From miRNA regulation to miRNA–TF co-regulation: computational approaches and challenges

Thuc Duy Le, Lin Liu, Junpeng Zhang, Bing Liu and Jiuyong Li

Submitted: 14th March 2014; Received (in revised form): 10th June 2014

Abstract
microRNAs (miRNAs) are important gene regulators. They control a wide range of biological processes and are involved in several types of cancers. Thus, exploring miRNA functions is important for diagnostics and therapeutics. To date, there are few feasible experimental techniques for discovering miRNA regulatory mechanisms. Alternatively, predictions of miRNA–mRNA regulatory relationships by computational methods have increasingly achieved promising results. Computational approaches are proving their ability as effective tools in reducing the number of biological experiments that must be conducted and to assist with the design of the experiments. In this review, we categorize and review different computational approaches to identify miRNA activities and functions, including the co-regulation of miRNAs and transcription factors. Our main focuses are on the recent approaches that use multiple data types for exploring miRNA functions. We discuss the remaining challenges in the evaluation and selection of models based on the results from a case study. Finally, we analyse the remaining challenges of each computational approach and suggest some future research directions.

Keywords: miRNA; transcription factor; causality discovery; data integration; miRNA target; co-regulation
INTRODUCTION
The human genome contains >20,000 genes. Most of these genes are expressed differentially to create proteins, the main actors in a living cell. The gene expression program is an extremely complex and well-organized procedure. While the expression program is made up of individual genes expressing on their own mechanisms, together they create a unified big picture to ensure that cells can function properly. Any interference to the expression program of genes would result in diseases. Therefore, understanding the mechanisms of regulating gene expression is crucial for preventing and curing diseases.

At the transcription level, transcription factors (TFs) are known to be the main gene regulators. TFs are proteins that control gene expression by directly binding to the DNA sequences of the target genes. The binding will result in an increase or decrease in the expression levels of the target genes. There has been extensive research on the regulation of TFs, and many experimental and computational methods have been proposed to elucidate the regulatory mechanisms of this type of regulators (see [1] for a review).

Recently, microRNAs (miRNAs) have been discovered to be the main regulators at post-transcriptional level. miRNAs are small nucleotide sequences of 21–25 bases [2], which are transcribed from the non-coding parts of the DNA. They recognize target genes by base pairing to complementary sequences in the 3'-untranslated region (3'UTR), 5'UTR and sometimes in the open reading frames of the target mRNA.

More and more evidence has emerged on miRNA involvements in biological processes and diseases. By regulating their target genes, miRNAs control a wide range of biological processes such as proliferation [3, 4], metabolism [5], differentiation [6], development [7], apoptosis [8], cellular signalling [9] and even cancer development and progression [2, 10]. Therefore, it is not surprising that miRNAs have been identified to be involved in several types of cancers including breast cancer [11], prostate cancer [12], lung cancer [13], colon cancer [14], ovarian cancer [15] and many other diseases [10, 16–19].

One of the challenges in miRNA research is that miRNAs have their own characteristics, making it difficult or impossible to apply the experimental and computational methods used for other gene regulators such as TFs, to miRNA research. Like TFs, one miRNA can regulate multiple target genes simultaneously, and multiple miRNAs may regulate their target genes cooperatively. A single miRNA can control up to hundreds of target genes [20], and most target genes in the genome can be regulated by multiple miRNAs [21]. Unlike TFs, however, miRNAs are very short in length (21–24 nucleotides), and the gene regulatory region of a miRNA is small in size, <1 kb compared with dozens of kbs in the case of a TF [22]. The small target regulatory region poses challenges for computational methods that search for the base pairing between miRNAs and the binding sites, as the complementary region is too short to be statistically significant. Therefore, novel and appropriate methods especially designed for miRNAs are required.

The significant roles of miRNAs have given rise to a fast growing body of research over the past decade (see Figure 1 for an illustration). In the early years of miRNA research, great effort was made to discover new miRNAs as well as to explore miRNA functions with wet experiments (see [23] for a review). These studies have significantly improved our understanding of miRNA functions. On the other hand, they also pose new challenges to scientists by producing a tremendous amount of data that needs to be analysed. Given the increasing number of novel miRNAs discovered (coupled with the large number of their regulated targets), it is infeasible to validate all the possible regulatory interactions by means of biological experiments. In response to the calls for discovering new knowledge from the vast amount of data available, computational methods have been developed and many of them have proved to be effective tools in assisting with the design of wet experiments, short-listing statistically significant regulatory interactions, thus making it feasible to conduct validation experiments [23, 24]. Although computational methods may never replace wet-lab experiments, they will assist with the design of the experiments.

A variety of computational methods have been proposed to explore miRNA functions, ranging from sequence based methods to methods incorporating expression data or combining multiple sources of data; correlation based methods to causality discovery based methods; methods for discovering single interactions to those finding modules of interacting molecules; analyses of interactions in a specific condition to differential analyses involving multiple conditions; and methods of studying miRNA...
regulation alone to those considering cooperative effects of miRNAs and TFs. These methods provide complementary views and approaches to exploring miRNA functions, and have their own advantages and limitations.

In this article, we review the new work for exploring (human) miRNA functions complementary to the previous reviews and discuss challenging issues in model evaluation. We categorize the methods into three main approaches: inferring miRNA–mRNA regulatory relationships, identifying miRNA–mRNA regulatory modules and discovering miRNA–TF co-regulatory relationships. Table 1 shows the overall picture of the computational approaches, including the topics that have been reviewed elsewhere [23, 24], and the topics that will be covered in this review. We discuss the remaining challenges of the evaluation and selection of models with a case study on three real world cancer data sets. Finally, we discuss some future research directions.

**INFERRING miRNA-mRNA REGULATORY RELATIONSHIPS**

Identifying miRNA targets is the first and foremost task to understand miRNA functions and regulatory mechanisms. Therefore, there have been several methods developed to predict putative target mRNAs at the sequence level [20, 25, 26]. The methods have contributed to the understanding of the regulatory functions of miRNAs, and have helped to narrow down the otherwise immeasurable amount of experimental work to be conducted. However, these approaches have a high rate of false-positive predictions, and do not take biological context into account. These approaches have been reviewed in [23, 24], and we do not cover them in this review.

