Proceedings of the 2005 IEEE
Engineering in Medicine and Biology 27th Annual Conference
Shanghai, China, September 1-4, 2005

Prediction and Analysis of Human microRNA Regulatory Modules

Sungroh Yoon, *Student Member, IEEE,* and Giovanni De Micheli, *Fellow, IEEE*

***Abstract*— MicroRNAs are a family of small, non-coding RNAs that regulate gene expression in a sequence-specific manner. We propose a computational method to predict *miRNA regulatory modules* or groups of miRNAs and target genes that are believed to participate cooperatively in post-transcriptional gene regulation. We tested our method with the human genes and miRNAs, predicting 431 miRNA regulatory modules. The validations of predicted modules with the literature and *Gene Ontology* show that the genes in the modules appear to be closely related in specific biological processes, such as gene regulation involved in various cancers.**

I. Introduction

MicroRNAs (miRNAs) are a novel class of gene products that repress mRNA translation or mediate mRNA degradation in a sequence-specific manner in animals and plants [1], [2]. Hundreds of different miRNAs have now been identified in complex eukaryotes, implying that they mediate a vast network of unappreciated regulatory interactions [3]. Normally, multiple miRNAs regulate one message, reflecting cooperative translational control. Conversely, one miRNA may have several target genes, indicative of target multiplicity [2]. This multiplicity of targets and cooperative signal integration on target genes are key features of the control of translation by miRNAs [4].

In this study, we mathematically formulate biological observations on the interactions of miRNAs and their targets and present a way to identify important patterns hidden in the complex interactions. Especially, we propose a computational method that can 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 was tested with the human miRNAs and genes, identifying 431 human MRMs. Some of the results obtained from this experiment will be presented in this paper.

Predicted miRNA regulatory modules can be useful for the reconstruction of gene regulatory networks. That is, 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.

This work was supported by a grant of Jerry Yang and Akiko Yamazaki.

S. Yoon is with the Computer Systems Laboratory, Stanford University, CA 94305, USA (corresponding author. phone: +1-650-724-5180; fax: +1-650-725-9802; e-mail: sryoon@stanford.edu).

G. De Micheli is with the Integrated Systems Center, EPFL, Lausanne, Switzerland (e-mail: giovanni.demicheli@epfl.ch).

Section II provides the formal definition of miRNA regulatory modules and presents an algorithm to predict them. In Section III, we show our experimental results, focusing on the analysis of a predicted module through a literature review and annotation with *Gene Ontology* (GO) [5].

II. Method

The input to our method is the human miRNA and gene sequences. The output are miRNA regulatory modules. The method consists of the following steps, and each step will be detailed in this section.

1) Pre-processing: target genes for each miRNA are identified (Section II-A).
2) Relation graph representation: interactions between miRNAs and their targets are represented by a data structure called *relation graph* (Section II-B).
3) Seed finding: a *seed* or a set of miRNAs that bind a common target with similar binding strength is identified (Section II-C).
4) Merging seeds: the seeds found are collected and merged to produce candidates for miRNA regulatory modules (Section II-D).
5) Post-processing: statistically significant miRNA regulatory modules are selected by computing the p-value [6] or the probability of finding a module by chance (Section II-E).

A. Pre-processing: identification of miRNA target sites

We identify individual miRNA-mRNA duplexes by the method described in [4] and [7]. Target selection is guided by the miRNA sequence [3]. Unlike plant miRNAs, animal miRNAs do not generally exhibit extensive complementarity to any endogenous transcripts. Thus, most algorithms to identify animal miRNA targets rely on three properties: (i) sequence complementarity using a position-weighted local alignment algorithm, (ii) free energies of miRNA-target duplexes, and (iii) evolutionary conservation of target sites in homologous genes.

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.

Various configurations for miRNA-target duplexes are possible. 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 [3]. This observation will be reflected in our mathematical formulation in Section II-B.

0-7803-8740-6/05/$20.00 ©2005 IEEE.

B. Relation graph representation

Based upon the results obtained in the pre-processing step, we can represent the many-to-many relationships between miRNAs and target genes by a data structure termed *relation graph*, which is defined as follows.

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 $w : E \to \mathbb{R}$.

In the definition, the weight function w is determined by performing *Principal Component Analysis* (PCA) [8] on the space spanned by s_A and s_E.

A miRNA regulatory module is modeled by a *biclique* or a complete subgraph in the relation graph. 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 II-A. Furthermore, to avoid redundancy, we find only *maximal* bicliques that are not contained by other bicliques as a proper subgraph.

For set A on $\mathbb{R}$, let $range(A)$ denote the absolute difference between the largest and the smallest elements of A. Then, we formally define a miRNA regulatory module as follows.

