IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 30, 2020, accepted August 14, 2020, date of publication August 20, 2020, date of current version September 1, 2020.

*Digital Object Identifier 10.1109/ACCESS.2020.3018105*

# Identification of MicroRNA Regulatory Modules by Clustering MicroRNA-Target Interactions

**YI YANG AND XUTING WAN**
College of Information Science and Engineering, Hunan Women's University, Changsha 410004, China
Corresponding author: Yi Yang (snryou@126.com)

This work was supported in part by the Double First-Class Discipline Construction Plan of Hunan Province, China, under Grant XiangJiaoTong (2018) 469, and in part by the Scientific Research Project of the Education Department of Hunan Province, China, under Grant 18A470.

**ABSTRACT** Identification of microRNA regulatory modules can help decipher microRNA synergistic regulatory mechanism in the development and progression of complex diseases, especially cancers. Experimentally validated microRNA-target interactions provide strong direct evidence for the analysis of microRNA regulatory functions. We here developed a novel computational framework named CMIN to identify microRNA regulatory modules by performing link clustering on such experimentally verified microRNA-target interactions. CMIN runs in two main steps: it first utilizes convolutional autoencoders to extract high-level microRNA-target interaction features from the expression profile data, and then applied affinity propagation clustering algorithm to interaction feature to obtain overlapping microRNA-target clusters. Clusters with significant synergy correlations are considered as microRNA regulatory modules. We tested the proposed framework and other three existing methods on three types of cancer data sets from TCGA (The Cancer Genome Atlas). The results showed that the microRNA regulatory modules detected by CMIN exhibit stronger topological correlation and more functional enrichment compared with other methods. Availability: The supplementary files of CMIN are available at https://github.com/snryou/CMIN.

**INDEX TERMS** MicroRNA regulatory module, MicroRNA-target interaction, convolutional autoencoder, affinity propagation, link clustering.

## I. INTRODUCTION

MicroRNAs (miRNAs) are a class of non-coding RNAs that bind to target messenger RNAs (mRNAs) to induce mRNA degradation or translational repression [1], [2]. Aberrantly expressed miRNAs are implicated in a variety of malignancies and function as either oncogenes or tumor suppressors [3], [4]. A large number of studies have monitored cancer progression by measuring the expression of miRNA [5]–[7]. This monitoring is typically based on the expression of individual miRNAs, but miRNAs are more likely to coordinate together to perform their functions [8], [9]. The identification of miRNA regulatory modules (MRMs) can help to decipher the synergy of miRNAs and provide a reasonable explanation for the combination of miRNAs [10]–[12].

In the past decade, many miRNA regulatory module identification algorithms have been developed. Based on the network elements used for identifying modules, the existing methods can be classified into two categories: node-based and structure-based [13]. The former finds a partition of network nodes to assign each node to one and only one module, whereas the latter assigns each specific substructure to one and only one module.

Specifically, node-based methods typically divide the network into several modules according to the topological characteristics and biological significance. For example, Jayaswal *et al.* [14] clustered miRNAs/mRNAs based on their microarray expression data and associated two types of clusters according to changes in miRNA/mRNA expression profiles to obtain the final miRNA regulatory modules. Li *et al.* proposed an approach called Mirsynergy [15] that first obtained miRNA clusters based on the miRNA-miRNA synergy, then added/removed genes to/from each miRNA cluster to form miRNA regulatory modules. Karim *et al.* identified MRMs by clustering miRNAs and mRNAs based on functional interaction similarities with common mRNAs (or miRNAs) [16]. These node-based methods demonstrate good performance under certain conditions. However, they

The associate editor coordinating the review of this manuscript and approving it for publication was Juan Wang.

are difficult to identify overlapping modules whose miRNAs or mRNAs may exist in multiple modules. Moreover, their results are greatly affected by some thresholds.

Structure-based methods usually treat modules as substructures in the network, so these methods usually mine some specific substructures from the network according to topological characteristics, and then expand or trim the substructures in combination with related biological characteristics. For instance, Derényi *et al.* defined a complete connected sub-graph of $k$ nodes as a $k$-clique and proposed an algorithm called Clique Percolation Method (CPM) for discovering modules [17]. Liang *et al.* developed a method BCM to identify modules by merging maximal bicliques [18]. Kalinka released the R language package linkcomm based on links to identify modules [19]. A node in a clique or link can belong to multiple modules. Hence, structure-based methods can naturally detect the overlapping modules. However, strictly defined substructures do not occur frequently in sparse miRNA regulatory networks. Therefore, it is difficult to expand these substructures to obtain moderate-size modules [20].

In this work, we develop a new computational framework CMIN (Clustering MiRNA-target INteractions) to identify miRNA regulatory modules from experimentally validated microRNA-target interactions. Unlike the existing methods, the present method does not directly apply the clustering algorithm to the miRNA-mRNA network, but first converts each miRNA-mRNA pair into a node, and then performs clustering on this basis to obtain overlapping miRNA regulatory modules. We apply the proposed method to bladder cancer (BLCA), breast cancer (BRCA), and liver hepatocellular cancer (LIHC) data sets from TCGA (The Cancer Genome Atlas). Compared with node-based Mirsynergy, structure-based BCM, and link-based linkcomm, the miRNA regulatory modules identified by CMIN have stronger internal correlation and more functional enrichment.

## II. METHODS

### A. WORKFLOW OVERVIEW

Figure 1 depicts the workflow of CMIN, which consists of two stages. In the first stage, we first apply convolutional autoencoder (CAE) to miRNA expression profile data to refine its high-level expression features, and the mRNA expression profile data does the same process as well. Then we match the miRNA-mRNA pairs contained in the miRNA-target interactions data [21] on the expression features, and merge them by Cartesian product [22] to construct a miRNA-mRNA interaction feature matrix. In the second stage, we employ Affinity Propagation (AP) clustering to the interaction matrix to automatically derive the miRNA-mRNA clusters, in which the default similarity measurement is negative Euclidean distance. Finally, the clusters are further filtered through synergy correlations, and the remaining ones are considered as miRNA regulatory modules.

