Computational Prediction of Competitive Endogenous RNA

Seunghyun Park[1,2], Soowon Kang[3], Hyeyoung Min[3], and Sungroh Yoon[2,*]

[1]School of Electrical Engineering, Korea University, Seoul 136-713, Republic of Korea
[2]Department of Electrical and Computer Engineering, Seoul National University, Seoul 151-744, Republic of Korea
[3]RNA Biopharmacy Laboratory, College of Pharmacy, Chung-Ang University, Seoul 156-756, Republic of Korea
*Email: sryoon@snu.ac.kr

***Abstract*—MicroRNAs (miRNAs) play an important role in the post-transcriptional regulation of gene expression by pairing target messenger RNAs (mRNAs). As the abnormal expression of miRNAs has been implicated in various diseases, there has been many studies on regulating the expression level of miRNA, including "miRNA sponges." miRNA sponges, which are artificial miRNA decoys, contain complementary binding sites to a target miRNA and regulate the expression level of target miRNAs. As competitive endogenous RNAs (ceRNAs) have been found in a recent study, there have been many efforts to find natural miRNA sponges. However, there are no related studies about the computational approach using the pairwise interactions of numerous mRNA-miRNA pairs. In this study, a computational approach to find candidates of natural miRNA sponges is proposed. Whole miRNA binding sites with query miRNA and the secondary structures of reference mRNA are predicted, followed by calulating the adjusted minimum free energy (AMFE) as the total score. We can quantitatively compare the interactions between miRNAs and target mRNAs by using this proposed approach. Thirty viral miRNAs and about 300 of thousands of human mRNAs are used in this study. As a results, the top 20 natural miRNA sponge candidates are recorded. The results are expected to provide appropriate knowledge before *in vivo* experiments to validate the identification of miRNA sponges.**

I. Introduction

MicroRNAs (miRNAs) are small noncoding RNAs of around 22 nucleotides that regulate gene expression by binding to the 3 UTR of their target mRNAs for translational suppression or cleavage [1]. Recent investigations demonstrate that miRNAs have unique expression profiles in different cancer types at different stages and play an important role in many diseases and viral infections [2], [3]. These results suggest that the regulation of miRNAs is very important in understanding the factors controlling biological processes.

An "miRNA sponge" is an artificial competitive inhibitor containing multiple repeated binding sites to a specific miRNA [4]. A sponge is able to bind and hold several miRNA copies, thereby decreasing the cellular levels of the target miRNA, which causes the deregulation of the mRNA regulated by the miRNA. Through various studies, miRNA sponges have been shown to be more effective than chemically modified antisense oligonucleotides and have been widely used for inhibition of miRNAs [4], [5].

In addition to these artificial miRNA sponges, it was thought that the small natural RNA fragment expressed from stable chromosomal insertions might also function as a sequence-specific miRNA blocker [6], [7]. The first such endogenous sponge RNA was detected in plants and was found to reduce miRNA-mediated alteration in response to environmental stress [8].

TABLE I. Datasets used in this study

Dataset		#Seq.	species	source
miRNA	Epstein-Barr virus (EBV)	17	virus	miRBase[a]
	Kaposi's sarcoma-associated herpesvirus (KSHV)	13		
mRNA	longer than 50 nt	313,517	human	GenBank[b]

[a] mature microRNA release 21
[b] Genome Reference Consortium Human Reference 38 (GRCh38)

Although a few natural miRNA sponges were validated in recent studies [9], [10], they are restricted as a circular type and their numbers are quite small. The study to predict circular RNA [11] does not serve the purpose of finding functionality as a sponge but to find structural characteristics. It is possible to derive enough information by investigating whole miRNA-mRNA interactions; however, a study of such a large scale has not yet been performed.

In this study, we propose a computational approach to find natural miRNA sponge candidates. Whole binding sites prediction with all combinations of miRNA-mRNA pairs was performed, and additional secondary structural information was used for a reasonable scoring scheme. Thirty viral miRNAs and 300 of thousands of human mRNAs were used for the experiment followed by representing the 20 miRNA-mRNA pairs with the highest scores. The results can be an appropriate guideline for *in vivo* experiments to validate the identification of miRNA sponges, and we anticipate that the proposed approach can provide useful insight for future research.

