BIOINFORMATICS

Vol. 21 Suppl. 2 2005, pages ii93–ii100
doi:10.1093/bioinformatics/bti1116

Genes and Genomes

Prediction of regulatory modules comprising microRNAs and target genes

Sungroh Yoon[1,*] and Giovanni De Micheli[2]

[1]Computer Systems Laboratory, Stanford University, Stanford, CA 94305, USA and [2]Integrated System Center, EPF Lausanne, Switzerland

ABSTRACT

Motivation: MicroRNAs (miRNAs) are small endogenous RNAs that can play important regulatory roles via the RNA-interference pathway by targeting mRNAs for cleavage or translational repression. We propose a computational method to predict miRNA regulatory modules (MRMs) or groups of miRNAs and target genes that are believed to participate cooperatively in post-transcriptional gene regulation.

Results: We tested our method with the human genes and miRNAs, predicting 431 MRMs. We analyze a module with genes: *BTG2*, *WT1*, *PPM1D*, *PAK7* and *RAB9B*, and miRNAs: *miR-15a* and *miR-16*. Review of the literature and annotation with Gene Ontology terms reveal that the roles of these genes can indeed be closely related in specific biological processes, such as gene regulation involved in breast, renal and prostate cancers. Furthermore, it has been reported that *miR-15a* and *miR-16* are deleted together in certain types of cancer, suggesting a possible connection between these miRNAs and cancers. Given that most known functionalities of miRNAs are related to negative gene regulation, extending our approach and exploiting the insight thus obtained may provide clues to achieving practical accuracy in the reverse-engineering of gene regulatory networks.

Availability: A list of predicted modules is available from the authors upon request.

Contact: sryoon@stanford.edu

1 INTRODUCTION

MicroRNAs (miRNAs) are endogenous 21–22 nt RNAs that can play crucial regulatory roles in animals and plants by targeting transcripts for cleavage or translational repression (Bartel, 2004). Hundreds of different miRNAs have now been identified in complex eukaryotes, implying that they mediate a vast network of unappreciated regulatory interactions (Lai, 2004).

Computational methods have been applied to the studies of miRNAs largely in two ways. First, techniques to identify miRNA host genes have been proposed (Ohler *et al*., 2004; Lim *et al*., 2003; Rodriguez *et al*., 2004; Lai *et al*., 2003). These methods rely upon the observation that miRNAs generally derive from phylogenetically conserved stem-loop precursor RNAs with characteristic features. Second, given that miRNA target gene selection is guided by the sequence, algorithms have been suggested to systematically identify miRNA targets *in silico* (Lewis *et al*., 2003; John *et al*., 2004; Rajewsky and Socci, 2003; Kiriakidou *et al*., 2004; Smalheiser and Torvik, 2004; Stark *et al*., 2003; Rehmsmeier *et al*., 2004; Enright *et al*., 2003).

Typically, multiple miRNAs regulate one message, reflecting cooperative translational control. Conversely, one miRNA may have several target genes, indicative of target multiplicity (Enright *et al*., 2003). This multiplicity of targets and cooperative signal integration on target genes are key features of the control of translation by miRNAs (John *et al*., 2004). However, this many-to-many relationship between miRNAs and target genes is often complicated (e.g. see Fig. 5c), and we thus need an automated analysis tool.

In this paper, we mathematically formulate the biological observations on the interactions of miRNAs and their targets and present a way to identify important patterns hidden in the complex interactions. In particular, we propose a computational method to predict miRNA regulatory modules (MRMs) or groups of miRNAs and their targets that are believed to participate cooperatively in post-transcriptional gene regulation. The proposed method provides groups of miRNAs and co-targeted genes automatically. We can thus avoid manually enumerating combinations of miRNAs and their target genes (or vice versa), which can be prohibitively time-consuming.

We apply our method to the prediction of human MRMs and here report a predicted module that contains the genes: *BTG2*, *WT1*, *PPM1D*, *PAK7* and *RAB9B*, and the miRNAs: *miR-15a* and *miR-16*. As will be detailed later, it has been reported that these genes are mostly regulators and their anomaly can be found in breast, renal and prostate cancers (Struckmann *et al*., 2004; Kawakubo *et al*., 2004; Ficazzola *et al*., 2001). Interestingly, *BTG2*, *WT1* and *PPM1D* have been shown to be directly associated with the function of *p53*, a tumor-suppressor gene (Vogelstein *et al*., 2000). Furthermore, the human miRNAs *miR-15a* and *miR-16* are clustered on chromosome 13q14, and this region has been shown to be deleted together in several types of cancer (Li *et al*., 2002; Saito-Ohara *et al*., 2003). The annotation of this module with the terms in the database Gene Ontology (GO) (The Gene Ontology Consortium, 2000) also suggests that the genes in this module indeed share some common roles in biological processes.

The MRMs predicted can further be useful in some important tasks including the reconstruction of gene regulatory networks as well as the biological validation of miRNA-target duplexes. Specifically, the regulatory interactions newly revealed by MRMs may provide a missing piece in the puzzle of gene regulation mechanisms, enabling us to reverse-engineer more accurate gene regulatory networks. In addition, the genes included in MRMs can be reasonable candidates for the experimental validation of miRNA targets, since these genes are detected multiple times by distinct miRNAs. Focusing on the genes

*To whom correspondence should be addressed.

included in MRMs may be an effective way to design an experiment for target validation.

The remainder of this paper is organized as follows. Section 2 formally defines MRMs and presents our approach to predict them. Section 3 provides the details of our analysis of a predicted module through a literature review and annotation with GO.

2 METHOD

Our method consists of five major steps, each of which will be detailed in this section.

(1) Target identification: given a set of miRNAs, their target genes are identified (Section 2.1).

(2) Relation graph representation: the relation between miRNAs and their targets are represented by a weighted bipartite graph called relation graph (Section 2.2).

