3rd National Conference on Innovative Research Trends in Computer Science and Technology (NCIRCST 2018) ISSN: 2454-4248
Volume: 4 Issue: 3 94 – 102

Discovery of Micrornas and Transcription Factors Co-Regulatory Modules by Integrating Multiple Types of Genomic Data

Rohith Babu
S.A.Engineering College
UG Scholar,IT

I.Rakesh
UG Scholar,IT
S.A.EngineeringCollege

Auxilia Osvin Nancy,
Assistant Professor,IT,
S.A.Engineering College

Abstract—It is well known that regulators known as microRNA (miRNA) and transcription factor (TF) have been found to play an important role in gene regulation. However, there are few researches of collaborative regulatory (co-regulatory) mechanism between miRNA and TF on system level (function level). Meanwhile, recent advances in high-throughput genomic technologies have enabled researchers to collect diverse large-scale genomic data, which can be used to study the co-regulatory mechanism between miRNA and TF. In this paper, we propose a novel method called SNCoNMF (Sparse Network regularized non-negative matrix factorization for Co-regulatory modules identification) which adopts multiple non-negative matrix factorization framework to identify co-regulatory modules including miRNAs, TFs and genes. This method jointly integrates miRNA, TF and gene expression profiles, and additional priori networks were added in a regularized manner. In addition, to avoid the sparsity of these networks, we employ the sparsity penalties to the variables to achieve modular solutions. The mathematical formulation can be effectively solved by an iterative multiplicative updating algorithm. We apply this method to multiple genomic data including the expression profiles of miRNAs, TFs and genes on breast cancer obtained from TCGA, priori miRNA-gene regulations, TF-gene regulations and gene-gene interactions. The results show that the miRNAs, TFs and genes of the co-regulatory modules are significantly associated and modules have a reasonable size distribution. Furthermore, the co-regulatory modules are significantly enriched in GO biological processes and KEGG pathways, respectively.

Index Terms—microRNA(miRNA), transcription factor, co-regulatory module, genomic data, non-negative matrix factorization

*****

1 Introduction

MicroRNAs (miRNAs) and transcription factors (TFs), as two vital gene regulatory molecules in multicellular organisms, share a common regulatory logic [1]. MicroRNAs are a family of small, non-coding RNAs that regulate gene expression in a sequence-specific manner, which participate in the regulation of numerous cellular process at the post-transcriptional level, such as cancer progression [2, 3]. TFs are proteins that control gene regulation by binding to cis-regulatory elements in the gene promoter region at the transcriptional level [4]. By activating or repressing their target genes, TFs can regulate the global gene expression program of a living cell, and form transcriptional regulatory networks [5]. However, it's still a challenge to elucidate coregulation mechanisms between miRNAs and TFs.

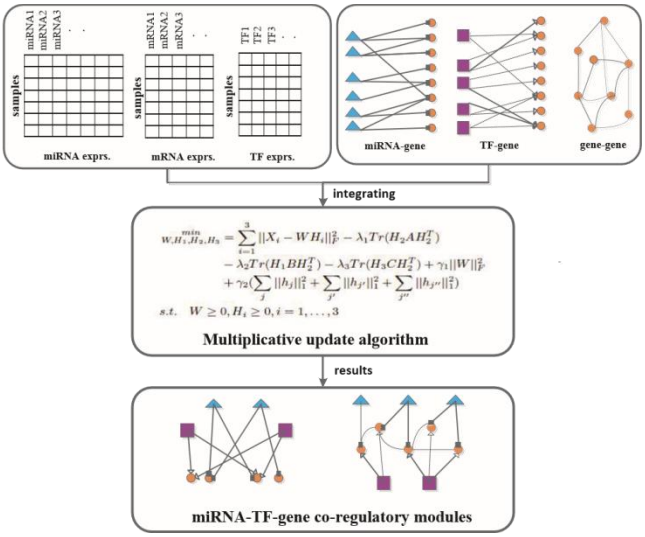
Recently, researchers studied the co-regulation of miRNAs and TFs by finding out their shared downstream targets [6, 7]. The method adopts probabilistic models and statistical tests to measure the significance of the shared targets between the regulators, and to remove the insignifinteractions that occurred by chance. Gene enrichment analysis was used in [8] to identify significant co-regulation between the transcriptional and posttranscriptional layers. They found that some biological processes emerged only in co-regulation and that the disruption of co-regulation may be closely related to cancers, suggesting the importance of the co-regulation of miRNAs and TFs. Tran et al. [9] 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. These work provides a good resource for exploring the regulatory relationships or identifying the network motifs. However, target prediction based on sequences have high rate of false discoveries, which affect the quality of the discoveries of the above mentioned methods. It would be ideal if expression data can be used to refine the discoveries.

Identification of modular structure of biological networks has greatly advanced our understanding of complex cellular systems [10]. However, little is known about the modules that exist in miRNA-TF-gene regulation systems,
and even less is known about these modules role in specific biological processes and key regulation assemblies. Several studies have made efforts to uncover how miRNAs and mRNAs interact on a system level [11, 12]. Thus, some methods are proposed to identify the miRNA-gene modules. Yoon *et al.* identified miRNA regulatory modules on sequencebased predicted miRNA-targets networks [13]. Peng *et al.* proposed a sequential integrative method that enumerate all maximal bi-cliques in miRNA-mRNA network [14]. Although these sequenced-based identification solutions are useful to some extent, it is impossible to detect highly credible miRNAs modules and accurately infer miRNAs functional regulation [12]. Meanwhile, these methods have not considered TFs regulation and the modules only contain miRNAs and genes.

With the development of next-generation highthroughout technology, numerous genomic data, including miRNA and mRNA expression profiling, somatic mutation and copy number variation, have substantially contributed to the comprehensive profiling of kinds of cancer [15]. In addition, databases like

TargetScan [16] (miRNA-mRNA interaction), TRANSFAC [17] (TF-gene interaction), BioGrid [18] (protein-protein interaction) *etc* are constantly updating. The availability of multiple types of functional genomic data provides a vast data platform for the miRNA targets prediction and miRNA-TF co-regulatory modules identification.

