Chapter 19

MicroRNAs, Regulatory Networks, and Comorbidities: Decoding Complex Systems

Francesco Russo, Kirstine Belling, Anders Boeck Jensen, Flavia Scoyni, Søren Brunak, and Marco Pellegrini

Abstract

MicroRNAs (miRNAs) are small noncoding RNAs involved in the posttranscriptional regulation of messenger RNAs (mRNAs). Each miRNA targets a specific set of mRNAs. Upon binding the miRNA inhibits mRNA translation or facilitate mRNA degradation. miRNAs are frequently deregulated in several pathologies including cancer and cardiovascular diseases. Since miRNAs have a crucial role in fine-tuning the expression of their targets, they have been proposed as biomarkers of disease progression and prognostication.

In this chapter we discuss different approaches for computational predictions of miRNA targets based on sequence complementarity and integration of expression data. In the last section of the chapter we discuss new opportunities in the study of miRNA regulatory networks in the context of temporal disease progression and comorbidities.

Key words microRNA, Regulatory network, Target prediction, Comorbidities, Systems biology, Data integration

1 Introduction

Gene expression regulation has increased in complexity since the recent discovery of noncoding RNAs (ncRNAs) [1, 2]. microRNAs (miRNAs) were identified as new, small ncRNAs with conserved sequences and functions involved in posttranscriptional regulation. miRNAs are widely distributed in animals, plants, and viruses and to date 2588 mature human miRNA transcripts are listed in miRBase v21 (http://www.mirbase.org/) [3]. By targeting specific mRNAs (usually in the 3′ untranslated region (UTR)), miRNAs lead to translational repression or promote the degradation of their target mRNAs [4, 5]. miRNA genes are under strong selective pressure to preserve the 5′ end of mature miRNAs as positions 2 to 7 are crucial for target recognition, termed the “seed” region. The further downstream nucleotides
(particularly nucleotide 8 and 13–16) also contribute to base pairing of miRNAs with their targets [5]. Further evidence suggests that miRNAs also can affect gene expression at translational levels by binding 5' UTR and coding regions of target mRNAs [6, 7] and under specific conditions induce translational upregulation of targets [8]. More than 60% of human protein-coding genes contain at least one conserved miRNA-binding site. A huge number of non-conserved sites also exist [9].

miRNAs have a strong impact on cellular functions (e.g., cell death and proliferation). Therefore, their biogenesis is carefully regulated at transcriptional and posttranscriptional levels [10] and their dysregulation is often associated with human diseases, including cancer [11].

It has also been proposed that miRNAs play a role in development of comorbidities (the presence of two diseases in a patient). miRNA–mRNA interactions can be the underlying causal effects for cases where one disease induces the development of another disease. Examples of such disease pairs are neurological and cardiovascular disorders [12], obesity and metabolic syndrome [13], and complex regional pain syndrome and headache [14]. Comorbidities have been studied in the context of disease associations for many years in epidemiologically research and recently more sophisticated data-driven approaches have linked molecular etiology to disease associations calculated from electronic patient records [15, 16]. The hypothesis of these studies is that strongly associated diseases have a high likelihood of sharing genetic etiology.

Evaluation of miRNA and mRNA expression profiles in blood and tissue samples from patients with comorbidities is of great importance for future studies of comorbidities. Data integration approaches with target prediction and miRNA–mRNA expression correlation analyses will lead to increased knowledge of how miRNA influence comorbidity development. Understanding the function of miRNAs in a single disease is challenging and the complexity is even higher when focus is expanded to comorbidities. Proper experimental design including computational and systems biology approaches can help to decode this complex scenario.

The identification of miRNA targets is a key step to understand the function of miRNAs themselves and for their potential role in diseases. In Fig. 1 we report main variables to take into consideration for a correct identification and interpretation of miRNA targets. The majority of miRNA targets are predicted computationally and several algorithms exist for carrying out the predictions [17–23]. They are based on different criteria using the information of sequence complementarity of miRNA to target. However, only few interactions have in fact been experimentally validated. Other approaches integrate in silico target prediction and use miRNA and mRNA expression data to reconstruct posttranscriptional regulatory networks [24, 30]. Such approaches increase the possibility of
discovering true positive miRNA–mRNA interactions reducing the number of false positives, which otherwise is a huge issue when solely using sequence complementarity.

In this chapter we discuss the in silico prediction of miRNA–mRNA interactions using sequence complementarity methods and afterwards focus on the more robust data integration approaches. In the last part of the chapter we discuss how to experimentally design and analyze and integrate miRNA and mRNA data in the complex scenario of disease comorbidities.