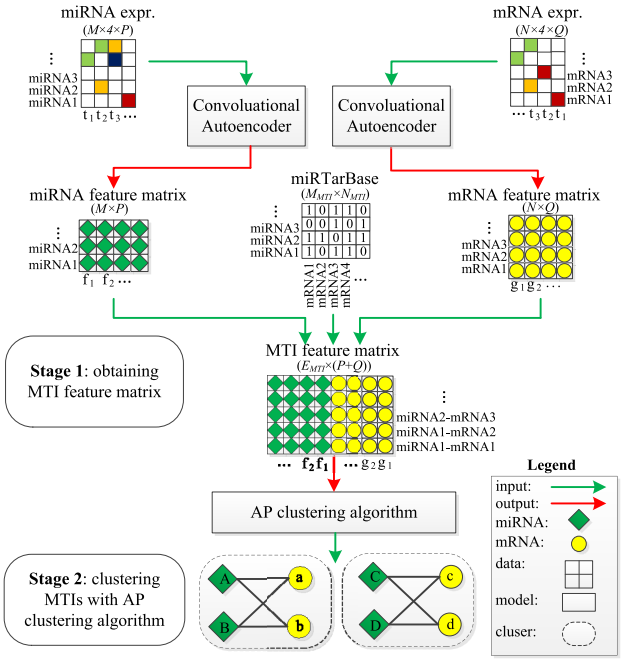


**FIGURE 1.** Illustration of CMIN workflow.

### B. FORMATION OF MTI MATRIX

Let $X$, $Y$ be miRNA and mRNA expression profiles, respectively. According to the properties of Cartesian product, $Z = X \times Y$ represents the relationship between miRNAs and mRNAs. Since miRNA and mRNA expression profiles are high-dimensional data, if the expression data of the two are directly combined to construct the miRNA-mRNA pair expression data, it is bound to obtain a higher dimensional data, which is not conducive to calculation and affects the accuracy of the results. Therefore, we chose to use convolutional autoencoders to reduce the dimensionality of miRNA and mRNA expression profiles before merging.

CAE has been widely applied to reduce data dimensions for higher computational performance in the recent years. For example, Serengil [23] applied CAE and K-means to classify the unlabeled pixel images. In bioinformatics, CAE has also been successfully applied to a variety of biomedical tasks such as risk prediction of tumors [24], miRNA-disease relationship detection [25], and finger-vein verification [26]. In this study, we proposed a convolutional autoencoder model to extract high-level features of miRNA/mRNA expression profiles.

In our work, the CAE network was treated as an data transformation with an encoder function $y = t(x, \alpha)$ and a decoder function $\hat{x} = t'(y, \beta)$, where $x$, $\hat{x}$, and y are the original interactions, reconstructed interactions, and the compressed representation of interactions, respectively. Here, $\alpha$ and $\beta$ are the parameter sets that need to be optimized in the functions. Since CAE networks have a small number of layers and neurons, their training and calculation speed are very fast. On the contrary, the networks with more layers and neurons may have higher reconstruction capacity and accuracy, but

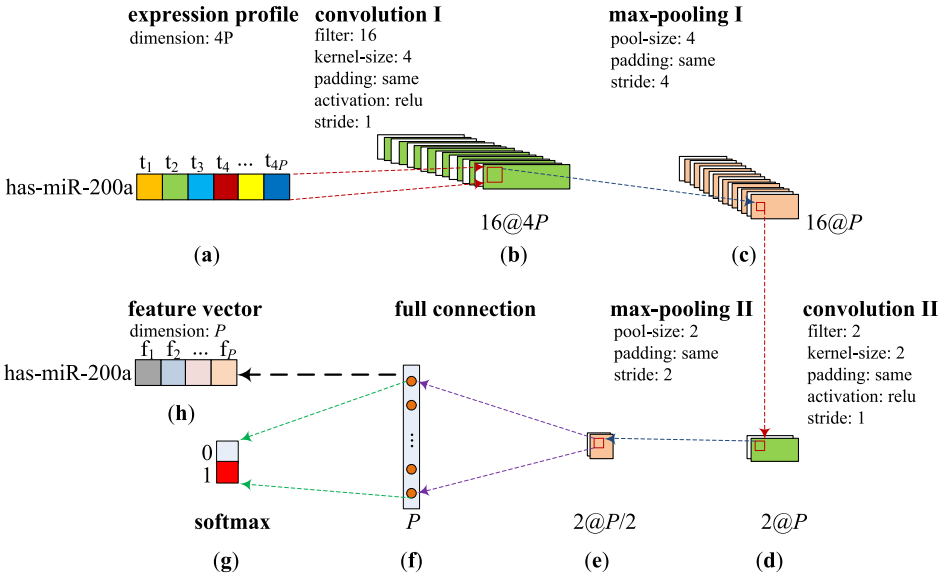


**FIGURE 2.** The structure of our proposed convolutional autoencoder. The model consists of two convolutional layers, two max-pooling layers, and a fully connected layer.

it is easy to cause overfitting. After trying CAE networks with different layers and neurons, we designed the current CAE network by balancing computation performance and reconstruction accuracy. The designed CAE was optimized to have high representation accuracy and fast computation speed.

Fig. 2 depicts the architecture of the designed CAE network that contains two convolution layers (*kernel-size* = *4* or *2*), two max-pooling layers (*pool-size* = *4* or *2*), one fully connected layer, and one softmax layer. Taking miRNA hsa-miR-200a as an example, we introduced the process of feature in detail. Here, let $P$ be an integer, we applied a $R^{4P}$ vector to represent the expression of the miRNA on $4P$ samples (Fig. 2a). After the first convolution of the input data, we obtained feature maps of $16@4P$ (16 $size = 4P$ feature matrices) (Fig. 2b). Due to unequal miRNA/mRNA expression on each sample, we took max-pooling operation with a $size = 4$ pool to obtain the most critical features. Once having completed the operations, we obtained the feature maps of $16@P$ (Fig. 2c). Similarly, we implemented the second *kernel-size* = *2* convolution and *pool-size* = *2* max-pooling to obtain the feature maps of $2@P/2$ (Fig. 2d and e). Finally, we obtained the final feature vector with $size = P$ by a fully connected layer (Fig. 2f and 2h). The result can be labeled, and the classification mode was set to sigmoid (Fig. 2g). In this way, the compressed representation only took four times less space than the original expression profile. The compression process of mRNA expression data is similar to the miRNAs.

After obtaining the high-level feature vectors of miRNA/mRNA expression profiles, we employed the Cartesian product of the two feature matrices to construct a miRNA-mRNA interaction feature matrix. However,

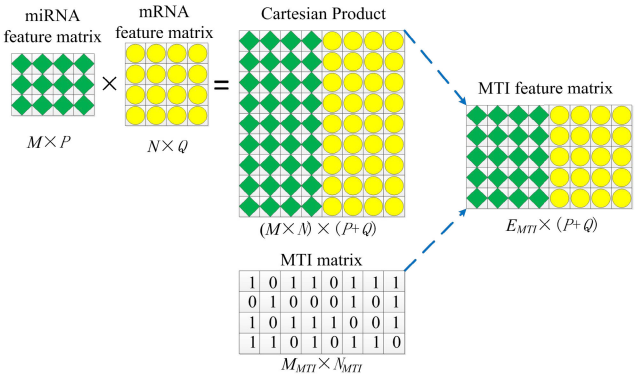


**FIGURE 3.** The formation of MTI feature matrix. Here, $M/N$ represents the number of miRNAs/mRNAs in the expression profiles. $P/Q$ is the length of the feature vectors of miRNAs/mRNAs. $M_{mti}/N_{mti}$ is the number of miRNA/mRNA in MTIs. $E_{mti}$ is the number of MTIs. Because of the sparsity of miRNA regulatory network, $E_{mt} \ll M_{mti} \times N_{mt} \ll M \times N$. Therefore, the dimension of MTI feature matrix is much less than the dimension of the Cartesian product.

as shown in Fig. 3, let $M/N$ be the number of miRNAs/mRNAs, $P/Q$ be the length of the feature vectors, the dimension of the Cartesian product was a great number $(M \times N) \times (P + Q)$. For instance, on BLCA, there are 432/17292 miRNAs/mRNAs and 212/212 high-level features of them. The dimension of their Cartesian product is $(432 \times 17292) \times (212+212)$. If we directly cluster MTIs on the Cartesian product, a large number of false-positive data from the product will lead to low clustering efficiency and unreliable results.

