A New miRNA Motif Protects Pathways' Expression in Gene Regulatory Networks

Alfredo Benso, Stefano Di Carlo, Gianfranco Politano, and Alessandro Savino

Politecnico di Torino, Control and Comp. Engineering Department, Torino (Italy)
E-mail: firstname.lastname@polito.it,
Home page: http://www.testgroup.polito.it/index.php/bio-menu-home

Abstract. The continuing discovery of new functions and classes of small non-coding RNAs is suggesting the presence of regulatory mechanisms far more complex than the ones identified so far. In our computational analysis of a large set of public available databases, we found statistical evidence of an inter-pathway regulatory motif, not previously described, that reveals a new protective role miRNAs may play in the successful activation of a pathway. This paper reports the main outcomes of this analysis.

Keywords: miRNA, regulatory networks, motif, pathways

1 Introduction

In the past few years, the quest for recurrent motifs in the complex world of gene regulation achieved several interesting results and identified the central role of non-coding RNAs in intra-pathway regulation [2]. In our research for higher-level mechanisms of transcriptional regulation, we identified an inter-pathway regulatory motif, not previously described, in which miRNAs seem to play a new role in the successful expression of a pathway. In particular, some intragenic miRNAs, co-expressed with one or more pathway genes, seem to act as *pathway protectors*. We observed that, in several situations, they target the transcription factors of a set of genes hosting miRNAs that could interfere with the pathway successful activation. We found statistical evidence that this inter-pathway regulatory motif, which we called Pathway Protection Loop (PPL), is very common in several classes of KEGG pathways [7].

Figure 1 shows the structure of a PPL. The successful full activation of a pathway not only depends on the correct expression of the pathway genes, but also on controlling the expression of other genes that could potentially dysregulate (down-regulate or silence) the pathway at some point. We call these genes the Pathway Antagonist Genes (PAGs). A possible way PAGs may interfere with the activity of a pathway is to express a set of miRNAs, in this context called Antagonist miRNAs, targeting and down-regulating some of the pathway's genes. Interestingly, we discovered that, in several situations, the pathway intragenic miRNAs target and silence the transcription factors of the PAGs, thus creating

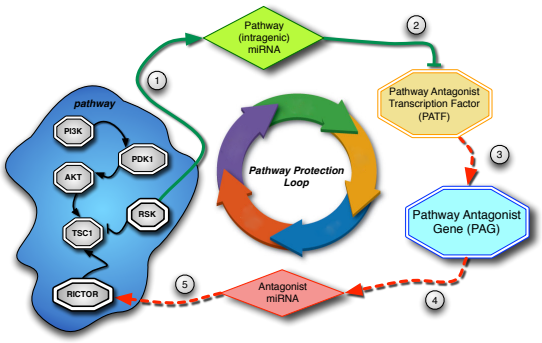

Fig. 1: Pathway Protection Loop (solid lines indicate active interactions; dashed lines indicate inhibited effects).

a loop that seems designed to prevent PAGs from interfering with the pathway expression process. When a PPL is present, its action is carried out in the following steps:

1. When the pathway is activated, one or more pathway genes co-express one or more intragenic miRNAs.
2. The intragenic miRNAs expressed by the pathway target one or more transcription factors of some of the PAGs. We call these transcription factors Pathway Antagonist Transcription Factors (PATFs).
3. The down-regulation of the PATFs has a repressive effect on the expression of the corresponding PAGs.
4. Down-regulated PAGs have lower ability to express the Antagonist miRNAs. miRNAs have a post-transcriptional regulation role [5]. Intragenic miRNAs that directly target the PAGs would not actually prevent the expression of the related Antagonist miRNAs since miRNAs are produced during transcription. The only way PPLs can form is therefore by mediating the PAGs down-regulation through their corresponding PATFs [19].
5. The reduced presence of Antagonist miRNAs contributes to the successful expression of the pathway genes, thus closing the protection loop.

This paper proposes an extensive bioinformatics study designed to analyze the existence and the characteristics of this new motif on a large set of public available pathways. Results will show statistical evidence that this inter-pathway regulatory motif is very common in several classes of KEGG homo sapiens pathways.