2 Materials

The majority of the tools described in this chapter are publicly available online or as R packages. R is a free software environment for statistical computing and graphics (https://www.r-project.org/).

3 Methods

3.1 Sequence Complementarity Based Methods

In the last two decades many computational methods have been developed to identify miRNA–target interactions. In the early years of this research area, algorithms were based on sequence complementarity using target recognition rules retrieved from published studies reporting experimentally validated interactions.
These approaches are still among the most popular algorithms nowadays and very useful for novel approaches that have been developed more recently. In the next subsections we briefly discuss the most popular algorithms used in this field based on sequence complementarity. We have selected methods that use different approaches to give the readers an overview.

### 3.1.1 TargetScan

One of the first computational target prediction algorithm was TargetScan, a rule-based method proposed in 2003 [17, 18]. TargetScan predicts biological targets of miRNAs by searching for the presence of canonical sites (8mer, 7mer, and 6mer sites) that match the seed region of each miRNA. “Canonical sites” are regions in the mRNA that contain an exact miRNA seed match. Some canonical sites are more effective at mRNA control than others. One option of the tool is to consider only conserved sites across species.

TargetScan is available for the species *H. sapiens*, *M. musculus*, *C. elegans*, *D. melanogaster*, and *D. rerio*. TargetScanHuman considers matches to human 3′ UTRs and their orthologues, as defined by UCSC whole-genome alignments (http://genome.ucsc.edu). Recently, a new version of TargetScan has been published (version 7.0) [18] that includes an improved quantitative model of canonical targeting using a compendium of experimental datasets. The proposed novel model not only considers the miRNA binding site type but also other 14 features to predict the most effectively targeted mRNAs using multiple linear regression models. The resulting models, one for each binding site type, were collectively called the context++ model. The considered features included characteristics of miRNAs, features of the sites (including their contexts and positions within the mRNAs), and features of the mRNAs. Furthermore, authors showed that TargetScan performed significantly better than existing models and was as informative as the best high-throughput in vivo cross-linking approaches such as HITS-CLIP [31] and CLASH [32]. TargetScan is available online at http://www.targetscan.org/vert_70/ as precompiled predictions or as perl script.

### 3.1.2 miRanda-mirSVR

The miRanda algorithm computes optimal sequence complementarity between a set of mature miRNAs and a given mRNA using a weighted dynamic programming algorithm [19]. Weights are position-dependent and reflect the relative importance of the 5′ and 3′ regions of miRNAs. In addition, miRanda uses an estimate of the free energy of miRNA–mRNA duplex as a filter.

Sequence conservation at and near miRNA binding sites is a strong indication of functional constraints in evolution. miRanda filters out less-conserved predicted target sites using the PhastCons conservation score. This score measures the evolutionary conservation of sequence blocks across multiple vertebrates using a
phylogenetic *Hidden Markov Model* [33]. To improve the prediction, authors proposed another algorithm called mirSVR [20] for scoring and ranking the efficiency of miRanda-predicted miRNA target sites. This approach uses a *Support Vector Regression (SVR)* to train on a wide range of features, including secondary structure accessibility of the site and conservation. mirSVR is also able to identify a significant number of experimentally determined noncanonical and nonconserved sites. MiRanda is available online at http://www.microrna.org/microrna/home.do. It is also possible to use the script of the algorithm that has been written in C.

### 3.1.3 PITA

The Probability of Interaction by Target Accessibility (PITA) algorithm is based on the evidence that target accessibility has a critical role in miRNA–mRNA interactions for a wide range of target types, and PITA is able to accurately capture these effects. It uses initial seeds for each miRNA in 3′ UTRs and then applies a model to each putative site [21]. It computes an energy-based score for miRNA–mRNA interactions, $\Delta \Delta G$, equal to the difference between the free energy gained by the binding of the miRNA to the target, $\Delta G_{\text{duplex}}$, and the free energy lost by unpairing the target-site nucleotides, $\Delta G_{\text{open}}$. Further, the method combines sites for the same miRNA to obtain a total interaction score for the miRNA and UTR.

PITA has been applied to all 3′ UTRs of fly, worm, mouse, and human, resulting in catalogues of target predictions for these organisms that are available online at http://genie.weizmann.ac.il/pubs/mir07/mir07_prediction.html. Through a web interface the user is able to upload both miRNA and mRNA sequences or to browse pre-calculated interactions. Moreover, it is possible to download the script version of the program for large scale predictions.

