Inferring Condition-Specific miRNA-Gene Modules from miRNA and mRNA Profiling Data

Abstract

Recently, a class of small RNA molecules, microRNAs or miRNAs, has attracted interest from researchers for their unique role in post-transcriptional regulation. Due to their distinct cell-type/tissue-specific expression patterns, it is of high importance to identify condition-specific miRNA-gene modules for a complete depiction of gene regulatory networks. In this paper, we propose a novel method to integrate miRNA and mRNA data to identify condition-specific miRNA-gene modules. Specifically, a Consistency-based Masking Nonnegative Matrix Factorization (CMNMF) method is developed to incorporate existing biological constraints (like the repression of miRNAs on potential target genes) with simultaneous miRNA and mRNA profiling data for an improved performance in module identification. The experimental results on simulation data show that the condition-specific modeling framework improves the performance in predicting miRNA-gene relationships. More importantly, application of CMNMF to human colon cancer data revealed a biologically significant miRNA-gene module, which contains four up-regulated miRNAs (miR-182, miR-183, miR-221 and miR-222) and six down-regulated target genes annotated as cytotoxicity mediated by nature killer cells. The proposed method can also be applied to various biological conditions, even with limited number of samples, to elucidate miRNA-involved gene networks.

1. Introduction

microRNAs (miRNAs) are novel regulators of gene networks to modulate mRNAs stability and translation post-transcriptionally [1, 2]. Recent studies also show that miRNAs may have an intrinsic function in tumor suppression [3]. One of the important observations on miRNAs is their cell-type/tissue-specific expression patterns [4, 5], which makes it possible to classify different types of human tumors according to miRNAs profiling [6].

It is important to discover miRNA-gene regulatory networks through profiling studies. The first key step is to identify miRNA-gene modules with a set of miRNAs and their target genes. Although several target prediction tools are now available [7], there are many false positives in the predicted miRNA targets (such as those in miRBase database [8]) solely based on the sequence match between miRNAs and their targets. To reduce the false positives, a common strategy is to combine the predicted target information with simultaneous profiling of miRNA and mRNA [9, 10].

Recent progress on computational biology highlights the use of condition-specific modeling on gene regulatory networks, as a complement to the previous wave of genome-level studies [6, 11-13]. Evidently, approaches with careful consideration of condition-specific regulation mechanisms often lead to more biologically relevant hypotheses for further validation. In our case, considering the highly cell-type/tissue-specific expression patterns of miRNAs, we propose to identify tissue-specific miRNA-gene modules with up- and down-regulated mechanisms, respectively. The rationales behind developing such condition-specific modules are summarized as follows:

1. There are contradictory results on miRNAs expression: most miRNAs are down-regulated in several types of solid tumors [6], but are up-regulated in another independent study [14]. The authors rightly note in their study [14] that contradictions are partially caused by the inappropriate application of data analysis across different types of tissues in the previous paper [6]. Therefore, instead of mixing all different tissue types, we prefer to develop specific models for each type of tissue.

2. Although the traditional view of miRNAs regulation role is mainly on their post-transcriptional repression of genes, a recent report claimed that miRNAs can also up-regulate translation via a different mechanism [15]. Thus the regulation mechanism of miRNAs can be either mRNA degradation, translation repression, or translation activation. Since the effect of miRNAs on translation may not be traceable solely by mRNA profiling (some post-transcriptional events may not be evident in the transcriptome profile), the simultaneous profiling of miRNAs and mRNAs might mainly reflect the mRNA degradation mechanism. Therefore in the present study we focus on the repression of miRNAs on target genes.
Besides above-mentioned condition-specific constraints, there are several data related issues that need to be carefully considered. These issues exist mainly due to the initial stage of miRNA profiling technology. Some typical issues are: 1) only a few data sets include miRNA and mRNA values measured concurrently on a genome-wide scale, 2) a smaller fold change of miRNAs is detected in some conditions (e.g., human cancer) when compared with that of mRNAs, and 3) low signal-to-noise ratio exists in most miRNA profiles.