For the above reasons, we utilized the experimentally validated miRNA-target interactions to obtain a subset of the Cartesian product. By this way, the dimension of MTI feature matrix was reduced to $E_{mti} \times (P + Q)$. Here $E_{mti}$ is the number of MTIs. For example, there are 107 MTIs on BLCA.

Accordingly, the dimension of MTI feature matrix is reduced to $107\times(212+212)$, which is about $4\times17292$ times less space than the original Cartesian product. Therefore, the expense of computation of subsequent clustering algorithm is greatly reduced.

### C. CLUSTERS DETECTION

AP clustering is an exemplar-based clustering algorithm [27]. Compared with other clustering methods such as K-means, it can process large-scale data faster and obtain more accurate clustering results. Another advantage of AP is that it is applicable to any given meaningful similarity measure. In addition, AP does not require a predetermined number of clusters. These advantages make AP suitable for the identification of miRNA regulatory modules.

In this work, we considered each miRNA-mRNA interaction as a node. Given the miRNA-mRNA interaction feature matrix, we employed Euclidean distance to evaluate the similarity between the interactions. Let $S = \{s(i, k)|i, k \in MTIs\}$ be interaction similarity matrix. The closer the $i$-th and $k$-th interaction are, the greater the value of $s(i, k)$ is. The responsibility $r(i, k)$, sent from data point $i$ to candidate exemplar $k$, reflects the accumulated evidence that the $k$-th interaction is to be the exemplar for the $i$-th interaction. The availability $a(i, k)$, sent from candidate exemplar $k$ to data point $i$, denotes the accumulated confidence that how suitable it is for the $i$-th interaction to choose the $k$-th interaction as its exemplar (see Fig. 4).

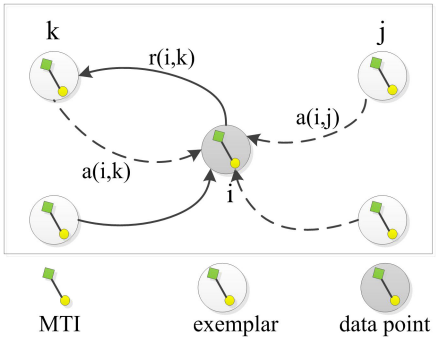


**FIGURE 4.** The process of transmitting messages.

In the clustering process, AP firstly takes the similarities $S$ as its input and all interactions as candidate exemplars. Then starting with $a(i, k) = 0$, AP evaluates and updates $r(i, k)$ and $a(i, k)$ based on the equation 1 and 2, respectively. As the iteration progresses, AP continuously removes candidate exemplars. The iteration process is terminated until the "final exemplars" are determined and other interactions are assigned to the corresponding exemplars.

$$r(i,k) = \begin{cases} s(i,k) - \max\limits_{j\neq k}\{a(i,j) + s(i,j)\}, & (i \neq k) \\ s(i,k) - \max\limits_{j\neq k}\{s(i,j)\}, & (i = k). \end{cases} \quad (1)$$

$$a(i,k) = \begin{cases} \min\{0, r(k,k) + \sum\limits_{j\neq i,k} \max\{0, r(j,k)\}\}, & (i \neq k) \\ \sum\limits_{j\neq k} \max\{0, r(j,k)\}\}, & (i = k). \end{cases} \quad (2)$$

AP clustering requires a predetermined optimization parameter "input preference" $p$, which defines the probability of interactions as exemplars. The greater the preference, the greater the probability that interactions become exemplars, and the more final clusters. Limited by the number of experimentally validated miRNA-target interactions, a larger preference will make interactions difficult to form clusters with biological meanings. On the contrary, AP with a smaller "input preference" can collect more interactions to form clusters. Here, we set $p$ to the default value of $-93$ to ensure that more interactions are being clustered together.

After obtaining the miRNA-mRNA clusters from the AP algorithm, we found a few clusters cover star structures such as one miRNA or one mRNA, which was inconsistent with the mechanism by which miRNAs synergistically regulate target mRNAs, so it should be discarded. As illustrated in Fig. 5, the left interaction cluster as the miRNA regulatory module is retained.

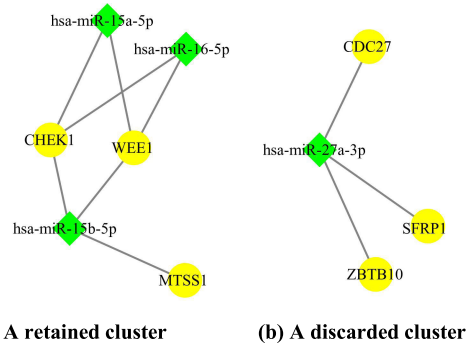


**FIGURE 5.** Two examples of interaction clusters. Based on the synergic mechanism of miRNA, we only keep clusters containing two or more miRNAs/mRNAs. Therefore, the interaction cluster on the right is discarded.

## III. EXPERIMENT AND RESULTS

We tested the four methods on three identical real cancer datasets. For CMIN, we set the parameters of CAE model and AP clustering algorithm according to the description in Section II throughout the test. For the other three counterparts, we adopted the default settings suggested by the authors. The major metrics for evaluating the performance of different algorithms are topological performance and biological significance. The former includes the number of modules, size, density, and node coverage, while the latter refers to the enrichment of miRNA family and target gene function (i.e. GO terms and KEGG pathways). Moreover, we validated the prognostic power of CMIN-MRMs.

**TABLE 1.** The collected data for identifying MRMs.

| Cancer | #miRNA | #mRNA | #Sample | #MTI | #GGI |
|---|---|---|---|---|---|
| BLCA | 432 | 17,292 | 424 | 107 | 680 |
| BRCA | 368 | 16,832 | 848 | 428 | 5351 |
| LIHC | 432 | 15,064 | 424 | 49 | 200 |

### A. DATA COLLECTION AND PREPROCESSING

As presented in Table 1, we collected three kinds of cancer data for the identification of miRNA regulatory modules.