(3) Seed finding: a seed or a set of miRNAs that bind a common target with similar binding strength is identified (Section 2.3).

(4) Merging seeds to find candidate modules: the seeds found in the previous step are collected and merged to produce candidates for MRMs (Section 2.4).

(5) Post-processing: statistically significant MRMs are selected by computing the P-value or the probability of finding a module by chance (Sokal and Rohlf, 1994) (Section 2.5).

2.1 Identification of miRNA target sites

Target selection is guided by the miRNA sequence, as informally shown in Figure 1 (Lai, 2004). In plants, probable targets of most miRNAs can be found simply by searching for highly complementary sequences in mRNA coding sequences or untranslated regions (UTRs). In contrast, animal miRNAs do not generally exhibit extensive complementarity to any endogenous transcripts (Fig. 1a). Various configurations for miRNA–target duplexes are possible, as presented in Figure 1b. In particular, when multiple binding sites exist on a target, the strength of each binding is not too strong or weak but modest and similar, according to Lai (2004, page 115.2). This observation will be reflected in our mathematical formulation in Section 2.2.

We first brief the reader on the existing target identification techniques, upon which the first step of our method depends. Most algorithms to identify animal miRNA targets rely on three properties: (1) sequence complementarity using a position-weighted local alignment algorithm, (2) free energies of miRNA-target duplexes and (3) evolutionary conservation of target sites in homologous genes. In particular, the conservation filter tends to be the most predictive criterion for accurate target detection (Enright *et al.*, 2003). The complementarity displayed by a miRNA and its binding site is usually not enough to be statistically significant, since a miRNA is only 21–22 nt long. Thus, this conservation filter plays a crucial role in reducing the number of false positives.

In the first step of our method, we identify miRNA–mRNA duplexes by the method described in Lewis *et al.* (2003) and John *et al.* (2004). (Other methods can also be used as long as they can quantify the strength of miRNA–target binding, as is usually the case.) We refer to the local alignment score and the free energy of a miRNA–target duplex as s_A and s_E, respectively. The scores s_A and s_E are (negatively) correlated in most cases, because a duplex with a high local alignment score tends to have a low free energy and vice versa.

2.2 Relation graph representation

In the second step of our method, we represent the many-to-many relation between miRNAs and target genes by a weighted bipartite graph termed relation graph.

DEFINITION 1. *Let M denote a set of miRNAs and T a set of targets (typically $|M| \ll |T|$). The relation graph is a weighted bipartite graph $G = (V, E, w)$ with the vertex set $V = M \cup T$, the edge set $E = \{\{m,t\} | \text{miRNA } m \in M \text{ binds target } t \in T\}$, and the weight function $w : E \to \mathbb{R}$.*

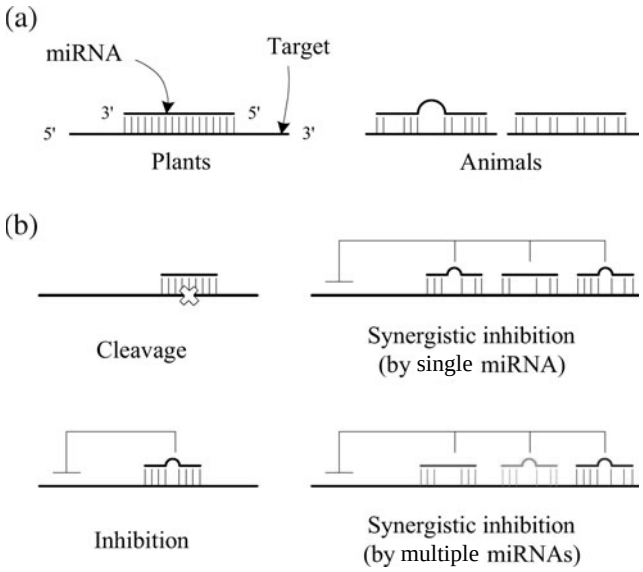


Fig. 1. miRNAs and targets (Lai, 2004). (**a**) Plant miRNAs exhibit extensive complementarity to their targets, but animal miRNAs generally do not. (**b**) Various configurations for miRNA–target duplexes: one near-perfect binding site for one miRNA (upper left), one strong site for one miRNA (lower left), multiple 'modest' sites for one miRNA (upper right), and multiple 'modest' sites for multiple miRNAs (lower right).

We determine the weight function w by performing Principal Component Analysis (PCA) (Jolliffe, 2002) on the space spanned by s_A and s_E. After making the populations of s_A and s_E have a zero mean, we find the unit vector u so that when the data is projected onto the direction corresponding to u the variance of the projected data is maximized. This unit vector u is equivalent to the principal eigenvector of Σ, the empirical covariance matrix of the data, defined as

$$\Sigma = \frac{1}{N} \sum_{i=1}^{N} \left(\begin{bmatrix} s_A \\ s_E \end{bmatrix}_i \cdot \begin{bmatrix} s_A \\ s_E \end{bmatrix}_i^{\mathsf{T}} \right), \tag{1}$$

where N represents the number of edges,

$$\begin{bmatrix} s_A \\ s_E \end{bmatrix}_i$$

is a score vector for the i-th edge, and $^{\mathsf{T}}$ means the transpose operator. Finally, for each $e \in E$, its weight $w(e)$ is calculated as the projection of a score vector onto u, namely,

$$w(e) = \begin{bmatrix} s_A \\ s_E \end{bmatrix}^{\mathsf{T}} u. \tag{2}$$

2.2.1 Modeling MRMs We model the MRM by a biclique or a complete subgraph in a bipartite graph (Aho *et al.*, 1983). In particular, we search only those bicliques in which, for each target vertex t, the edges incident on t have similar weights, following the biological observation explained in Section 2.1. To avoid redundancy, we find only maximal bicliques that are not contained by other bicliques as a proper subgraph.