Definition 2: Let $G = (M \cup T, E, w)$ be the relation graph and $\tau \geq 0$ be given as a parameter. Graph $G' = (M' \cup T', E', w)$ is called a *miRNA regulatory module* (MRM), if G' is a maximal biclique in G, and for each $t \in T'$, $range(\{w | w = w(\{m, t\}),\ \forall m \in M'\}) \leq \tau$.

C. Finding seeds

After having constructed the relation graph, we find *seeds* for each target gene. A seed is formally defined as follows.

Definition 3: 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) $range(S_t) \leq \tau$, and (ii) there is no $M' \supset S_t$ such that $M' \subseteq M_t$ and $range(M') \leq \tau$.

Algorithm 1 presents our approach to generate a seed for a given target transcript. The key idea of this algorithm 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 therefore sorted in ascending order by their binding strength to the target. In order to keep track of the first and the last elements, two variables $begin$ and end are maintained in Lines 7–17. The **while** loop in Lines 8–17 is (i) to handle multiple instances of the seed for a single target and (ii) to find only maximal seeds. The worst-case complexity of the algorithm is polynomial in $|M_t|$.

D. Deriving MRMs from seeds

We collect all the seeds found in the previous step and derive MRMs from this collection of seeds. To store and manage seeds in a systematic and effective manner, we exploit the *trie*, a compact data structure to represent sets of character strings [9]. The seeds are stored in the nodes of the trie and then merged to form MRMs as the trie is traversed.

Algorithm 1: Find seeds for each target gene

```
   input : t, target transcript
   input : M_t, set of all miRNAs binding t
   input : τ, threshold
   output: S_t ⊆ M_t, seed for target t
1  i := 1;
2  foreach 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 := 1; end := 2;
8  while (end ≤ |M_t|) do
9      if (s[end].w − s[begin].w ≤ τ) then
10         end := end + 1;
11         if (end > |M_t|) then
12             Report S_t = {s[begin].id, ..., s[end − 1].id};
13     else
14         Report S_t = {s[begin].id, ..., s[end − 1].id};
15         repeat
16             begin := begin + 1;
17         until (begin = end) or (s[end].w − s[begin].w ≤ τ);
```

Algorithm 2 details our approach. In addition to the seeds found by Algorithm 1, Algorithm 2 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). For each seed S_t of target t, we then sort its elements with respect to this 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