In recent years, with the advances in experimental technology, gene expression data (GE) have emerged as important and promising resources for exploring miRNA functions. Various computational methods

---

**Table 1:** Summary of computational approaches for inferring miRNA functions

<table>
<thead>
<tr>
<th>Sequence data</th>
<th>Expression data</th>
<th>Multiple data sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>miRNA–mRNA regulatory relationships</td>
<td>See [23, 24] for reviews</td>
<td>Classical approaches: see [23] for a review</td>
</tr>
<tr>
<td>miRNA–mRNA regulatory modules</td>
<td>See [24] for a review</td>
<td>Emerging approaches</td>
</tr>
<tr>
<td>miRNA-TF co-regulation</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

The grey areas are the topics have been reviewed elsewhere. The areas with ticks are the topics will be covered in this review.
have been devised to incorporate gene expression profiles into the study of miRNA–mRNA regulatory relationships. Figure 2 shows a common framework used by these existing methods. In this framework, matched samples of miRNA and mRNA expression data are integrated into a data set. To reduce the computational complexity, researchers apply some constraints such as differentially expressed gene analysis to reduce the number of variables (genes) in the data set, or use sequence-based target predictions to reduce the number of miRNA–mRNA interactions considered by the method. The input expression data set is then analysed by statistical/machine learning models to predict miRNA targets. Recent advances focus on integrating multiple expression data sets, or using heterogeneous data such as overexpression data and sequence data in building a model. There are also some newly emerged methods for inferring miRNA targets, such as the methods based on causality discovery approach. In this section, we review these new works (summarized in Table 2) for inferring miRNA–mRNA regulatory relationships.

**Classical approaches with GE and beyond**

Classical approaches such as correlation, regression analysis and Bayesian parameter learning have achieved significant results in inferring miRNA–mRNA regulatory relationships. The principle of these methods is that if a gene is regulated by a miRNA, a correlation should show between the expression levels of the gene and the miRNA. Some of these methods also take into consideration the available miRNA target information previously predicted using sequence data, which has proved
Table 2: Summary of methods to infer miRNA–mRNA regulatory relationships

<table>
<thead>
<tr>
<th>Method</th>
<th>Brief description</th>
<th>Data sources</th>
<th>Notes</th>
<th>Software tool</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chen et al. [27]</td>
<td>Combine different gene expression data sets into a model to infer miRNA–mRNA interactions.</td>
<td>Multiple matched miRNA and mRNA expression data sets</td>
<td>Suitable for analysing multiple expression data sets of the same condition, e.g. cancer.</td>
<td><a href="http://bioinformatics.med.yale.edu/group/">http://bioinformatics.med.yale.edu/group/</a></td>
</tr>
<tr>
<td>Jacobsen et al. [28]</td>
<td>Use regression analysis to analyse miRNA–target interactions across diverse cancer types.</td>
<td>DNA copy number, promoter methylation, miRNA and mRNA expression data</td>
<td>A good resource for biologists to explore the miRNA–mRNA interactions that occur in multiple cancer types.</td>
<td><a href="http://cancerminer.org">http://cancerminer.org</a></td>
</tr>
<tr>
<td>Li et al. [29]</td>
<td>Integrate miRNA-overexpression data and target information to predict miRNA-mRNA interactions.</td>
<td>miRNA-overexpression, target information based on sequence data</td>
<td>A good resource of miRNA overexpression data When new overexpression data is available, users can apply the method to predict miRNA targets.</td>
<td><a href="http://www.bioconductor.org/packages/devel/bioc/html/TargetScore.html">http://www.bioconductor.org/packages/devel/bioc/html/TargetScore.html</a></td>
</tr>
<tr>
<td>Liu et al. [30]</td>
<td>Use Bayesian network learning in split samples to learn miRNA–mRNA interactions.</td>
<td>Target information based on sequence data, matched miRNA and mRNA expression data</td>
<td>Suitable for data sets with multiple conditions</td>
<td>Upon request</td>
</tr>
<tr>
<td>Le et al. [31]</td>
<td>Use a causality discovery-based method to infer the causal effect that a miRNA has on an mRNA.</td>
<td>Matched miRNA and mRNA expression data</td>
<td>A good method of simulating the gene knockdown experiments Computational complexity is high</td>
<td><a href="http://bioinformatics.oxfordjournals.org/content/29/6/765/suppl/DC1">http://bioinformatics.oxfordjournals.org/content/29/6/765/suppl/DC1</a></td>
</tr>
<tr>
<td>Liang et al. [32]</td>
<td>Explore miRNA activities in different conditions of the data sets</td>
<td>miRNA and mRNA expression data</td>
<td>Suitable for finding active miRNAs in expression data sets with multiple conditions</td>
<td><a href="http://sysbio.ustc.edu.cn/software/mirAct">http://sysbio.ustc.edu.cn/software/mirAct</a></td>
</tr>
<tr>
<td>Amar et al. [33]</td>
<td>Identify a group of differential co-expression genes</td>
<td>miRNA and mRNA expression data</td>
<td>Can be used to find a group of biomarkers for a condition of interest.</td>
<td><a href="http://ag.cs.tau.ac.il/dicer/">http://ag.cs.tau.ac.il/dicer/</a></td>
</tr>
</tbody>
</table>
to reduce the false discoveries compared with sequence-based approaches. Details of these approaches have been reviewed elsewhere [23]. Recent advances in these approaches concern the problem of integrating heterogeneous data sets.

Chen et al. [27] proposed a method to combine different GE sets into a model for exploring miRNA–mRNA regulatory relationships. The main argument of the method is that analysing the relationships from multiple data sets at the same time may improve the identification of miRNA–target interactions. The method firstly used Pearson correlation coefficient method to calculate the correlation of miRNA–mRNA pairs in each individual data set. It then used the empirical Bayes approach to incorporate shared information between data sets for the identification of miRNA–mRNA interactions. Experimental results on both simulation studies and real world cancer data sets proved that the method is better than the ones that use only one data set or simple aggregated data set.

Meanwhile, Jacobsen et al. [28] used regression analysis to analyse miRNA–target interactions across diverse cancer types. The method firstly inferred the miRNA targets in individual cancer types using a multivariate linear model. Apart from miRNA and mRNA expression profiles, the multivariate linear model takes into consideration the effects of DNA copy number and promoter methylation at the mRNA gene locus. Specifically, the mRNA expression is represented as a linear function of miRNA expression, DNA copy number changes (log2 tumor/normal ratio) and promoter methylation, i.e. methylation beta-value of the selected methylation probe. The selected methylation probe is the one that shows the strongest negative correlation between its beta value and the gene mRNA expression across all samples in a cancer type. It is found that the miRNA–mRNA pairs predicted by the method largely overlap with those predicted by miRanda [34] and TargetScan [20]. To explore the recurrence of target association across cancer types, the authors proposed a statistical score (REC ((recurrence) score) to rank the miRNA–mRNA expression associations. The top-ranked miRNA–mRNA interactions are those that have strong association across different cancer types. They applied the method to 11 different cancer types in The Cancer Genome Atlas (TCGA, https://tcga-data.nci.nih.gov/tcga/tcgaHome2.jsp) and reported all the results as an online resource at http://cancerminer.org.

In contrast to the above methods which used expression data, Li et al. [29] developed a method called TargetScore to integrate miRNA-overexpression data and target information based on sequence data. The method firstly compiled 113 miRNA-transfected experiment data from 84 Gene Expression Omnibus (GEO) data sets and calculated the gene expression fold-changes in each experiment. The authors then used the Variational Bayesian–Gaussian Mixture Model to integrate the scores from sequence-based prediction methods and the miRNA-overexpression data. The prediction results were validated using mirTarBase [35], proteomic data in Baek et al. [36] and gene functional enrichment analysis. The transfection data and source codes are available at Bioconductor: http://www.bioconductor.org/packages/devel/bioc/html/TargetScore.html.