2 Results

We analyzed a set of 158 pathways from the KEGG database [7], representing the full set of available Homo Sapiens maps with the exception of human diseases and a few pathways not actually containing a regulatory network. Selected

pathways have been clustered in two main groups based on the KEGG BRITE hierarchy. The first group contains 107 metabolic pathways (KEGG metabolism classification) while the second group contains 51 non-metabolic pathways (9 from KEGG cellular processes classification, 16 from KEGG environmental information processes classification, 6 from KEGG genetic information processes classification and 20 from KEGG organismal systems classification). PPLs in KEGG pathways have been statistically compared to those identified in a population of 107 randomly generated pathways. Genes have been randomly shuffled among the 158 KEGG networks obtaining a set of 158 random pathways. 107 out of the 158 generated networks have been then randomly selected. The number of random pathways selected for the statistical analysis is directly connected to the numerosity of the metabolic pathway group that represents the largest group of our dataset. Using this randomization approach the frequency in which each gene occurs in the data set has been preserved while its position in a given network has been randomly modified.

2.1 Statistical analysis

Table 1 summarizes the overall statistics of our experimental setup. We first investigated statistical differences in the frequency in which PPLs appear among the three considered groups of pathways (metabolic, non-metabolic, random). Table 1-a reports the obtained contingency matrix. A pathway is counted in the L column if it manifests at least one PPL, otherwise it is counted in the NL column. Since miRNAs play a pivotal role in the formation of a PPL, we performed the full analysis by considering high-score miRNA target predictions, only (mirSVR $<$ -0.3, see section 3). PPLs appear in about 55% of non-metabolic pathways while they appear only in about 9% of metabolic pathways. This is in accordance with previous publications that highlight the deep involvement of miRNAs in the regulation of signaling pathways, which represent a large portion of the non-metabolic pathways group [15]. Pearson's Chi-squared test among the three groups confirms this qualitative observation and highlights significant statistical dependence between rows and columns of the contingency table (p = 5.48e-09). To better understand where differences among groups lie, post-hoc analysis has been performed. We performed a chi-squared test considering all possible pairs of groups (i.e., metabolic vs. random, non-metabolic vs. random, metabolic vs. non-metabolic) adjusting resulting p-values using Holms correction. All tests highlighted significant statistical differences among the groups as reported in Table 1-c.

We also investigated if we can observe statistical difference in the number of PPLs per pathway among the different groups. Due to lack of Normality in this variable (Kolmogorov-Smirnov test, p-value $<$ 2.2e-16) we resorted to a Kruskal-Wallis test. The test highlights that there is statistical difference in the number of loops among the three groups of pathways (p-value = 2.633e-09). Again to better understand the differences among the different groups we performed a post-hoc analysis running a set of Mann-Whitney U tests among pairs of different groups

applying a Holms p-value adjustment. All tests confirmed significant statistical difference among the considered groups (p-values are reported in Table 1-e).

From the statistical analysis we can therefore conclude that the presence of the PPL motifs in the set of considered KEGGs is not a random event.

To understand the influence of the miRNA target predictions on the statistics we repeated the statistical analysis considering an increased cut-off value of the miRNA target prediction scores (mirSVR < -0.6) thus reducing the number of predicted targets. Tables 1-b, 1-d and 1-f summarize the overall statistical results for this additional analysis. Also these results confirm the statistically significant presence of the PPLs. This outcome is particularly interesting since it highlights that the identified PPLs mainly involve high-score miRNA gene predictions, thus adding reliability to our findings.

Table 1: Statistical analysis summary.

(a) Pathway frequency contingency table.

	L	**NL**
Metabolic	10 (9.3%)	97 (90.7%)
Non-Metab.	28 (54.9%)	23 (45.1%)
Random	33 (30.8%)	74 (69.2%)

(b) Pathway frequency contingency table.