### 3.1.4 RNA22

RNA22 is a pattern-based approach for discovery of miRNA binding sites and it can be applied to any genome without requiring genome-specific model retraining [22]. It does not use a cross-species sequence conservation filter allowing the discovery of nonconserved miRNA binding sites. Instead the method is based on the *Teiresias* algorithm [34] that is used to discover variable-length motifs (“patterns”) in mature miRNA sequences followed by a *second-order Markov chain* to estimate statistical significance of each pattern. After the pattern discovery step follows a step of “target islands” identification. Target islands refer to any regions of the UTR that is reverse complement of the mature miRNA. The Vienna package [35] is used to predict the structure of each miRNA/island-segment duplex and its Gibbs free energy (“folding energy”). RNA22 recognizes a binding site based on the presence of multiple, distinct, statistically significant patterns that have been discovered by processing known mature miRNA sequences.
RNA22 is available online at https://cm.jefferson.edu/rna22/ and it allows uploading user defined miRNA and mRNA sequences and setting several parameters, e.g., the folding energy cutoff. Moreover, the user is also able to download a program that runs locally allowing to submit batch queries to the RNA22 server. Precomputed predictions are also available for fly, worm, mouse, and human.

3.1.5 RNAhybrid

RNAhybrid predicts multiple potential miRNA binding sites [23]. The program finds the energetically most favorable hybridizations of small RNAs to large RNAs. The tool computes an optimization of miRNA–mRNA duplexes by an accurate statistical analysis of Minimum Free Energies (MFEs). Then, it normalizes MFEs by sequence lengths of miRNA and targets and models the normalized MFEs as distributions. A second program, RNAcalibrate [23], evaluates the predictions for every miRNA and assigns p-values to normalized MFEs. The significance of multiple binding sites in a single target is evaluated with Poisson statistics.

The web page (http://bibiserv.techfak.uni-bielefeld.de/rnahybrid/) allows users to upload miRNA and mRNA sequences. It is also possible to download the RNAhybrid binary package for Windows, src package for almost every Unix like system and binary package for OSX.

3.2 Data Integration Approaches

All available algorithms based on sequence complementarity produce a significant number of false positives. This is mainly caused by the current lack of knowledge of miRNA–mRNA interaction and the miRNA action itself. Integration of target predictions with gene expression profiles improves the detection of true functional interactions. Several novel methods integrate gene expression assuming that miRNAs downregulate target mRNAs, thus, looking for anti-correlation either upregulation of miRNA and downregulation of target or vice versa. In this section we discuss methods that make use of expression correlation, i.e., the strength and direction of a relationship between two random variables such as Pearson correlation (parametric) and Spearman correlation (non-parametric). Other methods use Mutual Information, a measure of mutual dependence of two variables. Mutual information is sensitive to any functional relationship and not just to linear dependencies. Finally, recent methods try to infer miRNA–mRNA causal regulatory relationships starting from observational data (e.g., gene expression profiles). The aim of these methods is to discover causal effects of miRNAs on mRNAs. In fact, it is important to highlight that Correlation or Association are not Causality for an observed event [36]. This key concept has been recently underlined by Altman and Krzywinski [36]. Correlation implies association, but not causation. Conversely, causation implies association, but not correlation.
MiRNA and genes integrated analysis (MAGIA) is a web tool for integrative analysis of target predictions, miRNA and mRNA expression data [24, 25]. The tool allows several kinds of analyses: (a) retrieve and browse miRNA target predictions for drosophila, mouse, rat, and human using target prediction algorithms discussed above such as TargetScan and PITA with the possibility of combining them with Boolean operators; (b) integrate gene expression data and use different functional measures: parametric and nonparametric correlation indexes, a variational Bayesian model, mutual information, and a meta-analysis approach based on P-value combination. The results of the analyses are bipartite regulatory networks of the best miRNA–mRNA interactions. Moreover, in the second version of the tool [25] authors have introduced integration of transcription factor (TF) binding sites giving the possibility to discover new regulatory circuits of miRNA–mRNA–TF.

MAGIA allows the user to follow several steps in the analysis in a clear way. The aim of the tool is to refine target predictions using gene expression data. First, the user can select the organism of interest and the gene or transcript annotation (EntrezGene, RefSeq or ENSEMBL annotations). Second, the user can select target prediction algorithms, either single or multiple algorithms, to find intersection of predictions. Finally, the user can upload normalized expression profiles for mRNA and miRNA.

A powerful aspect of MAGIA is that it takes into account whether data has been collected from experiments by measuring expression profiles of miRNAs and targets in exactly the same biological samples. The tool allows use of a meta-analysis approach based on a P-value combination in the case of different biological samples or Spearman and Pearson correlation, mutual information, and a variational Bayesian model if expression profiles come from the same biological samples.

