BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 6 2012, pages 771–776
doi:10.1093/bioinformatics/bts043

Sequence analysis

Advance Access publication January 27, 2012

Functional microRNA targets in protein coding sequences

Martin Reczko[1,2,*], Manolis Maragkakis[1,3,4], Panagiotis Alexiou[1,4], Ivo Grosse[3] and Artemis G. Hatzigeorgiou[1,*]

[1]Institute of Molecular Oncology, Biomedical Sciences Research Center 'Alexander Fleming', Vari, Greece, [2]Synaptic Ltd, Heraklion, Greece, [3]Institute of Computer Science, Martin Luther University Halle-Wittenberg, 06120 Halle, Germany and [4]Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, 19104 Philadelphia, USA

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: Experimental evidence has accumulated showing that microRNA (miRNA) binding sites within protein coding sequences (CDSs) are functional in controlling gene expression.

Results: Here we report a computational analysis of such miRNA target sites, based on features extracted from existing mammalian high-throughput immunoprecipitation and sequencing data. The analysis is performed independently for the CDS and the 3′-untranslated regions (3′-UTRs) and reveals different sets of features and models for the two regions. The two models are combined into a novel computational model for miRNA target genes, DIANA-microT-CDS, which achieves higher sensitivity compared with other popular programs and the model that uses only the 3′-UTR target sites. Further analysis indicates that genes with shorter 3′-UTRs are preferentially targeted in the CDS, suggesting that evolutionary selection might favor additional sites on the CDS in cases where there is restricted space on the 3′-UTR.

Availability: The results of DIANA-microT-CDS are available at www.microrna.gr/microT-CDS

Contact: hatzigeorgiou@fleming.gr; reczko@fleming.gr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on August 31, 2011; revised on December 16, 2011; accepted on January 11, 2012

1 INTRODUCTION

MicroRNAs (miRNAs) are small endogenous RNA molecules that play a key role in development and diseases through post-transcriptional regulation of gene expression. They are part of the RNA-induced silencing complex (RISC) and guide it to specific miRNA recognition elements (MREs) on the mRNA molecules of target genes. This leads either to translational repression and/or messenger RNA (mRNA) degradation (Bartel, 2009).

Although most of the MREs have been found in 3′-UTRs of protein coding genes (Papadopoulos *et al*., 2009), there are individual reports of MREs located in protein coding sequences (CDSs) of target genes with evidence for their relation to biological function (Tay *et al*., 2008). In Duursma *et al*. (2008), it is shown that miR-148 represses specific splice variants of DNA methyltransferase 3b (Dnmt3b) by targeting its coding sequence, and that this mechanism might play a role in determining the relative abundance of different splice variants. Forman *et al*. (2008) demonstrate that four let-7 miRNA target sites within the CDS of the miRNA-processing enzyme Dicer establish a mechanism for a miRNA/Dicer autoregulatory feedback loop. In Elcheva *et al*. (2009) it is shown that the coding region of β-transducin repeat containing protein 1 is regulated by miR-183. Takagi *et al*. (2010) show that Hepatocyte nuclear factor 4 α(HNF4α) is downregulated by miR-24 targeting its CDS. The expression of miR-24 is regulated by cellular stress, thus affecting metabolism and cellular biology. Abdelmohsen *et al*. (2010) show that, based on miRNA targeting in the CDS, miR-519 represses the translation of the RNA-binding protein Hu antigen R (HuR), which in turn reduces HuR-regulated gene expression and cell division. Wang *et al*. (2011) measure the effect of four human miRNAs and find that miR-107 tends to target the CDS, but not the 3′-UTR. Finally, Schnall-Levin *et al*. (2011) show that miR-181 targets the repeat-rich CDS of the well-known tumor suppressor retinoblastoma protein (RB1) and RB-associated, Kruppel-associated-box zinc finger (RBAK).