II. Materials and Methods

A. Dataset

Among a total of 35,828 mature miRNA sequences (release 21) provided by miRBase [12], we used 30 types of viral miRNA: 17 types of Epstein-Barr virus miRNA sequences and 13 types of Kaposi sarcoma-associated herpesvirus miRNA sequences.

mRNA sequences were obtained from the Genome Reference Consortium Human Reference 38 (GRCh38) pub-

978-1-4799-7234-0/15/$31.00 ©2015 IEEE

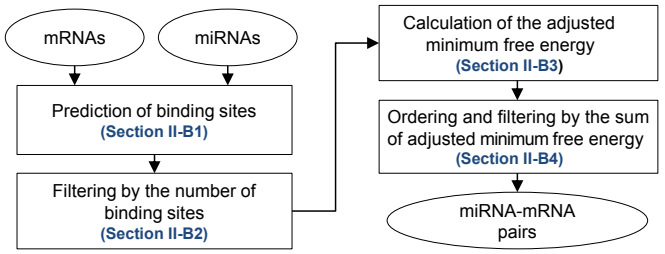

Fig. 1. Overview of the proposed approach

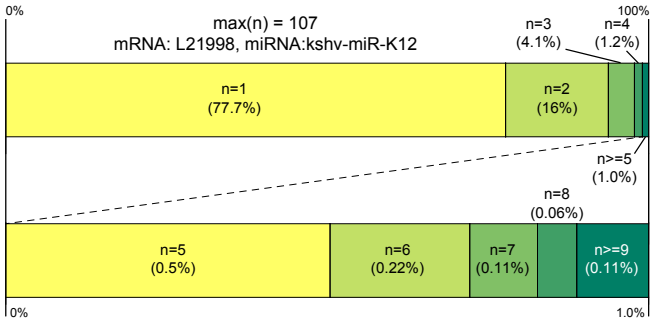

Fig. 2. The statistics of the number of binding sites (n) in an mRNA-miRNA pair.

lished by GenBank [13]. The total number of mRNA sequences was 2,009,256 but 84% of the given sequences (or 1,680,000/2,009,256) have about 20 nt, such as siRNA and piRNA. However, these short sequences were excluded from our consideration due to their low possibility of being mRNA sponges. Thus, only 313,517 mRNA sequences longer than 50 nt were included as potential candidates for being miRNA sponges. The details of the dataset used in our experiment are explained in Table I.

B. Algorithm

The overall flowchart of the proposed algorithm is explained in Fig. 1.

1) Prediction of the target binding sites: For the first step, potential binding sites for every mRNA-miRNA pair were detected by miRanda [14], a widely used miRNA target prediction tool. miRanda finds optimal miRNA target sites searched by dynamic programming. The level of minimum free energy (MFE), an optimal measure of strand-strand interaction, at each potential target site was calculated by the Vienna RNA package [15].

2) Filtering by the number of binding sites: Given the number of binding sites detected in every mRNA-miRNA pair, we filtered out mRNA sequences when the total number of detected binding sites for a given mRNA-miRNA pair is smaller than the user-specific threshold value, denoted by θ. We empirically chose θ as 2, which is intuitively a limited number to function as a miRNA sponge.

3) Calculation of the adjusted minimum free energy: A score is measured for each mRNA-miRNA pair remaining after filtering. Note that, the detection of binding sites based on the tertiary structure of mRNA may be more reliable than those on the primary or secondary structure, but the prediction of the tertiary structure of RNA is currently challenging. In this paper, we propose a scoring scheme based on information about RNAs' secondary structure. Among various types of binding processes, binding to the loop structure has been reported as the most reliable model [16]. In the proposed algorithm, we first predict the secondary structure of mRNA by RNAfold [17] (phase 1), then we give weights of the value of MFE for each miRNA binding site (phase 2). The sum of the adjusted minimum free energy (AMFE) $E(r, q)$ for a pair of mRNA r and miRNA q with k binding sites is defined as follows:

$$E(r,q) = \sum_{i=1}^{|r|} g_i \cdot \omega_i \quad (1)$$

$$g_i = \begin{cases} \frac{e_j}{|\boldsymbol{b_j}|}, & \text{if } i \in \boldsymbol{b_j}, j = \{1, \cdots, k\} \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