MAGIA is available online at http://gencomp.bio.unipd.it/magia2/start/.

mirConnX combines sequence information with gene expression data to create a disease-specific, genome-wide regulatory network [26]. The aim of this tool is to provide a unique environment for transcriptional and posttranscriptional regulation considering miRNA–mRNA–TF regulatory networks integrating mRNA and miRNA expression data measured under the same set of conditions. Expression profiles are preprocessed to remove lowly expressed miRNAs and mRNAs followed by the construction of an association network using statistical approaches including Pearson, Spearman and Kendall correlations. This undirected network is then compared to a prior pre-compiled, species-specific static network, which is derived from TF–gene binding, miRNA–mRNA prediction and literature evidence. The result is a directed, weighted graph. An integration function allows obtaining a final directed network containing robust interactions.
mirConnX web tool allows the visualization and exploration of the network and identifies network motifs. The tool is available at http://www.benoslab.pitt.edu/mirconnx/.

### 3.2.3 **GenMiR++**

The Generative model for miRNA regulation (GenMiR++) [27] is a Bayesian model and learning algorithm that includes gene expression data and a set of candidate miRNA targets obtained from target prediction algorithms. The model assumes that downregulation of target mRNA is related to the action of multiple miRNAs. High expression of one or many miRNAs causes a modification in the expression of mRNA that is negatively shifted with respect to an estimated background expression level. For any transcript, regulating miRNAs will be selected using a set of unobserved binary indicator variables. Given the expression data and the predicted interactions, the problem of finding functional miRNA targets consists of inferring which indicator variables are turned on and which are turned off.

GenMiR++ is available as MATLAB code at http://www.psi.toronto.edu/genmir/.

### 3.2.4 **miRNet**

Recently, a new web-based tool called miRNet has been published [28]. It offers statistical, visual, and network-based approaches to study miRNA regulatory networks. miRNet includes a comprehensive knowledge base integrating miRNA–target interaction data from 11 databases: miRTarBase [37], TarBase [38], miRecords [39], SM2miR [40], Pharmaco-miR [41], miR2Disease [42], PhenomiR [43], StarBase [44], EpimiR [45], miRDB [46], and miRanda [19, 20]. Moreover, it allows differential expression analysis of data from microarray, RNA sequencing, quantitative PCR as well as is a powerful network visualization system and provides enrichment analysis. miRNet currently supports eight species: *S. mansoni*, *C. elegans*, *D. rerio*, *D. melanogaster*, *M. musculus*, *R. norvegicus*, *B. taurus*, and *H. sapiens*.

To facilitate the analysis, authors have implemented an interactive flowchart to allow users to choose the appropriate setup based on input data. The input can be a list of miRNA, mRNA, long ncRNA IDs or a data table containing expression values. It is also possible to select miRNA, disease, or gene names from a list of available database entries. miRNet supports differential expression analysis using limma [47], edgeR [48] and HTqPCR methods. For functional enrichment analysis miRNet use standard hypergeometric test and unbiased random sampling.

miRNet is available online at http://www.mirnet.ca/.

### 3.2.5 **miRLab**

miRLAB is an R package to discover miRNA–mRNA regulatory relationships using several tools including popular target prediction and ensemble methods [29]. miRLAB contains both miRNA and mRNA expression datasets, a pipeline to retrieve datasets from
The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov/), and computes differential expression. Users can also provide their own datasets and compare the prediction results with those predicted by the embedded benchmark methods.

The authors have re-implemented several computational methods including correlation, regression, and causal inference approaches. The correlation methods included are: Pearson, Spearman, Kendall, distance correlation, Hoeffding’s D measure, and randomized dependence coefficient. In this context miRNA–mRNA pairs are ranked based on the correlation coefficient where negative correlations are in the top ranking. The authors also included Mutual Information for nonlinear relationship discovery. Other commonly used methods are based on regression models of which miRLAB includes the high-dimensional techniques Lasso [49] and Elastic-net [50], which both can be used to infer associations between variables. Another interesting method implemented in miRLAB is a causal inference algorithm called Intervention calculus when the Directed Acyclic Graph is Absent (IDA) [51, 52], which estimates the causal effect that a variable has on others. miRNAs are considered causes while mRNAs are considered effects. IDA follows two main steps consisting of learning a causal structure from observational data, then inferring the causal effects and scoring them. miRLAB is one of the few tools that integrate this novel approach, but it is also computationally time consuming. Future validation is needed in order to understand whether causal based methods perform better than standard correlation in the context of miRNA–mRNA prediction, but it is considered a growing and promising field [53].

miRLAB further provides an option to incorporate the target information from different sequence based prediction algorithms (e.g., TargetScan) obtaining strong miRNA–mRNA relationships based on expression data and physical interactions.