DEFINITION 2. *For set A on $\mathbb{R}$, range(A) denotes the difference between the largest and the smallest elements of A.*

DEFINITION 3. *Let $G = (M \cup T, E, w)$ be the relation graph and $\delta \geq 0$ be given as a parameter. Graph $G' = (M' \cup T', E', w)$ is called*

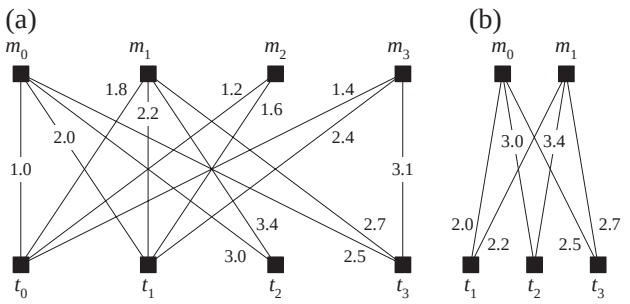


Fig. 2. (**a**) Example relation graph $G = (M \cup T, E, w)$, where $M = \{m_0, m_1, m_2, m_3\}$, and $T = \{t_0, t_1, t_2, t_3\}$ with some hypothetical weights. (**b**) An MRM found in G with the parameter $\delta = 0.5$.

a MRM, if G' is a maximal biclique in G, and for each $t \in T'$, range$(\{w|w = w(\{m,t\}),\ \forall m \in M'\}) \leq \delta$.

EXAMPLE 1. Figure 2 shows an example of the relation graph and an MRM found in this relation graph.

2.3 Finding seeds

The third step of our method is to find seeds for each predicted target gene.

DEFINITION 4. *Let t be a target gene and M_t be a set of miRNAs that binds the target gene t. A* seed *for t, denoted by $S(t)$, is a subset of M_t such that (i) range$(S(t)) \leq \delta$, and (ii) there is no $M' \supset S(t)$ such that $M' \subseteq M_t$ and range$(M') \leq \delta$.*

Algorithm 1. Find a seed for each target gene

```
   input  : t, a target transcript
   input  : M_t, a set of all miRNAs binding t
   input  : δ, a threshold
   output: {S(t)}, a set of seeds
 1 i := 0;
 2 for each m ∈ M_t do
 3     s[i].w := w(t, m);
 4     s[i].id := m;
 5     i := i + 1;
 6 sort array s in ascending order with respect to the w field;
 7 begin := 0;
 8 end := 1;
 9 S = ∅;
10 while (end < |M_t|) do
11     if (s[end].w − s[begin].w ≤ δ) then
12         end := end + 1;
13         if (end = |M_t|) then
14             S := GetOneSeed(begin, end, s);
15             S := S ∪ {S};
16     else
17         S := GetOneSeed(begin, end, s);
18         S := S ∪ {S};
19         repeat
20             begin := begin + 1;
21         until (begin = end) or (s[end].w − s[begin].w ≤ δ);
22 return S;
23 procedure GetOneSeed(begin, end, s)
24 begin
25     S := ∅;
26     for i = begin to end − 1 do
27         S := S ∪ {s[i].id};
28     return S;
29 end
```

Algorithm 1 presents our approach to generate a seed set for a given target transcript. This algorithm takes as input a target gene and a set of all the miRNAs binding the target gene regardless of binding strength. The output is a seed for the target gene or a maximal set of miRNAs whose binding strength to the target gene is similar in the sense that the difference between the maximum and the minimum strength is less than given δ.

The key idea of Algorithm 1 is simple: when the elements of set A are sorted and arranged in the corresponding order, range(A) is simply the absolute difference between the first and the last elements of A.

In Lines 1–6, miRNAs are sorted in ascending order by their binding strength. The variables 'begin' and 'end' in Lines 7–8 are to point to the first and the last elements of the sub-array under consideration at some point. The set of seeds $\mathcal{S}$, which is to be returned as output, is initialized in Line 9. Inside the while loop in Lines 10–21, seeds are generated as the variables begin and end are incremented. Since the miRNAs are sorted, we only need to compare the first element (s[*begin*]) and the last element (s[*end*]), as is done in Line 11, in order to see if all the elements in the sub-array are similar. In Lines 11–12, the variable end is extended as long as $s[end].w - s[begin].w \leq \delta$. A seed is found and collected in Lines 14–15 and Lines 17–18. Lines 19–21 are to adjust the variable begin after a seed is found.

Note that multiple seeds can exist for a single target gene, and thus a set of all the seeds for the given target gene is returned as output. Also notice that two distinct seeds for the same target gene can overlap. The worst-case complexity of the algorithm is polynomial in $|M_t|$.

Table 1. The seeds generated by Algorithm 1 from the relation graph in Figure 2a with the parameter $\delta = 0.5$

t: target gene	$S(t)$: seed for target gene t	No. of seeds
t_0	$\{m_0, m_2, m_3\}$, $\{m_1, m_3\}$	2
t_1	$\{m_0, m_1, m_3\}$, $\{m_0, m_2\}$	2
t_2	$\{m_0, m_1\}$	1
t_3	$\{m_0, m_1\}$, $\{m_1, m_3\}$	2

EXAMPLE 2. Table 1 shows all the seeds generated by Algorithm 1 from the relation graph in Figure 2a with the parameter $\delta = 0.5$.

2.3.1 Related data mining tasks Before describing the next step of our method, we show how the process of finding MRMs is related to several data mining techniques, in order to put the description in proper context.