1) MiRNA/mRNA expression profiles. We downloaded miRNAs and their target gene expression profiles from Firebrowse (http://firebrowse.org/), which provide purified TCGA data. For expression profile data, we first filtered out miRNAs and genes with missing values exceeding 10% of the sample size, and then performed log2 transform on the filtered data. As a result, we retained 432, 368, and 432 miRNAs for BLCA, BRCA, and LIHC, along with 424, 848, and 424 samples for each miRNA, respectively. Similarly, we obtained 17,292, 16,832, and 15,064 genes for BLCA, BRCA, and LIHC, along with 424, 848, and 424 samples for each gene, respectively.

2) Experimentally validated microRNA-target interactions. We downloaded the experimentally validated microRNA-target interactions from the miRTarBase website (http://mirtarbase.cuhk.edu.cn/cache/download/8.0/hsa_MTI.xlsx). There are totally 502,652 microRNA-target interactions in the downloaded file. Here, we only retained 107, 428, and 49 microRNA-target interactions related to BLCA, BRCA, and LIHC, respectively.

3) Gene-gene interactions. Since gene-gene interactions (GGI) are required for the Mirsynergy and BCM clustering methods, we collected the cancer-related gene-gene interactions from the STRING website (https://string-db.org/).

In addition, to evaluate the prognostic value of miRNA regulatory modules, we collected the clinical data of 408 bladder cancer patients and 566 breast cancer patients from the TCGA website.

### B. TOPOLOGY ANALYSIS OF MRMs

The results produced by the four methods are presented in Table 2. CMIN identified 4/8/1 modules on the BLCA/BRCA/LIHC datasets. Although less than the number of modules identified by Mirsynergy (9/21/1), the proposed method discovered more modules than BCM (1/2/0) and linkcomm (7/5/0).

For BLCA and BRCA, the average number of miRNA/mRNA per CMIN-module is 6.3/8.0 (8.0/20.3), respectively, which is significantly more than 2.0/2.0 (3.0/3.0) of BCM and 2.7/3.1 (2.0/2.6) of linkcomm. Compared with Mirsynergy, the average mRNA number per module is roughly comparable to 7.8/21.3 of Mirsynergy, but the average microRNA number per module is more than 3.7/4.1 of Mirsynergy.

**TABLE 2.** Performance of CMIN, Mirsynergy, BCM, and linkcomm.

| Cancer | Method | #MRM | #miRNA | #mRNA | #MF | Density |
|---|---|---|---|---|---|---|
| BLCA | CMIN | 4 | 6.3 | 8.0 | 1 | 0.650 |
| | Mirsynergy | 9 | 3.7 | 7.8 | 0 | 0.127 |
| | BCM | 1 | 2.0 | 2.0 | 0 | 1.000 |
| | linkcomm | 7 | 2.7 | 3.1 | 0 | 0.708 |
| BRCA | CMIN | 8 | 8.0 | 20.3 | 2 | 0.494 |
| | Mirsynergy | 21 | 4.1 | 21.3 | 0 | 0.088 |
| | BCM | 2 | 3.0 | 3.0 | 0 | 0.917 |
| | linkcomm | 5 | 2.0 | 2.6 | 0 | 0.850 |
| LIHC | CMIN | 1 | 2 | 1 | 0 | 0.667 |
| | Mirsynergy | 1 | 2 | 1 | 0 | 0.667 |
| | BCM | 0 | 0 | 0 | 0 | 0 |
| | linkcomm | 0 | 0 | 0 | 0 | 0 |

**Note: #MRM: module number; #miRNA and #mRNA: average miRNA and mRNA per module; #MF: number of enriched miRNA family; Density: average density of per module.**

For LIHC, only one miRNA regulatory module was identified by CMIN and Mirsynergy, respectively, which may be attributed to the sparseness of the data set. However, BCM and linkcomm failed to detect any module. This indicates that CMIN and Mirsynergy are more suitable for sparse network than the other two methods. Since there is only one module, there is no need to compare the performance of these four methods on the LIHC dataset.

Module density reflects the correlation between nodes within the module. As shown in Fig. 6, except for Mirsynergy, most of the modules detected by the other three methods have high internal correlations. The average density of CMIN-MRMs on BLCA and BRCA data sets are 0.650 and 0.494, respectively (Table 2). Although the average density of CMIN-MRMs is less than that of BCM-MRMs (1.000/0.917) and linkcomm-MRMs (0.708/0.850), the densities are significantly higher than Mirsynergy-MRMs (0.127/0.088). BCM-MRMs have the greatest density among three methods, but it detected only 1/2 modules on the two data sets, respectively. This implies that BCM is not suitable for the BLCA and BRCA data sets. Therefore, in the subsequent sections, we would not compare BCM with the other three methods anymore.

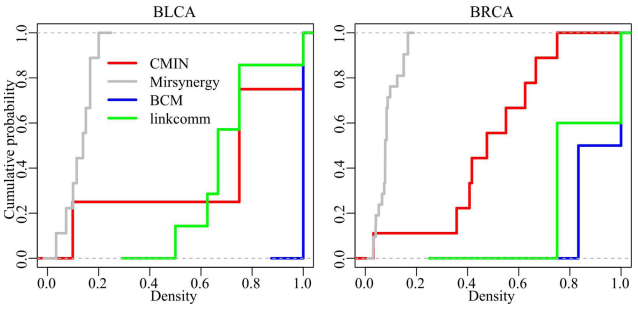


**FIGURE 6.** The cumulative distribution curve of the density of modules.

**TABLE 3.** Overlap of miRNAs/mRNAs.

| Cancer | Method | Community Coverage | Overlap Coverage |
|---|---|---|---|
| BLCA | CMIN | 0.445 | 0.468 |
| | Mirsynergy | 0.695 | 0.805 |
| | linkcomm | 0.328 | 0.430 |
| BRCA | CMIN | 0.454 | 0.497 |
| | Mirsynergy | 0.594 | 1.066 |
| | linkcomm | 0.048 | 0.066 |

Studies have shown that a number of miRNAs/mRNAs may simultaneously belong to multiple modules, which play different biological roles in different contexts [28] as a part of overlapping miRNA regulatory modules. To evaluate the overlapping performance of the identified modules, we introduced two coverage measures proposed by Ahn *et al.* [29]: community coverage and overlap coverage. The community coverage describes how much of the network is divided by each method, and the overlap coverage calculates how much overlap is identified. They can be defined as

$$\text{Comm}Coverage = \frac{CNode}{NNode}. \tag{3}$$

and

$$\text{Over}Coverage = \frac{ONode}{NNode}. \tag{4}$$

Here, *NNode* denotes the total number of nodes in the network, *CNode* is the number of nodes assigned from the network to MRMs, and *ONode* is the total number of nodes of MRMs.