These methods effectively apply classical approaches to different types of data, which increasingly become available, to explore miRNA–target relationships. In another direction, researchers have also explored novel approaches for elucidating miRNA functions by designing and applying the emerging data mining and machine learning techniques.

**Emerging approaches**

The first emerging approach is to infer miRNA–target relationships using causality discovery-based methods. The main argument of this approach over the classical association approach is that correlations or associations are not necessarily causality. For instance, the expression values of a miRNA and an mRNA may be strongly correlated across a set of samples, but it is not sufficient to conclude that the miRNA regulates the mRNA. The strong correlation between the miRNA and the mRNA may be a result of the mRNA regulating the miRNA, or a third molecule such as a TF regulating both the miRNA and the mRNA.

The gold standard for tackling the causality discovery problem is randomized control experiments. For example, we can use gene knockdown experiments to knock down miRNAs one by one while measuring the changes (i.e. causal effects) in the expression level of mRNAs. However, such experiments are time-consuming, expensive and not necessarily definitive. In computer science, the main theory of discovering causality from data is the theory of causal Bayesian networks. However,
the computational complexity of Bayesian network learning algorithms is very high, and thus learning the causal network becomes impractical in high-dimensional data sets like gene expression data sets. In application, researchers usually integrate some biological knowledge into the Bayesian network learning process to reduce the complexity.

For instance, Bayesian network structure learning was used in [30] to discover the regulatory relationships between miRNAs and mRNAs. Bayesian network learning algorithm searches for all possible networks in the form of directed acyclic graphs and uses a scoring function to score each graph based on observational data. In this work, the authors assumed that there was a bipartite of interactions between the group of miRNAs and group of mRNAs (miRNAs regulate mRNAs) and there were no interactions between molecules in the same group. Target information predicted with sequence-based methods was used to initialize the network (bipartite). The gene expression profiles of miRNAs and mRNAs were then discretized and input into the Bayesian network learning algorithm [37–39]. The innovation of this work is that the authors used target information to restrict the search space for the computational expensive Bayesian network learning algorithm. Additionally, they split the samples according to their conditions (i.e. normal and cancer) and infer the top interactions under each condition. The final results are the merge of the interactions in all conditions.

However, the method assumes the bipartite of interactions between miRNAs and mRNAs. This assumption may limit the ability to interpret the results and may not necessarily hold in reality. For example, a TF may regulate other mRNAs and even may regulate miRNAs [40]. Therefore, it is necessary to consider different kinds of interactions, e.g. TF regulates miRNA. Ideally, we should assume that every molecule could interact with each other when building a computation model.

Recently, Le et al. [31] adapted a causal discovery approach, IDA (intervention-calculus when the DAG is absent) [41, 42], to uncover the causal regulatory relationship between miRNAs and mRNAs. The method firstly learnt the causal structure from the expression profiles of miRNAs and mRNAs using Peter & Clark (PC) algorithm [43, 44]. do-calculus [38, 42] was then used to estimate the causal effect that a miRNA has on a mRNA. The estimated causal effects simulate the effects of randomized controlled experiments. The method tackles two drawbacks of current miRNA regulatory relationships research. Firstly, the method discovers causal relationships between miRNAs and mRNAs, not just the statistical relationships. Secondly, the method assumes that miRNAs and mRNAs interact with each other in a complex system; for instance, a miRNA can causally regulate mRNAs as well as other miRNAs. This assumption is more reasonable than the assumption from commonly used approaches that considers only the bipartite of interactions between miRNAs and mRNAs. For example, Zisoulis et al. [45] shows that let-7 can regulate other non-coding RNAs including miRNAs. However, the method has high computational complexity and can only infer the lower bounds of the real causal effects.

Another emerging approach is using differential analyses. To understand the causes of a disease, it often requires analysing the differences between normal and disease samples. For example, differentially expressed analysis identifies genes that express differently in different conditions, and thus reveal possible biomarkers for a disease. Recent advances have seen several forms of differential analyses for different purposes, such as differential co-expression and differential networking. Differential co-expression analysis identifies the groups of co-expressed genes that differ markedly between disease and control samples [33, 46]. Meanwhile, differential networking goes a step further to identify the differences in the gene regulatory networks between diseased and healthy conditions.

In this research stream, researchers infer miRNA activity changes in two biological conditions. Highlights in this direction are miReduce [47], DIANA-mirExTra [48], Sylamer [49] and MIR [50]. The common feature of these methods is that they firstly infer the differences in gene expression levels in the two biological conditions, then correlate those alterations with the miRNA binding motifs predicted based on sequence data.

Contrastingly, Liang et al. [32] proposed a method (mirAct) to explore the miRNA activity in a sample and then analyse the overall behavior of the miRNA activity in samples with different biological conditions. Meanwhile, Amar et al. [33] proposed a method called DICER to detect differential co-expression in disease and control samples. They hypothesized that changes in co-expression may be the result of changes in regulatory patterns, and thus
the discovered differential co-expression may be the targets of specific miRNA families. To test the approach, they identified miRNA families whose targets are enriched in the gene groups detected by their method, and tested whether those miRNAs are associated with the relevant diseases.

As suggested in [46], the differential networking may be a promising approach to elucidate the causes of diseases. An example of such approach is a work in protein–protein interaction networks [51]. The method identifies the differences of the hubs of protein–protein interactions between two conditions of interest. They hypothesized that the hubs of interactions play an important role in the regulation networks and the differences in the interaction behaviors may provide some clues for the causes of diseases. However, these differential networking techniques are rare in miRNA research.

DISCOVERING miRNA-mRNA MODULES BY INTEGRATING HETEROGENEOUS DATA SOURCES

It is important to know the modular organization of the regulatory networks, as the recognition of these structures advances our understanding of the complex systems [52–54]. However, this is a challenge in the case of miRNA–gene interactions, as each miRNA can regulate a large number of genes, and multiple miRNAs can regulate the same gene. Therefore, researchers aim to search for a set of miRNAs and their co-regulated genes [55].

In the first stream of research, researchers identify the groups of co-expressed miRNAs and mRNAs using sequence data, or integrate sequence and expression data with or without taking the biological conditions of the data sets into consideration (see [24] for a review). However, these methods use only one or two data resources (sequence and expression data), and are thus sensitive to the data noise as pointed out in [55]. As each data type provides different but complementary information, recent studies (Table 3) integrate multiple sources of data into inferring miRNA-gene regulatory modules.