High-throughput CLIP data now allow for the direct identification and localization of MREs on the target genes (Chi *et al*., 2009; Hafner *et al*., 2010). Hafner *et al*. (2010) show through immunoprecipitation of the miRNA containing ribonucleoprotein complexes and sequencing of the associated RNA fragments (PAR-CLIP) that miRNAs tend to bind in approximately equal proportions on the 3′-UTR as well as on the protein coding sequences (CDSs) of target mRNAs. Hafner *et al*. (2010) also suggest that miRNA targeting in the CDS has a measurable effect on miRNA-mediated mRNA degradation. The same observation has been made by two more groups after computational analysis of previously published high-throughput studies regarding miRNA targets. Forman and Coller (2010) analyze the dataset derived from the measurements of protein and mRNA level changes after the transfection of five miRNAs in HeLa cells as provided by (Selbach *et al*., 2008) and detect a functional role of miRNA binding sites in the CDS. Fang and Rajewsky (2011) analyze the same dataset and additionally the protein and mRNA level measurements after over- and underexpression of five miRNAs in mouse neutrophils provided by (Baek *et al*., 2008). They find that genes containing target sites both in the CDS and the 3′-UTR exhibit significantly stronger regulation than genes targeted in the 3′-UTR only and that this effect is stronger for conserved CDS sites with longer binding sites. Schnall-Levin *et al*. (2010) developed an algorithm to predict CDS target sites in Drosophila genes based

*To whom correspondence should be addressed.

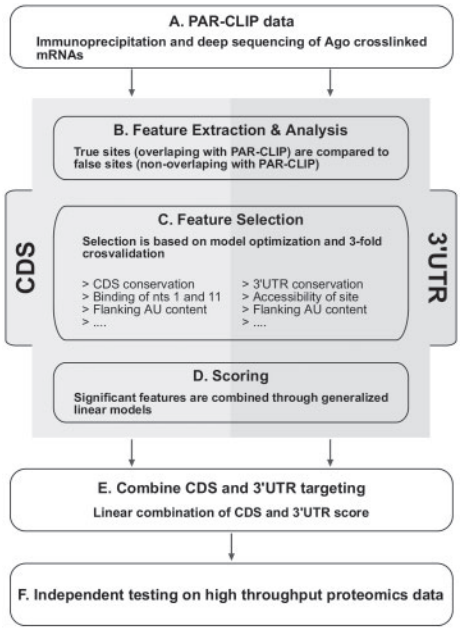

Fig. 1. Flowchart of the analysis on the PAR-CLIP data. MREs specified by the PAR-CLIP data are divided in two categories according to the genomic region in which they lie (**A**). For these two datasets, several features are extracted and the most informative of them are selected by comparing true MREs with false MREs (**B**). The selection is performed through a three-fold cross-validation (**C**). For each identified miRNA MRE, the selected features (depending on the gene region it lies in) are combined into an MRE score through generalized linear models (**D**). For each gene, the CDS score and the 3′-UTR score is defined by summing the MRE scores that lie in CDSs and 3′-UTRs, respectively. These two scores are linearly combined into a final score (**E**). To test for the overall performance of this scoring approach, an independent test on the high-throughput proteomics data of Selbach *et al*. is performed (**F**).

only on conservation, and they validate five of their top seven predictions. Most miRNA target prediction programs nevertheless limit their search for MREs only within the 3′-UTR (Alexiou *et al*., 2009).

Here we describe an algorithm for the prediction of miRNA targets in both 3′-UTRs and CDSs that are trained on a positive and a negative set of MREs defined by PAR-CLIP data of Hafner *et al*. (2010).

The analysis is performed independently for MREs in CDSs and 3′-UTRs, which enables the identification of miRNA:mRNA binding features specific for CDS or UTRs (Fig. 1). For each of these regions, a separate prediction model is built and the models are combined for computing a final miRNA:gene interaction score.

This algorithm permits the identification of a large number of protein-coding genes that are only targeted in their CDSs and provides a model of the interaction between the CDS and the UTR targeting mechanism.

2 METHODS

2.1 Datasets

PAR-CLIP data: the PAR-CLIP data (Fig. 1A) are downloaded from the Supplementary Material of Hafner *et al*. (2010).