Accordingly, several studies have suggested that integrating miRNA and mRNA expression profiles could play an important role in miRNA function identification [19–22]. Recently, Zhang *et al.* proposed a framework of sparse network-regularized multiple non-negative matrix factorization (SNMNMF) to identify miRNA-mRNA modules [23]. Specifically, SNMNMF utilizes the expression and target-site information as well as the gene-gene interaction (GGI) and transcription factor binding sites (TFBS). SNMNMF was applied to ovarian cancer samples and achieved great performance in identifying numerous miRNA-mRNA modules with biological significance. However, although the SNMNMF considered the TFs effect, they are treated as genes, instead of regulators which regulate genes. Soon after the SNMNMF, Zhang *et al.* developed a method to identify multi-dimensional modules by joint non-negative matrix factorization (jNMF) [24].



2 Materials And Methods

2.1 Datasets and preprocessing

We performed a joint analysis of miRNA,TF and gene variables to identify miRNA-TF-gene co-regulaotry modules associated with breast cancer. The expression profiles of miRNAs, TFs and genes were downloaded from TCGA Project [25]. For breast cancer, by employing R package *limma* [26], we performed differential gene expression analysis on expression profiles to extract differentially expression miRNAs and mRNAs at significant level (adjusted pvalue <0.05, adjusted by BH methods) between tumor and normal samples. As a result, we obtained 1046 miRNAs and 20502 mRNAs which were filtered with differential expression from a common set of 813 breast cancer samples (see supplementary material). Then we separated TFs expression data from gene expression profile and obtained 81 TFs expression profiles.

The protein-protein interaction network was downloaded from BioGrid database (http://thebiogrid.org/) [27]. Meanwhile, we obtained miRNA-gene interatcions and TF-gene interactions from TargetScan (http://www.targetscan.org/) and ENCODE Project (http://encodenets.gersteinlab.org/) [28], respectively. Here, we only kept regulations and interactions involving miRNAs and mRNAs which were presented in the expression data. This process resulted in a network with 21483 gene-gene interactions, 57582 miRNA-gene regulations and 7995 TF-gene regulations. Moreover, We downloaded the disease-related miRNA sets from HDMM [29] and disease-related gene sets from SemFunSim [30] and DisGeNET(http://www.disgenet.org/) for the cancer-specific miRNAs and mRNAs analysis.

To satisfy the property of non-negative, we adopt the method proposed by Kim and Tidor [31] to ensure the nonnegativity constraint. More specifically, given a expression matrix X of size s by l, we create a matrix X' of size s by $2l$. For each element of the original matrix $X(i,j) \geq 0$, we set $X'(i,j) = X(i,j)$ and $X'(i,l+j) = 0$, where $i = 1,2,...,s$ and $j = 1,2,...,l$ denote the row index and column index, respectively. Similarly, we set $X'(i,j) = 0$ and $X'(i,l + j) = -X(i,j)$ for each element $X(i,j) \leq 0$. In short, we create a new matrix which doubles the column size of the original matrix. There are two columns to represent each variable(miRNA, TF or gene) in X'. In other words, the first column contains the positive of that variable or a zero value, the second contains the absolute value of its negative value or a zero value. The transformed non-negative expression matrices are our input matrices X_1, X_2 and X_3.

2.2 Problem formulation

To identify miRNA-TF-gene co-regulatory modules, we design an objective function with three components (redFigure 1). As mentioned above, the optimization function consists of a joint NMF, a regularized term for three prior networks and a sparse penalized term. Here we provide the final optimization function:

$$\min_{W,H_1,H_2,H_3} = \sum_{i=1}^{3} ||X_i - WH_i||_{F2} - \lambda_1 Tr(H_2 A H_2{}^T) - \lambda_2 Tr(H_1 B H_2{}^T) - \lambda_3 Tr(H_3 C H_2{}^T) + \gamma_1 ||W||^2_F + \gamma_2 (\sum_j ||h_j||_{12} + \sum_{j'} ||h_{j'}||_{21} + \sum_{j''} ||h_{j''}||_{12}),$$

$$s.t. \quad W \geq 0, H_i \geq 0, i = 1,...,3$$

where expression matrix $X_{1,2,3}$ is decomposed by basic matrix W with size of $N \times K$ and coefficient matrix $H_{1,2,3}$ with size of $K \times M$. The following subsections will provide the details as well as the solution of objective function.

2.2.1 NMF and joint NMF

Non-negative matrix factorization (NMF) are a powerful method for data reduction and clustering that has widely been used to analyse high-throughput genomic data [32, 33]. Given a non-negative data matrix $X_{N\times M}$ NMF aims to find a non-negative factorization WH of rank K that best approximates X, typically in terms of the Frobenius norm
[34]:

$$min_{W,H} = ||X - WH||^2_F, \qquad s.t. \qquad W \geq 0, H \geq 0.$$

where H is an $K \times M$ matrix containing the basic components of the X, while W is an $N \times K$ matrix containing the basis vectors and the elements of $W_{N\times K}$ can be thought of as latent factors associated with these components. Each observation (row of X) is approximated by a linear combination of components (rows of H) with weights given by each row of W. The entire data are explained by a sum of additive parts. It is intuitive in biological contexts because biological entities and mechanisms can be naturally described with a signal that either present or absent [35].

Joint NMF (jNMF) is proposed as an extension to NMF for integrating multiple datasets with a common set of observations [24]. To extract multi-dimensional modular structures across multiple data matrices, joint NMF shares a common basis matrix W with different coefficient matrices. Given I data matrices $(X_1)_{N\times M_1},\ldots,(X_I)_{N\times M_I}$, the problem formulation is:

$$\min_{W,H_1,\ldots,H_I} = \sum_{i=1}^{I} ||X_i - WH_i||_{F2},$$

$$s.t. \qquad W \geq 0, H_i \geq 0, i = 1,\ldots,I$$

where W is an $N \times K$ matrix, and each column of W represents a basis vector of the reduced system. H_i is a matrix of size $K \times M_i$. The jNMF is found to well detect coordinated activity across multiple genomic variables in the form of multi-dimensional modules [24]. In jNMF, each module represents a biclustering of both observations and variables, which can be visualized as a block in the data matrix after appropriate rotation.