Zhang et al. [55] proposed a framework to identify miRNA-gene regulatory modules by integrating miRNA target predictions based on sequence data, miRNA and gene expression profiles, protein–protein interaction and DNA–protein interaction networks. In their work, sequence-based miRNA target predictions were considered as a static prior network, and expression profiles were used subsequently to identify the active miRNA–gene interactions. These active interactions were further refined by the gene–gene interaction networks (protein–protein and DNA–protein networks). Applied to the human ovarian cancer samples from TCGA, the method discovered several miRNA-gene regulatory modules. The results are then validated against miRNA cluster from miRBase (http://www.mirbase.org/), and the miRNAs are validated using gene functional enrichment analysis.

Similarly, Le et al. [56] proposed a regression-based method to integrate sequence, expression and protein interactions data for identifying modules of miRNAs and mRNAs for a specific condition. The authors firstly used a regression model to link the expression profiles of miRNAs and mRNAs. The assumption was that the expression level of an mRNA can be represented as a linear function of expression profiles of all predicted miRNAs. The predicted miRNA regulators for an mRNA were taken from sequence-based prediction databases. To assign miRNAs and mRNAs into a module, they designed a function to measure the strength of the predicted miRNA–mRNA interactions based on the information from miRNA target information and protein–protein interactions. Specifically, the assigning function is based on the logistic-sigmoid function with parameters to adjust the contributions of the two types of interaction data. The higher the probabilities of interactions are, the more chances the interacting entities are assigned into the same module. The method was applied to multiple cancer data sets from TCGA to explore the regulators (miRNAs) that are common for all cancer types and the specific active regulators for each cancer type. The results were then validated against knowledge from literature and by gene functional enrichment analysis.

Zhang et al. [57] proposed a method to integrate cancer genomic data from different platforms, including DNA methylation (DM), GE and miRNA expression data (ME). The method adopted a joint matrix factorization technique to integrate different data sources for ovarian cancer samples from TCGA. The method firstly identified the subsets of miRNAs, mRNAs and methylation markers that show the similar patterns across a subset or all of the samples. The aims were to reduce the complexity of the data and provide a global overview of the data.
### Table 3: Summary of methods to discover miRNA-mRNA modules

<table>
<thead>
<tr>
<th>Method</th>
<th>Brief description</th>
<th>Data sources</th>
<th>Notes</th>
<th>Software tool</th>
</tr>
</thead>
</table>
| Zhang et al. [55] | A framework to identify miRNA-gene regulatory modules by integrating multiple data sources  
- Use sequence-based predictions to build the prior network.  
- Use expression data and known gene–gene interactions to refine the findings |  
- Sequence-based miRNA target predictions,  
- miRNA and gene expression profiles,  
- protein–protein interactions,  
- DNA–protein interactions |  
- Can be applied when we have matched data sets for multiple data types.  
- The outputs are groups of miRNAs and mRNAs that are important in the biological condition of the data set | [http://zhoulab.usc.edu/SNMNMF/](http://zhoulab.usc.edu/SNMNMF/) |
| Le et al. [56] | A regression-based method for identifying modules of miRNAs and mRNAs for a specific condition by integrating multiple data sources  
- Use sequence-based predictions to assign regulator–target relationships.  
- Use expression data to build a regression model between regulators and targets  
- Use protein–protein interactions to design a function for measuring the strength of the predicted miRNA–mRNA relationships. |  
- Sequence-based miRNA target predictions,  
- miRNA and mRNA expression data,  
- protein–protein interactions |  
- Provide a list of miRNA-mRNA regulatory modules from multiple cancer data set from TCGA. This could be a good resource for further exploration | NA |
| Zhang et al. [57] | A joint matrix factorization method to integrate cancer genomic data from different platform  
- Integrate multiple data types using matrix factorization technique  
- Apply to Ovarian data sets from TCGA |  
- DM,  
- miRNA and GE |  
- Suitable when matched samples of DM data available  
- May exclude important genes | [http://nar.oxfordjournals.org/content/40/19/9379/suppl/DC1](http://nar.oxfordjournals.org/content/40/19/9379/suppl/DC1) |
| Li et al. [58] | The sparse Multi-Block Partial Least Squares method to integrate multiple data sources  
- Reduce the dimension of each data type and represent as a vector  
- Use regression to integrate the summarised vectors each data type |  
- CNV,  
- DM,  
- miRNA and GE |  
- Prove that using multiple data types generates better results than using single data type  
- May miss important genes in the modules | [http://zhoulab.usc.edu/sMBPLS/](http://zhoulab.usc.edu/sMBPLS/) |
from different platforms. The authors applied the method to the ovarian cancer data sets to find the so-called multi-dimensional modules. They validated the method by investigating the genes in the modules that have been confirmed as ovarian cancer-related genes and by gene functional enrichment analyses.

In a similar fashion but targeting different data types, Li et al. [58] proposed a method to integrate multiple data sources, including copy number variation (CNV), DM, GE and ME for inferring multi-layer gene regulatory modules. The proposed method is called sparse Multi-Block Partial Least Squares regression method (MBPLS) and is employed to identify multi-dimensional regulatory modules from the data. The assumption was that CNV, DM and ME all regulate the gene expression. The method projected each data type into a summary vector, and maximizes the covariance between the summary vectors of source data (CNV, DM, ME) and the response data (GE). Finally, it used the weighted sum of the summary vectors of source data to represent the unique input source data, and again maximizes the covariance between the input data and the response data. The method was tested on simulated data as well as the ovarian cancer data sets from TCGA and was capable of identifying the modules that have significant functional and transcriptional enrichments. The results predicted from this method were proved to be better than the results from those methods that use only one type of data.

The results from the above studies encourage future work of integrating more types of genomic data for elucidating the causes of diseases.

**DISCOVERING miRNA AND TF CO-REGULATORY RELATIONSHIPS**

TFs and miRNAs are primary gene regulators, and identifying their functions is a challenging and important research topic. Currently, there are still no feasible experimental techniques to discover miRNA and TF co-regulatory mechanisms. Meanwhile, computational methods have mainly focused on exploring the functions of miRNA and TF separately in the past decade.

A unified picture of regulatory relationships of the two main regulators and target genes would provide useful insights into the causes of diseases. The combined regulations of miRNAs and TFs are important but difficult to explore, as miRNAs and TFs can regulate each other in addition to regulating target genes. Recently, there are some studies constructing the gene regulatory networks with the presence of both TFs and miRNAs based on sequence data. Few other works use both sequence-based target predictions of miRNAs and TFs and expression profiles to learn the complex regulatory network and inferring network motifs. Table 4 shows the basic features of the methods in this category.

**Sequence-based approaches**

A common framework of exploring miRNA–TF co-regulatory relationships is to integrate the putative target information of both TFs and miRNAs to obtain an interaction network with the three components, miRNAs, TFs and miRNAs. Statistical tests, network inference algorithms or gene functional enrichment analyses are then used to infer gene regulation knowledge from the combined network. Figure 3 shows the procedure of this framework.