As shown in Table 3, the coverage values of CMIN are 0.445/0.468 and 0.454/0.497 for BLCA and BRCA, respectively, which are much less volatile than linkcomm (0.328/0.430, 0.048/0.066). It means that the overlapping performance of linkcomm-MRMs depends on the structure of the tested network, whereas CMIN is independent of that. Among the three methods, Mirsynergy tried to assign each node to a certain cluster by maximizing the synergy value between nodes, and obtained the largest coverage value (0.695/0.805 and 0.594/1.066). However, the high coverage of Mirsynergy came at the expense of internal correlation of modules. Among the three methods, Mirsynergy modules are the sparsest.

From the above analysis of number, size, density, and node coverage of the identified modules, we can observe that although CMIN is not the winner in most individual aspects of the performance of structure, it is the overall leader.

### C. SYNERGISTIC ANALYSIS FOR miRNA REGULATORY MODULES

Studies have shown that microRNAs from the same family are more probably to coordinate one or several common target genes and perform certain functions [30]. Hence, we employed microRNA family enrichment analysis to evaluate the synergy of microRNAs in microRNA regulatory modules. We downloaded the microRNA hairpin sequence family classification file from miRBase (http://www.mirbase.org/) and applied a hyper-geometric test to verify whether the detected modules are significantly enriched in the miRNA family. The hyper-geometric test is shown in equation 5.

$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}. \tag{5}$$

where $N$ is the number of background miRNAs, $M$ is the number of miRNAs in a family, $n$ is the number of miRNAs in an identified module and $k$ is the number of overlaps between the identified module and the miRNA family.

By hyper-geometric testing ($q$-value $< 0.05$), we counted the number of modules enriched in one or more microRNA families. For the BLCA and BRCA datasets, one and two modules detected by CMIN are enriched in microRNA families, respectively, but none of the modules identified by Mirsynergy and linkcomm (Table 2). For instance, two miRNAs from CMIN-MRM 4 on the BRCA, hsa-miR-26a-5p and hsa-miR-26b-5p, are enriched in the MIPF0000043 family (Table 4). These two microRNAs have been confirmed by many literatures that they co-regulate target genes and play an important role in the production, formation, and development of cancers [31]–[33].

**TABLE 4.** miRNA family enrichment for CMIN-MRMs in the BRCA dataset.

| No. | miRNAs | miRNA Family | $q$-value |
|---|---|---|---|
| 4 | hsa-miR-26a-5p<br>hsa-miR-26b-5p | MIPF0000043 | 0.03028 |
| 5 | hsa-miR-27a-3p<br>hsa-miR-27b-3p | MIPF0000036 | 0.04410 |

### D. FUNCTIONAL ENRICHMENT ANALYSIS OF GENES

For comparing the biological significance of microRNA regulatory modules detected by CMIN, Mirsynergy, and linkcomm, we calculated the functional enrichment in the GO terms and the KEGG pathways. Note that we have not obtained the enrichment of GO terms and KEGG pathways of the linkcomm-MRMs at $p$-value $< 0.05$. We here only compared the functional enrichment of CMIN-MRMs and Mirsynergy-MRMs. Among 4/8 CMIN-MRMs on BLCA/BRCA, 4/8 and 2/4 modules are enriched in GO terms and KEGG pathways, respectively. Among 9/21 Mirsynergy-MRMs on BLCA and BRCA, the corresponding numbers are 6/18 and 3/9, respectively. As shown in Fig. 7, we can observe that the proportion of functional enrichment modules detected by CMIN is higher than that of Mirsynergy.

Moreover, we counted the number of enriched GO/KEGG for each module by hyper-geometric testing ($p$-value $<0.005$). As shown in Table 2 and Fig. 8, we found that although the mRNA number per module identified by CMIN and Mirsynergy is approximately the same, the number of

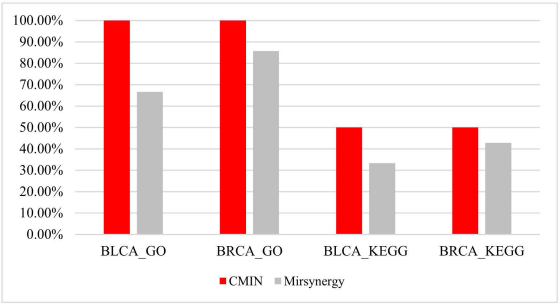


**FIGURE 7.** The comparison of functional enrichment modules between CMIN and Mirsynergy. The percent of the module is defined in the form of dataset item. For instance, BLCA_GO represents the percent of GO term enriched MRM on BLCA.

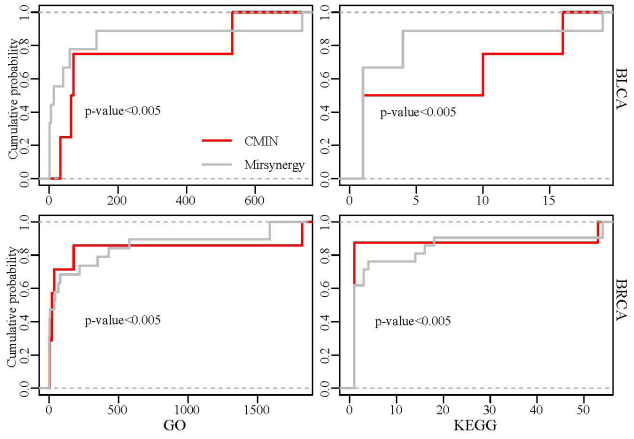


**FIGURE 8.** The cumulative distribution curve of the number GO term and KEGG pathway of modules. The abscissa is the number of GO/KEGG entries enriched in each module.

meaningful GO terms and KEGG pathways per CMIN-MRM is slightly more than that of Mirsynergy-MRM.

In terms of the miRNA family enrichment and gene functional enrichment analysis, we can conclude that CMIN modules are of higher functional quality than the Mirsynergy modules.

### *E. THE PROGNOSTIC SIGNATURES OF MRMs*

To validate the prognostic signature of the CMIN-modules, we applied the ''risk score model'' developed by Shukla *et al.* [34]. The model follows three steps: (1) The risk score is calculated by estimating the coefficients of risk factors (i.e. microRNA) related with survival time by multivariate Cox proportional hazard regression. (2) The samples are divided into two groups (i.e. low risk and high risk) based on their risk scores below or above the median risk score. (3) The survival characteristics of the two groups are compared by Kaplan-Meier survival analysis. Here, we employed this model to analyze all modules identified by CMIN on the two data sets. The result demonstrated that 3 of 4 BLCA-MRMs and 7 of 8 BRCA-MRMs have significant prognostic signatures.