First, the problem of frequent itemset mining (Agrawal *et al.*, 1993) is to find a group of items that occur together frequently in a database. Formally, let I be a set of all items in database D. A set, I', is called an *itemset* if $I' \subseteq I$. A *transaction* is pair (tid, I'), where *t.i.d.* is the transaction identifier and I' is an itemset. The transaction (tid, I') is said to *support* itemset I_s if $I_s \subseteq I'$. The *cover* of an itemset is the set of the identifiers of transactions that support the itemset. That is, for itemset I_s,

$$cover(I_s) = \{tid | (tid, I') \in D, I_s \subseteq I'\}. \quad (3)$$

The itemset I_s is called frequent if $|cover(I_s)| \geq \beta$, where β is a given threshold.

In the current problem, the set of miRNAs and the set of targets forming an MRM are similar to a frequent itemset and its cover, respectively. One difference is that a target can have multiple seeds whereas a transaction has only one itemset in a typical setup.

Second, the term biclustering (or co-clustering) (Madeira and Oliveira, 2004) refers to a class of clustering techniques that perform simultaneous clustering of rows and columns in a data matrix. The objective is to discover biclusters or patterns appearing in the form of overlapping submatrices in the matrix.

A matrix can be converted to a weighted bipartite graph, and vice versa. Thus, the relation graph $G = (M \cup T, E, w)$ can be converted to a matrix of weights with the row set M and the column set T. Then, a MRM is similar to

Algorithm 2. Find miRNA regulatory modules from the seeds

input : All the seeds generated by Algorithm 1
input : min_T, the minimum number of target genes in MRMs
input : min_M, the minimum number of miRNAs in MRMs
output: miRNA regulatory modules

```
1  /* Represent seeds by a trie */
2  for each seed S(t) do
3      Sort the elements in S(t);
4      n := the node whose path is specified by (sorted) S(t);
5      n.T := n.T ∪ {t};
6      n.S := S(t);

7  /* Merge the seeds */
8  for each node n in the post-order traversal of the trie do
9      for each node n' s.t. |n'.S| = |n.S| − 1 ≥ min_M do
10         n'.T := n'.T ∪ n.T;

11 /* Prune the trie and collect candidates */
12 for each node n in the pre-order traversal of the trie do
13     if |n.S| ≥ min_M then
14         if |n.T| < min_T then
15             Remove n and its subtrie rooted at n;
16         else
17             Collect (n.T, n.S) as a candidate MRM;
18 Return maximal candidates as MRMs;
```

a bicluster with constant values on columns, according to the classification of biclusters by Madeira and Oliveira (2004).

An issue is that the relation graph is usually sparse, and converting it to a matrix will result in many empty entries in the matrix (Figure 5c), thus often making it inappropriate to use matrix-based biclustering algorithms. Bigraph-based biclustering algorithms may be more effective, but certain assumptions some algorithms rely on, such as the maximum degree constraints, can be less meaningful in the present problem.

2.4 Deriving MRMs from seeds

The fourth step of our method is to collect all the seeds and derive MRMs from the seed collection. To collect seeds in a systematic and effective manner, we exploit a trie, a compact data structure to represent sets of character strings (Aho *et al*., 1983). Many overlaps often occur between the seeds, and a trie can provide compact representations. The seeds stored in the nodes of the trie are then merged to form MRMs as the trie is traversed. This technique bears some similarities to that used by some biclustering methods (Madeira and Oliveira, 2004; Yoon *et al*., 2005), but is more suitable to handle the sparsity of data.

Algorithm 2 details our approach. In addition to the seeds found by Algorithm 1, the algorithm takes as input two parameters, $\min_T$ and $\min_M$, to specify the minimum size of MRMs to find.

In Lines 2–6, each seed is inserted into a trie. To decide the location of the node into which a seed is inserted, we first assume a total order among the elements of M (the set of all miRNAs in Definition 1), of which every seed is a subset. For each seed $S(t)$ of target t, we then sort its elements with respect to the total order. The sorted seed can now be inserted into the node whose path is specified by the ordered elements.

To keep track of the seeds and associated target genes represented by the trie efficiently, two sets $n.S$ and $n.T$ are associated with each node n, as seen in Lines 5–6. Suppose that $S(t)$, a seed for target t, is inserted into node n. Then the set $n.S$ stores $S(t)$ proper, and the set $n.T$ contains the target gene t. Later in Line 10, the set $n.T$ will be expanded in such a way that $n.T = \{\tau \in T | n.S \subseteq S(\tau)\}$.

EXAMPLE 3. The trie in Figure 3a collectively represents the seeds in Table 1.

In Lines 8–10, the algorithm expands the trie to systematically merge the seeds and find candidates for MRMs. For each node n encountered in the post-order traversal of the trie, the set $n.T$ is distributed to every node n' in

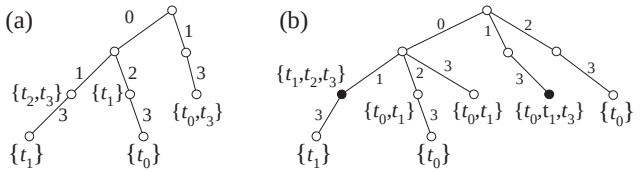


Fig. 3. (**a**) The trie representation of the seeds in Table 1. The edge labeled i represents miRNA m_i. (**b**) The seeds merged by Algorithm 2 with the parameters $\min_T = 3$, and $\min_M = 2$. The solid-circled vertices represent a candidate for MRMs.

which $|n'.M| = |n.M| - 1$ and $|n'.M| \geq \min_M$. The node n' is a node such that the number of elements in $n'.S$ is one smaller than n, but not less than $\min_M$.

In Lines 14–15, every node n in which $|n.T| < \min_T$ is deleted. This step can be performed efficiently by a pre-order traversal of the trie. Target genes were distributed in post-order in Lines 8–10. Consequently, node n in the trie always has a superset of the genes its children have. Thus, if the node n has less than $\min_T$ target genes, then none of its children can have more. For this reason, we can safely remove the entire subtree whose root is located at the node n without visiting its child nodes.