Since the aim of our study is to discover the coregulatory modules across multiple genomic data, and its inputs involve in three expression matrices, considering the concept of jNMF we can define the original objective function as follows:

$$\min_{W,H_1,H_2,H_3} = \sum_{i=1}^{3} ||X_i - WH_i||_{F2},$$

$$s.t. \qquad W \geq 0, H_i \geq 0, i = 1,2,3$$

There are several algorithms proposed for above optimization problem [36]. However, the solutions to it tend to be a local minimum and may be sensitive to noise in the expression data. Thus, we adopt the optimization method proposed in Zhang *et al.* [23], which guide the optimization process toward reasonable biological solutions by incorporating prior knowledge into the objective function.

2.2.2 Network-regularized constraints

Recent studies suggest that incorporating biological priori knowledge, such as gene-gene interaction networks, as a regularization term can help to more accurately identify new genes in the existing pathways [37]. Thus, the priori knowledge consists of predicted miRNA-gene regulations, TF-gene regulations and gene-gene interactions. The main reason we choose these regulatory networks is to improve the performance that any variables linked in these three networks are more likely to be placed into the same module. In other words, such constraints can greatly facilitate the discovery of co-regulatory modules by narrowing down the large search space, apart from improving the biological relevance of the results [23].

Let A,B,C denote $N \times N, M \times N$ and $P \times N$ adjacency matrices representing gene-gene interaction network, bipartite miRNA-gene network and bipartite TF-gene network, respectively. For gene-gene interactions network, we employ "*must-link*" constraints by maximizing the following objective function:

$$\Theta_1 = \sum_{ij} a_{ij}(h_{2i})^T h_{2j} = Tr(H_2 A H_2^T),$$

This term denotes all the "*must-link*" constraints in genegene interaction, which ensures that genes with known interactions have similar coefficient profiles.

Similarly, we can get the corresponding objective functions of miRNA-gene network and TF-gene network:

$$\Theta_2 = \sum_{ij} b_{ij}(h_{1i})^T h_{2j} = Tr(H_1 B H_2^T),$$

$$\Theta_3 = \sum_{ij} c_{ij}(h_{3i})^T h_{2j} = Tr(H_3 C H_2^T),$$

Θ_2 term can be considered as all the "*must-link*" constraints in miRNA-gene network, which Θ_3 term can be regarded as all the "*must-link*" constraints in TF-gene network. By combining the three "*must-link*" constraints with the former jNMF framework, we define a new optimization function as follows:

$$\min_{} \quad 3$$

$$_{W,H_1,H_2,H_3} = \sum_{i=1} ||X_i - WH_i||_{F2} - \lambda_1 Tr(H_2 A H_2{}^T) - \lambda_2 Tr(H_1 B H_2{}^T) - \lambda_3 Tr(H_3 C H_2{}^T),$$

$$s.t. \quad W \geq 0, H_i \geq 0, i = 1,\ldots,3$$

2.2.3 Sparse representations

As NMF naturally gives rise to parsimonious solutions and generates sparse representations of the data, we adopt a method similar to Zhang *et al.* [23] as penalization which makes the coefficient matrices H sparse. The objective function after adding sparse constraints is formulated as follows:

$$\min_{W,H_1,H_2,H_3} = \sum_{i=1}^{3} ||X_i - WH_i||_{F2} - \lambda_1 Tr(H_2 A H_2{}^T) - \lambda_2 Tr(H_1 B H_2{}^T) - \lambda_3 Tr(H_3 C H_2{}^T) + \gamma_1 ||W||^2_F + \gamma_2 (\sum_j ||h_j||_{21} + \sum_{j'} ||h_{j'}||_{21} + \sum_{j''} ||h_{j''}||_{21})$$

$$s.t. \quad W \geq 0, H_i \geq 0, i = 1,\ldots,3$$

where h_j, $h_{j'}$ and $h_{j''}$ are the j-th, j'-th and j''-th columns of H_1, H_2 and H_3, respectively. The term $\gamma_1 ||W||^2_F$ limits the growth of W, while $\gamma_2(\sum_j ||h_j||_1^2 + \sum_{j'} ||h_{j'}||_1^2 + \sum_{j''} ||h_{j''}||_1^2)$ encourage sparsity.

2.3 miRNA-TF-gene modules assignment

After obtaining coefficient matrices H_1, H_2 and H_3 produced by the above SNCoNMF algorithm, we use them to identify miRNA-TF-gene modules. In the general applications of NMF [32], researchers have used the maximum of each column of H(or row of W)to determine memberships. However, this approach supposes that each gene or sample can belong to one and only one module, while the fact is that some genes/miRNAs/TFs may be active in multiple modules or may be active in multiple modules with multiple functions. In this paper, we refer to Zhang *et al.* [23] which adopts a z-score to determine the module assignment.

Given to the matrices of H_1, H_2 and H_3, we calculate the z-socre for each element in each row of H_i by:

$$z_{ij} = \frac{x_{ij} - \mu_i}{\sigma_i}$$

where μ_i is the average value for variable j(gene/miRNA/TF) in $H_i (i = 1,2,3)$ and σ_i is the standard deviation. We assign variable j as a member of module k, if z_{ij} is greater than a given threshold T. In other words, each gene/miRNA/TF may be assigned to multiple modules which can reflect the biological realities.

2.4 Evaluation

2.4.1 Modules characteristic analysis

Each miRNA-TF-gene co-regulatory module in this article is composed of miRNAs, TFs, genes and the interactions between them, which determines the specificity of each module. Thus, it is necessary to analyze the modules distributions and its topological characteristics. Accordingly, we analyze the modules densities to observe the tightness of modules. The density function is defined as $\frac{2l}{n(n-1)}$, where l is the number of edges in the module, n represents the number of nodes including miRNAs, TFs and genes.

We further employe miRNA-gene expression correlation (MiMEC) for each module to compute the sum value of (anti)-correlations between miRNAs and genes. The MiMEC is defined as $\sum corr(x,y)$, where *corr* is a function that calculates the Pearson correlation coefficients for the pair (x,y), and x, y represents a miRNA and a gene respectively. Here, only the pair with connected between them in matrix W is considered, because the network is constructed when the modules are determined. Accordingly, the Ave-MiMEC κ is defined as $\frac{1}{K}\sum_i \sum |corr(x,y)|$, where K represents the number of detected modules. As important regulators of the expression of mRNAs, MiMECs between miRNAs and genes are supposed to be a reasonable value. Similar to the MiMEC between miRNAs and genes, we also calculated the TfMEC between TFs and genes within modules.