In the first stream of research, researchers studied the co-regulation of TFs and miRNAs by discovering their shared downstream targets [59, 60]. Shalgi et al. [59] built the network that involves miRNAs, TFs and mRNAs using sequence data. They used evolutionary conserved binding sites of miRNA targets to construct the interactions between miRNAs and genes (including TFs). Meanwhile, conserved binding sites of TFs in promoters were used to uncover the interactions between TFs and mRNAs and the interactions between TFs and miRNAs. The combined network was then analysed to identify the shared targets of the regulators. It was found that the hub of interactions is usually TFs and also discovered some network motifs that involve miRNAs, TFs and miRNAs using sequence data. They used evolutionary conserved binding sites of miRNA targets to construct the interactions between miRNAs and genes (including TFs). Meanwhile, conserved binding sites of TFs in promoters were used to uncover the interactions between TFs and mRNAs and the interactions between TFs and miRNAs. The combined network was then analysed to identify the shared targets of the regulators. It was found that the hub of interactions is usually TFs and also discovered some network motifs that involve miRNAs, TFs and miRNAs. Similar to Shalgi’s approach, Zhou et al. [60] used PicTar [26] as the miRNA putative targets and Transfac [70] as TF target information to build the network of miRNAs, mRNAs and TFs. They then used Fisher’s Exact Test to measure the significance of the shared targets of the regulators. It was found that the hub of interactions is usually TFs and also discovered some network motifs that involve miRNAs, TFs and miRNAs. Similar to Shalgi’s approach, Zhou et al. [60] used PicTar [26] as the miRNA putative targets and Transfac [70] as TF target information to build the network of miRNAs, mRNAs and TFs. They then used Fisher’s Exact Test to measure the significance of the shared targets of the regulators. It was found that the hub of interactions is usually TFs and also discovered some network motifs that involve miRNAs, TFs and miRNAs. Similar to Shalgi’s approach, Zhou et al. [60] used PicTar [26] as the miRNA putative targets and Transfac [70] as TF target information to build the network of miRNAs, mRNAs and TFs. They then used Fisher’s Exact Test to measure the significance of the shared targets of the regulators. It was found that the hub of interactions is usually TFs and also discovered some network motifs that involve miRNAs, TFs and miRNAs. Similar to Shalgi’s approach, Zhou et al. [60] used PicTar [26] as the miRNA putative targets and Transfac [70] as TF target information to build the network of miRNAs, mRNAs and TFs. They then used Fisher’s Exact Test to measure the significance of the shared targets of the regulators. It was found that the hub of interactions is usually TFs and also discovered some network motifs that involve miRNAs, TFs and miRNAs.
<table>
<thead>
<tr>
<th>Method</th>
<th>Brief description</th>
<th>Data sources</th>
<th>Notes</th>
<th>Software tool</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shalgi et al. [59]</td>
<td>Build the regulatory network that involves miRNAs, TFs and mRNAs. Use sequence-based predictions to build the network. Identify the shared targets of a pair of regulators.</td>
<td>Sequence-based target predictions</td>
<td>The hubs of interactions are usually TFs. The results may involve a high rate of false discoveries.</td>
<td>No software tool available. The procedure is repeatable. Users can replace the target information with recent prediction programs.</td>
</tr>
<tr>
<td>Zhou et al. [60]</td>
<td>Build the miRNA-TF-mRNA regulatory network. Use sequence-based predictions to build the network. Use Fisher's exact test to identify the significant shared targets of regulators.</td>
<td>Sequence-based target predictions</td>
<td>Shared targets of TFs are much more abundant than that of TF-miRNA. Shared targets in FRL motifs with TFs as master regulators are statistically significant.</td>
<td>As above</td>
</tr>
<tr>
<td>Chen et al. [61]</td>
<td>Explore the miRNA–TF co-regulatory relationships. Use sequence-based predictions to build the network. Use gene functional enrichment analysis to find functional profiles of shared target genes. Identify significant shared targets of regulators.</td>
<td>Sequence-based target predictions</td>
<td>Some biological processes emerged only in co-regulation. Did not consider the relationships between the regulators.</td>
<td>As above</td>
</tr>
<tr>
<td>Tran et al. [62]</td>
<td>Discover gene regulatory modules that consist of miRNAs, TFs and mRNAs using a rule based method. Build the regulatory network based on target information. Define the network based on a set of rules, e.g. binding score is significant.</td>
<td>Sequence-based target predictions</td>
<td>Output a set of modules that involve miRNAs and mRNAs. Results are sensitive to the defined rules.</td>
<td>As above</td>
</tr>
<tr>
<td>Le Bechec et al. [63]</td>
<td>Provide a web tool for miRNA-TF co regulation analysis. Use available target prediction databases to build the network. Infer network motifs based on the built network.</td>
<td>Sequence-based target predictions</td>
<td>A good resource for exploring network motifs that involve miRNA and TF. Did not use expression data, so the network may be static.</td>
<td><a href="http://mironton.uni.lu">http://mironton.uni.lu</a></td>
</tr>
<tr>
<td>Roqueiro et al. [64]</td>
<td>Identify the key regulators (miRNAs or TFs) of pathways. Curate known pathways that involve miRNAs and TFs. Use Bayesian inference to integrate sequence based target predictions.</td>
<td>KEGG pathways, Sequence-based target predictions</td>
<td>Can be used to identify key regulators of a disease. Good resources for validating computational predictions</td>
<td>NA</td>
</tr>
<tr>
<td>Sun et al. [65]</td>
<td>Uncover miRNA-TF regulatory network in GBM. Curate a list of miRNAs, TFs and genes related to GBM. Build a regulatory network based on sequence data. Use gene expression to gene–gene interactions, and network motifs.</td>
<td>Sequence-based target predictions,</td>
<td>Specific for GBM. Good resource for exploring network motifs involved in GBM.</td>
<td>NA</td>
</tr>
<tr>
<td>Jiang et al. [66]</td>
<td>Identify active miRNA-TF regulatory pathways in Alzheimer's disease. Use curated databases to build the regulatory network. Use gene expression to identify differentially expressed miRNAs and genes between disease and control samples.</td>
<td>Curated miRNA and TF target databases, GE</td>
<td>Specific for Alzheimer's disease. A good resource for validating computational predictions related to Alzheimer's disease</td>
<td>NA</td>
</tr>
</tbody>
</table>
more statistically significant than other types and feed-forward loops.

Tran et al. [62] proposed a rule-based method to discover the gene regulatory modules that consist of miRNAs, TFs and their target genes based on the available predicted target binding information. Sequence-based target prediction database was used to construct putative targets of miRNAs and TFs. The obtained network was refined by retaining only the genes regulated by at least two miRNAs and two TFs with a significant binding score ($P$-value < 0.05). The method then searched for the module that consists of genes, miRNAs and TFs such that the miRNAs and TFs regulate the genes with a stringent $P$-value. The validity of the modules was assessed using gene functional enrichment analysis for the target genes.