In Lines 17–18, candidates for MRMs are collected, and the maximal ones are returned as MRMs.

EXAMPLE 4. Figure 3b shows the trie after the seeds have been merged. Table 2 lists two candidate MRMs predicted from our running example.

The problem of enumerating maximal bicliques is inherently intractable (Madeira and Oliveira, 2004), and the worst-case complexity of Algorithm 2 is exponential in the number of miRNAs in the relation graph. However, the execution time of the algorithm on typical benchmarks is practical (see Section 3). This is because a seed seldom contains all the miRNAs in the relation graph, and the trie-based representation of seeds helps to prevent unnecessary enumeration of intermediate results.

2.5 Post-processing

Out of the MRMs found by Algorithm 2, we select those with a low P-value. We estimate the P-value of an MRM, or the probability of finding it by chance, on top of the statistical framework by Califano *et al*. (2000). They calculated the probability that a random submatrix of a gene expression data matrix has near-constant values on rows. They also reported that the distribution of the number of such matrices can be well approximated by the Poisson distribution. As previously stated, an MRM can be viewed as a matrix in which the values of each row are similar. Thus, we take advantage of the result by Califano *et al*. (2000) with minor modifications in order to approximate the P-values of MRMs. More precise assessment of their statistical significance will be possible as more exact mechanisms of miRNA–target interaction are revealed.

We assume that the number of MRMs with m miRNAs and t targets in the relation graph is a Poisson random variable denoted by $X_{m\times t}$. That is,

$$P(X_{m\times t} = k) = \frac{\lambda^k \mathrm{e}^{-\lambda}}{k!}, \quad k = 0, 1, 2, \ldots \tag{4}$$

The parameter λ corresponds to the average number of the MRMs with m miRNAs and t targets in the relation graph, namely,

$$\lambda = \binom{|M|}{m}\binom{|T|}{t} P_{m\times t}, \tag{5}$$

where $P_{m\times t}$ is the probability that a random $(m \times t)$ biclique in the relation graph satisfies the condition to be a $(m \times t)$ MRM. Based upon the result by Califano *et al*. (2000), $P_{m\times t}$ can be approximated by

$$P_{m\times t} \simeq \zeta^t [1-\zeta]^{|T|-t} [1 - (1 + m^{-1})^t \delta^t]^{|M|-m}, \tag{6}$$

Table 2. miRNA regulatory modules predicted with the parameters $\delta = 0.5$, $\min_T = 3$ and $\min_M = 2$

MRM #	Targets in the module	miRNAs in the module
1	$\{t_1, t_2, t_3\}$	$\{m_0, m_1\}$
2	$\{t_0, t_1, t_3\}$	$\{m_1, m_3\}$

where

$$\zeta = m\delta^{m-1} - (m-1)\delta^m. \quad (7)$$

The P-value of the MRM with m miRNAs and t targets is then defined to be the probability that one or more such MRMs occur by chance in the relation graph, namely,

$$P(X_{m\times t} \geq 1) = 1 - P(X_{m\times t} = 0) = 1 - \mathrm{e}^{-\lambda}. \quad (8)$$

Finally, we choose those MRMs whose P-value computed by Equation (8) is less than a certain threshold, highlighting statistically significant modules.

3 RESULTS AND DISCUSSION

We tested our method with the miRNAs and genes in *Homo sapiens* and predicted 431 MRMs. On average, an MRM consists of 3.58 miRNAs and 6.74 target genes. Among these predicted modules, here we present a cancer-related module and analyze it at length as an attempt to reveal the biological meanings implied. The analysis of the other modules is omitted due to the space limitation but can be performed in a similar manner.

Our data showed that a set of genes *PAK7*, *BTG2*, *WT1*, *PPM1D* and *RAB9B* are candidate targets for human *miR-15a*, *miR-16* and *miR-195*. Table 3 lists more details of this module. In what follows, we consider *miR-15a* and *miR-16* only, since *miR-195* is a predicted miRNA based on homology to a verified miRNA from mouse (Lagos-Quintana *et al*., 2003), and the expression of this miRNA has not been verified in human.

3.1 Validation with Gene Ontology

Using GO (The Gene Ontology Consortium, 2000) has become a standard way to validate the functional coherence of genes in a list. Typically, this type of validation is accompanied by a statistical significance analysis.

Figure 4 shows the annotation of the genes in this module with the terms in Biological Process category of GO. In particular, Figure 4a shows the distribution of the GO terms over the genes, and Figure 4b presents how these terms are related in the GO dag. We observe that the abundant terms include GO:0007582 (physiological process), GO:0008152 (metabolism), GO:0050875 (cellular physiological process), GO:0008151 (cell growth and/or maintenance) and GO:0008283 (cell proliferation).

Furthermore, we used the tool GO::TermFinder (Boyle *et al*., 2004) to find significantly over-represented GO terms. This tool calculates a P-value relative to the hypergeometric distribution and also performs the multiple comparison correction. For example, Table 4 presents some more details on one of the enriched GO term shown in Figure 4b.

3.2 Supporting evidence from the literature

BTG2 is a negative regulator of cell cycles, and impaired expression of *BTG2* has been found in breast, renal and prostate cancers in human (Struckmann *et al*., 2004; Kawakubo *et al*., 2004; Ficazzola *et al*., 2001). *WT1* is a gene encoding zinc-finger transcription factor, and defects in *WT1* are a cause of Wilms' tumor (WT), an embryonal malignancy of the kidney (Loeb and Sukumar, 2002). *PPM1D* is a *p53*-inducible protein phosphatase and its overexpression has been reported to cause breast cancer and neuroblastoma in human (Li *et al*., 2002; Saito-Ohara *et al*., 2003).

