Computational discovery of miR-TF regulatory modules in human genome

Dang Hung Tran*[1,3], Kenji Satou[1,2] , Tu Bao Ho[1], Tho Hoan Pham[3]

[1]School of Knowledge Science, Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan [2]Kanazawa University, Kakuma-machi, Kanazawa 920-1192, Ishikawa, Japan [3]Hanoi National University of Education, 136 Xuanthuy, Caugiay, Hanoi, Vietnam. Dang Hung Tran- Email: D. H. Tran - hungtd@jaist.ac.jp; K;*Corresponding author

Abstract:

MicroRNAs (miRNAs) are small non-coding RNAs that regulate gene expression at the post-transcriptional level. They play an important role in several biological processes such as cell development and differentiation. Similar to transcription factors (TFs), miRNAs regulate gene expression in a combinatorial fashion, i.e., an individual miRNA can regulate multiple genes, and an individual gene can be regulated by multiple miRNAs. The functions of TFs in biological regulatory networks have been well explored. And, recently, a few studies have explored miRNA functions in the context of gene regulation networks. However, how TFs and miRNAs function together in the gene regulatory network has not yet been examined. In this paper, we propose a new computational method to discover the gene regulatory modules that consist of miRNAs, TFs, and genes regulated by them. We analyzed the regulatory associations among the sets of predicted miRNAs and sets of TFs on the sets of genes regulated by them in the human genome. We found 182 gene regulatory modules of combinatorial regulation by miRNAs and TFs (miR-TF modules). By validating these modules with the Gene Ontology (GO) and the literature, it was found that our method allows us to detect functionally-correlated gene regulatory modules involved in specific biological processes. Moreover, our miR-TF modules provide a global view of coordinated regulation of target genes by miRNAs and TFs.
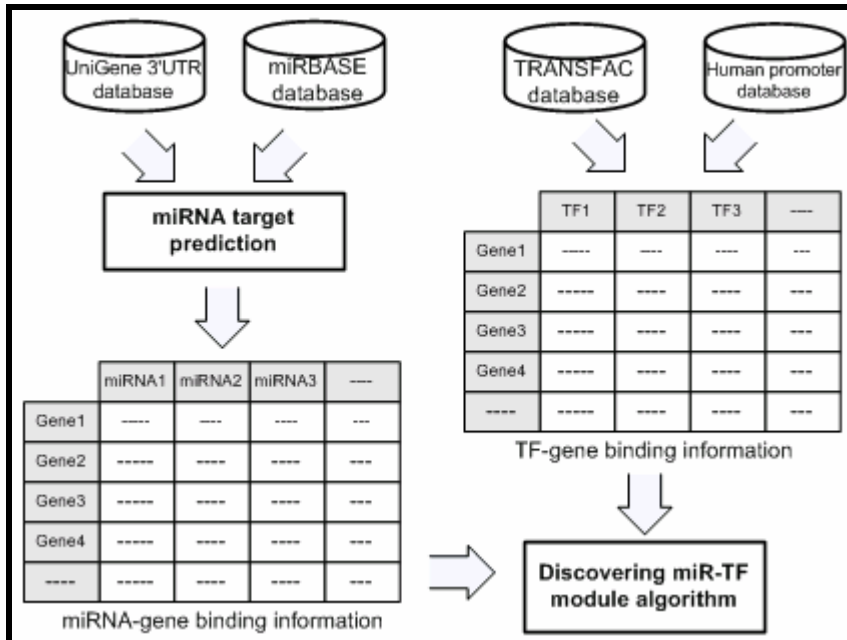
Background:

MicroRNAs (miRNAs) are a class of endogenous non-coding small RNAs. They negatively regulate gene expression at the post-transcriptional level through binding to complementary sites in the 3'-untranslated regions (3'-UTR) of target genes [1, 2]. The first miRNA was found by Victor Ambros and his colleagues in 1993. During the last several years, hundreds of miRNAs and their target genes have been identified in mammalian cells. Recent studies suggest that miRNAs play critical roles in multiple biological processes, including cell growth, cell differentiation, embryo development and so on [2, 3]. In cellular organisms, the regulation of gene expression is an important mechanism to control biological processes. To understand how differential gene expression is controlled at the genome-wide level, it is important to identify all factors involved, and how and when they affect gene expression. To date, differential gene expression can be regulated at the transcriptional and at the post-transcriptional levels by transcription factors (TFs) and miRNAs [4]. TFs can activate or repress transcription by binding to *cis*-regulatory elements located in the upstream regions of genes. Many TFs are well characterized, and corresponding binding motifs have been identified in various species. The combinatorial interactions between TFs have been well studied [5, 6]. MiRNAs repress gene expression by physically interacting with *cis*-regulatory elements located in the 3'-UTR of their target messenger RNAs (mRNAs). Compared with the work on transcriptional regulation by TFs, the study of biological regulatory networks mediated by miRNAs is just beginning. There are a few studies which have investigated the regulatory functions of miRNAs in the gene regulation network [7, 8].
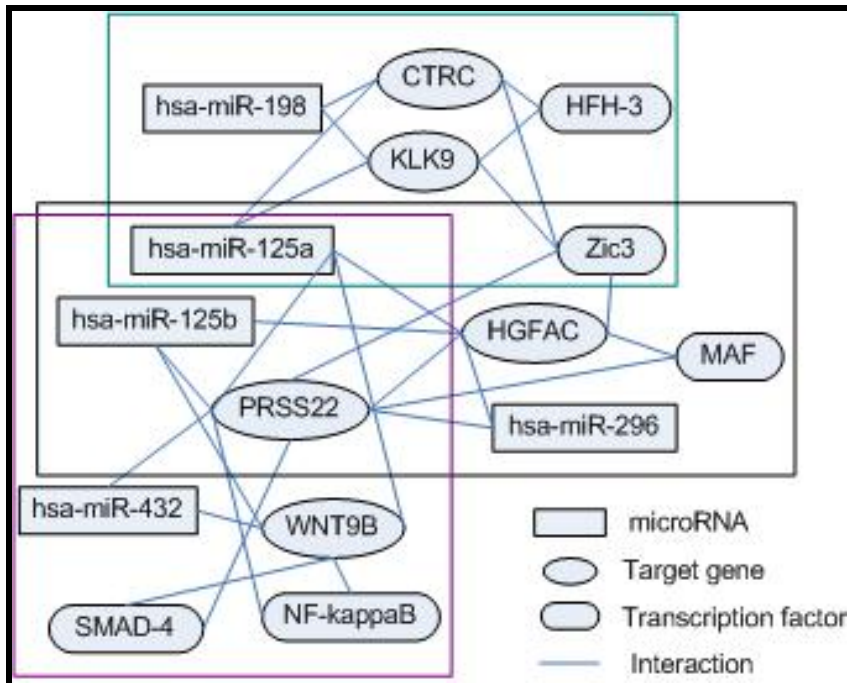
Similar to TFs, miRNAs regulate gene expression in a combinatorial fashion, i.e., an individual miRNA can regulate multiple genes, and an individual gene can be regulated by multiple miRNAs. We also know that a gene is regulated through multiple mechanisms before producing its products. Thus, an attractive possibility is that miRNAs and TFs may cooperate in regulating the same target genes. This cooperation of miRNAs and TFs reflects the collaborative regulation at the transcriptional and post-transcriptional levels. The combined complexity of miRNAs and TFs may be relevant to creating cellular complexity in a developing organism [9]. However, how TFs and miRNAs function together in the gene regulatory network has not yet been examined. Investigating coordinated regulation of genes by miRNAs and TFs is an effort that may elucidate some of their roles in various biological processes.