2.4.2 Biological significance analysis

In order to analyze the functional significance of modules, we make biological significance analysis for the genes lists in the identified modules. Biological significance analysis is conducted by seeking for the enrichment in GO-BP terms and KEGG pathways. We here employ R package *GOstats* [39] to test gene lists for GO term associations and KEGG pathways analysis. In order to visually display the level of function enrichment, we implement GO enrichment score (GOES) and KEGG pathways enrichment score (KEGGES) to measure GO-BP enrichment and KEGG pathways enrichment degree for each module, respectively. The GOES of one module is defined as $\frac{1}{I}\sum_i^I -\log_{10} p_i$, where I represents the number of GO-BP terms that the genes enriched in and p_i represents p-value of the i^{th} GO-BP terms. Accordingly, we define the KEGGES of one module as GOES does.

3 Results

We applied the SNCoNMF algorithm to identify miRNATF-gene modules by integrating multiple independent data sources(the six matrices described in the Section 2.2).

3.1 Choose of parameters

The proposed SNCoNMF algorithm requires setting of several parameters as described in the pseudo code. Here it's important to decide the value of the reduced dimension of matrix factorization *K*. Similar to the method described by Zhang [23], we conducted a miRNA cluster analysis which required miRNA cluster data from the miRBase website(http://www.mirbase.org/). As a result, we obtained about 20 clusters containing miRNAs range from 2 to 50. So in this paper we set the K to 20, approximately equals to the number of miRNA clusters represented in our data. Meanwhile, we set parameters λ_1,λ_2,λ_3,γ_1 and γ_2 to 0.01, 0.01, 0.01, 20, 10, respectively. Duo to the lack of TFs, we set the threshold T to 2 for TFs by conducting a series of tests, while T is set to 3 for miRNAs and genes.

3.2 Module character and size distribution

We performed the proposed SNCoNMF algorithm on breast cancer dataset and obtained 20 miRNA-TF-gene coregulatory module which are composed of by a set of miRNAs, TFs and genes that are denoted as miRNA modules, TF modules and gene modules, respectively. Specifically, there is not any TFs in module 12, which means that module 12 is just a miRNAs-genes regulatory module. The 20 miRNA-TF-gene modules identified in this paper have an average of 5.5 miRNAs, 2.5 TFs and 28.15 genes per module. The average density of all module is 0.0176. Meanwhile, we calculated the average miRNA-gene expression correlation and TF-gene expression correlation among all modules as described in section **2.4.1**.

In addition, to verify feasibility of our method, we run the SNMNMF algorithm on our datasets which TFs are treated as genes. In SNMNMF, the average miRNAs, genes number and TFs are 5.6 26.4 and 0.55, respectively, which genes and TFs are less than ours, especially the TTs. It demonstrates that SNCoNMF can effectively discover miRNA-TF-gene co-regulatory modules. Meanwhile, due to more genes in per module, SNCoNMF bears a less average module density. The result details are shown as Table 1:

TABLE 1
Performance summary of SNCoNMF

Methods	M	Ave(mi)	Ave(g)	Ave(T)	Ave-den	MiMEC	TfMEC
SNCoNMF	20	5.5	28.15	2.5	0.0176	0.0975	0.2110
SNMNMF	20	5.6	26.4	0.55	0.025	0.0119	0.0

Note: M: is the number of modules; Ave(mi),Ave(g) and Ave(T) are the average number of miRNAs,genes and TFs per module; Ave-den is the average density of identified modules; MiMEC is the average miRNA-gene expression correlation across all modules; TfMEC is the average TF-gene expression correlation across all modules.

To analyze the module size distribution, we calculated the size distribution of miRNA modules, TF modules and gene modules, respectively(see Fig.2). As can be seen from the figure, miRNA module size distribution has a relatively uniform distribution within a size range of 2 to 8. Differently, TFs module size distribution enjoys a similar normal distribution which most distributed in 2 and 3. Finally, the size distribution of genes centers in 10 to 40. From the results, we can speculate that miRNAs, TFs and genes have a different size distribution which denotes the effect difference of the three variables.

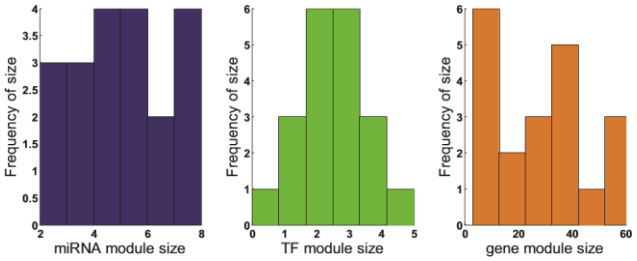


Fig. 2. *Module size distribution for miRNA modules, TF modules and gene modules, respectively.*

3.3 Analysis of density and expression correlation within modules

In the section **3.2**, we just briefly summarized the densities of modules on average (consider TFs as genes in SNMNMF), we herein conduct comparative analysis to the density for each module in details. As illustrated in Fig.3, due to SNMNMF achieves higher densities in some modules, the densities of modules from SNCoNMF are slightly less than SNMNMF. However, SNCoNMF has a more average genes and TFs per modules.

