PeerJ Preprints NOT PEER-REVIEWED

Detecting significant features in modeling microRNA-target interactions

Claudia Coronnello[1,2,*], Giovanni Perconti[1], Patrizia Rubino[1], Flavia Contino[3], Serena Bivona[3], Salvatore Feo[1,3], Agata Giallongo[1]

[1] Istituto di Biomedicina ed Immunologia Molecolare (IBIM) CNR, Palermo, Italy;
[2] Fondazione Ri.MED, Palermo Italy;
[3] Dipartimento di scienze e Tecnologie Biologiche Chimiche e Farmaceutiche, Università degli Studi di Palermo, Italy
[*] Corresponding author: Claudia Coronnello email: ccoronnello@fondazionerimed.com

Introduction

MicroRNAs (miRNAs) are small non-coding RNA molecules mediating the translational repression and degradation of target mRNAs in the cell [1]. Mature miRNAs are used as a template by the RNA-induced silencing complex (RISC) to recognize the complementary mRNAs to be regulated. Up to 60% of human genes are putative targets of one or more miRNAs. Several prediction tools are available to suggest putative miRNA targets, however, only a small part of the interaction pairs has been validated by experimental approaches. In addition, none of these tools does take into account the network structure of miRNA-mRNA interactions, which involves collaboration and competition [2] effects that are crucial to efficiently predict the miRNA regulation effects in a specific cellular context. A first solution to consider collaboration effects is given by the web tool ComiR [3], which predicts the targets of a weighted set of miRNAs, provided the miRNA expression profile of the samples/tissues of interest. The analysis of the expression profile of the RNA fraction immunoprecipitated (IP) with the RISC proteins is an established method to detect which genes are actually regulated by the RISC machinery. In fact, genes that result over-expressed in the IP sample with respect to the whole cell lysate RNA, are considered as involved in the RISC complex, then miRNA targets. Here, we aim to find the features useful to predict which genes are overexpressed in IP, i.e. miRNA targets, without actually performing the IP experiments. To this purpose, we compiled and analyzed a novel high throughput data set suitable to unravel the features involved in the miRNA regulatory activities.

Methods

We used the MCF-7 human breast cancer cells as test bed collecting by IP the RNA associated with two RISC proteins, AGO2 and GW182. Three independent IP experiment for each RISC protein were performed and in each experiment, we collected three samples: total RNA, representing the input (IN) sample, the IP, enriched in miRNA targets, the flow-through (FT), deprived of miRNA targets. Each sample was analyzed by using the Agilent microarray platforms, to obtain both the mRNA and miRNA expression profiles. The expression profiles of the IN were used to compute scores (features) to be assigned to each mRNA. On the other hand, we compared IP *vs* FT samples to detect the mRNA involved in the miRNA regulatory network. We used SAMR algorithm to detect the over- and under-expressed genes in IP *vs* FT, with the three independent replicates. The obtained gene set was used to train and test a Support Vector Machine regression model (leave one out cross validation), able to predict the IP/FT ratio in each single experimental replica. Several strategies to compute the mRNA scores were tested, in order to discover the features relevant to model the entire interaction network. We used the miRNA and mRNA expression profiles of the IN samples, together with other biological features, to set up an algorithm to discriminate the genes enriched in IP (miRNA targets) from the depleted genes (not targets). The performance in discriminating miRNA targets from not targets was computed by using ROC-AUC analysis.

Results

We analyzed separately the mRNA targets co-precipitated with AGO2 and GW182 proteins. Despite the fact that AGO2 and GW182 are both involved in the RISC complex, the computed lists of mRNA over-expressed in AGO2 IP and in GW182 IP show a null overlap (Figure 1A).
We tested several features involved in miRNA-mRNA binding activity. We considered 1) the number of binding sites predicted by TargetScan [4] for each miRNA in the mRNA 3'UTR and coding region; 2) the length of 3'UTR and coding region; 3) the mRNA expression profile; 4) the

miRNA expression profile; 5) formulas using the previous features, as described in Figure 1B. For each of the computed feature, we tested the performance in discriminating miRNA targets from not targets. We also tested the performance of the SVM regression models trained with any pair of the analyzed features. Figure 1C summarizes the obtained results, showing the ROC-AUC values of each performed test.