	L	**NL**
Metabolic	7 (6.5%)	100 (93.5%)
Non-Metab.	24 (47.0%)	27 (53.0%)
Random	25 (23.3%)	82 (76.7%)

(c) Pearson's Chi-Square test p-value on the frequency of loops with Holms correction (mirSVR < -0.3)

	Metabolic	**Non-Metab.**
Non-Metab.	1.318e-09	-
Random	1.746e-4	6.339e-3

(d) Pearson's Chi-Square test p-value on the frequency of loops with Holms correction (mirSVR < -0.6)

	Metabolic	**Non-Metab.**
Non-Metabolic	7.395e-09	-
Random	1.119e-3	4.705e-3

(e) Mann-Whitney U test p-value on the loop numerosity with Holms correction (mirSVR < -0.3)

	Metabolic	**Non-Metab.**
Non-Metab.	1.0e-09	-
Random	2.9e-4	9.2e-4

(f) Mann-Whitney U test p-value on the loop numerosity with Holms correction (mirSVR < -0.6

	Metabolic	**Non-Metab.**
Non-Metab.	9.5e-09	-
Random	1.4e-3	1.6e-3

2.2 An example of a PPL in the Aminoacyl-tRNA biosynthesis pathway

The *Aminoacyl-tRNA biosynthesis* pathway (KEGG map #hs-00970), is "Genetic Information Processing pathway describing the activity of precisely matching amino acids with tRNAs containing the corresponding anticodon [14].

The simplicity of the PPL regulatory mechanisms in this pathway can be appreciated looking at Fig.2. *hsa-mir-215* represents a sort of hub for the creation of PPLs. Interestingly, we observed that this "hierarchical" structure is typical of the PPL motif, where a very small number of intragenic miRNAs (usually one or two) is able to create loops targeting a large number of pathway genes. In this

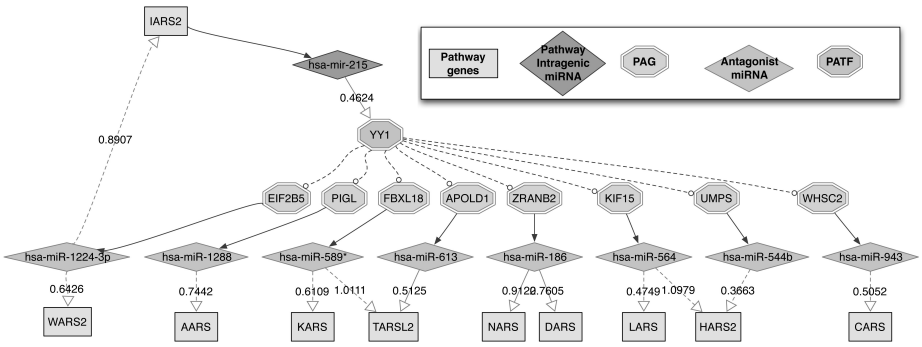

Fig. 2: PPLs identified in the KEGG Aminoacyl-tRNA biosynthesis pathway (hs-00970). Only edges forming PPLs have been depicted.

example, the co-expression of a single miRNA allows the pathway to protect the expression of 25.6% of its genes.

The involvement of YY1 in the formation of PPLs in this pathway is interesting. YY1 is well known as regulator of the expression of numerous genes that are mostly involved in cancers. Its up-regulation has been observed in different types of cancers and it has therefore been proposed as a potential prognostic marker [18]. The fact that this specific gene is also involved in the formation of several PPLs in this pathway and therefore controlled by the regulatory effects of the loop is in accordance with our hypothesis of the protective role this motif has on the correct expression of a pathway. A complete list of the identified PPL with their graphical representation inside each pathway is available at `http://www.testgroup.polito.it/index.php/component/k2/item/184-ppl-list`.

3 Materials and Methods

To study characteristics and properties of PPLs, we designed a PHP-MySQL pipeline integrating pathway and Micronome related data in order to search for miRNA mediated interactions at the pathway level, thus searching for the existence of PPLs. The source code of the pipeline is available at `http://www.testgroup.polito.it/index.php/component/k2/item/185-pathway-protection-loops-finder`.

Pathways' information and structures have been retrieved through Pathway API [17], an aggregated database that combines three major sources of information: (1) the WikiPathway database, the (2) Ingenuity database and the (3) KEGG. One of the main advantages of Pathway API is the normalization of the network nodes that are all consistently translated and named using the corresponding NCBI Gene ID, thus enabling an easy data integration with the other data sources considered in this work.

Figure 3 highlights the data sources and computations steps performed in our pipeline. The pipeline basically consists of two information-retrieval flows

applied to each gene of each target pathway. The outcomes of these two flows are then intersected to identify the presence of PPLs.