Contrastingly from the above works that use statistical tests as the filter (Figure 3), Chen et al. [61] used gene functional enrichment analysis (second filter in Figure 3) to explore the co-regulatory relationships. They also used target information to construct the co-regulation network as the first step. The authors then applied Gene Ontology (GO) for gene functional enrichment analysis for the shared target genes to find the functional profiles for these co-regulation pairs. To calculate the significant levels of the shared targets, they compare their method with the randomly pick method and use the hypergeometric distribution to calculate the $P$-values of the findings. It was found that some biological processes emerged only in co-regulation and that the disruption of co-regulation might be closely related to cancers, suggesting the importance of the co-regulation of miRNAs and TFs.

Apart from the methods for inferring miRNA-TF co-regulations, there are some works focusing on providing resources for exploring miRNA and TF shared targets, network motifs involving miRNAs and TFs and known pathways related to miRNAs and TFs. For instance, Le Bechec et al. [63] provided a web tool for miRNA-TF co-regulation analysis, MIR@NT@N, which is available at http://miron-ton.uni.lu. They integrated available target prediction databases for both TFs and miRNAs to construct a regulatory network that involves miRNAs, TFs and mRNAs. This work provides a web resource for facilitating the retrieval of regulatory relationships and network motifs. Users can explore the shared targets of miRNAs and TFs, and query a list of Feed-Forward Loops (FFLs) and Feed-Back Loops (FBLs) that involve miRNAs and TFs. Differently,
Roqueiro et al. [64] proposed a method to identify the key regulators (miRNAs or TFs) of pathways. The method used Bayesian inference on known pathway structures to infer a set of regulators in the pathway network. The Bayesian network in this method was constructed manually using the known KEGG pathways by removing the cycles in the pathways and applying some filtering criteria. The method drew findings based on existing knowledge and provided a good resource for other methods to validate their results. However, it is not fit for application in further exploratory studies.

The common feature standing out from the above methods is their employment of sequence-based putative target information. However, the networks constructed from sequence-based predictions involve a high rate of false negatives and false positives [71]. Therefore these methods are only the first step of exploring the complex relationships between the three components, miRNAs, TFs and mRNAs. It would be ideal if expression data can be incorporated to refine the discoveries.

Expression data and other data-based approaches
Recently, there has been some work in the second stream of research into miRNA and TF co-regulations that incorporates GE into the studies. These methods also use sequence-based target prediction programs to initialize the network at the first step, after which the expression data are used to refine the findings. This procedure is similar to the framework in Figure 2 of the previous section. For instance, Sun et al. [65] proposed a method to uncover miRNA and TF regulatory networks in Glioblastoma (GBM). They firstly filtered the miRNAs, TF and genes related to GBM based on existing knowledge from literature. They then integrated the target prediction of miRNA and TF, which are based on sequence data for constructing the regulatory network. Only the gene expression was used to infer the gene–gene interactions by assuming that the interaction occurs when they are co-expressed. The authors then infer the 3-node FFL and 4-node motifs, which involve miRNA–TF interactions. These motifs were integrated into a so-called GBM miRNA–TF mediated network. The authors then conducted the signalling pathway and gene functional enrichment analyses to validate the results.

Similarly, Le et al. [69] proposed a framework to learn from heterogeneous data the three-component regulatory networks, with the presence of miRNAs, TFs and mRNAs. They firstly used
target information of miRNA and TF to define the bipartite of interactions between regulators and target genes. They then used Bayesian network structure learning to construct a regulatory network from gene expression profiles of miRNAs, TFs and mRNAs. Then, in order to produce more meaningful results for further biological experimentation and research, the method searched the learnt network to identify the interplay between miRNAs and TFs, and the FFLs that involve miRNAs and TFs from the learnt network.

In another direction, researchers used target information to build the TF, miRNA and mRNA regulatory networks as the first step. The expression data was then used to identify active pathways that involve miRNAs and TFs [66], or to identify active regulators in different biological conditions of the data sets [68]. For example, Jiang et al. [66] proposed a method to identify active miRNA-TF regulatory pathways in Alzheimer’s disease. In the first instance, the curated databases, including TransmiR [40], TRANSFAC [70], miRecords [72], TarBase [73] and miRTarBase [35], were used to create the network that included miRNAs, TFs and genes. The authors then integrated miRNA and GE from different sources and identified the differently expressed genes and miRNAs between disease and control samples. They defined these genes as active seeds (nodes) in the curated network, and found the subnetwork by connecting all the active nodes with their immediate neighbours. Furthermore, the authors searched for the active pathway in the subnetwork by searching for the directed acyclic paths from nodes without parents to nodes without children. The results were validated using gene functional enrichment analysis.

Differently, to create a tool for clinical scientists to generate hypotheses and for explorations, Huang et al. [67] develops a web tool (mirConnX) for constructing the regulatory networks that include miRNAs, TFs and mRNAs. They firstly created the prior network based on sequence-based miRNA and TF target prediction programs as well as the experimentally confirmed target databases. The expression data were then used to build the association network where the edges represent the significant correlation between the expression levels of the two genes (nodes). They then integrated the association network based on expression data and the prior network based on sequence data into the final network. Using this tool, users can input the expression data and receive the regulatory network that includes miRNA, TFs and mRNAs. The built networks can be further analysed to identify network motifs. However, an edge in this network simply shows the statistical association in expression levels between a regulator and a target gene, which may not indicate a regulatory relationship.

It is challenging and interesting to design a new class of methods that integrate multiples types of data for exploring miRNA-TF-mRNA regulatory complex relationships.

CHALLENGES IN EVALUATION AND SELECTION OF MODELS

Evaluation of models

A question raised even before designing any computational methods of identifying miRNA–mRNA interactions is how to validate the predictions. Although there has been some progress in tackling the validation problem, the challenge remains due to the sparse number of experimentally confirmed miRNA–mRNA interactions. Therefore, there is no complete ground-truth for evaluating and comparing different computational methods. In this section, we review the current methods of validating or enriching miRNA target predictions.

Currently, the common methods for validating computational results about miRNA targets are as follows:

- Using experimentally validated target databases such as TarBase [73], miRecords [72] and miRTarbase [35]
- Using miRNA transfection experiment data
- Performing enrichment analysis using proteomic data
- Calculating the number of target genes appearing in known pathways
- Applying functional and/or pathway enrichment analyses to investigate the relevance of the target genes to the biological conditions of the data set

Tarbase and miRecords are manually curated experimental interaction databases and are commonly referred to when validating miRNA target predictions. These databases contain the collection of the experimentally confirmed interactions, and provide interfaces for facilitating the information retrieval process. miRTarBase stores miRNA targets validated...
by high-confidence low-throughput assays such as western blot. Meanwhile, Transmir [40] stores the curated experimentally confirmed interactions between miRNAs and TFs. These databases provide good tools to validate miRNA-mRNA regulatory predictions, although the number of confirmed targets is still small.

A common approach to validating miRNA-mRNA regulatory predictions is to examine the overlap between the predictions and the experimentally confirmed databases. However, the small number of experimentally validated targets would make the validation of a new method difficult, as we do not know whether a predictive target that is not in those databases is a false discovery or a novel true target.