$$\omega_i = \begin{cases} 1, & \text{if } i \text{ is on the loop} \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

where g_i and ω_i represents the normalized minimum free energy and its weight value at position i ($1 \leq i \leq |r|$) of a given mRNA, respectively. $\boldsymbol{b_j}$ is the position of j-th binding site, and $\boldsymbol{B} = \{\boldsymbol{b_1}, \cdots, \boldsymbol{b_k}\}$ refers to a set of $\boldsymbol{b_j}$. e_j and $|\boldsymbol{b_j}|$, which are the MFE and the length of the binding site, respectively. The output ω_i will be 1 when the binding site is on the loop structure of RNA; otherwise, it is 0.

4) Sorting and filtering by adjusted minimum free energy: Total energy scores from the section II-B3 were sorted in ascending order. Then, only the mRNA-miRNA pairs whose score is higher than the value of threshold γ were recorded.

III. Results

The number of miRNA-mRNA interactions after the procedure of prediction of target binding sites is 1,586,771. The distribution of the number of mRNAs via the number of binding sites is depicted in Fig. 2. Among these, the number of miRNA-mRNA pairs that have a binding site is 1,233,678 (77.7%) and the number of those that have two binding sites is 253,629 (16.0%). The number of miRNA-mRNA pairs of $n \geq 5$ and $n \geq 9$ is 99,464 (1%) and 1,728 (0.11%) respectively. The miRNA-mRNA pairs that have one or two binding sites are removed because they are too low to be a miRNA sponge. A total of 99,464 (1%) whole miRNA-mRNA pairs moved on to the next procedure.

The top 20 miRNA-mRNA pairs are shown in Table II. The pair of mRNA *AB600271* and miRNA *ebv-miR-BART12* has the highest score of 664.48. The top three highest-scoring mRNAs have a similar number of binding sites and close scores because they are similar mRNA sequences. *L21998* (No.5) has 107 binding sites, which has the largest number of binding sites among the mRNAs considered; however, its total score is not the highest because most of its binding sites are on the helix region. On the other hand, *AK308816* (No.12) has a low number of binding sites (13 sites), but

TABLE II. THE TOP 20 MRNA-MIRNA PAIRS THAT HAVE THE HIGHEST SUM OF ADJUSTED MINIMUM FREE ENERGY

No.	mRNA label	miRNA label	#binding	mRNA length (nt)	Score† (kcal/mole)	Description
1	*AB600271*	*ebv-miR-BART12*	49	6,019	664.48	*PBMUCL1* mRNA type1
2	*AB600272*	*ebv-miR-BART12*	49	5,970	656.02	*PBMUCL1* mRNA type2
3	*AB560770*	*ebv-miR-BART12*	49	6,005	654.53	*PBMUCL1* mRNA
4	*AK131380*	*ebv-miR-BART12*	30	2,230	575.09	cDNA *FLJ16449* fis
5	*L21998*	*kshv-miR-K12-10b*	107	15,720	557.18	intestinal mucin (*MUC2*) mRNA
6	*AK123111*	*ebv-miR-BART3-5p*	84	3,619	496.29	cDNA *FLJ41116* fis
7	*AJ606308*	*ebv-miR-BHRF1-3*	40	14,094	399.20	mRNA for secreted mucin *MUC17* (*MUC17* gene)
8	*AJ606307*	*ebv-miR-BHRF1-3*	40	14,246	396.01	mRNA for membrane mucin *MUC17* (*MUC17* gene)
9	*M64594*	*ebv-miR-BART12*	27	1,830	381.12	Human tracheo-bronchial mucin (*MUC4*) mRNA
10	*L11672*	*kshv-miR-K12-3-5p*	35	3,839	305.84	Kruppel related zinc finger protein (*HTF10*) mRNA
11	*AB209162*	*ebv-miR-BART3-5p*	18	5,189	300.32	mRNA for PR domain
12	*AK308816*	*ebv-miR-BART12*	13	1,292	299.91	*FLJ98857*
13	*AK291256*	*kshv-miR-K12-3-5p*	34	3,851	285.06	cDNA, *FLJ98857*
14	*X06290*	*ebv-miR-BART12*	30	13,938	269.39	mRNA for apolipoprotein(a)
15	*AK128822*	*ebv-miR-BART2-5p*	18	1,761	265.14	cDNA *FLJ46108* fis
16	*AK300294*	*kshv-miR-K12-3-5p*	35	3,572	263.23	cDNA *FLJ57591*
17	*AB051895*	*kshv-miR-K12-6-5p*	45	15,534	255.98	mRNA for epiplakin 1 (*EPPK1*)
18	*AK128811*	*kshv-miR-K12-10b*	22	3,249	253.40	cDNA *FLJ46008* fis
19	*AB209086*	*kshv-miR-K12-3-5p*	26	4,368	238.66	mRNA for zinc finger protein 493
20	*AK131380*	*kshv-miR-K12-10b*	20	2,230	236.65	cDNA *FLJ16449* fis