In this paper, we propose a novel method to find condition-specific miRNA-gene modules. The term "condition-specific" reflects that the modules are identified from one given state (tumor vs. normal) in a given type of tissues (e.g., colon). The proposed method, namely Consistency-based Masking Non-negative Matrix Factorization (CMNMF), integrates biological constraints, like the information on potential target genes of miRNAs, with simultaneous miRNA and mRNA profiling data to infer condition-specific miRNA-gene modules. We have applied the CMNMF approach to a public simultaneous miRNA and mRNA profiling data set of multiple types of human cancers [6]. The experimental results show that our method can be used to identify important condition-specific miRNA-gene modules. For example, one of the modules in colon tumor is related to cytotoxicity mediated by nature killer cells, which is generally considered as the first defense line on tumorigenesis.

The paper is organized as follows. In Section 2, we give a detailed description of the proposed approach for the identification of condition-specific miRNA-gene modules (CMNMF approach). In Section 3, we present the experimental results on simulation data and real data, and demonstrate the biological significance of an identified miRNA-gene modules. Finally we conclude the paper with some remarks and future directions in Section 4.

2. Method

Linear modeling has been widely used to model transcriptional regulation [16], in which expression variation can be regarded as a linear combination of transcription factor activities:

\[ E = AT, \]  

where \( E \) is the log-ratio of gene expression and \( T \) is the log-ratio of transcription factor activity. Connection matrix \( A \) indicates the regulation strength from a specific transcription factor to its downstream targets. Because a transcription factor can either up-regulate or down-regulate the expression of its targets, the regulation strength \( A \) can be either positive or negative.

As mentioned previously, we only focus on the repression of miRNAs on target genes and the regulation strength of miRNAs is always negative. We denote the strength of the negative regulation as \( A^- \). We then separate the downstream targets into two non-overlapped groups: up-regulated group \( E^- \) and down-regulated group \( E^+ \). Signals in the up-regulated group \( E^- \) might be caused by down-regulation of miRNAs (\( Mr^- \)); similarly, members of the down-regulated group \( E^- \) might be caused by up-regulation of miRNAs (\( Mr^+ \)).

\[ E = \begin{bmatrix} E^+ \\ E^- \end{bmatrix} = \begin{bmatrix} A^- Mr^- \\ A^+ Mr^+ \end{bmatrix}. \]  

Fig. 1. Illustration of the matrices used for miRNA-gene module discovery.

To model the repression effects of active miRNAs on mRNAs, we focused on the up-regulated miRNAs group and the corresponding down-regulated mRNAs group:

\[ E^- = A^- Mr^+. \]  

Apparently, all the items in \( E^- \) and \( A^- \) are negative, and all the items in \( Mr^+ \) are positive. If we multiply a negative sign on both side of (3), and denote \( G = -E^- \), \( B = -A^- \) and \( M = Mr^+ \), we have a non-negative decomposition form. That is, \( G \) is a product of two non-negative matrices \( B \) and \( M \):

\[ G = BM \]  

To recover \( B \) from Equation (4), Non-negative Matrix Factorization (NMF), a well-known numerical decomposition method, can be used [17]. However, direct application of NMF to obtain a solution is unlikely to make \( B \) biologically meaningful. To overcome this difficulty, we propose to introduce a concept called regulatory pattern to regularize the solution space. Given a matrix \( B \) and a set \( R_0 = \{(i, j) : \forall i \in [1, \ldots, N], \forall j \in [1, \ldots, N]\} \),
we say that $B$ is characterized by a regulatory pattern $R_0$ (and we denote such property as $B \Downarrow R_0$) if and only if:

$$b_{ij} \equiv 0 \text{ for } (i, j) \notin R_0.$$  

(5)

The regulatory pattern provides us biologically relevant constraints to make the non-negative decomposition result meaningful for miRNA-gene modules inference. We will name this method as Masking Non-negative Matrix Factorization (MNMF). To obtain the solution to our MNMF problem, we can formulate it as an optimization problem:

$$\min_{B,M, s.t. \ B \sim R_0} \| G - BM \|^2 .$$  

(6)

An appropriate regulatory pattern can be directly obtained through miRNA target prediction algorithms or related database [8]. The optimization procedure is quite straightforward as summarized in Algorithm 1, with a plug-in step to enforce $B$ following regulatory pattern $R_0$.