The two investigated protein, AGO2 and GW182, revealed two different behaviors regarding the strategies to discriminate the mRNA over-expressed in the respective IP sample. The most relevant features associated by the analysis of the RNA co-precipitated with AGO2 protein are the number of miRNA binding sites, predicted in both the 3'UTR and coding region of the target, weighted by miRNA expression. This behavior is well explained by the dynamics of the miRNA activity: the higher is the number of binding sites a miRNA finds on a target sequence, the higher is the chance for the miRNA machinery of regulating the mRNA. In addition, higher expressed miRNAs plays a more relevant role in miRNA regulation. Interestingly, both 3'UTR and coding region are relevant in detecting miRNA targets. Regarding GW182, we find that the most relevant feature is the coding region length. More investigation will be performed to understand whether the different behavior observed regarding AGO2 and GW182 proteins is due to the fact that GW182 is involved in other cellular activities, in addition to miRNA-mediated gene silencing, or it is correlated to the two different miRNA-based activities, i.e. inhibition of translation or mRNA degradation. The obtained information regarding relevant features in detecting over-expressed genes in RISC protein Immunoprecipitated fractio will be used to learn how best to model the miRNA-mRNA interaction network to improve the miRNA targets prediction algorithm ComiR.

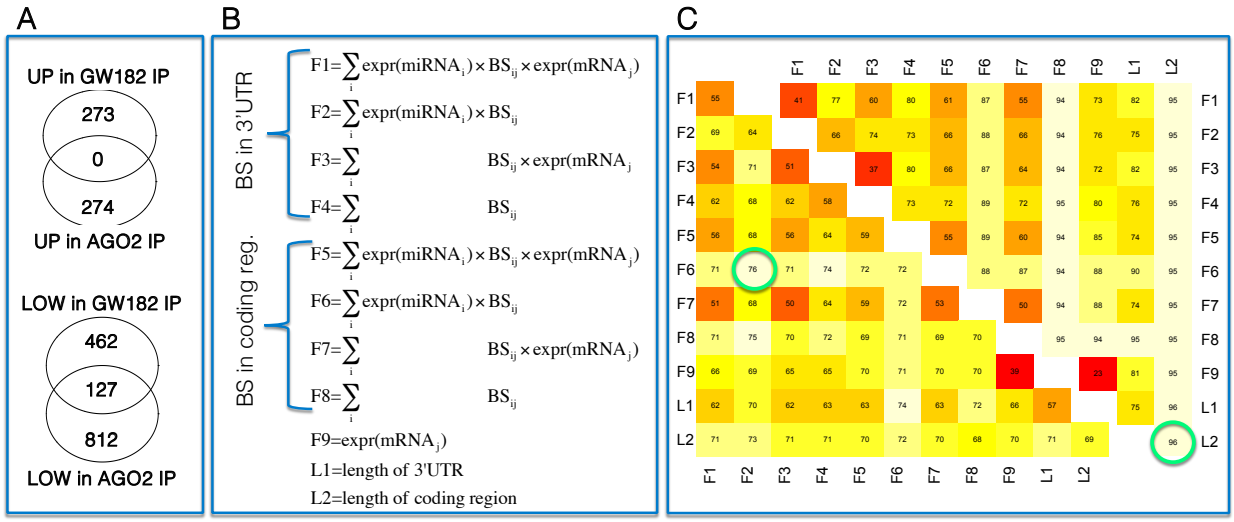

Figure 1: A) Venn plots of differentially expressed genes in AGO2 IP and GW182 IP experiments. B) Details on the computation of the analyzed features. BS_{ij} = number of binding sites of $miRNA_i$ in $mRNA_j$, F1-F4 are computed by using the BS in the 3'UTR, F5-F8 are computed by using the BS in the coding region; $expr(miRNA_i)$ = expression level of $miRNA_i$; $expr(mRNA_j)$ = expression level of $mRNA_j$. C) ROC-AUC values (%) showing the performance in predicting IP/FT ratio of differentially expressed genes. The left triangular matrix refers to AGO2 experiments; the right triangular matrix refers to GW182 IP experiments. In the diagonals, are reported all the single feature performances. In the rest of the matrix is reported the performance of a SVM regression model trained with the correspondent two features. Green circles indicate the best ROC-AUC values.

References

1. Bartel, D.P. (2004) *MicroRNAs: genomics, biogenesis, mechanism and function.* Cell, **116**, 281-297
2. Sumazin, P., *et al.* (2011), *An Extensive MicroRNA-Mediated Network of RNA-RNA Interactions Regulates Established Oncogenic Pathways in Glioblastoma.* Cell. **147**(2): p. 370-381.
3. Coronnello, C et al, (2013) *Novel Modeling of Combinatorial miRNA targeting Identifies SNP with Potential Role in Bone Density*. Plos Computational Biology, **8**, 12