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  foreach 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  foreach node n in the post-order traversal of the trie do
9      foreach 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 foreach 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;
```

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 $n.T = \{t' \in T | n.S \subseteq S_{t'}\}$.

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 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.

The problem of enumerating maximal bicliques is inherently intractable [10], 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 III). 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.

E. Post-processing: selecting modules with low p-values

Out of the miRNA regulatory modules found, statistically significant ones are selected. To this end, we estimate the p-value of an MRM, or the probability of finding it by chance, on top of the statistical framework by [11].

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

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

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

$$\lambda = \binom{|M|}{m}\binom{|T|}{t} P_{m\times t}, \quad (2)$$

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 [11], $P_{m\times t}$ can be approximated by

$$P_{m\times t} \simeq \zeta^t [1-\zeta]^{|T|-t} [1 - (1 + m^{-1})^t \tau^t]^{|M|-m}, \quad (3)$$

where

$$\zeta = m\tau^{m-1} - (m-1)\tau^m. \quad (4)$$

TABLE I

THE PARAMETERS USED FOR THE EXPERIMENT AND SOME STATISTICS OBTAINED. († THE STANDARD DEVIATION OF THE WEIGHT DISTRIBUTION.)

Parameters/statistic	Value/reference
Parameters (s_A cutoff, s_E cutoff)	$(91, -17$ kcal/mol$)$
Parameters (min_T, min_M, τ)	$(3, 3, 2\sigma^\dagger = 2.40)$
Size of the relation graph $(\|T\|, \|M\|, \|E\|)$	$(2888, 156, 7886)$
Total number of modules found	431
Average module size (# targets, # miRNAs)	$(6.74, 3.58)$

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,

$$P(X_{m\times t} \geq 1) = 1 - P(X_{m\times t} = 0) = 1 - e^{-\lambda}. \quad (5)$$

Finally, we choose those MRMs whose p-value computed by (5) is less than a certain threshold, highlighting statistically significant modules.

III. EXPERIMENTS

A. Procedure

The proposed method was tested with the human miRNA sequences (http://www.sanger.ac.uk/Software/Rfam/mirna) and the human gene sequences (http://www.ensembl.org/EnsMart). The methods described in [4] and [7] were first used to identify 7,886 human miRNA-mRNA duplexes. 2,888 genes and 156 miRNAs were found to participate in forming a duplex (see Table I). After the relation graph was constructed, Algorithms 1 and 2 were invoked with the parameters listed in Table I. Statistically significant MRMs were selected with the p-value threshold of 0.01. The tool GO::TermFinder [5] was used to annotate the genes in selected modules. 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.

B. Results and validation

431 miRNA regulatory modules were predicted from the human miRNAs and genes. On average, an MRM consists of 3.58 miRNAs and 6.74 target genes, as seen in Table I. Among these predicted modules, here we present a cancer-related module and analyze it at length. The analysis of the other modules is omitted due to the space limitation but can be performed in a similar manner. (A complete list of the predicted modules is available from the authors upon request.)

1) Prediction of an oncogenic module: 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 II lists more details of this module. The first column of the table 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 PCA. The parameters used are listed in Table I. In what

TABLE II

A PREDICTED HUMAN MIRNA REGULATORY MODULE. († HAS NOT BEEN VERIFIED EXPERIMENTALLY IN HUMAN.)

Target (HUGO ID)	Ensemble ID	Description	*hsa-miR-15a*	*hsa-miR-16*	*hsa-miR-195*†
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

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 [12], and the expression of this miRNA has not been verified in human.

2) Validation with GO: For the genes included in the module in Table II, the GO terms annotating these genes repeatedly 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). For more quantitative analysis, we used the tool GO::TermFinder [5] 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 III presents an enriched GO term for the genes *BTG2* and *PPM1D* and related information.

3) Supporting evidence from the literature: The genes *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 [13]. 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 [14]–[16]. A recent study by [16] demonstrated that *miR-15a* and *miR-16* are located within a 30-kb region of loss in CLL, and both genes are deleted or down-regulated 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.

TABLE III

ENRICHED GO TERM OBTAINED BY THE TOOL GO::TERMFINDER [5].

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

IV. CONCLUSION

We developed a computational method to predict a module of miRNA and targets. This method was tested with the human genome and identified biologically meaningful miRNA regulatory modules.

REFERENCES

[1] D. P. Bartel, "MicroRNAs: Genomics, biogenesis, mechanism, and function," *Cell*, vol. 116, no. 2, pp. 281–297, 2004.

[2] A. J. Enright *et al*., "MicroRNA targets in *Drosophila*," *Genome Biology*, vol. 5, no. 1, p. R1, 2003.

[3] E. C. Lai, "Predicting and validating microRNA targets," *Genome Biology*, vol. 5, no. 9, pp. 115.1–6, 2004.

[4] B. John, *et al*., "Human microRNA targets," *PLoS Biology*, vol. 2, no. 11, p. e363, 2004.

[5] E. I. Boyle *et al*., "GO::TermFinder," *Bioinformatics*, vol. 20, no. 18, pp. 3710–3715, December 2004.

[6] R. R. Sokal and F. J. Rohlf, *Biometry*. WH Freeman and Co., 1994.

[7] B. P. Lewis, I. Shih, M. W. Jones-Rhoades, D. Partel, and C. B. Burge, "Prediction of mammalian microRNA targets," *Cell*, vol. 115, no. 7, pp. 787–798, 2003.

[8] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. New York: Springer-Verlag, 2002.

[9] A. V. Aho, J. E. Hopcroft, and J. D. Ullman, *Data Structures and Algorithms*. Reading, Massachusetts: Addison-Wesley, 1983.

[10] S. C. Madeira and A. L. Oliveira, "Biclustering algorithms for biological data analysis: A survey," *IEEE Transactions on Computational Biology and Bioinformatics*, vol. 1, no. 1, pp. 24–45, 2004.

[11] A. Califano, G. Stolovitzky, and Y. Tu, "Analysis of gene expression microarrays for phenotype classification," in *Proc. Int Conf Intell Syst Mol Biol*, 2000, pp. 75–85.

[12] M. Lagos-Quintana, R. Rauhut, J. Meyer, A. Borkhardt, and T. Tuschl, "New microRNAs from mouse and human," *RNA*, vol. 9, no. 2, pp. 175–179, 2003.

[13] B. Vogelstein, D. Lane, and A. J. Levine, "Surfing the p53 network," *Nature*, vol. 408, no. 6810, pp. 307–310, 2000.

[14] S. Stilgenbauer *et al*., "Expressed sequences as candidates for a novel tumor suppressor gene at band 13q14 in B-cell chronic lymphocytic leukemia and mantle cell lymphoma," *Oncogene*, vol. 16, no. 14, pp. 1891–1897, 1998.

[15] A. Migliazza *et al*., "Molecular pathogenesis of B-cell chronic lymphocytic leukemia: analysis of 13q14 chromosomal deletions," *Curr Top Microbiol Immunol.*, vol. 252, pp. 275–284, 2000.

[16] G. A. Calin *et al*., "Frequent deletions and down-regulation of micro-RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia," *Proc Natl Acad Sci USA*, vol. 99, no. 24, pp. 15 524–15 529, 2002.