Brief Communication

Identifying the target mRNAs of microRNAs in colorectal cancer

Shinuk Kim a, Minsoo Choi a,b, Kwang-Hyun Cho b,*

a Interdisciplinary Program in Bioinformatics, Seoul National University, 599 Gwanakno, Gwanak-gu, Seoul 151-747, Republic of Korea
b Department of Bio and Brain Engineering and KI for the Bio Century, Korea Advanced Institute of Science and Technology (KAIST), 335 Gwahangno, Yuseong-gu, Daejeon 305-701, Republic of Korea

A R T I C L E   I N F O

Article history:
Received 4 March 2008
Received in revised form 7 July 2008
Accepted 7 July 2008

Keywords:
MicroRNAs
Target mRNAs
Optimization
Linear system equations

A B S T R A C T

MicroRNAs (miRNAs) play an important role in gene regulatory networks by inhibiting the expression of target mRNAs. Various experimental studies have been carried out to unravel the role of miRNAs in cancer development (Cummins et al., 2006; Lim et al., 2005; Zhang et al., 2007) and some relationships between miRNAs and their target mRNAs were identified (Joung et al., 2006; Yoon and Micheli, 2005; Kirakiidou et al., 2004). However, it still remains as a challenging issue to develop a computational method for identification of such relationships in cancer. Computational methods have been developed in two ways so far: identifying miRNAs conserved in different species or stem loop prosecutors (Krek et al., 2005), or identifying the relationship between miRNA and its target mRNAs by using sequence homologues. Such methods provide us with useful information, but they result in too many false positives with which we have difficulties in studying particular miRNAs in cancer. Experimentally, a novel miRNA microarray method was recently proposed, called a bead-based detection, to obtain high throughput expression profiles including 217 miRNA genes in primary tumors (Lu et al., 2005). With such a large volume of expression profiles, we can apply a numerical optimization method. Previous studies showed that multiple miRNAs can target one mRNA, or conversely, one miRNA can also target multiple mRNAs. Yoon and Micheli (2005) suggested that a method of identifying multi-to-multi-relationships between miRNAs and mRNAs. However, such computational methods do not make use of experimental data. So, in this paper, we present a new mathematical formulation and computational method to identify the multi-to-multi-relationships by using microarray profiles.

The relationships between miRNAs and their target mRNAs are modeled by linear system equations and the parameter identification problem of this linear matrix equation is then formulated as a multidimensional mathematical optimization problem. We apply Broyden–Fletcher–Goldfarb–Shannon (BFGS) (Press et al., 1992) optimization method to this problem since it converges faster than other methods as it is based on an approximated Hessian matrix. In particular, we considered colorectal cancer and formulated a linear system model on the relationships between 22 miRNAs and 22 mRNAs involved in colorectal cancer. The proposed method properly identified 207 relationships out of 484 in total. Among those, the number of major relationships was 16 out of 30 which were verified through previous experimental evidences. Moreover, the most valuable findings were all 8 out of the 16 predictions in that they have not been predicted by any other previous computational method.

The rest of the paper is organized as follows. Section 2 describes the mathematical formulation and the proposed identification method. Section 3 shows the identification results of the proposed method and presents the confirming experimental evidences from literature. Finally, conclusions and the future research directions are described in Section 4.

1. Introduction

MicroRNAs (miRNAs) composed of 19–22 nt play important regulatory roles in post-transcriptional gene regulation by targeting mRNAs for translational repression. Recently, various studies have been carried out to unravel the role of miRNAs in cancer development (Cummins et al., 2006; Lim et al., 2005; Zhang et al., 2007) and some relationships between miRNAs and their target mRNAs were identified (Joung et al., 2006; Yoon and Micheli, 2005; Kirakiidou et al., 2004). However, it still remains as a challenging issue to develop a computational method for identification of such relationships in cancer. Computational methods have been developed in two ways so far: identifying miRNAs conserved in different species or stem loop prosecutors (Krek et al., 2005), or identifying the relationship between miRNA and its target mRNAs by using sequence homologues. Such methods provide us with useful information, but they result in too many false positives with which we have difficulties in studying particular miRNAs in cancer. Experimentally, a novel miRNA microarray method was recently proposed, called a bead-based detection, to obtain high throughput expression profiles including 217 miRNA genes in primary tumors (Lu et al., 2005). With such a large volume of expression profiles, we can apply a numerical optimization method. Previous studies showed that multiple miRNAs can target one mRNA, or conversely, one miRNA
2. Model and Method

The mathematical formulation of the proposed identification method is described in this section.