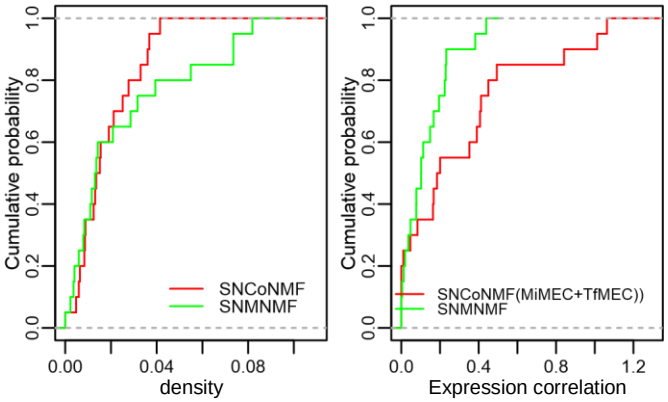


Fig. 3. *The cumulative distribution function (CDF) for the density and expression correlation(miRNA-gene and TF-gene) of modules identified by SNCoNMF and SNMNMF on breast dataset.*

As mentioned above, we employ MiMEC and TfMEC to compute the averaged correlation between miRNAs and genes and the averaged correlation between TFs and genes in the modules identified by each approach, respectively. Exhilaratingly, SNCoNMF achieves a comparable MiMEC and a better TfMEC than SNMNMF. Regretfully, the TfMEC of SNMNMF is zero which may denote that SNMNMF can not discover miRNA-TF-gene co-regulatory modules effectively. As illustrated in Fig.3, the cumulative distribution function(CDF) for total expression correlation (MiMEC and TfMEC) of SNCoNMF is almost overall outperform than SNMNMF. The observations confirmed that relations within modules are also strong correlated. Therefore, the modules identified by our method are stronger regulation than SNMNMF, especially for regulations between TFs and genes.

3.4 Functional Enrichment Analysis

To evaluate function enrichment of the identified modules, we calculated their enrichment in GO-BPs and KEGG pathways by employing *GOstats* [39] (using hypergeometric test with q-value <0.05). Accordingly, we adopt a functional enrichment comparative analysis just for the gene lists involved in the modules identified by SNCoNMF and SNMNMF. The GOES and KEGGES of cumulative probability for all modules detected by SNCoNMF and SNMNMF is shown in Fig.4. The averaged value of GOES and KEGGES from SNCoNMF are 2.4845 and 1.7081, respectively, while the respective averaged value of GOES and KEGGES from SNMNMF are 2.3602 and 1.8129. As illustrated in Fig.4, our method shares a comparable performance with SNMNMF as a whole which show slightly superiority in GOES.

We further explore the distribution between density and GOES/KEGGES for the identified modules from SNCoNMF and SNMNMF. As illustrated in Fig.5, our modules' GOES mainly distribute between 2.0 and 2.8 of the density value, and the slope of the fitting curve is positive. By contrast,

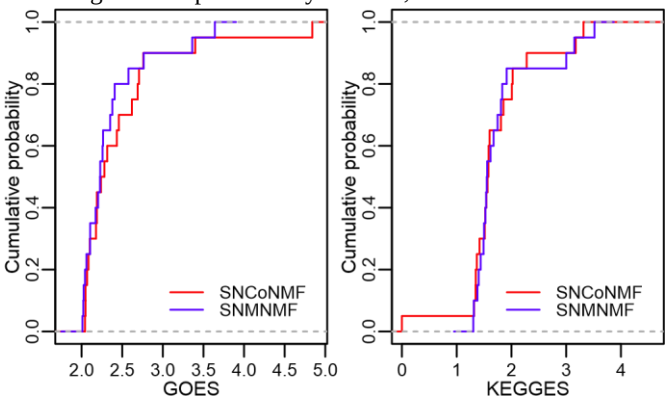


Fig. 4. *The GOES and KEGGES of cumulative probability for all modules detected by SNCoNMF and SNMNMF.* the fitting curve of SNMNMF shares a more plain trend. However, intuition tells us that large density module may share a larger GOES. Meanwhile, the distribution between density and KEGGES from SNMNMF shares a more messy state due to some outliers. In conclusion, the observations are in agreement with the viewpoint that our modules are significantly enriched.

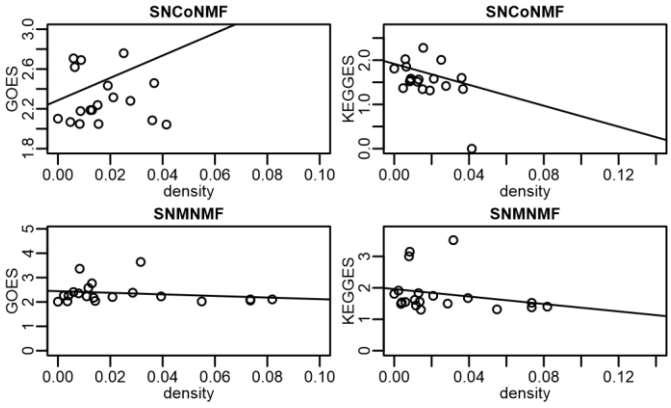


Fig. 5. *The distributions between density and GOES/KEGGES for modules detected by SNCoNMF and SNMNMF.*

In addition, experimental results show that all of modules from SNCoNMF are enriched in at least one GO-BP term and only one module (Module 9) is not enriched in any pathway (see supplementary material). In Table 2, we list the details of top 3 enriched Gene Ontology biological process term clusters for genes in each modules having 5 genes at least. We found that 45% (9/20) modules are enriched at least 5 genes in top 3 enriched GO-BP terms. Further, these GO-BP terms are enriched with similar functions or similar genes. Taking module 20 for example, the top 3 GO terms are GO:0002376, GO:0006955 and GO:0002682 which all related to immune system or immune response. Meanwhile, they are enriched some similar genes, like CD7,CD38.

In order to analyze pathways enrichment, similar to functional enrichment, we calculated the top pathways that enriched by more than 3 genes. There are total 121 pathways enriched by all modules which 24.8% (30/121) have genes more than 3. Taking module3 LAMA3, LAMB3, MET and LAMC3 are markedly enriched in the pathways related to cancer. In module 20, TNFRSF17, CCR6, LTA, LTB, CCL13 and TNFSF14 are mainly enriched in Cytokine-cytokine receptor interaction (p-value=8.6e-6). (see supplementary material)

3.5 Cancer-related modules

We investigated the details of the miRNA-TF-gene modules identified by SNCoNMF for finding out onco-miRNAs and oncogenes, and for discovering some cancer-related processes or pathways. We expect that the identified modules are related to corresponding cancers because our datasets included miRNA, TF and gene expression profiles of breast samples. To validate relations to diseases, we referred to the HMDD database [29] to find the experimentally supported human miRNA and disease associations. Simultaneously, we consulted corresponding literatures to find cancerrelated TFs or genes.