For instance, Fig. 9 illustrates the prognostic signatures of CMIN-MRM 4 on BLCA data set. This module consists of 2 microRNAs and 4 target genes, whose network topology is shown in Fig. 9(a). As shown in Fig. 9(b), two microRNAs in the module tend to exhibit the same expression patterns under the same conditions. In particular, these two microRNA expression values are significantly higher in low-risk samples than in high-risk samples. The difference in expression is also reflected in the different survival time. As illustrated in Fig. 9(c), samples with low expression are at greater risk of survival ($HR = 0.602$, 95% $CI = 0.407$ to 0.891). In fact, in previous studies, the two microRNAs have been reported to be tumor suppressor microRNAs that often suppress bladder cancer development by inhibiting expression of some oncogenes [35]–[37]. We also analyzed the GO and KEGG enrichment of mRNAs in MRM 4. As shown in Fig. 9(d), we found that ''nitric-oxide synthase biosynthetic process'', ''synaptic vesicle transport'', and ''telencephalon cell migration'' are the three most common GO-BPs enriched in target genes in BLCA-MRM 4. More impressively, this module also involves some carcinogenic pathways, such as ''Pathways in cancer'', ''Cell cycle'', and ''Bladder cancer'' (Fig. 9e). The defects of these pathways are reflected in the pathogenesis of various types of tumors including bladder cancer [38]–[40].

The remaining CMIN-modules have significant prognostic signatures except BLCA-MRM 1 and BRCA-MRM 5. For instance, BRCA-MRM 6 consisting of tumor suppressor hsa-miR-24-3p, hsa-miR-27a-3p, and hsa-miR-27b-3p has similar characteristics as BLCA-MRM 4 ($HR = 0.595$, 95% $CI = 0.345$-0.943). The three microRNAs are low expressed at high-risk and high expressed at low-risk group. While BRCA-MRM 2 containing miR-182-5p and miR-96-5p has different expression patterns from the above modules ($HR = 2.519$, 95% $CI = 1.029$-5.436). The two microRNAs are high expressed at high-risk and low expressed at low-risk. This is because hsa-miR-182-5p and hsa-miR-96-5p are oncogenic miRNAs, which are frequently up-regulated in cancer and inhibit the expression of tumor suppressor genes [41], [42]. Thus, these two types of modules exhibit significantly different prognostic characteristics.

## IV. CONCLUSION

In this article, we propose a novel link clustering method CMIN that provides a new insight for the identification of microRNA regulatory modules. The novelty of our method resides in the two aspects. On the one hand, the presented method benefits from the utilization of convolutional autoencoders to refine the expression profile data, which improves the accuracy of similarity and provides a guarantee for the success of clustering. On the other hand, the proposed novel link clustering framework employs a traditional clustering algorithm to obtain overlapping microRNA regulatory modules, which is consistent with the fact that microRNAs participate in multiple regulatory processes at the same time.

Compared with the existing methods like Mirsynergy, BCM, and linkcomm, the advantage of the proposed framework lies in the ability to automatically determine overlapping modules and the ability to cluster non-neighbor links. Unlike BCM and linkcomm, the performance of the provided

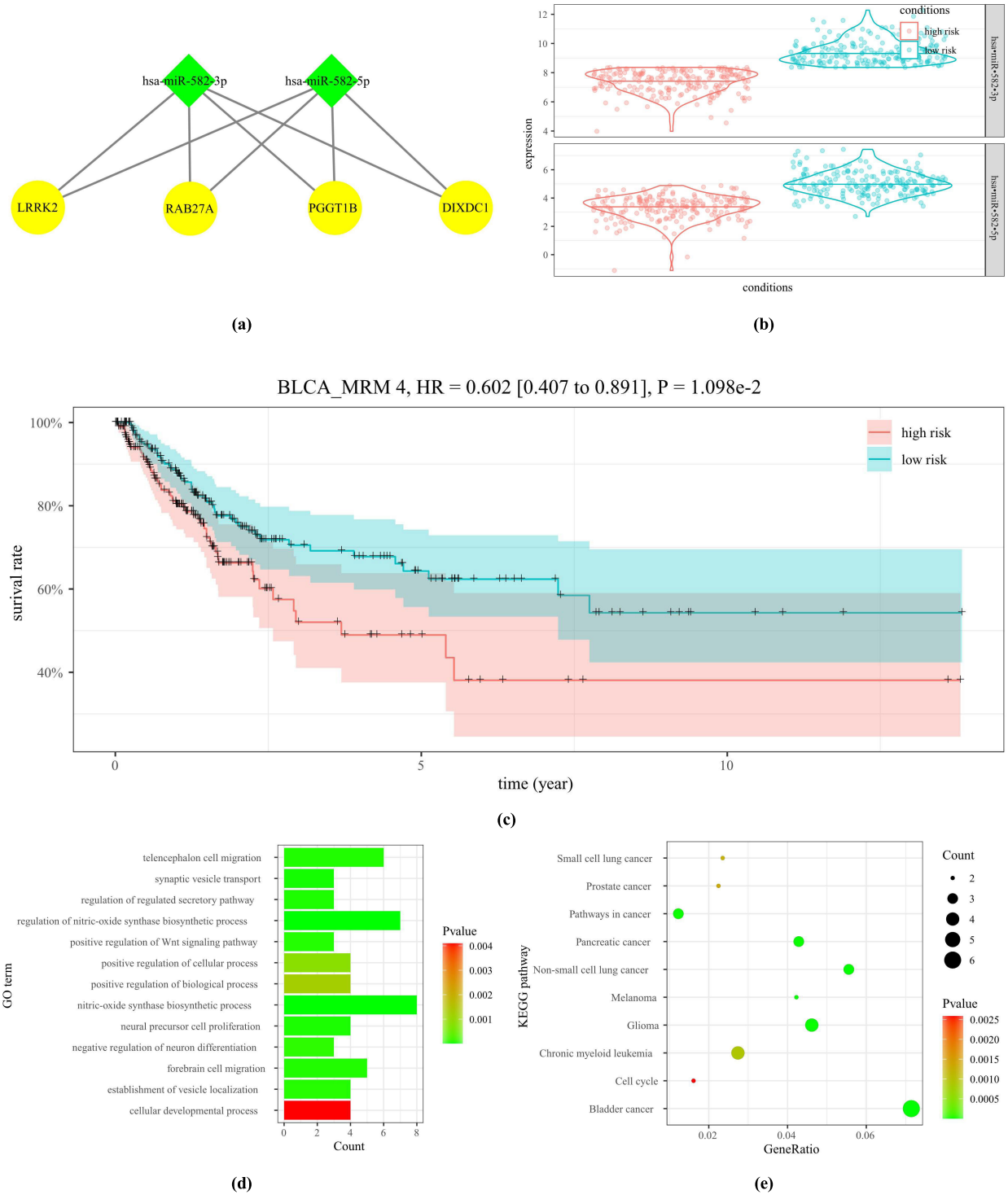


**FIGURE 9.** The survival analysis of BLCA-MRM 4. (a) The topological structure of MRM 4. (b) Differential expression of 2 miRNAs in high-risk and low-risk tumor samples. (c) Kaplan-Meier curves for overall survival. The statistically significant GO-BPs (d) and KEGG pathway (e) enriched by the module target genes.

method is not affected by network sparsity. More importantly, microRNA regulatory modules detected by CMIN exhibit stronger topological correlation and more functional enrichment compared with other methods. With more experimentally validated microRNA-target interactions becoming available, we believe that CMIN has the potential to serve as a powerful tool for identifying miRNA regulatory modules in cancers.