Interestingly, *BTG2*, *WT1* and *PPM1D* have been shown to be directly associated with the function of *p53*, a tumor-suppressor gene whose activation results in cell cycle arrest and apoptosis upon DNA damage, viral infection and oncogene activation (Vogelstein *et al*., 2000). Since inactivation of *p53* by deletion or mutation can cause tumor, it is also possible that the impaired function of *p53* by dysregulation of *BTG2*, *WT1* or *PPM1D* mediated by *miR-15a* and *miR-16* might develop tumor in an indirect way.

Several lines of evidence suggest that miRNAs may be related with leukemia and other cancers. For example, the human *miR-15a* and *miR-16* are clustered within 0.5 kb on chromosome 13q14, and this region has been shown to be deleted in B cell chronic lymphocytic leukemia (B-CLL), mantle cell lymphoma, multiple myeloma and prostate cancer cases (Stilgenbauer *et al*., 1998; Migliazza *et al*., 2000; Calin *et al*., 2002). A recent study by Calin *et al*. (2002) demonstrated that *miR-15a* and *miR-16* are located within a 30 kb region of loss in CLL, and both genes are deleted or downregulated in more than two-thirds of CLL cases, strongly suggesting the involvement of miRNA genes in human cancers.

Given that *miR-15a* and *miR-16* are detected together and found to regulate a set of genes that are actively involved in tumorigenesis by the use of our method, further studies should be focused on elucidating the direct role of *miR-15a* and *miR-16* in many types of cancer through dysregulation of their target gene expression.

3.3 Discussion

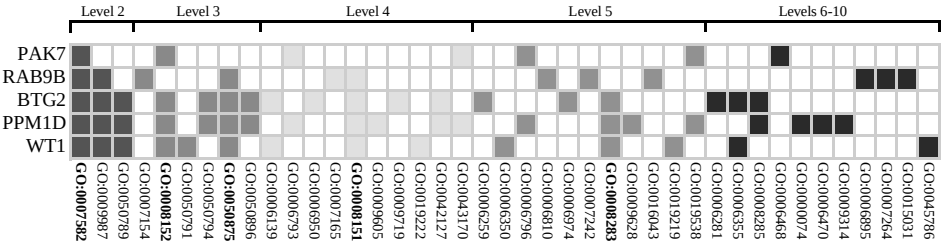
3.3.1 Extension of our computational method As the understanding of *in vivo* miRNA target selection mechanisms deepens, more advanced methods to computationally identify miRNA targets will emerge. New findings on miRNA-target interactions will need to be represented by a new relation graph. Our method can then be applied to the augmented relation graph without further modifications. In fact, identifying animal miRNA targets is computationally difficult and there is much room for improvements. This is because animal miRNAs are short and only partially complementary to their targets. Enhanced methods may consider interactions involving RNA binding proteins, conservation filtering through sophisticated phylogenetic profiling techniques and handling for some unusual structures in targets. For example, a very long loop structure in the target sequence cannot easily be detected without adversely affecting the rate of false positive detection (Enright *et al*., 2003).

The MRM defined in this work consists of miRNAs and their targets. Since computational methods have been proposed to identify the gene that encodes miRNAs (Ohler *et al*., 2004; Lim *et al*., 2003; Rodriguez *et al*., 2004; Lai *et al*., 2003), it is possible to redefine the MRM as a group of host genes, the miRNAs encoded by the host genes, and the target genes bound by the miRNAs. This will complete the regulatory chain of 'host gene → miRNA → target

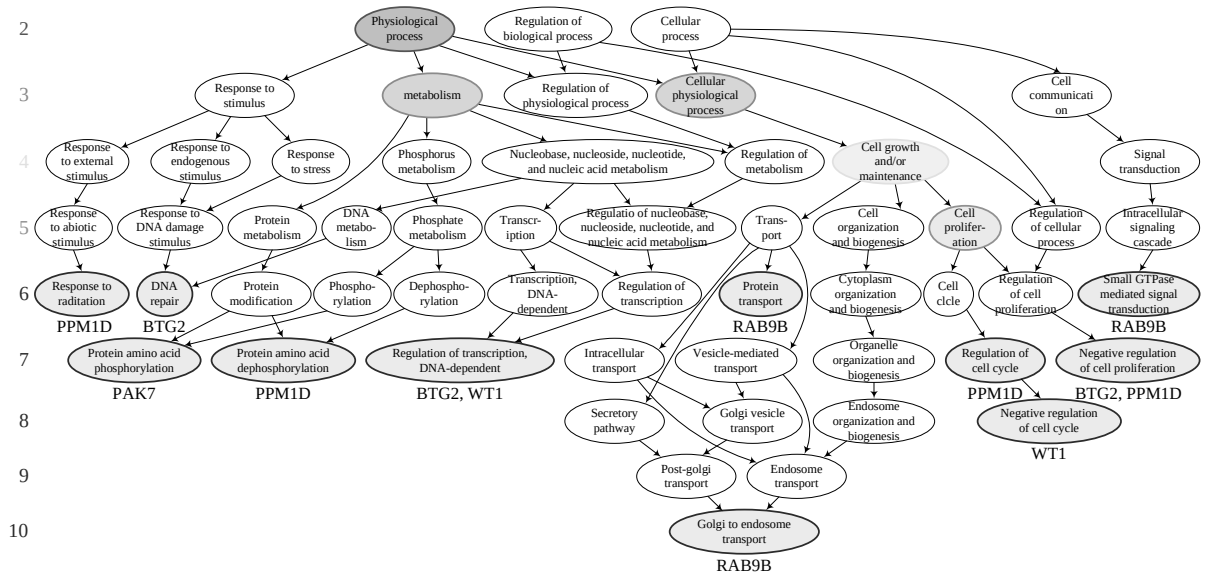
Table 3. A predicted human MRM. The first column represents the genes in the module, and the last three columns show the miRNAs with their binding strength to each target in terms of the weight calculated by Equation (2). The parameters used are listed in Table 5

Target (HUGO ID)	Ensemble ID	Description	*hsa-miR-15a*	*hsa-miR-16*	*hsa-miR-195*[a]
PAK7	ENSG00000101349	p21-activated kinase 7	1.609	−0.789	0.676
RAB9B	ENSG00000123570	Ras-associated oncogenic protein 9b	1.303	−0.746	−0.956
BTG2	ENSG00000159388	B cell translocation gene 2	−0.162	−0.816	−1.259
PPM1D	ENSG00000170836	protein phosphatase 1D Mg-dependent, delta isoform	−0.487	−0.817	−1.143
WT1	ENSG00000184937	Wilms' tumor	0.275	1.019	−0.514

[a] Has not been verified experimentally in human.