† The sum of absolute values of adjusted minimum free energy (AMFE)

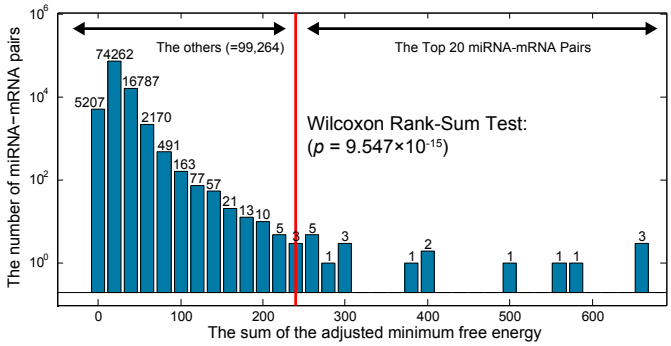

Fig. 3. The distribution of the total scores of miRNA-mRNA pairs.

it has a relatively high total score considering its sequence length. Therefore, these results imply that the total scores are greatly affected by the secondary structure of the RNA. Fig. 3 shows the distribution of the total scores of the miRNA-mRNA pairs. According to the total scores, we could divide the pairs into two groups: the top 20 highest-scoring pairs and the rest. We performed a non-parametric Wilcoxon rank-sum test to statistically assess the difference between the average values of the two groups and then could reject the null hypothesis that they were equivalent (p-value = 9.547×10^{-15}).

There are many high scoring mRNAs whose sequence length is over 10,000 nt, such as *L21998*, *AJ606308*, *AJ606307*, *X06290*, and *AB051895*. It is possible that the length of the mRNA and the total score are highly correlated; therefore, we performed a correlation analysis on their relationship. Given miRNA-mRNA pairs, the correlation coefficient between their total scores and the lengths of the mRNAs in the pairs is only 0.135 (p-value $< 10^{-15}$), suggesting no significant correlation between the total score and the mRNA length.

IV. DISCUSSION

In various papers, it has been consistently reported that there are interactions between virus-encoded miRNAs and host mRNAs [18], [19]. Moreover, mammals have developed natural sponges or ceRNA with the purpose of neutralizing the action of miRNAs [20]–[22]. Through these two results, it is likely that cellular endogenous sponges against viral miRNAs have been naturally developed in the host defense system and that they might play important roles to protect the host during virus infection. In addition, in virus-host interactions, it was reported that host cellular miRNA could be regulated by virus non-coding RNA; however, there were no reports about the reverse case, the regulation of viral miRNA by host-originated non-coding RNA [23].

Many miRNA target prediction algorithms such as miRanda, RNAhybrid, TargetScan(S), DIANAmicroT, PITA, PicTar, and RNA22 were developed [24], [25]. Most of these algorithms proposed to predict binding sites to the 3' URT region of target mRNAs; however, the general tools used to find the interactions between miRNAs and any RNAs were needed in this study. miRanda, RNAhybrid, and RNA22 met our needs. We chose miRanda because the user interface of RNA22 is inconvenient and requires frequent human interventions especially for large datasets, and RNAhybrid has relatively low prediction accuracy [26].