## ACKNOWLEDGMENT

*(Yi Yang and Xuting Wan are co-first authors.)* The authors would like to thank Chu Pan for his kind assistance. They

would also like to thank the anonymous reviewers for their constructive suggestions.

## REFERENCES


[1] L. He and G. J. Hannon, "MicroRNAs: Small RNAs with a big role in gene regulation," *Nature Rev. Genet.*, vol. 5, no. 7, pp. 522–531, Jul. 2004.

[2] A. Gomes, I. V. da Silva, C. M. P. Rodrigues, R. E. Castro, and G. Soveral, "The emerging role of microRNAs in Aquaporin regulation," *Frontiers Chem.*, vol. 6, pp. 1–8, Jun. 2018.

[3] R. Garzon, G. A. Calin, and C. M. Croce, "MicroRNAs in cancer," *Annu. Rev. Med.*, vol. 60, no. 1, pp. 167–179, Feb. 2009.

[4] X. Liao, G. Zhu, R. Huang, C. Yang, X. Wang, K. Huang, T. Yu, C. Han, H. Su, and T. Peng, "Identification of potential prognostic microRNA biomarkers for predicting survival in patients with hepatocellular carcinoma," *Cancer Manage. Res.*, vol. 10, pp. 787–803, Apr. 2018.

[5] H. Shen, S. Lu, L. Dong, Y. Xue, C. Yao, C. Tong, C. Wang, and X. Shu, "Hsa-miR-320d and hsa-miR-582, miRNA biomarkers of aortic dissection, regulate apoptosis of vascular smooth muscle cells," *J. Cardiovascular Pharmacol.*, vol. 77, no. 5, pp. 275–282, May 2018.

[6] Y. Liu, Y. Yao, R. Liu, C. Gao, T. Zhang, L. Qi, G. Liu, W. Zhang, X. Wang, J. Li, J. Li, and C. Sun, "Identification of prognostic biomarkers for breast cancer based on miRNA and mRNA co-expression network," *J. Cellular Biochem.*, vol. 120, no. 9, pp. 15378–15388. Sep. 2019.

[7] J. Luo, C. Pan, G. Xiang, and Y. Yin, "A novel cluster-based computational method to identify miRNA regulatory modules," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 16, no. 2, pp. 681–687, Mar. 2019.

[8] T. Shao, G. Wang, H. Chen, Y. Xie, X. Jin, J. Bai, J. Xu, X. Li, J. Huang, Y. Jin, and Y. Li, "Survey of miRNA-miRNA cooperative regulation principles across cancer types," *Briefings Bioinf.*, vol. 20, no. 5, pp. 1621–1638, Sep. 2019.

[9] Q. Zhao, W. Song, D. Y. He, and Y. Li, "Identification of key gene modules and pathways of human breast cancer by co-expression analysis," *Breast Cancer*, vol. 25, no. 2, pp. 213–223, Mar. 2018.

[10] S. Yoon and G. De Micheli, "Prediction of regulatory modules comprising microRNAs and target genes," *Bioinformatics*, vol. 21, no. 2, pp. ii93–ii100, Sep. 2005.

[11] X. Q. Dai, L. Ding, H. Liu, Z. Xu, H. Jiang, S. K. Handelman, and Y. Bai, "Identifying interaction clusters for miRNA and mRNA pairs in TCGA network," *Gene*, vol. 10, p. 702, Sep. 2019.

[12] L. Ding, Z. Feng, and Y. Bai, "Clustering analysis of microRNA and mRNA expression data from TCGA using maximum edge-weighted matching algorithms," *BMC Med. Genomics*, vol. 12, no. 1, pp. 1–27, Aug. 2019.

[13] S. Lim, S. Ryu, S. Kwon, K. Jung, and J.-G. Lee, "LinkSCAN*: Overlapping community detection using the link-space transformation," in *Proc. IEEE 30th Int. Conf. Data Eng. (ICDE)*, Mar. 2014, pp. 292–303.

[14] V. Jayaswal, M. Lutherborrow, D. D. Ma, and Y. H. Yang, "Identification of microRNA-mRNA modules using microarray data," *BMC Genomics*, vol. 12, no. 1, p. 138, Mar. 2011.

[15] Y. Li, C. Liang, K.-C. Wong, J. Luo, and Z. Zhang, "Mirsynergy: Detecting synergistic miRNA regulatory modules by overlapping neighbourhood expansion," *Bioinformatics*, vol. 30, no. 18, pp. 2627–2635, Sep. 2014.

[16] S. M. M. Karim, L. Liu, T. D. Le, and J. Li, "Identification of miRNA-mRNA regulatory modules by exploring collective group relationships," *BMC Genomics*, vol. 17, no. S1, p. 7, Jan. 2016.

[17] I. Derényi, G. Palla, and T. Vicsek, "Clique percolation in random networks," *Phys. Rev. Lett.*, vol. 94, no. 16, 2005, Art. no. 160202.

[18] C. Liang, Y. Li, and J. Luo, "A novel method to detect functional microRNA regulatory modules by bicliques merging," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 13, no. 3, pp. 549–556, May 2016.

[19] A. T. Kalinka and P. Tomancak, "Linkcomm: An R package for the generation, visualization, and analysis of link communities in networks of arbitrary size and type," *Bioinformatics*, vol. 27, no. 14, pp. 2011–2012, May 2011.

[20] L. Huang, G. Wang, Y. Wang, E. Blanzieri, and C. Su, "Link clustering with extended link similarity and EQ evaluation division," *PLoS ONE*, vol. 8, no. 6, Jun. 2013, Art. no. e66005.

[21] H. Y. Huang *et al.*, "miRTarBase 2020: Updates to the experimentally validated microRNA–target interaction database," *Nucleic Acids Res.*, vol. 48, no. D1, pp. 1–8, Oct. 2019.

[22] J. Cáceres, C. Hernando, M. Mora, I. M. Pelayo, M. L. Puertas, C. Seara, and D. R. Wood, "On the metric dimension of Cartesian products of graphs," *SIAM J. Discrete Math.*, vol. 21, no. 2, pp. 423–441, Jan. 2007.