Another approach of validating miRNA-mRNA regulatory predictions is to use miRNA transfection data. Transfection data presents the changes in gene expression between the control and miRNA-transfected conditions. The differentially expressed genes identified from the control and miRNA-transfected samples are considered as targets of the miRNA. Although this approach cannot differentiate between direct and indirect miRNA-mRNA interactions, it provides a tool to evaluate the real effect that a miRNA has on mRNAs.

Khan et al. [74] collected the miRNA transfection data from 151 published transfection experiments in seven different human cell types. Luo et al. [75] provided the fold-change in the gene expression levels between control and transfected samples in MDA-MB-231 cells for miR–200c, miR–375 and miR–205. Le et al. [31] presented the transfection data for miR–200 family in MDA-MB-231 cells. Li et al. [29] compiled 84 data sets from GEO for 113 transfected miRNAs. Table 5 shows the available miRNA transfection data sources.

In the similar fashion, protein expression data in miRNA transfection experiments can be used to validate computational predictions. Although measuring protein expression data generates higher cost compared with GE, protein expression data better reflect the effect of miRNA regulations. For instance, Baek et al. [36] reported both gene and protein expression data for transfection experiments of hsa-miR-1, 124 and 181a. They then use the data to validate and compare different target prediction methods.

In another direction, researchers use pathway and/or gene functional enrichment analyses to investigate the relevance of the functions of predicted targets to the biological conditions of the data sets used for the prediction [23, 30, 83]. The assumption is that a good prediction method will generate a set of miRNA target genes which are involved in the pathways and processes relevant to the biology behind the used data set. There are several software tools designed for this purpose. The common ones are GO enrichment analysis, KEGG pathway analysis [85], GeneCodis [86], Ingenuity Pathway Analysis (IPA, Ingenuity Systems, www.ingenuity.com), GeneGo Metacore from GeneGo Inc. The first three are free for research use, while the last two are commercial software tools with a limited trial time.
To date, the validation tools for evaluating miRNA-mRNA regulatory predictions are still limited and there is no unique ground-truth for assessing the performance of different computational methods. A possible remedy for the problem is to combine different validating methods. For example, a good computational method should have a significant number of miRNA–mRNA interactions confirmed by transfection experiments and have the target genes highly relevant to the biological condition of the data sets. However, the differences in biological conditions between the training data sets and the transfection experiments may result in biased validation results. It is desirable to have follow-up experiments with the same biological conditions of the data set used for validating the results of a computational prediction method.

Selection of models

As we do not have the ground-truth for evaluating the models, it is impossible to conclude which method is better than the other. Therefore, selecting a model for assisting with the experiment design is a difficult task. To investigate whether we can select one model over another based on its relative performance validated by the above-mentioned evaluation methods, in this section we conduct a case study in three different cancer data sets.

In this study, we choose Pearson correlation [87], maximal information coefficient (MIC) [88], Lasso [89], Elastic-net [90] and the method which is based on IDA in [31] for the comparison. We apply them to the matched miRNA and mRNA expression profiles from epithelial-mesenchymal transition (EMT) [91, 92], multi-class cancer (MCC) [93, 94] and 51 cell lines breast cancer (BR51) [95] data sets. We then use experimentally validated miRNA targets and miRNA transfection data for validating the predictions.

For EMT data sets, the miRNA expression profiles are from [92] (data available at http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE26375). They were profiled from the 60 cancer cell lines of the drug screening panel of human cancer cell lines at the National Cancer Institute (NCI-60). The mRNA expression profiles of EMT for NCI-60 were obtained from ArrayExpress http://www.ebi.ac.uk/arrayexpress accession number E-GEOD-5720. There are 11 samples of epithelial and 36 samples of mesenchymal. For MCC data sets, the miRNA expression profiles were obtained from [94] (data available at http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE2564). The mRNA expression profiles of MCC are from [93]. They can be downloaded at http://www.broad.mit.edu/cancer/pub/migcm. Samples of the MCC data classified as normal (21 samples) and tumour (67 samples). Finally, for BR51 data sets, the miRNA data are from [95] and the mRNA GE are available at http://www.ncbi.nlm.nih.govgeoqueryacc.cgiacc=GSE41313. There are 27 samples in the luminal group and 23 samples in the basal group for this data set.

We perform differential gene expression analysis on the gene expression profiles to identify differentially expressed miRNAs and mRNAs between the two conditions in each data set. In this work, we use limma package [96] of Bioconductor for the analysis. As a result of the analysis, 46 probes of miRNAs and 1635 probes of mRNAs for the EMT data set, and 62 probes of miRNAs and 1363 probes of mRNAs for the MCC data set have been identified to be differentially expressed at significant level (adjusted $P$-value < 0.05, adjusted by BH method). Meanwhile, 92 miRNAs (adjusted $P$-value < 0.2) and 2354 mRNAs (adjusted $P$-value < 0.0001) are identified to be differentially expressed in the BR51 data set. To cover the important miRNAs mentioned in the analysis of [95] and to have a manageable number of mRNAs for the computational method, we choose different thresholds of adjusted $P$-values in the differential gene expression analysis for this data set.

Figure 4 shows the performance of the methods validated by experimentally confirmed target databases. We firstly use the union of three databases, Tarbase 6.0 [97], miRecords [72] and miRWalk [98], as the ground-truth. We use the miR-200a transfection data from [31] to validate the predictions of the methods. The comparison results are shown in Figure 5.

In general, there is no superior method that outperforms all other methods in any validation method. Association methods such as MIC and Pearson perform well when validating against experimentally
confirmed target databases. Meanwhile, the causality discovery-based method, IDA, performs well when we use miRNA transfection data as the ground-truth.

It is important to note that the validation results, however, do not imply that all the methods are of equal merits. Figure 6A shows the overlap in miR-200a predicted targets by IDA, Pearson and MIC. There is a significant number of targets predicted by a method, but other methods fail to discover. This suggests different methods may infer different sets of the miRNA targets. Moreover, Figure 6B, 6C

Figure 4: Number of interactions validated by experimentally confirmed target databases for IDA, Pearson correlation coefficient, MIC, Lasso and Elastic-net in EMT, MCC and BR5I data sets.

<table>
<thead>
<tr>
<th>p-values</th>
<th>IDA</th>
<th>MIC</th>
<th>PEARSON</th>
<th>LASSO</th>
<th>ELASTIC-NET</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMT (Experimentally validated)</td>
<td>4.00E-03</td>
<td>2.60E-04</td>
<td>8.50E-04</td>
<td>6.20E-03</td>
<td>4.00E-03</td>
</tr>
<tr>
<td>MCC (Experimentally validated)</td>
<td>1.90E-01</td>
<td>5.10E-01</td>
<td>2.20E-01</td>
<td>4.60E-01</td>
<td>6.20E-01</td>
</tr>
<tr>
<td>BR5I (Experimentally validated)</td>
<td>7.20E-06</td>
<td>2.10E-12</td>
<td>4.50E-10</td>
<td>1.00E-10</td>
<td>1.90E-09</td>
</tr>
</tbody>
</table>

Figure 5: Number of interactions validated by miR-200a transfection data for IDA, Pearson correlation coefficient, MIC, Lasso and Elastic-net in EMT, MCC and BR5I data sets.