(a) Distribution of the GO terms over the genes in the module in Table 3



(b) A subgraph of the GO dag to show the relationship among the GO terms used

Fig. 4. Annotation with GO terms. (**a**) Each row represents a target gene, and each column a GO term in Biological Process. A colored box exists at row i and column j if target i has GO term j. The abundant terms are GO:0007582 (physiological process), GO:0008152 (metabolism), GO:0050875 (cellular physiological process), GO:0008151 (cell growth and/or maintenance) and GO:0008283 (cell proliferation). (**b**) The blue vertices are for the terms in levels 6–10 associated with the targets in the predicted MRM. The ancestor vertices are also included, where the most abundant terms in each level are colored in red (level 2), orange (level 3), yellow (level 4) and green (level 5). Further analysis of each enriched GO term is possible. For example, Table 4 presents some detailed analysis of the term negative regulation of cell proliferation.

gene'. This new piece of information can be incorporated into the modeling of gene regulatory networks.

3.3.2 Detailed experiment procedure The input to our method was the human miRNA sequences (http://www.sanger.ac.uk/Software/Rfam/mirna) and the human gene sequences (http://www.ensembl.org/EnsMart). The output was a list of MRMs. The methods described in Lewis *et al*. (2003) and John *et al*. (2004) were first used to identify 7886 human miRNA-mRNA duplexes. 2888 genes and 156 miRNAs were found to participate in forming a duplex (Table 5). After scalar weights were calculated by Equation (2), the relation graph was constructed. Figure 5 shows the distributions of s_A and s_E

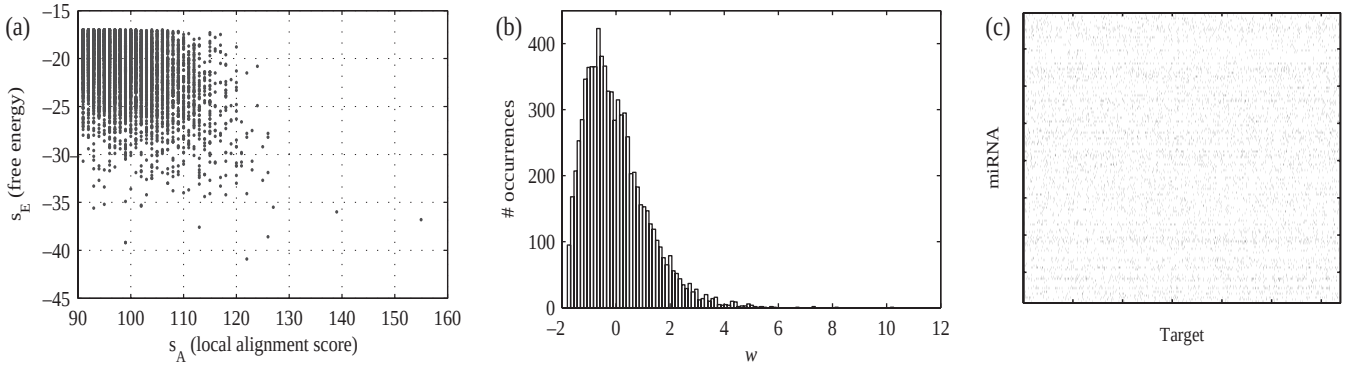


Fig. 5. (**a**) The distributions of the scores s_A and s_E. (**b**) The edge weight distribution with $\mu = 0$, $\sigma = 1.20$, min $= -1.79$ and max $= 10.26$. (**c**) The relation graph represented by a $156 \times 2,888$ matrix. A dot exists at row i and column j, if target j has a binding site for miRNA i. This plot visualizes the initial raw dataset, clearly showing a need for an automated tool to identify important patterns underlying the complex interactions between miRNAs and targets.

Table 4. Further details on an enriched GO term in Figure 4b, obtained by the tool GO::TermFinder (Boyle *et al*., 2004)

Item	Value
GO ID	GO:0008285
Term	Negative regulation of cell proliferation
P-value	0.000259
Corrected P-value	0.0184
Annotated genes	*BTG2*, *PPM1D*
Genome frequency of use	134 out of 23531 genes

Table 5. The parameters used for the experiment and some statistics obtained

Parameters/statistic	Value/reference
Parameters (s_A cutoff, s_E cutoff)	(91, −17 kcal/mol)
Parameters ($\min_T, \min_M, \delta$)	$(3, 3, 2\sigma^{a} = 2.40)$
Size of the relation graph ($\|T\|, \|M\|, \|E\|$)	(2888, 156, 7886)
Weights in the relation graph	Figure 5
Total number of modules found	431
Average size of modules (# targets, # miRNAs)	(6.74, 3.58)

[a]The standard deviation of the weight distribution in Figure 5b.

and a matrix representation of the relation graph. Algorithms 1 and 2 were invoked with the parameters listed in Table 5. Statistically significant MRMs were selected with the P-value threshold of 0.01. The annotation of selected modules with the terms in Gene Ontology (http://www.geneontology.org) was finally performed. The computation ran on a 3.06 GHz Linux machine with 4 GB RAM, and the response time for Algorithms 1 and 2 was in the order of minutes.