In our experimental results, TFs in 50% modules are related to corresponding disease while the fact is that there are only few TFs in per module. Due to the numerous genes in per module, 100% modules include cancer-related genes. For instance, module 1 contains three gene members CCL20, CLDN1 and CST6, all of which are oncogenes related to breast cancer [40–42]. Moreover, CCL20 is described as a essential role in inducing migration and proliferation on breast epithelial cells [40]. Meanwhile, miR-577, miR-224-5p and miR-934 in module 1 are cancer-related regulators [43– 45], especially miR-934, which may involve in breast cancer in Hispanic women [46]. We further analyze all three TFs in module 1, which shows that BCL11A, FOSL1 and POU5F1 are crucial regulators related to breast cancer [47–49]. And BCL11A is reported as a triple-negative breast cancer gene with critical functions in stem and progenitor cells [47]. In table 3, we list a few typical modules related to cancers.

We further investigate the interactions of miRNA-gene, TF-gene and gene-gene in the predicted modules. There are 45%(9/20) modules containing all three types interactions. Taking module 1 for example, specifically, there are four important regulations: miR-767-5p to CST6, miR-934 to KCNN4, miR-934 to PDZK1IP1 and FOSL1 to PRAME, which

CST6, miR-934 and FOSL1 are related to breast cancer. We then consult the regulation of FOSL1 to PRAME. There are not experimental evidence about the regulation. Interestingly, confirming modules identified by SNCoNMF were significantly enriched. We further analyzed the relationship between cancer and identified modules using HDMM database and literature survey,

TABLE 2
The top 3 enriched Gene Ontology biological process term clusters for genes in each modules having 5 genes at least

Module	Term Num	Term	Description	P-value	Count	Genes
1	457	GO:0060429	epithelium development tissue	1.91e-11	21	CA9;CDH3;CST6;NKX2-5;DLX5 etc
		GO:0009888	development	5.57e-10	24	FOXC1;SFN;GSTA2;IL6;KRT4 etc
		GO:0052548	regulation of endopeptidase activity	4.08e-08	11	CST6;SFN;IL6;SERPINB2;SERPINA1 etc
3	457	GO:0060429	epithelium development tissue	3.57e-13	22	BMP7;KLF5;CDH3;COL2A1;COL7A1 etc
		GO:0009888	development organ	8.73e-11	24	FOXC1;FOXC2;GSTA1;NRG1;KRT5 etc
		GO:0009887	morphogenesis	1.52e-09	17	HOXA2;MET;PTHLH;RYR1;EDAR etc
5	515	GO:0014074	response to purine-containing compound	4.19e-11	10	CDO1;EGR1;EGR2;FOS;FOSB etc
		GO:1901700	response to oxygen-containing compound	1.94e-10	22	FOS;FOSB;GPD1;KLF15;CXCL2 etc
		GO:0033993	response to lipid	2.46e-10	17	FOS;FOSB;GPD1;KLF15;CXCL2 etc
8	72	GO:0000956	nuclear-transcribed mRNA catabolic process	1.74e-10	6	RPL5;RPLP0;RPS2;RPS4X;RPS18;EXOSC5
		GO:0006402	mRNA catabolic process translational	2.68e-10	6	RPL5;RPLP0;RPS2;RPS4X;RPS18;EXOSC5
		GO:0006414	elongation	2.84e-10	6	EEF1G;RPL5;RPLP0;RPS2;RPS4X;RPS18
11	465	GO:0044767	single-organism developmental process	1.06e-9	29	CAV1;CD36;EGR1;ELK3;ELN etc
		GO:0032502	developmental process system	1.62e-9	29	FHL1;FLNC;GRIN2A;HIP1;IGF2 etc
		GO:0048731	development	5.38e-9	25	PROX1;SIM1;SVIL;TBX15;NR2F1 etc
16	148	GO:0002376	immune system process immune	8.88e-9	17	CD7;CD79A;CD79B;CLU;IFI6 etc
		GO:0006955	response	5.31e-7	13	IFIT1;LTB;PDCD1;TNF;MAP3K14 etc
17	27	GO:0007059	chromosome segregation nuclear	2.12e-5	5	CDC6;INCENP;SMC4;SMC1B;RACGAP1
		GO:0000280	division organelle fission	7.19e-5	6	CDC6;INCENP;CKAP5;SMC4;SMC1B etc
		GO:0048285		1.02e-4	6	CDC6;INCENP;CKAP5;SMC4;SMC1B etc
18	165	GO:0010817	regulation of hormone levels	8.81e-5	8	FOXA1;INHA;SNAP25;SYT5;DHRS2 etc
		GO:0065008	regulation of biological quality regulation	4.03e-5	20	COL2A1;IFI6;GATA4;FOXA1;INHA etc
		GO:0044057	of system process	6.39e-5	7	GATA4;GNAO1;INHA;KCNH2;TNNI3 etc
20	174	GO:0002376	immune system process immune	2.89e-18	30	TNFRSF17;BLK;CD7;CD19;MS4A1 etc
		GO:0006955	response	2.81e-16	24	BLK;CD7;CD19;MS4A1;CD38 etc
		GO:0002682	regulation of immune system process	2.94e-14	21	LBP;LTA;PDCD1;PRKCB;SELL etc

PRAME is reported that is critical for breast cancer growth and metastasis [50].

4 Discussion

The co-regulation mechanisms between miRNAs and TFs has become an important issue recently. However, it still remains a challenge to figure out the principle, especially in terms of system level (i.e. module) [6–9]. The advances in high-throughout enable us use expression profiles to discover modular structure in a broader biological context.

In this article, we develop an effective approach SNCoNMF to identify miRNA-TF-gene co-regulatory modules by integrating multiple types of genomic data. A previous method SNMNMF proposed by Zhang et al [23] only considers the miRNA-gene regulation, which means miRNAgene modules are found in a bipartite networks. It is therefore unrealistic to apply SNMNMF to detect co-regulatory modules which involves in two regulators. We build on the concept of data integration from SNMNMF and further add TF-gene regulation as well as the TF expression profile. To quantify the integration process, we propose a objective function as well as a iteratively solving algorithm.