Note that in Algorithm 1, .* is the operator of the element-by-element product of the two matrices of the same dimension, ./ is the operator of the element-by-element division of the two matrices of the same dimension, superscript $T$ denotes matrix transpose, and $\sigma$ is a very small positive constant to keep numerical iteration stable.

**Algorithm 1:**

Randomly initialize $B$ and $M$;

for $i = 1 : \text{max\_iteration}$

$M = M . \ast (B^T G . / (B^T BM + \sigma))$ ;

$B = B . \ast (GM^T . / (BMM^T + \sigma))$ ;

Enforce $B \sim R_0$ .

end

When simultaneous profiling measurements of miRNAs and mRNAs are available, linear regression can be utilized to directly infer the regulation relationship $B$, provided that both measurements are of good signal quality. However, in the data set we used here, the signal-to-noise ratio of miRNA measurements is much lower than that of mRNA (data not shown). Direct application of linear regression will result in a poor estimation of the regulation relationship. With the awareness of quality limitation in miRNA measurements and uncertainty or incompleteness in knowledge of the available potential regulation relationship of miRNAs, we believe that the integrated utilization of both miRNA measurements and miRNA target information can improve the estimation of regulation relationship $B$. In particular, we introduce a practical assumption that the estimated miRNA expressions should not be too far away from the measured miRNA expressions. With such an assumption, we can formulate the estimation of regulation relationship $B$ into a modified form of MNMF, called Consistency-based Masking Non-negative Matrix Factorization (CMNMF):

$$\min_{B,M, s.t. \ B \sim R_0} \left\{ \| G - BM \|^2 + \lambda \| M - M_0 \| \right\} .$$  

(7)

where $M_0$ is a matrix representing miRNA measurements and $\lambda$ is a parameter to control the tradeoff between the two items in (7). Intuitively, the first item of (7) is used to estimate the regulation relationship $B$ to explain the mRNA expression variation; the second item is used to force the estimated miRNA expression values to be reasonably close to the experimental measurements. The controlling parameter $\lambda$ can be determined by the L-curve method [18]. When $\lambda$ is given, the optimization procedure can be summarized as in Algorithm 2. Note that in Algorithm 2, $\sigma$ is a very small positive constant to keep numerical iteration stable, and $\alpha$ and $\beta$ are sufficiently small positive constants.

**Algorithm 2:**

Randomly initialize $B$ and $M$;

for $i = 1 : \text{max\_iteration}$

$M = M . \ast (B^T E . / (B^T BM + \beta M + \sigma))$ ;

$B = B . \ast (EM^T . / (BMM^T + \alpha B + \sigma))$ ;

Enforce $B \sim R_0$ .

end

3. Experimental Results

3.1 Performance evaluation on simulation data

To test the proposed MNMF and CMNMF algorithms, we generated a simulation data set with 5 miRNAs regulating 1000 genes in 5 samples. To further make the simulation close to the real scenario, we purposely introduced 10% false-positives and 20% false-negatives in regulatory pattern $R_0$, and added noise to make the signal-to-noise ratio of miRNA measurements ($M_0$) at ~5dB level. This simulation set-up reflects the fact that both miRNA target prediction information and miRNA measurements are noisy. With the simulated data, we conducted performance evaluation of the proposed algorithms (i.e., MNMF and CMNMF) as well as the traditional NMF algorithm. A performance index is designed for performance comparison using averaged absolute Pearson correlation between estimated $B$ and true regulation relationship. The performance indices on the
simulation data are summarized in Table 1. It can be seen from Table 1 that both MNMF and CMNMF outperformed conventional NMF, which demonstrated the importance of biologically relevant constraints. The performance of CMNMF is slightly better than that of MNMF as shown in Table 1, although theoretically CMNMF would have more significant improvement over MNMF if $R_0$ is noisier or $M_0$ is more precise.