2.1. Mathematical Formulation

In general, one miRNA degrades or represses the translation of either one or multiple mRNAs during post-transcription while one mRNA has an effect on several miRNAs ($x_1, x_2, \ldots, x_m$). So, we have a linear equation model of this mechanism as follows:

$$y_i = a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{im}x_m, \quad i = 1, \ldots, n$$  \hspace{1cm} (1)

where the matrix elements, $a_{ij}$s, describe the influence of the $j$th miRNA on $i$th mRNA, $x_j$ represents the expression level of the $j$th miRNAs, and $y_i$ represents the expression level of the $i$th mRNA. For the whole relationships of $n$ mRNAs, we have the following matrix representation:

$$ \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}. \hspace{1cm} (2)$$

From experiments, we take measurements of the $n$ different miRNAs simultaneously, denoted by $(x_1, \ldots, x_m)^T$. We also have the measurements of mRNA expression levels, denoted by $(y_1, \ldots, y_n)^T$.

By repeating this procedure $K$ times, we have results in the form as follows:

$$X_{M \times K} = \begin{bmatrix} x_1^1 & x_2^1 & \cdots & x_m^1 \\ x_1^2 & x_2^2 & \cdots & x_m^2 \\ \vdots & \vdots & \ddots & \vdots \\ x_1^K & x_2^K & \cdots & x_m^K \end{bmatrix}, \quad Y_{N \times K} = \begin{bmatrix} y_1^1 & y_2^1 & \cdots & y_n^1 \\ y_1^2 & y_2^2 & \cdots & y_n^2 \\ \vdots & \vdots & \ddots & \vdots \\ y_1^K & y_2^K & \cdots & y_n^K \end{bmatrix}, \hspace{1cm} (3)$$

Fig. 1. A flow chart of the proposed numerical scheme. (1) Two sets of microarray data ($x$ and $y_{real}$) are required for initial inputs. (2) Numerical mRNA data sets are generated from the computational algorithm $y(x, a)$ where $a$'s are the initial guess on parameter estimates. (3) Error function gives us the criterion for whether new parameter estimation is required or to stop here. (4) Optimization method gives us new parameter estimates. The illustration is adopted from (Hanselman and Littlefield, 1996)
with the relation matrix $A$,

$$A_{N \times M} = \begin{bmatrix}
a_{11} & a_{12} & \cdots & a_{1m} \\
a_{21} & a_{22} & \cdots & a_{2m} \\
\vdots & \vdots & \ddots & \vdots \\
a_{n1} & a_{n2} & \cdots & a_{nm}
\end{bmatrix}. \tag{4}$$

We can rewrite (1) for $K$ times experiments as follows:

$$\sum_{k=1}^{K} y_N^k = \sum_{k=1}^{K} A_{N \times M} x_M^k. \tag{5}$$

2.2. Computational Scheme

In this section, we present a computational identification scheme of the interaction relationships between miRNAs and their target mRNAs. The proposed scheme is composed of three main components: the direct solver, the optimization routine, and the objective function. The role of direct solver is to generate computational data by solving the linear system equations. The optimization routine is to find a new set of parameter estimates using the generated computational data. In particular, we employ the BFGS method for optimization which is a quasi-Newton method based on an approximated Hessian matrix composed of second derivatives of the objective function. The approximated Hessian is positive definite as well as symmetric, and thereby the objective function converges to a minimum within quadratic Newton steps. The applicability and performance of BFGS in global optimization were proven by the previous study (Kim et al., 2007). Finally, the objective function provides us with the criterion for further processing to the next iteration based on integrated error norms. For the error norm $f(a_{ij})$, we employed $f(a_{ij}) = |y(x, a_{ij}) - y_{real}|$ since $L_1$ norm gave the best results in our previous studies (Kim and Kreider, 2006; Kim et al., 2007).