To demonstrate the effectiveness of our method SNCoNMF, we compared SNCoNMF with SNMNMF on breast cancer dataset. We showed that modules detected by SNCoNMF shared a more reasonable size distribution of miRNA, TF and gene, especially for TF, which validated the necessity of considering TF as an important regulator. Moreover, functional enrichment of gene sets from modules were analyzed using GOES and KEGGES, which validated many cancer-related genes, including breast cancer. Despite the superiority for detecting coregulatory modules, we noticed that SNCoNMF needs prior setting of some parameters of algorithm, especially for dimension K. Meanwhile, the quality of the co-regulatory modules may suffer from the expression profiles with much noise.

SNCoNMF is a flexible framework that incorporates multiple types of data. Apart from expression profiles (better with differential expression processing to denoise), other genomic data including copy number variation, DNA methylation can be applied to this method to identify complicated regulatory patterns.

5 Conclusion

In this paper, we propose a novel computational approach SNCoNMF to discover miRNA-TF-gene co-regulatory modules by integrating the miRNA/TF/gene expression profiles, target-site information (miRNA-gene and TF-gene regulations) as well as the protein-protein interactions. Firstly, SNCoNMF achieves great performance in identifying miRNA-TF-gene module, especially for TFs quantity. Furthermore, the discovered modules are significantly enriched in abundant GO biological processes and KEGG pathways. By literature survey, we validate that modules are biological meaningful. In conclusion, SNCoNMF is an effective method to integrate multiple types of genomic data to discover miRNA-TF-gene co-regulatory modules. In future,

we will focus on the application to other cancer type and genomic data.

Acknowledgment

This work has been supported by the National Natural Science Foundation of China (Grant NO.61572180)

References

[1] O. Hobert, "Gene regulation by transcription factors and micrornas," *Science*, vol. 319, no. 5871, pp. 1785–1786, 2008.

[2] L. He and G. J. Hannon, "Micrornas: small rnas with a big role in gene regulation," *Nature Reviews Genetics*, vol. 5, no. 7, pp. 522–531, 2004.

[3] J. LU, G. Getz, and E. A. Miska, "Microrna expression profiles classify human cancers," *nature*, vol. 435, no. 9, pp. 834–838, Jun 2005.

[4] J. M. Vaquerizas, S. K. Kummerfeld, S. A. Teichmann, and N. M. Luscombe, "A census of human transcription factors: function, expression and evolution," *Nature Reviews Genetics*, vol. 10, no. 4, pp. 252–263, 2009.

[5] T. I. Lee, N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, and I. Simon, "Transcriptional regulatory networks in saccharomyces cerevisiae," *science*, vol. 298, no. 5594, pp. 799–804, 2002.

[6] R. Shalgi, D. Lieber, M. Oren, and Y. Pilpel, "Global and local architecture of the mammalian micrornactranscription factor regulatory network," *PLoS Comput Biol*, vol. 3, no. 7, p. e131, 2007.

[7] Y. Zhou, J. Ferguson, J. T. Chang, and Y. Kluger, "Inter-and intra-combinatorial regulation by transcription factors and micrornas," *BMC genomics*, vol. 8, no. 1, p. 1, 2007.

[8] C.-Y. Chen, S.-T. Chen, C.-S. Fuh, H.-F. Juan, and H.-C. Huang, "Coregulation of transcription factors and micrornas in human transcriptional regulatory network," *BMC bioinformatics*, vol. 12, no. 1, p. 1, 2011.

[9] D. H. Tran, K. Satou, T. B. Ho, and T. H. Pham, "Computational discovery of mir-tf regulatory modules in human genome," *Bioinformation*, vol. 4, no. 8, pp. 371–377, 2010.

[10] Y. Qi and H. Ge, "Modularity and dynamics of cellular networks," *PLoS Comput Biol*, vol. 2, no. 12, p. e174, 2006.

[11] Q. Cui, Z. Yu, E. O. Purisima, and E. Wang, "Principles of microrna regulation of a human cellular signaling network," *molecular systems biology*, vol. 2, p. 46, September 2006.

[12] X. Yang, M. Feng, X. Jiang, Z. Wu, Z. Li, M. Aau, and Q. Yu, "mir-449a and mir-449b are direct transcriptional targets of e2f1 and negatively regulate prbce2f1 activity through a feedback loop by targeting cdk6 and cdc25a," *Genes Dev.*, vol. 23, no. 20, pp. 2388–2393, October 2009.

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

[14] X. Peng, Y. Li, K.-A. Walters, E. R. Rosenzweig, S. L. Lederer, L. D. Aicher, S. Proll, and M. G. Katze, "Prediction of regulatory modules comprising micrornas and target genes," *BMC Genomics*, vol. 10, p. 373, 2009.

[15] S. Zhang, C.-C. Liu, W. Li, H. Shen, P. W. Laird, and X. J. Zhou, "Discovery of multi-dimensional modules by integrative analysis of cancer genomic data," *Nucleic Acids Research*, vol. 40, no. 19, pp. 9379–9391, October 2012.

[16] B. P. Lewis, C. B. Burge, and D. P. Bartel, "Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microrna targets," *Cell*, vol. 120, no. 1, pp. 15–20, January 2005.

[17] W. E, C. X, H. R, K. H, L. I, M. V, M. T, P. M, R. I, and S. F, "Transfac: an integrated system for gene expression regulation," *Nucleic Acids Research*, vol. 28, no. 1, pp. 316–319, January 2000.

[18] C. Stark, B.-J. Breitkreutz, A. Chatr-aryamontri, L. Boucher, R. Oughtred, M. S. Livstone, J. Nixon, K. V. Auken, X. Wang, X. Shi, T. Reguly, J. M. Rust, A. Winter, K. Dolinski, and M. Tyers, "The biogrid interaction database: 2011 update," *Nucleic Acids Research*, vol. 39, pp. D698–D704, Novermber 2011.

[19] J. C. Huang, T. Babak, T. W. Corson, G. Chua, S. khan,IEEE TRANSACTION ON NANOBIOSCIENCE , VOL.16 , NO.1 , JAN.2017
B. L. Gallie, T. R. Hughes, B. J. Blencowe, B. J. Frey, and Q. D. Morris, "Using expression profiling data to identify human microrna targets," *Nature Methods*, vol. 4, pp. 1045– 1049, 2007.