<table>
<thead>
<tr>
<th>p-values</th>
<th>IDA</th>
<th>MIC</th>
<th>PEARSON</th>
<th>LASSO</th>
<th>ELASTIC-NET</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMT (Transfection data)</td>
<td>1.00E-12</td>
<td>1.00E-12</td>
<td>4.90E-12</td>
<td>1.00E-03</td>
<td>2.00E-02</td>
</tr>
<tr>
<td>MCC (Transfection data)</td>
<td>6.60E-04</td>
<td>1.00E-02</td>
<td>1.00E-02</td>
<td>4.60E-02</td>
<td>4.60E-02</td>
</tr>
<tr>
<td>BR5I (Transfection data)</td>
<td>1.10E-04</td>
<td>4.50E-05</td>
<td>1.10E-04</td>
<td>4.50E-05</td>
<td>4.50E-05</td>
</tr>
</tbody>
</table>
and 6D show a large number of confirmed genes predicted by a method but not by the other. There is a strong implication that these methods discover results that are complementary to each other. Therefore, it is not simple to claim that a method is outstanding over the others using existing validation methods.

CONCLUSIONS AND OUTLOOKS

As more and more evidence has suggested the important roles of miRNAs in the development of several diseases, identifying miRNA functions will reveal further insights into the causes of fatal diseases such as cancer. The huge amount of data available in different types poses opportunities and challenges for computational approaches to exploring miRNA functions. These methods assist with the designs of the wet lab experiments. In this review we have discussed different computational approaches to inferring miRNA functions. These approaches provide different views on how to elucidate the complex regulatory mechanism of miRNAs and their relationships with target genes as well as other regulators.
The approaches that are based on sequence data provide a list of potential target genes. However, the results from these approaches involve a high rate of false discoveries. There is a need to design new accurate methods for using sequence data. Recently, such computational and experimental methods have started to emerge. For instance, Wang et al. [99] proposed the mirTarPri method to reduce the false discovery rates of commonly used target prediction methods by ranking the predicted targets to select optimal results. In another direction, experimental methods such as HITS-CLIP [100] and PAR-CLIP [101] have achieved better accuracy rates compared with prior prediction programs. These new approaches may have potential to enhance our understanding in miRNA functions.

Complementary to sequence-based approaches, which discover static miRNA targets, approaches that use GE can infer the miRNA activities and miRNA–mRNA relationships in a specific condition. These approaches have long been studied based on the foundation of correlation or association. However, it is well known that association is not causation, and thus the correlation between a miRNA and target genes may not imply the real regulatory relationships, which are causal relationships. Recently, causality discovery-based methods have emerged and provided an alternative approach to inferring miRNA–mRNA causal regulatory relationships. In saying so, these methods usually have a set of assumptions on the distribution of the data and those assumptions may be mostly violated in some real-world data sets. As a result, the causality discovery methods may not perform consistently in all data sets. Therefore, it is crucial to design a new class of methods that can infer causal relationships as well as produce stable results across different data sets.

On another note, the approaches of integrating multiple sources of data have been shown to be effective in improving the predictive power of a computational method. Each data type provides complement information, and thus integrating more sources of data would provide a clearer picture of gene regulations. However, different data sources often come from different labs with different experiment settings and different biological conditions. The question raised here is how to uniformly transfer the knowledge learnt from one lab to another.

As the ultimate research goal is to elucidate the causes of a disease, differential analysis approaches are promising. We have seen differential analysis techniques in identifying gene differentially expressed analysis, and recent methods of identifying differential co-expression. As suggested in [46], the next step would be identifying differential networking, i.e. identifying the differences between gene regulatory networks in different conditions. However, it is still unclear how to achieve the goal.

As discussed in the above section, it is still challenging to evaluate and select a model for assisting with the experimental design. It is necessary to create tools for systematically evaluating a model based on current knowledge, and visualizing prediction results from different methods. As different methods may infer complementary results and have their own advantages, ensemble approaches such as in [102] can be promising in producing stable and high accurate results. Although we used human cancer data sets in the case study, the methods discussed in this review generally are not limited to human cancer data sets. The methods can be applied to different organisms and biological conditions.

Gene regulatory networks involve several classes of regulators, and thus constructing the unified picture of the network with the presence of important regulators and genes is crucial. Recent methods have constructed the gene regulatory networks to present the relationships between miRNAs, TFs and genes. However, most of the methods use the sequence-based target prediction to build the network, which may involve a high rate of false discoveries. There is a potential to integrate multiple sources and/or types of data to learn such complex networks. Furthermore, differential analyses can be used to explore the changes in regulatory patterns, e.g. the changes in FFLs and FBLs that involve miRNAs, TFs and/or genes. The FBLs and FFLs have been found to play important roles in cancers and other diseases [103]. Please refer to [103] for a recent review of the TF-miRNA motif roles in biological processes and diseases. Understanding the roles of such motifs in disease development would better assist with the diagnosis process and the design of pharmaceutical products for treatment.

Together with the two major gene regulators, miRNAs and TFs, long non-coding RNAs (lncRNAs) play important roles in biological processes and diseases [104]. More emerging evidence has shown that lncRNAs, miRNAs and TFs are important nodes in the signaling networks that regulate vital biological processes and diseases, including cancer [104, 105]. Recent evidence also reveals the
interactions between miRNAs, TFs and lncRNAs [106]. With more and more data available, there is a strong possibility for computational methods to tap into the area of learning gene regulatory networks with the existence of miRNAs, TFs, lncRNAs and genes. It would pose interesting implications for future work as these methods can help elucidate the complex gene regulatory relationships and the causes of diseases.

Key Points

- microRNAs and transcription factors are important regulators, but few feasible experimental techniques are available for exploring all of their functions. Computational methods can generate hypotheses about miRNA and TF targets, narrowing down the vast amount of possibilities to be tested by wet lab experiments.
- Integrating multiple sources of data and/or multiple types of data enhances the capability to identify miRNA functions.
- Exploring miRNA and TF co-regulation gives more insights into the causes of diseases. Some recent methods use both target information and gene expression profiles to construct the regulatory network with the presence of miRNAs, TFs and genes.
- Although several computational methods are presented and made available, it is challenging to evaluate them and decide which method is better than the others as they are more often complementary to one another.

FUNDING

This work has been partially supported by Australian Research Council Discovery Project DP130104090 and DP140103617.

References


Pearson K. Mathematical contributions to the theory of evolution—on a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proc R Soc L* 1896;60:489–98.