Detection of microRNA regulatory modules within biological meanings has received much attention recently. Tran *et al.* [10] presented a rule-based learning method to find miRNA regulatory modules that contain coherent miRNAs and mRNAs. Joung and Fei [11] used an author-topic model, which is a family of probabilistic graphical models for identification of miRNA regulatory modules. Recently, Liu *et al.* [12] tried to identify negatively regulated patterns of miRNAs and mRNAs which are associated with specific conditions. These studies, however, only concentrated on finding miRNA functions at the post-transcriptional level. Two other recent studies also showed that miRNA mediated regulatory circuits are prevalent in the human genome [8, 13]. Thus, the question regarding to the combined roles of miRNAs and other factors in gene regulatory network still remains unanswered. Unlike previous studies as discussed above, in this paper, we propose a new computational method to discover the gene regulatory modules consisting of miRNAs, TFs, and the genes regulated by them. The method combines TF-gene binding information and miRNA target gene data for module discovery. Specifically, we analyzed the regulatory associations among the sets of predicted miRNAs and sets of TFs on the sets of regulated genes by them in the human genome. We found 182 gene regulatory modules of combinatorial regulation by miRNAs and TFs, named miR-TF regulatory modules (miR-TF modules in short). By validating these modules with the Gene Ontology (GO) and the literature, it was found that our method allows us to detect functionally-correlated gene regulatory modules involved in specific biological processes. Moreover, our miR-TF modules provide a global view of coordinated regulation of target genes by miRNAs and TFs. Results suggest that combinatorial regulation at the transcriptional and post-transcriptional levels is more complicated than previously thought. Taken together, our results may provide insights into how miRNAs and TFs function in the complex regulatory network of the human genome.

Figure 1: Schematic description of the data set construction. The miRNA-gene binding information is constructed from miRNA and 3'-UTR databases by using a miRNA target prediction, and the TF-gene binding information is constructed by TF binding site prediction integrating promoter and TRANSFAC databases.

Figure 2: An illustration of three miR-TF modules discovered by our algorithm Inside green rectangle, black rectangle, and cyan rectangle are modules 68, 66, and 67, respectively. The line from a miRNA/TF to a gene indicates that miRNA/TF regulates that target gene. MiRNAs repress target genes, while TFs may activate or repress target genes.

Methodology:
Datasets:
We obtained the miRNA regulatory signatures and TF regulatory signatures from CRSD database [14]. CRSD utilizes and integrates six well-known, large-scale databases, including the human UniGene, TRANSFAC, mature miRNAs, putative promoter, pathway and GO databases. The miRNA regulatory signature database was constructed

using mature human miRNA from miRBase [15] and 3'-UTR sequences, and TF regulatory signature database was constructed using TRANSFAC [16] and the promoter database [17]. From these databases, we build a data set that contains three components, miRNAs, TFs, and genes regulated by them. To avoid trivial cases of binding associations, each gene in our data set is regulated by at least two miRNAs and two TFs with a significant binding score (p-value <

5 x 10^{-2} for miRNA-target gene binding, and *p*-value < 5 x 10^{-2} for TF-target gene binding). Specifically, our data set contains 267 mature miRNAs, 483 TFs, and 1253 genes regulated by miRNAs and TFs. On average, each miRNA binds to 22 genes and each TF binds to 17 genes. In addition, each gene was regulated by 5 miRNAs and 6 TFs on average. A schematic description of the data set construction is illustrated in **Figure 1.**

Algorithm:
In this section we present our algorithm for discovering miR-TF regulatory modules. As mentioned above, each module consists of genes, miRNAs, and TFs that all these miRNAs and TFs regulate these genes with a stringent *p*-value. Our method exhaustively and efficiently searches on the entire combinatorial space of subsets of miRNAs and subsets of TFs to discover what factors control what genes. *Algorithm:* Let A be a matrix of miRNA-gene binding *p*-values, where rows correspond to genes and columns correspond to miRNAs, so that A*ij* denotes the binding *p*-value of gene *i* with miRNA *j*. And, let B be a matrix of TF-gene binding *p*-values, where rows correspond to genes and columns correspond to TFs, so that B*ij* denotes the binding *p*-value of gene *i* with TF *j*. Let M(*i, p*) denotes the set of all miRNAs that bind to gene *i* with a *p*-value less than *p*. Let T(*i, p*) denotes the set of all TFs that bind to gene *i* with a *p*-value less than *p*. Let G(*X, p*) be a set of all genes to which all the factors (i.e. miRNAs or TFs) in *X* bind with a given significance threshold *p*. Our algorithm begins by going over all genes *i*. Using a strict *p*-value *p*1, we search all subsets of M'$\subseteq$ M(*i, p*1) that not have been explored in previous steps (i.e. the function *IsAvailable(M')* is true). For each M', we search all subsets of T'$\subseteq$ T(*i, p*2) that not have been explored in previous steps (i.e. the function *IsAvailable(T')* is true). Then, we find the set of all genes G that are bound by all miRNAs in M' and all TFs in T' with given significance thresholds *p1* and *p2*. Finally, if the number of genes in G is greater than 1, we obtain a module and mark all the subsets of M' and T', so that they are not considered next time. All steps of the algorithm are presented in Algorithm 1. Basically, our algorithm works with a dataset containing more than one hundred miRNAs and TFs. Thus, there are potentially exponential numbers of larger-size subsets of miRNAs (or TFs). However, as we showed in the previous section, each gene is bound by only 5 miRNAs and 6 TFs on average. In addition, we mark all the subsets of M' and T' when they are involved in the same module, so that the number of subsets of T and M to be considered is reduced in the algorithm. Therefore, the number of actual subsets we are searching is much smaller.