Microarray data: microarray data are downloaded from ArrayExpress (http://www.ebi.ac.uk/microarray-as/ae) and from Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo). The datasets used are from Gennarino *et al*. (2009): E-GEOD-12091 (mir-26b), E-GEOD-12092 (mir-98); from Wang and Wang (2006): E-GEOD-6207 (miR-124), E-GEOD-9586 (miR-335); from Linsley *et al*. (2007): GSM155604 (miR-106b); from Grimson *et al*. (2007): GSM210897 (miR-7), GSM210898 (miR-9), GSM210901 (miR-122a), GSM210903 (miR-128a), GSM210904 (miR-132), GSM210909 (miR-142), GSM210911 (miR-148b), GSM210913 (miR-181a).

Proteomics data: changes in protein levels resulting from overexpressing miRNAs hsa-mir-1, hsa-mir16, hsa-mir30a, hsa-mir155 and hsa-let-7b as estimated in Selbach *et al*. (2008) are downloaded from http://psilac.mdc-berlin.de. RefSeq protein IDs are converted to corresponding Ensembl Gene IDs (Ensembl release 54). There are only 120 RefSeq protein IDs that correspond to multiple Ensembl IDs, 20 of which correspond to multiple Ensembl IDs with different 3′-UTR lengths. For these 20 cases, the Ensembl ID corresponding to the longest 3′-UTR is used. In total, 16 164 measurements for potential miRNA:mRNA interactions are identified, of which 2447 have a logarithmic protein downregulation exceeding 0.2 and are considered true targets and 13 717 are considered false targets (see also Supplementary Fig. S4).

HITS-CLIP data: the HITS-CLIP data are downloaded from the Supplementary Material of Chi *et al*. (2009).

miRNA sequences: the miRNAs used are downloaded from miRBase build 13. CDSs and 3′-UTRs are downloaded from Ensembl build 54. In case of multiple CDSs or 3′-UTRs per gene, the longest annotated transcript is used.

Multiple alignments: multiple genome alignments are downloaded from UCSC Genome Browser. Human (hg18) alignments to the following 16 vertebrate genomes are used: panTro1, rheMac2, rn4, mm8, oryCun1, bosTau2, canFam2, dasNov1, loxAfr1, echTel1, monDom4, galGal2, xenTro1, tetNig1, fr1 and danRer3. Mouse (mm9) alignments to the following 16 vertebrate genomes are used: rn4, oryCun1, hg18, panTro2, rheMac, canFam, bosTau3, dasNov1, loxAfr1, echTel, monDom4, galGal3, xenTro2, tetNig, fr2 and danRer5.

miRNA target prediction of other programs: the predictions of all miRNA target prediction programs are obtained as discussed in Alexiou *et al*. (2009). Briefly, flat files of miRanda target prediction data are downloaded (January 2008) from: http://www.microrna.org/microrna/getDownloads.do. For Pictar, the target results are downloaded from the Pictar web page (http://pictar.org/) following the link for 'Predictions in vertebrates, flies and nematodes'. The four species conservation is used. For RNA22, the target prediction data are downloaded from a collection of precompiled predictions dated November 11, 2006. Individual predictions can be calculated at http://cbcsrv.watson.ibm.com/rna22.html. For TargetScan 5.0, data are downloaded from http://www.targetscan.org/cgi-bin/targetscan/data_download.cgi?db=vert_50. Finally, for AnTar, the AnTar targets from http://servers.binf.ku.dk/antar/browse.php (miRNA transfection) are used, which contains target sites with a false positive rate < 0.25. The scores of multiple target sites on the same 3′-UTR are added to produce a total miRNA:gene interaction score.

2.2 Feature extraction

Alignment for putative MRE identification: a dynamic programming algorithm identifies the optimal alignment between the miRNA extended seed sequence [nucleotides 1–9 from the 5′-end of the miRNA] and every 9 nt window on the 3′-UTR or CDS. The alignment is initially restricted such that the pairing of the miRNA extended seed with the 9 nt window begins at position 1 or 2 of the miRNA extended seed. A minimum of four

consecutive Watson–Crick (WC) binding nucleotides is required starting at position 1 or 2 of the miRNA extended seed. A single G:U wobble pair is allowed for binding sites with more than six WC binding nucleotides. Either a single bulge or a single mismatch is allowed for binding sites with eight WC binding nucleotides.