[20] Y. Lu, Y. Zhou, W. Qu, M. Deng, and C. Zhang, "A lasso regression model for the construction of microrna-target regulatory networks," *Bioinformatics*, vol. 27, no. 17, pp. 2406–2413, July 2011.

[21] C. Cheng and L. M.Li, "Inferring microrna activities by combining gene expression with microrna target prediction," *PLoS one*, vol. 3, no. 4, p. e1989, April 2008.

[22] C. H. Ooi, H. K. Oh, H. Z. Wang, A. L. K. Tan, J. Wu,
M. Lee, S. Y. Rha, H. C. Chung, D. M. Virshup, and P. Tan, "A densely interconnected genome-wide network of micrornas and oncogenic pathways revealed using gene expression signatures," *PLoS genetics*, vol. 7, no. 12, p. e1002415, December 2011.

[23] S. Zhang, Q. Li, J. Liu, and X. J. Zhou, "A novel computational framework for simultaneous integration of multiple types of genomic data to identify microrna-gene regulatory modules," *Bioinformatics*, vol. 27, no. 13, pp. i401–i409, 2011.

[24] S. Zhang, C.-C. Liu, W. Li, H. Shen, P. W. Laird, and X. J. Zhou, "Discovery of multi-dimensional modules by integrative analysis of cancer genomic data," *Nucleic acids research*, p. gks725, 2012.

[25] R. McLendon, A. Friedman, and D. Bigner, "Comprehensive genomic characterization defines human glioblastoma genes and core pathways," *nature*, vol. 455, pp. 1061–1068, 2008.

[26] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth, "limma powers differential expression analyses for rna-sequencing and microarray studies," *Nucleic acids research*, p. gkv007, 2015.

[27] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, "Biogrid: a general repository for interaction datasets," *Nucleic acids research*, vol. 34, no. suppl 1, pp. D535–D539, 2006.

[28] E. P. Consortium, "The encode (encyclopedia of dna elements) project," *Science*, vol. 306, no. 5696, pp. 636–640, 2004.

[29] Y. Li, C. Qiu, J. Tu, B. Geng, J. Yang, T. Jiang, and Q. Cui, "Hmdd v2. 0: a database for experimentally supported human microrna and disease associations," *Nucleic acids research*, p. gkt1023, 2013.

[30] L. Cheng, J. Li, P. Ju, J. Peng, and Y. Wang, "Semfunsim: A new method for measuring disease similarity by integrating semantic and gene functional association," *PLoS ONE*, vol. 9, no. 6, p. e99415, June 2014.

[31] P. M. Kim and B. Tidor, "Subsystem identification through dimensionality reduction of large-scale gene expression data," *Genome Research*, vol. 13, pp. 1706–1718, 2003.

[32] J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, "Metagenes and molecular pattern discovery using matrix factorization," *Proceedings of the national academy of sciences*, vol. 101, no. 12, pp. 4164–4169, 2004.
[33] K. Devarajan, "Nonnegative matrix factorization: an analytical and interpretive tool in computational biology," *PLoS Comput Biol*, vol. 4, no. 7, p. e1000029, 2008.
[34] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.
[35] Z. Yang and G. Michailidis, "A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data," *Bioinformatics*, p. btv544, 2015.
[36] F. Shahnaz, M. W. Berry, V. P. Pauca, and R. J. Plemmons, "Document clustering using nonnegative matrix factorization," *Information Processing and Management*, vol. 42, no. 2, pp. 373–386, 2006.
[37] T. H. Hwang, G. Atluri, R. Kuang, V. Kumar, T. Starr,K. A. Silverstein, P. M. Haverty, Z. Zhang, and J. Liu, "Large-scale integrative network-based analysis identifies common pathways disrupted by copy number alterations across cancers," *BMC genomics*, vol. 14, no. 1, p. 440, 2013.
[38] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, Conference Proceedings, pp. 556–562.
[39] S. Falcon and R. Gentleman, "Using gostats to test gene lists for go term association," *Bioinformatics*, vol. 23, no. 2, pp. 257–258, November 2007.
[40] S. Marsigliante, C. Vetrugno, and A. Muscella, "Ccl20 induces migration and proliferation on breast epithelial cells," *Journal of cellular physiology*, vol. 228, no. 9, pp. 1873– 1883, 2013.
[41] Y. Myal, E. Leygue, and A. A. Blanchard, "Claudin 1 in breast tumorigenesis: revelation of a possible novel claudin high subset of breast cancers," *BioMed Research International*, vol. 2010, 2010.
[42] Z. C. D'Costa, C. A. Higgins, C. W. Ong, G. Irwin, D. Boyle, D. G. McArt, K. McCloskey, N. E. Buckley, N. T. Crawford, and L. Thiagarajan, "Tbx2 represses cst6 resulting in uncontrolled legumain activity to sustain breast cancer proliferation: a novel cancer-selective target pathway with therapeutic opportunities," *Oncotarget*, vol. 5, no. 6, pp. 1609–1620, 2014.
[43] W. Zhang, C. Shen, C. Li, G. Yang, H. Liu, X. Chen, D. Zhu, H. Zou, Y. Zhen, and D. Zhang, "mir-577 inhibits glioblastoma tumor growth via the wnt signaling pathway," *Molecular carcinogenesis*, vol. 55, no. 5, pp. 575–585, 2016.
[44] F. Zhou, S. Li, H.-M. Meng, L.-Q. Qi, and L. Gu, "Microrna and histopathological characterization of pure mucinous breast carcinoma," *Cancer biology & medicine*, vol. 10, no. 1, p. 22, 2013.
[45] M. n. Castilla, M. n. Lpez-Garcła, M. R. Atienza, J. M. RosaRosa, J. Dłaz-Martłn, M. L. Pecero, B. Vieites, L. RomeroPrez, J. Benłtez, and A. Calcabrini, "Vgll1 expression is associated with a triple-negative basal-like phenotype in breast cancer," *Endocrine-related cancer*, vol. 21, no. 4, pp. 587–599, 2014.