3.3 miRNAs and Comorbidities

Given the increased evidence of the role of miRNAs in diseases and biological processes, it is likely that miRNAs are involved in driving the co-occurrence of diseases (i.e., comorbidities) [12–14]. miRNAs are expressed across tissues and can affect molecular players in several pathways, and thus, miRNA deregulation might cause co-occurrence of two or more diseases. Therefore, including miRNA regulation in comorbidity analysis can uncover new knowledge about disease associations. In this context many variables such as environmental and genetic factors also play a role. These are important factors to take into consideration in the experimental design of gene expression measurements and later downstream analyses. In fact, each factor could change miRNA and mRNA expressions (Fig. 1) resulting in misinterpretation of biological results.
Recently, several studies have explored disease association in a data-driven manner where the association between all diseases are calculated based on electronical health records (EHR) or registry data [16, 54, 55]. These studies have demonstrated that genetic factors can be derived from EHR data analysis [55] and that significantly associated diseases share genetic etiology than random pairs of diseases [54, 56]. A crucial factor to be considered in such data-driven studies of disease associations is the temporal order that the diagnoses develops in. Recently, a data-driven approach of temporal disease progression has been proposed using EHR data covering the whole population of Denmark, considering 14.9 years of registry data on 6.2 million patients [16]. This approach mapped how patients progress through different diseases. Such maps can be utilized to identify groups of patients with different disease development that might be explained by genetic factors such as miRNAs. These types of approaches open up for new opportunities in the discovery of molecular biomarkers. If coupled with expression data, they can serve as a powerful tool for uncovering novel association between miRNA and disease. miRNAs have been proposed to be novel candidate biomarkers for disease progression for cancer and other pathologies [57, 58]. Moreover, it has been shown that miRNAs circulate in the human bloodstream complexed in vesicles such as exosomes, microvesicles, high-density lipoproteins (HDLs), and Ago2 protein [59–62]. Since circulating miRNAs are stable in body fluids, they can serve as fast and noninvasive novel candidate biomarkers for the early detection and progression of diseases. With these recent discoveries, it is possible to design proper experiments and computational analyses to better integrate miRNA and mRNA expressions in the context of disease comorbidities (Fig. 2). Considering a large cohort of patients with different risk factors for a specific primary disease, an ideal experimental design would consist of follow-up of patients where blood samples are taken at each time point. The blood samples can be used to extract RNAs and then quantify miRNA and mRNA expressions using Next Generation Sequencing technologies or microarrays. Following patients from the time of the first diagnosis of the primary disease through months or years can allow us to monitor the onset of secondary diseases that show strong correlation with the primary disease. At each time point, miRNAs and mRNAs are likely to have fluctuating expression levels due to risk factors, the presence of secondary disease or because the age is an important parameter to take into account (Fig. 1). The data integration methods described above should be considered in a dynamic way where crucial changes in miRNA–mRNA interactions are observed. This approach will elucidate the role of miRNAs and their targets in disease progression and comorbidities, allowing the discovery of novel molecular players.
In this chapter we discuss computational target prediction approaches for miRNA–mRNA interactions. Recently, it has been shown that miRNAs can bind other molecules such as long
noncoding RNAs, circular RNAs and pseudogenes [63]. By attenuating shared miRNAs the different kinds of RNAs could cross talk and regulate each other. These RNAs are known as competing endogenous RNAs (ceRNAs) [63]. Recently, a new generation of miRNA target prediction analysis has been proposed [64, 65]. These algorithms do not only include complementary sequence comparisons and gene expression, but also consider other novel variables in the context of the ceRNA cross talk. In this complex scenario the miRNA inhibition depends on several aspects: (1) number of miRNA binding sites, (2) miRNA binding affinity, (3) unbound miRNA expression level, and (4) target expression level [64, 65]. Future methods will integrate more data such as proteins, methylation, and copy number variation, increasing our understanding of miRNA regulation.

Acknowledgments

Francesco Russo has been supported by a fellowship sponsored by Progetto Istituto Toscano Tumori-Grant 2012 Prot. A00GRT. Novo Nordisk Foundation (grant agreement NNF14CC0001).

References

7. Lytle JR, Yario TA, Steitz JA (2007) Target miRNAs are repressed as efficiently by microRNA-binding sites in the 5′ UTR as in the 3′ UTR. Proc Natl Acad Sci U S A 104:9667–9672