Initially, we use miRNA expression data $x_M^k$, mRNA expression data $y_N^k$, and initial parameter guess $a_{ij}$ as inputs. Here, we set $a_{ij}$ to zero since we have no a priori information. In the next step, we generate computational mRNA data from (4) and then compare these with real expression mRNA data $y_N^k$. If the computational data well estimate the real data, then we can predict the relationship between miRNAs and their target mRNAs by analyzing $a_{ij}$; if not, then we need new $a_{ij}$ generated after further optimization. Fig. 1 describes the overall scheme of the proposed computational algorithm. The algorithm iterates until the tolerance of $f(a_{ij})$ is less than $10^{-3}$.

3. Results

As we are interested in miRNAs and their target mRNAs that are involved in the regulation of colorectal cancer, we consider the miRNAs reported from colorectal cancer cells (Cummins et al., 2006). Among those, we select 22 miRNAs whose target mRNAs were experimentally known. So, we consider 22 miRNAs and 22 mRNAs with 484 relationships among which 30 experimentally verified relations are included. As a true solution, we add 30 more relations obtained from sequence binding pairs (Sethupathy et al., 2006).

The proposed identification method is applied to the colorectal cancer microarray expression profiles (Lu et al., 2005). Since miRNAs interrupt or regress their target mRNAs during a post-transcriptional process (Miranda et al., 2006; Lim et al., 2005; Mattick and Makunin, 2005), we only need to consider negative relations in the miRNA–mRNA interaction pairs.

3.1. Analysis of the Identified Relations

Among the total 484 relations, the proposed method identified 207 multiple miRNA–mRNA relations. With the identified parameter estimates, we reconstructed mRNA profiles. The comparison of real microarray profiles and numerically computed data that are generated from obtained relations by using the proposed method are exactly matched with each other, which implies that the parameters are well defined in the system model.

Fig. 2 shows the number of identified target mRNAs for each miRNA. Each miRNA has 9.4 targets on the average ranging from 4 (e.g., let-7) to 15 (e.g., miR-24). Each of the 8 miRNAs including miR-24, miR-223, miR-1, miR-30a, miR-16, miR-192, miR-19a, and miR-34a, inhibits more than 12 genes and occupy 51% target interactions out of 207 in total. This result implies that those

![Fig. 2. Estimation of the number of target mRNAs for each miRNA ranging from 4 to 15.](image-url)
miRNAs have more significant effects on colorectal cancer than others.

Fig. 3 shows the distribution of the number of miRNAs for each target gene. To identify more specific relations, we have incorporated 60 true relations—30 relations obtained from sequence analysis and 30 relations from previous experimental reports. Then, we have 44 identified relations. Fig. 4 illustrates the receiver operating characteristic (ROC) (Schroeder et al., 2006) curves showing the reliability of the proposed identification method with respect to different thresholds. With the proposed system model, we confirm that our method has consistently identified the true relations. The sensitivity 0.8 is the largest ratio which implies that true positive rate is 4 times that of false negative and the true negative rate is 1.5 times that of false positive.

Note that the proposed method has identified 16 relations out of 30 experimentally verified relations. Among these, let us focus on the eight relations which were verified experimentally but could not be identified through the previous sequence analysis. The eight relations are shown in Table 1. These findings exemplify the need for a new computational approach that can predict the relations using microarray expression profiles like the proposed method.

