonenberg, and W. Filipowicz, *Regulation of mRNA Translation and Stability by microRNAs.* Annual Review of Biochemistry, 2010. **79**(1): p. 351-379.

206. Huntzinger, E. and E. Izaurralde, *Gene silencing by microRNAs: contributions of translational repression and mRNA decay.* Nature Reviews Genetics, 2011. **12**(2): p. 99-110.

207. John, B., et al., *Human MicroRNA targets.* PLoS Biology, 2004. **2**(11): p. e363.

208. Krek, A., et al., *Combinatorial microRNA target predictions.* Nature Genetics, 2005. **37**(5): p. 495-500.

209. Maragkakis, M., et al., *DIANA-microT web server: elucidating microRNA functions through target prediction.* Nucleic Acids Research, 2009. **37**(Web Server issue): p. W273-6.

210. Zhou, Y., R. Qureshi, and A. Sacan, *Data simulation and regulatory network reconstruction from time-series microarray data using stepwise multiple linear regression.* Network Modeling Analysis in Health Informatics and Bioinformatics, 2012. **1**(1-2): p. 3-17.

211. Sumazin, P., et al., *An Extensive MicroRNA-Mediated Network of RNA-RNA Interactions Regulates Established Oncogenic Pathways in Glioblastoma.* Cell, 2011. **147**(2): p. 370-381.

212. Liu, B., J. Li, and M.J. Cairns, *Identifying miRNAs, targets and functions.* Briefings in Bioinformatics, 2014. **15**(1): p. 1-19.

213. Huang da, W., B.T. Sherman, and R.A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.* Nat Protoc, 2009. **4**(1): p. 44-57.

214. Naume, B., et al., *Detection of isolated tumor cells in bone marrow in early-stage breast carcinoma patients: comparison with preoperative clinical parameters and primary tumor characteristics.* Clin Cancer Res, 2001. **7**(12): p. 4122-9.