Results and Discussion:
miRNA-TF regulatory module detection:
We have applied our module discovery algorithm to genome-wide binding data for 267 mature miRNAs and 483 TFs. In order to determine the appropriate values for miRNA-gene interaction threshold *p*1 and TF-gene interaction threshold *p*2, we have tested binding data at a number of different confidence levels (0.05*,* 0.01*,* 0.005*,* and 0.001). **Table 1 (supplementary material)** shows the number of modules found by our algorithm with different *p*-value thresholds. As can be seen, when we relax the binding thresholds, the number of modules found increases. For further analysis, we selected the values of *p*1 and *p*2 equal to 0.01. Consequently, we obtained 182 miR-TF modules. These modules contain the relationships between 107 miRNAs, 176 TFs, and 261 target genes regulated by them **(Table 1 - supplementary material)**. Among 182 miR-TF modules found, we take fifteen modules as examples **(Table 2 - supplementary material)**, the remaining module data can be found in the supplement files.

As mentioned above, each module consists of three components, miRNAs, TFs, and genes regulated by all miRNAs and TFs in the module. And the relationship between one gene and one miRNA or TF represents the strength of binding score. We found that some of our modules share a subset of miRNAs or TFs on regulation of the target genes. This is because of the complexity of the miRNA regulation network. Consequently, our results provide a global view of the combinatorial coregulation of two main factors on gene regulation network. A part of the network constructs from miR-TF modules is shown in **Figure 2**. Three modules (68, 66, and 67) share a common miRNA hsa-miR-125b; two modules (66 and 67) share two miRNAs: hsa-miR-125a and hsa-miR-125b. Also, modules 68 and 66 have the same TF Zic3. This suggests that the coordinated regulation of target genes by miRNAs and TFs is more complicated than previously thought.

Validation using Gene Ontology annotations:
To test if the regulated target genes and the host genes of TFs for each miR-TF module might be enriched functionally based on arbitrary GO terms, we performed GO annotation and significance analysis using GOstat **[18]**. We observed terms associated significantly with the target genes included in the GO gene-association database. To find significantly overrepresented GO terms associated with a given gene set, GOstat counts the number of appearances of each GO term for the genes in this group, as well as in the reference group. Then, for each GO term, a *p*-value is calculated, representing the probability that the observed numbers of counts could have resulted from randomly distributing this GO term between the tested group and the reference group. The GO terms most specific to the analyzed list of genes will have the lowest *p*-values.

Table 3 (supplementary material) shows fifteen significant GO terms directly related to the set of regulated target genes and host genes of TFs in miR-TF modules. Some of them are annotated as regulating specific biological processes (GO:0019219, GO:0019222, GO:0045449, GO:0031323). Other terms have function in organ morphogenesis (GO:000987), gland development (GO:0048732), and organ development (GO:0048513). In addition, GO terms associated with the gene set of miR-TF modules are not so general, for example: GO:000987, GO:0019222, and GO:0019219 are located at levels 8, 6, and 7 in the hierarchical structure of Gene Ontology. Thus we speculate that our miRNA-TF modules may be functional biological modules in gene regulation and molecular development.