<table>
<thead>
<tr>
<th>Performance index</th>
<th>NMF</th>
<th>MNMF</th>
<th>CMNMF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Averaged absolute Pearson correlation</td>
<td>0.5737</td>
<td>0.7225</td>
<td>0.7371</td>
</tr>
</tbody>
</table>

Table 1. The performance comparison of the proposed algorithms

3.2 Results on real human cancer data

We applied the MNMF and CMNMF approaches to a real data set on human cancer [6]. Here we focused on the colon samples including 7 tumor samples and 4 normal samples. To initialize matrix $B$, we used pre-ensemble predicted targets of all human miRNAs in the miRBase database [8]. We first calculated the expression ratio of each miRNA between a tumor sample and normal sample by dividing the median of the miRNA expression across all normal samples, followed by a log transformation. The transformed values (log-ratios) reflect the change in each miRNA's expression values in tumor samples compared with normal samples.

Taking into account the relatively low quality of miRNA profiling data, we made a detailed examination of the signal-to-noise ratios of all miRNAs. In this study, only up-regulated miRNAs with signal-noise-ratios $>1.5$ were selected for further analysis. As a result, there are four miRNAs (miR-182, miR-183, miR-221 and miR-222) that satisfy the above criteria. The significant over-expression of these miRNAs in tumors is consistent with the finding that miR-221/222 are over-expressed and can be regarded as a new family of oncogenes, directly targeting the tumor suppressor gene p27 [19-21].

To properly determine the tradeoff parameter $\lambda$ for minimizing both items in Equation (7), we used the L-curve method, which was originally proposed to select regularization parameter for solving ill-conditioned linear equation [18]. Fig. 2 shows the L-curve result on the colon tumor data, in which both items in Equation (7) were plotted as x-axis and y-axis, and the horizontal and vertical lines indicates over-regularization and under-regularization respectively. Therefore the best tradeoff point is the corner, where both items in Equation (7) achieve reasonable small values in order to fulfill the minimization requirement.

After obtaining $B$ and $M$ with our CMNMF approach, we defined a Contribution Score ($CS$) to reflect the relative contribution of each miRNA on regulation of a given gene. Specifically, for the $i$-th gene in the $k$-th sample, we calculate the contribution of the $j$-th miRNA on the regulation of this gene, ($B_{ij}^*_M$), then calculate its ratio to the change of the $i$-th gene. Further, we choose the significant miRNA-gene modules from those that satisfy the following condition: the $CS$ mean value divided by its standard deviation across different samples is greater than two. The miRNA-gene modules identified by our approach are illustrated in Fig.3, which shows four miRNA-gene modules by the edges from the miRNAs to their target genes, respectively, in the given condition (human colon cancer). Several genes are linked by multiple miRNAs, which is consistent with the known biological understanding that one gene may be regulated by several miRNAs.

Fig. 3. The identified miRNA-gene modules in human colon tumor.
3.3 Biological interpretation of the identified miRNA-gene modules

To evaluate the biological significance of the identified miRNA-gene modules, we conducted Gene Ontology (GO) enrichment analysis using DAVID [22]. Especially we used original candidate target genes generated by mirBase as background rather than using the whole human genome, serving as a stricter test to focus on the enriched gene functions identified by our method.