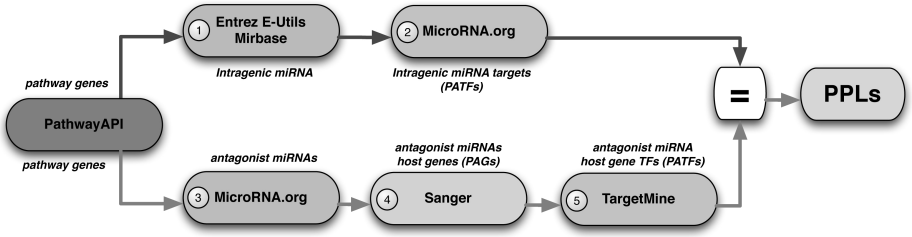

Fig. 3: Computational pipeline for the identification of pathway protection loops.

The external data sources employed in the proposed pipeline are the Entrez E-Utils web API (e-Utils) [13], the miRBase database release 18 (miRBase) [8], the microRNA.org database (microRNA.org) [10], the TargetMine database [3], the Transcription Factor encyclopedia (TFe) [20], and the Sanger Genecode database release 9 (Sanger) [12].

To identify miRNAs co-expressed with the pathway genes we restrict our search to the set of intragenic miRNA (Step 1, Figure 3). Intragenic miRNA represent around 50% of the mammalian miRNAs [11]. We assume that intragenic miRNA are in general co-expressed with their related host-genes as supported by previous studies [16]. Recently Chunjiang et al. [6] also suggested that evolutionary conserved intragenic miRNA tend to be co-expressed with their host genes more likely than poorly conserved ones. This consideration could further refine the outcome of our analysis, however at the current stage it has not yet be implemented in our pipeline.

Intragenic miRNAs are retrieved through the miRBase. To identify intragenic miRNA of a given host gene we first search for the coordinate of the gene using the e-Utils. Once obtained the gene coordinates we search for all miRNAs with coordinates embedded in the ones of the gene resorting to miRBase.

We search for potential targets of each identified intragenic miRNA (Step 2, Figure 3) resorting to microRNA.org, which searches for miRNA targets applying the miRanda algorithm [4]. Other miRNA target databases such as TargetScan [9] use different prediction algorithms that aim at filtering many false positives from the beginning of the prediction process. However, the availability of the mirSVR score in microRNA.org provided us an additional degree of freedom to investigate the robustness of our prediction when changing the way microRNA targets are filtered (see Table 1). To work with reliable predictions and limit the amount of returned miRNA-gene interactions, during the analysis we restricted our search to the microRNA.org "Good mirSVR score, Conserved miRNA" and "Good mirSVR score, Non-conserved miRNA" with negative mirSVR score lower

than -0.3/-0.6. Given the selected intragenic miRNA name, searching for the targets simply requires a SQL query into the microRNA.org database.

Antagonist miRNAs (Step 3, Figure 3) are miRNAs that target one of the genes of the pathway and whose targets, similarly to the Intragenic miRNA, can be retrieved through microRNA.org. Given the NCBI GeneID we query the microRNA.org database to identify the set of miRNAs targeting the gene. Query to microRNA.org at this step follows the same filtering rules on the mirSVR score applied for the identification of the intragenic miRNA targets.

The identification of an antagonist miRNA host gene (Step 4, Figure 3) follows an inverted flow compared to the one employed to identify the pathway intragenic miRNAs. For each antagonist miRNA we identify the related coordinates using miRBase, and, given the coordinates, we search into Sanger for a gene whose coordinates embraces the one of the considered miRNA. Genes identified at this step represent potential PAGs.

As already mentioned in the introduction of this paper, miRNAs act as post-transcriptional regulators. The expression of miRNAs can be activated or repressed by transcription factors of the related host genes, which therefore can serve as upstream regulators of miRNA [19]. For each antagonist miRNA host gene we search for the related transcription factors (Step 5, Figure 3) . Due to the limited availability of information about Transcription Factors in public databases, we aggregated the data from two databases: (1) TargetMine and (2) TFe.

4 Conclusions

The identification of recurrent motifs in regulatory networks is a crucial step in a more comprehensive understanding of the complex gene regulation system. Motifs allow to functionally cluster together groups of genes, and to reduce the complexity of the networks allowing a higher-level view of the regulatory mechanisms. In this work we discussed the presence of an interesting inter-pathway protection loop motif that seems designed to "protect" the pathway activation and expression by an indirect down-regulation of potential antagonist genes.