Functional validation from the literature:
We did a literature review and found that some miRNAs and TFs in miR-TF modules were discovered to have associations with cancer development. Interestingly, several miRNAs have been confirmed to be related to lung and other human cancers. For example, hsa-let-7d acted as a tumor suppressor in human lung cancer; hsa-miR-15a was downregulated in B-cell chronic and lymphocytic leukemia **[19]**. Hsa-miR-125b was downregulated and hsa-miR-25 was upregulated in human breast cancer. These facts suggest that these miRNAs may potentially act as tumor suppressors. Several transcription factors in miR-TF modules have roles in the development of several types of cancers. NF-κB kappa prevents apoptosis in several untransformed and tumor cell types. AP transcription factor family (AP-1 and AP-2) is encoded from a gene which is known to regulate a variety of cellular processes including cellular proliferation, differentiation, and oncogenic transformation **[20].** In addition, ETS family regulates numerous genes, and is involved in stem cell development, and tumorigenesis **[21].** Taken together with evidence shown in Table 4, it is reasonable for us to think that the miR-TF modules we have discovered may be related to human cancers.

Conclusion:
Transcription factors and microRNAs are two main and important factors regulating gene expression. TFs play the functions at the transcriptional level by binding to promoter regions of the target genes. In other way, miRNAs regulate the messenger RNAs at the post-transcriptional level by binding to their 3'-UTR. The functions of miRNAs and TFs in biological regulatory networks recently have been quite well explored. However, the combinatorial regulation by miRNAs and TFs on the common target genes is less understood. In this paper, we introduced a computational method for discovering

functional miR-TF modules and applied this method to the data set in the human genome.

Using GO annotations and the literature review, we found that our algorithm allows us to detect functionally correlated gene regulatory modules involved in specific biological processes. Specifically, miRNAs and the host gene of TFs in our modules may be involved in some development processes, and may be involved in several types of cancer diseases. Moreover, our miR-TF modules provide a global view of coordinated regulation of target genes by miRNAs and TFs.

References:
[1]   A Ambros, *Nature*, **431:**350 (2004) [PMID: 15372042]
[2]   DP Bartel, *Cell*, **116:**281 (2004) [PMID: 14744438]
[3]   L He and GJ Hannon, *Nature Review*, **5:**522 (2004) [PMID: 15211354]
[4]   K Chen and N Rajewsky, *Nature Reviews*, **8:**93 (2007) [PMID: 17230196]
[5]   E Segal *et al.*, *Nature Genetics*, **34:**166 (2003)
[6]   H Yu, M Gerstein, *Proc Natl Acad Sci.*, **103**:14724 (2006) [PMID: 17003135]
[7]   Q Cui *et al., Mol Syst Biol.,* **2:**46 (2006) [PMID: 16969338]
[8]   J Tsang *et al.*, *Mol Cell*, **26:**753 (2007) [PMID: 17560377]
[9]   O Hobert, *TRENDS in Biochemical Sciences*, **29:**462 (2004) [PMID: 15337119]
[10] DH Tran *et al.*, *BMC Bioinformatics*, **9:**S5 (2008) [PMID: 19091028]
[11] JG Joung and Z. Fei, *Bioinformatics*, **25:**387 (2009) [PMID: 19056778]
[12] B Liu *et al.*, *J Biomedical Infor*, **42:**685 (2009) [PMID: 19535005]
[13] R Shalgi *et al.*, *PLoS Comput Biol.*, **3:**42 (2007) [PMID: 17630826]
[14] CC Liu *et al.*, *Nucleic Acids Res.*, **34:**W571 (2006) [PMID: 16845073]
[15] S Griffiths-Jones *et al.*, *Nucleic Acids Res.*, **34:**D140 (2006) [PMID: 16381832]
[16] V Matys *et al.*, *Nucleic Acids Res.*, **31:**374 (2003) [PMID: 12520026]
[17] AS Halees, Z Weng, *Nucleic Acids Res.*, **32:**W191 (2004) [PMID: 15215378]
[18] T Beissbarth, TP Speed, *Bioinformatics*, **20:**1464 (2004) [PMID: 14962934]
[19] GA Calin *et al.*, *PNAS*, **99**:15524 (2002) [PMID: 12434020]
[20] Q Shen *et al.*, *Oncogene*, **27:**366 (2008) [PMID: 17637753]
[21] J Dwyer *et al.*, *Ann N Y Acad Sci.*, **1114**:36 (2007) [PMID: 17986575]
[22] CD Johnson, *Cancer Res.,* **67:**7713 (2007) [PMID: 17699775]
[23] MV Iorio *et al.*, *Cancer Res.*, **65:**7065 (2005) [PMID: 16103053]
[24] L He *et al.*, *Nature*, **435:**828 (2005) [PMID: 15944707]
[25] S Volinia *et al.*, *PNAS*, **103**:2257 (2006) [PMID: 16461460]
[26] B Vincent *et al.*, *Bio Phar.*, **60:**1085 (2000) [PMID: 11007945]
[27] KA Donnell, *Nature*, **435:**839 (2005) [PMID: 15944709]