### Table 1

<table>
<thead>
<tr>
<th>miRNA</th>
<th>Target mRNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>miR-24</td>
<td>NOTCH1</td>
</tr>
<tr>
<td>miR-19a</td>
<td>PTEN</td>
</tr>
<tr>
<td>miR-192</td>
<td>SIP1</td>
</tr>
<tr>
<td>miR-17</td>
<td>E2F1</td>
</tr>
<tr>
<td>miR-30a</td>
<td>THBS1</td>
</tr>
<tr>
<td>miR-27b</td>
<td>NOTCH1</td>
</tr>
<tr>
<td>miR-34a</td>
<td>E2F3</td>
</tr>
<tr>
<td>miR34a</td>
<td>E2F1</td>
</tr>
</tbody>
</table>

These relations have never been predicted by previous computational methods based on sequence analysis, but are successfully identified by the proposed method and also be verified by previous experimental evidences.

3.2. Supporting Evidence from the Literature

The 16 predicted relations obtained by applying the proposed method are further supported from the experimental evidences found from various literatures. Fig. 6 shows the 16 predicted relations in colorectal cancer where 12 target genes are classified into two groups depending on their relation to the colorectal cancer. For instance, PTEN functions as a tumor suppressor while NOTCH1 and BCL2 are considered as oncogenes for colorectal cancer. PTEN is the target gene of miR-19a (Lewis et al., 2003) and known as a tumor suppressor (Delnatte et al., 2006; Jin et al., 2007). In particular, it is known that PTEN pathway plays an important role in colon carcinogenesis (Wang et al., 2007). Moreover, miR-24 and miR-27b regulate NOTCH1, the receptor protein of NOTCH signaling pathway (Fukuda et al., 2005), and NOTCH1 is known frequently activated in a wide range of human cancers (Gu et al., 2007). Furthermore, BCL2, the target gene of miR-16, is also identified in many types of human cancers including leukemias, lymphomas and carcinomas (Sanchez-Beato et al., 2003; Zhang et al., 2007). Overexpressed BCL2 proteins attenuate Sulindac sulfide-induced apoptosis in SW480 human colon cancer cells and constitutively suppress a novel proapoptotic function of p53 in colorectal cancer cells (Sinicrope and Penington, 2005; Jiang and Milner, 2003). Various studies have shown that E2F1 and E2F3 play either as oncogene or tumor suppressor depending on cellular context, but they are known to induce apoptosis in colorectal cancer. In particular, miR-17 negatively regulates the transcription factor E2F1 (O’Donnell et al., 2005) which inhibits the proliferation in human colon can-
In this paper, we proposed a computational identification method to unravel the relationships between miRNAs and their target mRNAs. The proposed method uses linear system equations to describe the multiple interaction relationships between miRNAs and mRNAs, and applies the BFGS optimization scheme to estimate the matrix elements. The method was applied to the colorectal cancer microarray gene expression profiles and could successfully uncover 16 relationships that can be verified through further experimental evidences from literature. Among those, eight relationships turned out very significant since they could not be discovered from sequence analysis alone. We might suggest that the newly unraveled miRNAs and their target mRNAs are deeply involved in the regulation of colorectal cancer. For instance, E2F1, E2F3, THBS1, PTEN, NOTCH1, MAPK14, HMGA2, BCL2, and SIP1 were already identified in colorectal cancer cell lines from previous studies. In particular, there is a recent report on that E2F1 and E2F3 are down-regulated after miR-34a transfection in colon cancer cell lines, HCT 116 and RKO (Tazawa et al., 2007). Unfortunately, no other experimental studies have yet been reported to confirm that both miRNAs and their target mRNAs are expressed in colorectal cancer.

The present study can be further extended to identify miRNAs that regulate certain mRNAs for a specific cancer, and can also be used to design a schematic model for particular molecular mechanisms of miRNA-involved cancer pathogenesis. We also note that we employed in this paper the miRNAs and mRNAs profiles conducted from independent experiments. So, if they are available from a same experiment under homogeneous environments, we can have better identification results. Moreover, we found that there are many false positives and negatives without sequence filtering. Therefore, to improve the identification result of the proposed numerical scheme, we further need a reinforced mathematical modeling and noise reduction in expression data measurements.

4. Discussion and Conclusions

References