The discovery of the Pathway Protection Loops is suggesting a possible regulatory mechanism at the pathway level not yet fully investigated. Studies conducted on specific miRNAs such as the one published by Barik [1] confirm the presence of this type of regulatory motif, but a high-level analysis such as the one proposed in this paper is still missing.

The understanding of this and other higher-level regulatory motifs could, for example, lead to new approaches in the identification of therapeutic targets because it could unveil new and "indirect" ways to activate or silence a target pathway.

References

1. Barik, S.: An intronic microrna silences genes that are functionally antagonistic to its host gene. Nucleic acids research 36(16), 5232–5241 (2008)

2. Beezhold, K., Castranova, V., Chen, F.: Microprocessor of microrna: regulation and potential for therapeutic intervention. Mol Cancer 9, 134 (2010)
3. Chen, Y.A., Tripathi, L.P., Mizuguchi, K.: Targetmine, an integrated data warehouse for candidate gene prioritisation and target discovery. PLoS ONE 6(3), e17844 (03 2011)
4. Enright, A., John, B., Gaul, U., Tuschl, T., Sander, C., Marks, D.: MicroRNA targets in drosophila. Genome Biology 5(1), R1+ (2003)
5. Filipowicz, W., Bhattacharyya, S., Sonenberg, N.: Mechanisms of post-transcriptional regulation by microrna: are the answers in sight? Nature Reviews Genetics 9(2), 102–114 (2008)
6. He, C., Li, Z., Chen, P., Huang, H., Hurst, L., Chen, J.: Young intragenic mirnas are less coexpressed with host genes than old ones: implications of mirna–host gene coevolution. Nucleic Acids Research (2012)
7. Kanehisa, M., Goto, S.: KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Research 28(1), 27–30 (2000)
8. Kozomara, A., Griffiths-Jones, S.: mirbase: integrating microrna annotation and deep-sequencing data. Nucleic Acids Research 39(suppl 1), D152–D157 (2011), http://nar.oxfordjournals.org/content/39/suppl_1/D152.abstract
9. Lewis, B., Shih, I., Jones-Rhoades, M., Bartel, D., Burge, C., et al.: Prediction of mammalian microrna targets. Cell 115(7), 787–798 (2003)
10. Memorial Sloan-Kettering Cancer Center: microrna.org - targets and expression database. [Online] http://www.microrna.org/ (2012)
11. Monteys, A., Spengler, R., Wan, J., Tecedor, L., Lennox, K., Xing, Y., Davidson, B.: Structure and activity of putative intronic mirna promoters. Rna 16(3), 495–505 (2010)
12. Sanger Institute: Sanger genecode database. [Online] ftp://ftp.sanger.ac.uk/pub/gencode/release9/gencode.v9.annotation.gtf.gz (2012)
13. Sayers, E.: E-utilities Quick Start. Bethesda (MD): National Center for Biotechnology Information (US) (2010)
14. Sheppard, K., Yuan, J., Hohn, M.J., Jester, B., Devine, K.M., Soll, D.: From one amino acid to another: tRNA-dependent amino acid biosynthesis. Nucleic Acids Res. 36(6), 1813–1825 (Apr 2008)
15. Shirdel, E.A., Xie, W., Mak, T.W., Jurisica, I.: NAViGaTing the micronome–using multiple microRNA prediction databases to identify signalling pathway-associated microRNAs. PLoS ONE 6, e17429 (2011)
16. Shomron, N., Levy, C.: Microrna-biogenesis and pre-mrna splicing crosstalk. Journal of biomedicine and biotechnology 2009 (2009)
17. Soh, D., Dong, D., Yike, G., Wong, L.: Pathwayapi. [Online] http://www.pathwayapi.com/ (2012)
18. Sui, G.: The regulation of yy1 in tumorigenesis and its targeting potential in cancer therapy. Mol Cell Pharmacol 1(3), 157–176 (2009)
19. Wang, J., Lu, M., Qiu, C., Cui, Q.: Transmir: a transcription factor–microrna regulation database. Nucleic acids research 38(suppl 1), D119–D122 (2010)
20. Wasserman Lab: Transcription factor encyclopedia (tfe). http://www.cisreg.ca/tfe (2012)