Supplementary Material:

Algorithm 1: The algorithm for discovering miR-TF regulatory modules

---

Input: Matrix A contains miRNA-target gene binding information; Matrix B contains TF-gene binding information
Output: miR-TF regulatory modules

foreach *gene i* **do**
    $M = M(i,p1)$;
    $T = T(i,p2)$;
    foreach $M' \subseteq M$ **do**
        if *IsAvailable(M')* **then**
            foreach $T' \subseteq T$ **do**
                if *IsAvailable(T')* **then**
                    $G1 = G(M', p1)$;
                    $G2 = G(T', p2)$;
                    $G = G1 \cap G2$;
                    if $|G| > 1$ **then**
                        ReportModule(*M', T' ,G*);
                        Mark all the subsets of *M'* and *T'* so that they are not considered;
                end
            end
        end
    end
end

Table 1: The number of miR-TF modules with different *p*-value thresholds

$p1^a$	$p2^b$	miRNAs[c]	TFs[d]	Genes[e]	miR-TF modules
$5 \times 10^{-2} <$	$5 \times 10^{-2} <$	107	171	253	187
$1 \times 10^{-2} <$	$1 \times 10^{-2} <$	107	176	261	182
$5 \times 10^{-3} <$	$5 \times 10^{-3} <$	105	177	238	170
$1 \times 10^{-3} <$	$1 \times 10^{-3} <$	94	146	196	139

[a] p1 is a miRNA-gene binding p-value; [b] p2 is a TF-gene binding p-value; [c] the number of microRNAs in all modules; [d] the number of transcription factors in all modules; [e] the number of target genes in all modules.

Table 2: Some miR-TF modules discovered

Module#	miRNAs	TFs	Regulated genes
4	hsa-let-7d hsa-let-7e	E2F-1 Nrf-1	GSG2 EZH1
24	hsa-miR-127 hsa-miR-198	SREBP XPF-1	RAPGEF3 OSBPL7
30	hsa-miR-103 hsa-miR-107 hsa-miR-222 hsa-miR-296	CIZ RP58	FOLR1 FOLR3
34	hsa-miR-134 hsa-miR-214 hsa-miR-324-5p	CDX RFX1 Tst-1 Whn	BTRC PORCN

38	hsa-miR-211 hsa-miR-483	AREB6 HeliosA	MRCL3 DNCLI1
40	hsa-miR-106a hsa-miR-125b	HNF-3 HNF3	FURIN PAK4
	hsa-miR-17-5p hsa-miR-423	HNF-3alpha IPF1	
45	hsa-miR-125a hsa-miR-212	AML GATA-1	SP110 STAT2
	hsa-miR-222 hsa-miR-25		
	hsa-miR-367		
60	hsa-miR-133a hsa-miR-133b	CCAATbox MAZR	MAP1A MID1IP1
73	hsa-miR-125a hsa-miR-15b	PITX2 POU1F1	MTL5 BCL2L2
	hsa-miR-193a		
86	hsa-miR-432 hsa-miR-485-5p	AP-2 Hmx3 SREBP-1	SLC41A1 SLC30A2
109	hsa-miR-133a hsa-miR-145 hsa-miR-18a hsa-miR-18b	CAC-bindingprotein SREBP	KNDC1 NR1I2
124	hsa-miR-214 hsa-miR-337	ATF ATF3 Bach2	RABL4 DIRAS1
	hsa-miR-338 hsa-miR-484		
	hsa-miR-338 hsa-miR-484		
135	hsa-miR-152 hsa-miR-22	GATA-1 GATA-2	SLC12A3 TP53
141	hsa-miR-145 hsa-miR-326	Egr-1 Egr-3 NGFI-C	RAB27A RGL1
167	hsa-miR-342 hsa-miR-485-5p	GATA-1 GATA-3	IREB2 AQP2