3.3.3 Update Lewis *et al*. (2005) recently revised their method for miRNA target identification. This updated method uses simplified detection rules and predicts more miRNA targets, which include most target genes already detected by the previous method (Lewis *et al*., 2003). Newly identified miRNA–target duplexes can be added into the relation graph. This expanded, more complex relation graph can provide our method with more opportunities to identify interesting modules for further analysis.

ACKNOWLEDGEMENTS

The authors thank Professor C. Z. Chen and Dr H. Min at Stanford Medical School for helpful discussions. This work was supported by a grant of Jerry Yang and Akiko Yamazaki.

Conflict of Interest: none declared.

REFERENCES

Agrawal,R., Imielinski,T. and Swami,A.N. (1993) Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, SIGMOD '93*, Washington, DC, May 26–28, 1993, ACM Press, pp. 207–216.

Aho,A.V., Hopcroft,J.E. and Ullman,J.D. (1983) *Data Structures and Algorithms*. Addison-Wesley, Reading, Massachusetts.

Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.

Boyle,E.I. *et al*. (2004) GO::TermFinder. *Bioinformatics*, **20**, 3710–3715.

Califano,A. *et al*. (2000) Analysis of gene expression microarrays for phenotype classification. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **8**, 75–85.

Calin,G.A. *et al*. (2002) Frequent deletions and down-regulation of micro-RNA genes *miR15* and *miR16* at 13q14 in chronic lymphocytic leukemia. *Proc. Natl Acad. Sci. USA*, **99**, 15524–15529.

Enright,A.J. *et al*. (2003) MicroRNA targets in *Drosophila*. *Genome Biol.*, **5**, R1.

Ficazzola,M.A. *et al*. (2001) Antiproliferative B cell translocation gene 2 protein is down-regulated post-transcriptionally as an early event in prostate carcinogenesis. *Carcinogenesis*, **22**, 1271–1279.

John,B. *et al*. (2004) Human microRNA targets. *PLoS Biol.*, **2**, e363.

Jolliffe,I.T. (2002) *Principal Component Analysis*, 2nd, edn, Springer-Verlag, New York.

Kawakubo,H. *et al*. (2004) Expression of the NF-kappaB-responsive gene BTG2 is aberrantly regulated in breast cancer. *Oncogene*, **23**, 8310–8319.

Kiriakidou,M. *et al*. (2004) A combined computational-experimental approach predicts human microRNA targets. *Genes Dev.*, **18**, 1165–1178.

Lagos-Quintana,M. *et al*. (2003) New microRNAs from mouse and human. *RNA*, **9**, 175–179.

Lai,E.C. (2004) Predicting and validating microRNA targets. *Genome Biol.*, **5**, 115.1–115.6.

Lai,E.C. *et al*. (2003) Computational identification of *Drosophila* microRNA genes. *Genome Biol.*, **4**, R42.1–R42.20.

Lewis,B.P. *et al*. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, **120**, 15–20.

Lewis,B.P. *et al*. (2003) Prediction of mammalian microRNA targets. *Cell*, **115**, 787–798.

Li,J. *et al*. (2002) Oncogenic properties of PPM1D located within a breast cancer amplification epicenter at 17q23. *Nat. Genet.*, **31**, 133–134.

Lim,L.P. *et al*. (2003) Vertebrate microRNA genes. *Science*, **299**, 1540.

Loeb,D.M. and Sukumar,S. (2002) The role of WT1 in oncogenesis: tumor suppressor or oncogene? *Int. J. Hematol.*, **76**, 117–126.

Madeira,S.C. and Oliveira,A.L. (2004) Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, **1**, 24–45.

Migliazza,A. *et al*. (2000) Molecular pathogenesis of B-cell chronic lymphocytic leukemia: analysis of 13q14 chromosomal deletions. *Curr. Top. Microbiol. Immunol.*, **252**, 275–284.

Ohler,U. *et al*. (2004) Patterns of flanking sequence conservation and a characteristic upsteam motif for microRNA gene identification. *RNA*, **10**, 1309–1322.

Rajewsky,N. and Socci,N.D. (2003) Computational identification of microRNA targets. *Dev. Biol.*, **267**, 529–535.

Rehmsmeier,M. *et al*. (2004) Fast and effective prediction of microRNA/target duplexes. *RNA*, **10**, 1507–1517.

Rodriguez,A. *et al*. (2004) Identification of mammalian microRNA host genes and transcription units. *Genome Res.*, **14**, 1902–1910.

Saito-Ohara,F. *et al*. (2003) PPM1D is a potential target for 17q gain in neuroblastoma. *Cancer Res.*, **63**, 1876–1883.

Smalheiser,N.R. and Torvik,V.I. (2004) A population-based statistical approach identifies parameters characteristic of human microRNA–mRNA interactions. *BMC Bioinformatics*, **5**, 139.

Sokal,R.R. and Rohlf,F.J. (1994) *Biometry*. W.H. Freeman and Co.

Stark,A. *et al*. (2003) Identification of *Drosophila* microRNA targets. *PLoS Biol.*, **1**, 397–409.

Stilgenbauer,S. *et al*. (1998) Expressed sequences as candidates for a novel tumor suppressor gene at band 13q14 in B-cell chronic lymphocytic leukemia and mantle cell lymphoma. *Oncogene*, **16**, 1891–1897.

Struckmann,K. *et al*. (2004) Impaired expression of the cell cycle regulator BTG2 is common in clear cell renal cell carcinoma. *Cancer Res.*, **64**, 1632–1638.

The Gene Ontology Consortium (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

Vogelstein,B. *et al*. (2000) Surfing the p53 network. *Nature*, **408**, 307–310.

Yoon,S. *et al*. (2005) Discovering coherent biclusters from gene expression data using zero-suppressed binary decision diagrams. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, **2**, in press.