The validation of scoring scheme and several parameters such as the length of mRNA, the number of binding sites, and thresholds, was not performed in this study because of the lack of validated natural miRNA sponges. Conventional miRNA sponges were produced in needs; therefore, the characteristics of natural sponges and artificial sponges are quite different. Artificial miRNA sponges are designed for relatively high binding density such that there are up to tens of binding sites in the length of hundreds of nucleotides [5]. In the case of miRNA sponges of a circular type, most candidates of sponges were not yet experimentally validated as well. Although there are a few circular-typed sponges such as *ciRS-7*, *Sry*, and *ZNF91* [10], existing prediction algorithms for the secondary structure were mainly designed for linear-type RNAs but not for circular-type ones. As more natural miRNA sponges were validated, parameter optimization would become possible.

We could derive more meaningful results if we extend the query miRNAs to whole viral or human miRNAs. However, algorithm parallelization for speed-up is needed. Many of the current algorithms used to predict binding sites and RNA secondary structures have a high time complexity due to the use of dynamic programming. The proposed approach can easily be parallelized since the amount of internode communication is low [27]. Thus, using grid computing or Hadoop will accelerate our method significantly.

V. Conclusion

It is important to identify natural miRNA sponges; however, there have been no large-scale studies or in-depth investigations about pairwise interactions of miRNA-mRNA pairs. In this study, interactions between 30 viral miRNAs and about 300 of thousands of human mRNAs were investigated to find natural miRNA sponge candidates. The binding sites of whole pairs of miRNAs and mRNAs were computed by miRNA target prediction tools, and additional information such as the secondary structure of mRNAs was then exploited. Finally, the 20 highest-scoring mRNAs were chosen as sponge candidates of interesting miRNA. The proposed approach could be useful as precedent research before *in vivo* experiments to validate the identification of miRNA sponges.

Acknowledgment

This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (Ministry of Science, ICT and Future Planning) [No. 2011-0009963, No. 2012M3A9D1054622, and No. 2014M3C9A3063541], and in part by Samsung Electronics Co., Ltd.

References

[1] D. P. Bartel, "MicroRNAs: target recognition and regulatory functions," *Cell*, vol. 136, no. 2, pp. 215–233, 2009.

[2] Y. W. Kong, D. Ferland-McCollough, T. J. Jackson, and M. Bushell, "microRNAs in cancer management," *The lancet oncology*, vol. 13, no. 6, pp. e249–e258, 2012.

[3] A. Roberts, A. P. Lewis, and C. L. Jopling, "The role of microRNAs in viral infection." *Progress in molecular biology and translational science*, vol. 102, pp. 101–139, 2010.

[4] M. S. Ebert, J. R. Neilson, and P. A. Sharp, "MicroRNA sponges: competitive inhibitors of small rnas in mammalian cells," *Nature methods*, vol. 4, no. 9, pp. 721–726, 2007.

[5] M. S. Ebert and P. A. Sharp, "MicroRNA sponges: progress and possibilities," *Rna*, vol. 16, no. 11, pp. 2043–2050, 2010.

[6] M. S. Ebert and P. A. Sharp, "Emerging roles for natural microRNA sponges," *Current Biology*, vol. 20, no. 19, pp. R858–R861, 2010.

[7] L. Salmena, L. Poliseno, Y. Tay, L. Kats, and P. P. Pandolfi, "A cerna hypothesis: the Rosetta Stone of a hidden RNA language?" *Cell*, vol. 146, no. 3, pp. 353–358, 2011.

[8] J. M. Franco-Zorrilla, A. Valli, M. Todesco, I. Mateos, M. I. Puga, I. Rubio-Somoza, A. Leyva, D. Weigel, J. A. García, and J. Paz-Ares, "Target mimicry provides a new mechanism for regulation of microRNA activity," *Nature genetics*, vol. 39, no. 8, pp. 1033–1037, 2007.

[9] T. B. Hansen, T. I. Jensen, B. H. Clausen, J. B. Bramsen, B. Finsen, C. K. Damgaard, and J. Kjems, "Natural RNA circles function as efficient microRNA sponges," *Nature*, vol. 495, no. 7441, pp. 384–388, 2013.