Table 3: Significant GO terms directly related to the regulated target genes and host genes of TFs in miR-TF modules

#	GOid	Biological process	*p-value*
1	GO:0009887	organ morphogenesis	2.09×10^{-5}
2	GO:0019219	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	7.32×10^{-4}
3	GO:0019222	regulation of metabolic process	6.48×10^{-4}
4	GO:0048732	gland development	1.1×10^{-4}
5	GO:0045449	regulation of transcription	1.15×10^{-3}
6	GO:0031323	regulation of cellular metabolic process	1.2×10^{-3}
7	GO:0006350	transcription	1.62×10^{-3}
8	GO:0048513	organ development	2.54×10^{-3}
9	GO:0051054	positive regulation of DNA metabolic process	2.56×10^{-3}
10	GO:0006351	transcription, DNA-dependent	2.71×10^{-3}
11	GO:0031325	positive regulation of cellular metabolic process	2.71×10^{-3}
12	GO:0030903	notochord development	2.71×10^{-3}
13	GO:0030900	forebrain development	2.71×10^{-3}
14	GO:0045740	positive regulation of DNA replication	2.71×10^{-3}
15	GO:0045944	positive regulation of transcription from RNA polymerase II promoter	4.67×10^{-3}

Biological processes of some significant GO terms were found by GOstat program. GOid is the identification of the Gene Ontology (GO) term. *p*-values were calculated upon assuming hyper-geometric distribution of annotated GO terms.

Table 4: Biological functions of miRNAs and TFs identified in miR-TF modules consistent with the literature

Name	Reported functions	References
microRNAs		
hsa-let-7d	Tumor suppressor (lung cancer)	**[22]**
hsa-miR-125a	Tumor suppressor (breast cancer)	**[23]**
hsa-miR-125b	Tumor suppressor (breast cancer and hodgkin lymphoma)	**[23]**

hsa-miR-17-5p	Oncogene (lung cancer and B-cell lymphomas)	[24]
hsa-miR-15a	Tumor suppressor (B-cell chronic lymphocytic leukemia)	[19]
hsa-miR-222	Oncogene (papillary thyroid carcinoma)	[24]
hsa-miR-25	Tumor suppressor (colon cancer)	[25]
Transcription factors		
NF-kappaB	Prevent apoptosis in several untransformed and tumor cell types	[26]
AP-1 and AP-2	Regulates cellular proliferation, differentiation, apoptosis, oncogene-induced transformation and cancer cell invasion	[20]
c-Myc	Regulates of cell proliferation, growth and apoptosis	[27]
ETS family	Involved in stem cell development, and tumorigenesis, oncogenes and tumor suppressor (prostate cancer)	[21]
GATA-1	Proliferation marker of breast cells, functions in proliferation and differentiation processes	[20]

Additional data

The data set used
http://www.jaist.ac.jp/~tran/miR-TF/dataset.txt; A plain text file, containing the relationships between 1253 target genes, 267 mature miRNAs, and 483 TF, which was used as the data set for this research. Each consecutive three rows describe the relation of one particular gene with miRNAs and TFs.

The list of miR-TF modules discovered
http://www.jaist.ac.jp/~tran/miR-TF/modules.txt; A plain text file containing the list of all 182 miR-TF modlues discovered by our algorithm. Each row describes one module.

Source code
A zip file that contains our source code written in C++, running on Linux operating system. (It can be distributed upon request.)