[23] S. I. Serengil. *Convolutional Autoencoder: Clustering Images With Neural Networks*. Accessed: Jun. 5, 2019. [Online]. Available: https://sefiks.com/2018/03/23/convolutional-autoencoder-clustering-images-with-neural-networks/

[24] T. Katsuki, M. Ono, A. Koseki, M. Kudo, K. Haida, J. Kuroda, M. Makino, R. Yanagiya, and A. Suzuki, "Risk prediction of diabetic nephropathy via interpretable feature extraction from EHR using convolutional autoencoder," *Stud. Health Technol. Informat.*, vol. 247, pp. 106–110, Jan. 2018.

[25] J. J. Peng, W. Hui, Q. Li, B. Chen, J. Hao, Q. Jiang, X. Shang, and Z. Wei, "A learning-based framework for miRNA-disease association identification using neural networks," *Bioinformatics*, vol. 35, pp. 1–8, Apr. 2019.

[26] B. Hou and R. Yan, "Convolutional autoencoder model for finger-vein verification," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 5, pp. 2067–2074, May 2020.

[27] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972–976, Feb. 2007.

[28] H. Li, B. Dai, J. Fan, C. Chen, X. Nie, Z. Yin, Y. Zhao, X. Zhang, and D. W. Wang, "The different roles of miRNA-92a-2-5p and let-7b-5p in mitochondrial translation in db/db mice," *Mol. Therapy-Nucleic Acids*, vol. 17, pp. 424–435, Sep. 2019.

[29] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann, "Link communities reveal multiscale complexity in networks," *Nature*, vol. 466, no. 7307, pp. 761–764, Aug. 2010.

[30] J. Xu, C.-X. Li, Y.-S. Li, J.-Y. Lv, Y. Ma, T.-T. Shao, L.-D. Xu, Y.-Y. Wang, L. Du, Y.-P. Zhang, W. Jiang, C.-Q. Li, Y. Xiao, and X. Li, "MiRNA–miRNA synergistic network: Construction via co-regulating functional modules and disease miRNA topological features," *Nucleic Acids Res.*, vol. 39, no. 3, pp. 825–836, Feb. 2011.

[31] X. Qin and Y. Song, "Bioinformatics analysis identifies the estrogen receptor 1 (ESR1) gene and hsa-miR-26a-5p as potential prognostic biomarkers in patients with intrahepatic cholangiocarcinoma," *Med. Sci. Monitor, Int. Med. J. Exp. Clin. Res.*, vol. 26, May 2020, Art. no. e921815.

[32] J. Wittenborn, L. Weikert, B. Hangarter, E. Stickeler, and J. Maurer, "The use of micro RNA in the early detection of cervical intraepithelial neoplasia," *Carcinogenesis*, vol. 41, pp. 1–6, May 2020.

[33] X. Zhu, G. Yang, J. Xu, and C. Zhang, "Silencing of SNHG6 induced cell autophagy by targeting miR-26a-5p/ULK1 signaling pathway in human osteosarcoma," *Cancer Cell Int.*, vol. 19, no. 1, pp. 1–11, Dec. 2019.

[34] S. Shukla, J. R. Evans, R. Malik, F. Y. Feng, S. M. Dhanasekaran, X. Cao, G. Chen, D. G. Beer, H. Jiang, and A. M. Chinnaiyan, "Development of a RNA-seq based prognostic signature in lung adenocarcinoma," *J. Nat. Cancer Inst.*, vol. 109, no. 1, Jan. 2017, Art. no. djw200.

[35] H. Wang, Z. Hu, and L. Chen, "Enhanced plasma miR-26a-5p promotes the progression of bladder cancer via targeting PTEN," *Oncol. Lett.*, vol. 16, no. 4, pp. 4223–4228, Jul. 2018.

[36] K. Wu, X. Mu, J. Jiang, M. Tan, R. Wang, W. Zhou, X. Wang, Y. He, M. Li, and Z. Liu, "MiRNA-26a-5p and miR-26b-5p inhibit the proliferation of bladder cancer cells by regulating PDCD10," *Oncol. Rep.*, vol. 40, pp. 3523–3532, Sep. 2018.

[37] K. Miyamoto, N. Seki, R. Matsushita, M. Yonemori, H. Yoshino, T. Sakaguchi, S. Sugita, H. Enokida, and M. Nakagawa, "Mp98-19 tumor-suppressive microrna-26a-5p/-26b-5p inhibit cancer cell migration and invasion through targeting PLOD2 that is a potential prognostic marker in bladder cancer," *J. Urol.*, vol. 197, no. 4, p. e1319, 2017.

[38] S. Nisar, S. Hashem, M. A. Macha, S. K. Yadav, S. Muralitharan, L. Therachiyil, G. Sageena, H. Al-Naemi, M. Haris, and A. A. Bhat, "Exploring dysregulated signaling pathways in cancer," *Current Pharmaceutical Des.*, vol. 26, no. 4, pp. 429–445, Mar. 2020.

[39] H. Zhu, S. Wang, H. Shen, X. Zheng, and X. Xu, "SP1/AKT/FOXO3 signaling is involved in miR-362-3p-mediated inhibition of cell-cycle pathway and EMT progression in renal cell carcinoma," *Frontiers Cell Develop. Biol.*, vol. 8, p. 297, May 2020.

[40] W. Wen, J. Gong, P. Wu, M. Zhao, M. Wang, H. Chen, and J. Sun, "Mutations in gliclazide-associated genes may predict poor bladder cancer prognosis," *FEBS Open Bio*, vol. 9, no. 3, pp. 457–467, Feb. 2019.

[41] K. Uhr, W. J. C. Prager-van der Smissen, A. A. J. Heine, B. Ozturk, M. T. M. van Jaarsveld, A. W. M. Boersma, A. Jager, E. A. C. Wiemer, M. Smid, J. A. Foekens, and J. W. M. Martens, "MicroRNAs as possible indicators of drug sensitivity in breast cancer cell lines," *PLoS ONE*, vol. 14, no. 5, May 2019, Art. no. e0216400.

[42] W. Qin, S. Feng, Y. Sun, and G. Jiang, "MiR promotes breast cancer migration by activating MEK/ERK signaling," *J. Gene Med.*, vol. 3, p. e3188, Mar. 2020.

**YI YANG** received the master's degree in computer science from the Wuhan University of Technology, in 2005. He has studied in Hunan University as a Visiting Scholar, in 2014. He is currently an Associate Professor with the College of Information Science and Engineering, Hunan Women's University. His research interests include computational biology, bioinformatics, data mining, and software agent.

**XUTING WAN** received the master's degree in computer science from Zhejiang University, in 2013. She is currently a Lecturer with the College of Information Science and Engineering, Hunan Women's University. Her research interests include data mining and software engineering.