Among the total 175 miRNA-gene pairs obtained, there are 161 target genes annotated in DAVID. There exist significant enrichments in "glycoprotein" (38 genes, p-value = 6.0E-3) and "oxidoreductase" (12 genes, p-value = 1.4E-2). It suggests that these miRNAs might inhibit immuno-recognition or cell adhesion (function of glycoprotein), and also have influence on oxide-reduction regulation. This result is consistent with the fact that oxidative stress is one of important mechanisms of tumorigenesis. And more interestingly, there are two pathways that are significantly enriched. The first pathway is "nature killer cells mediated cytotoxicity" (6 genes, p-value = 3.9E-2), and the second one is "calcium signaling pathway" (7 genes, p-value = 5.0E-2). Natural killer (NK) cells play a critical role in cytokine production and have been considered as a first line of defense against tumorigenesis [23, 24]. Triggering NK cell mediated cytotoxicity could serve as a novel therapeutic strategy for the treatment of human cancer [25, 26]. The related miRNAs and its target genes that are annotated by "natural killer cell mediated cytotoxicity" are listed in Table 2. The subnetworks of the functional miRNA-gene module (inhibition of NK cell mediated cytotoxicity) discovered by our method are illustrated in Fig. 4. It is worth noting that three different miRNAs (i.e., miR-182, miR-183 and miR-222) work cooperatively to contribute to the inhibition effect.

Table 2. Genes annotated by "nature killer cell mediated cytotoxicity"

<table>
<thead>
<tr>
<th>miRNA</th>
<th>Gene</th>
<th>Gene Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>hsa-miR-222</td>
<td>ICAM1</td>
<td>intercellular adhesion molecule 1 (CD54), human rhinovirus receptor</td>
</tr>
<tr>
<td></td>
<td></td>
<td>protein phosphatase 3 (formerly 2B), catalytic subunit, gamma isoform</td>
</tr>
<tr>
<td>hsa-miR-182</td>
<td>PPP3CC</td>
<td>tyrosine kinase binding protein</td>
</tr>
<tr>
<td>hsa-miR-183</td>
<td>TYROBP</td>
<td>TYRO protein tyrosine kinase binding protein</td>
</tr>
<tr>
<td>hsa-miR-182</td>
<td>LCP2</td>
<td>lymphocyte cytosolic protein 2 (SH2 domain containing leukocyte protein of 76kDa)</td>
</tr>
<tr>
<td>hsa-miR-222</td>
<td>NFATC3</td>
<td>nuclear factor of activated T-cells, cytoplasmic, calcineurin-dependent 3</td>
</tr>
<tr>
<td>hsa-miR-222</td>
<td>PIK3CD</td>
<td>phosphoinositide-3-kinase, catalytic, delta polypeptide</td>
</tr>
</tbody>
</table>

4. Conclusions

In this paper, we have presented a novel computational approach for the identification of condition-specific miRNA-gene modules. The proposed approach, namely Consistency-based Masking Non-negative Matrix Factorization (CMNMF), integrates the existing biological information (in particular, potential target genes of miRNAs) with miRNA and mRNA profiling data for an improved performance in module identification. The approach has been tested on a simulation data set to show its superior performance over traditional NMF approach. The experimental results on a real human colon data set have further demonstrated its capability for miRNA-gene module discovery. Specifically, one of the identified condition-specific miRNA-gene modules in human colon tumor is comprised of three oncogene-like miRNAs (miR-182, miR-183 and miR-222), which might play an important role in tumorigenesis via inhibiting genes related to nature killer (NK) cells activity. The inhibition of NK cell activity might affect early cell mediated immune surveillance of aberrant cells and also contribute to the escape from apoptosis or cell death by modifying the secretion of cytotoxic cytokines.

The current condition-specific modeling framework incorporated recent understanding on miRNAs, i.e., cell-type/tissue-specific expression patterns, different regulation mechanism (mRNA degradation, translation repression or translation activation). As more mechanisms of miRNAs are revealed, the framework could be adapted to discover miRNA-gene modules in different conditions and lead to a comprehensive overview of gene regulation networks.

Acknowledgments. This research was supported in part by NIH Grants (NS29525-13A, EB000830, CA109872 and CA096483) and a DoD/CDMRP grant (BC030280).
5. References