[10] J. U. Guo, V. Agarwal, H. Guo, and D. P. Bartel, "Expanded identification and characterization of mammalian circular RNAs," *Genome Biol*, vol. 15, no. 409, pp. 304–313, 2014.

[11] S. Ghosal, S. Das, R. Sen, P. Basak, and J. Chakrabarti, "Circ2Traits: a comprehensive database for circular RNA potentially associated with disease and traits," *Frontiers in genetics*, vol. 4, 2013.

[12] A. Kozomara and S. Griffiths-Jones, "miRBase: annotating high confidence microRNAs using deep sequencing data," *Nucleic acids research*, p. gkt1181, 2013.

[13] D. A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers, "GenBank," *Nucleic acids research*, vol. 41, no. D1, pp. D36–D42, 2013.

[14] A. J. Enright, B. John, U. Gaul, T. Tuschl, C. Sander, D. S. Marks *et al.*, "MicroRNA targets in Drosophila," *Genome biology*, vol. 5, no. 1, pp. R1–R1, 2004.

[15] R. Lorenz, S. H. Bernhart, C. H. Zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, I. L. Hofacker *et al.*, "ViennaRNA Package 2.0." *Algorithms for Molecular Biology*, vol. 6, no. 1, p. 26, 2011.

[16] M. Hariharan, V. Scaria, and S. K. Brahmachari, "dbSMR: a novel resource of genome-wide SNPs affecting microRNA mediated regulation," *BMC bioinformatics*, vol. 10, no. 1, p. 108, 2009.

[17] I. L. Hofacker, "Vienna RNA secondary structure server," *Nucleic acids research*, vol. 31, no. 13, pp. 3429–3431, 2003.

[18] A. Grundhoff and C. S. Sullivan, "Virus-encoded micrornas," *Virology*, vol. 411, no. 2, pp. 325–343, 2011.

[19] V. Nair and M. Zavolan, "Virus-encoded microRNAs: novel regulators of gene expression," *Trends in microbiology*, vol. 14, no. 4, pp. 169–175, 2006.

[20] Y. Tay, L. Kats, L. Salmena, D. Weiss, S. M. Tan, U. Ala, F. Karreth, L. Poliseno, P. Provero, F. Di Cunto *et al.*, "Coding-independent regulation of the tumor suppressor PTEN by competing endogenous mRNAs," *Cell*, vol. 147, no. 2, pp. 344–357, 2011.

[21] M. Cesana, D. Cacchiarelli, I. Legnini, T. Santini, O. Sthandier, M. Chinappi, A. Tramontano, and I. Bozzoni, "A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA," *Cell*, vol. 147, no. 2, pp. 358–369, 2011.

[22] P. Sumazin, X. Yang, H.-S. Chiu, W.-J. Chung, A. Iyer, D. Llobet-Navas, P. Rajbhandari, M. Bansal, P. Guarnieri, J. Silva *et al.*, "An extensive microRNA-mediated network of RNA-RNA interactions regulates established oncogenic pathways in glioblastoma," *Cell*, vol. 147, no. 2, pp. 370–381, 2011.

[23] C. Li, J. Hu, J. Hao, B. Zhao, B. Wu, L. Sun, S. Peng, G. F. Gao, and S. Meng, "Competitive virus and host RNAs: the interplay of a hidden virus and host interaction," *Protein & cell*, vol. 5, no. 5, pp. 348–356, 2014.

[24] H. Min and S. Yoon, "Got target?: computational methods for microRNA target prediction and their extension," *Experimental & molecular medicine*, vol. 42, no. 4, pp. 233–244, 2010.

[25] S. Yoon and G. De Micheli, "Computational identification of microRNAs and their targets," *Birth Defects Research Part C: Embryo Today: Reviews*, vol. 78, no. 2, pp. 118–128, 2006.

[26] Y. Zhang and F. J. Verbeek, "Comparison and integration of target prediction algorithms for microRNA studies," *J Integr Bioinform*, vol. 7, no. 3, p. 127, 2010.

[27] Y. Jeon and S. Yoon, "Multi-threaded hierarchical clustering by parallel nearest-neighbor chaining," *IEEE Transactions on Parallel and Distributed Systems*, vol. 26, no. 9, pp. 2534–2548, 2015.