Primary analysis of PAR-CLIP data and training set construction: the PAR-CLIP data produced by Hafner *et al*. (2010) consist of genomic coordinates specifying potential positions of MREs. Each putative MRE position is further refined through the existence of a T to C mutation in the sequenced tags as reported in Hafner *et al*. (2010). To identify the miRNA involved in each MRE, sequences of all identified genomic locations of the PAR-CLIP data are aligned against the miRNA sequence of the top 100 expressed miRNAs (Supplementary List 1). These aligned locations are putative MREs and are further filtered to keep only those located closer than 5 nt to the T to C mutation. In case there is more than one putative miRNA binding in the same region, only the MRE with the highest number of WC binding nucleotides is retained. This set of MREs is defined as the true set. On the other hand, the false set consists of all aligned locations that do not overlap with the PAR-CLIP data. To take into account the possibility that part of the false set corresponds to miRNAs or genes that are functional but not expressed in the particular tissue of the PAR-CLIP experiment, only aligned locations of the top 100 expressed miRNAs in the experiment and genes that already contained at least one true MRE are retained. Overall, out of the 17 310 PAR-CLIP peaks throughout the genome, 5075 overlap with MREs in 3′-UTRs and 6057 overlap with MREs in CDSs.

Detection of binding categories with significant PAR-CLIP reads enrichment: the binding category of a putative MRE is determined through the alignment procedures described above. All binding categories are then separated based on whether the mRNA nucleotide opposite the first nucleotide of the miRNA is an A or not and whether it is a matching nucleotide or not. This procedure defines 64 different binding categories that are then compared between the true and false set of MREs as defined in the PAR-CLIP dataset (Fig. 1B). This comparison is performed through a logistic regression (Venables and Ripley, 2002) between the binding categories and the presence or absence of the corresponding MRE in the true or false set of the PAR-CLIP data. The estimated regression coefficients (Supplementary Table S1) are then used as a feature denoted as the 'binding category weight' feature in a generalized linear model for characterizing the overall efficiency of each. An example category is labeled '8mer+3′-pairing 1st:mismatch+NotA' and corresponds to eight matches between the miRNA extended seed and the mRNA plus additional bindings in the 3′-end and the first nucleotide opposite the 5′-end of the mRNA is not a match nor is an Adenine.

Conservation measure of the MRE sequence in CDS: the CDS conservation scoring method is based on a recently proposed approach (Forman *et al*., 2008) of calculating excess sequence conservation above the one required for amino acid conservation. The underlying concept is that functional MREs in CDSs are expected to preferentially conserve those nucleotides that would have no effect on the amino acid outcome, but would interfere with miRNA targeting. The length of each predicted MRE is spanned by triplets that map fully or partially inside the MRE. For each of the triplets, the log of the conditional probability that the triplet sequence is conserved, given that the amino acid it codes for is conserved, is added to the 'CDS conservation' score of the MRE, using the 30 way genomic alignments (UCSC) for the CDSs of all mRNAs. The score for the final 'CDS conservation' feature is normalized by the maximum score that this MRE could have achieved had it been perfectly conserved in all species.

Conservation measure of the MRE sequence in 3′-UTRs: the 3′-UTR conservation score assesses the evolutionary conservation of a MRE based on 16 species. To compensate for the overall degree of conservation in the whole 3′-UTR, the conservation score for each MRE is defined as the ratio of the number of species in which the binding positions of the extended seed region are conserved and the respective number using the maximal number of species having any conservation in the whole 3′-UTR region. This feature is denoted as 'conservation'.

Detection of significantly accessible locations within MREs: logistic regression between the presence or absence of reads in the PAR-CLIP data and the accessibility of the 3′-UTR sequence as calculated with the Sfold algorithm (Ding *et al*., 2004) using each of the 40 nt upstream and 10 nt downstream of the start of each MRE as a feature is performed to identify any significant targeting feature related to accessibility. The largest region with a $P < 0.1$ (Wald test) and consistent direction of the contribution at all positions extends across positions −1, 1 and 2. The sum of accessibilities in this region, denoted as 'MRE accessibility (−1 to 2)', is used as a feature.

Other MRE features: two of the three features identified in Grimson *et al*. (2007), the MRE flanking AU content denoted as 'flanking AU content' and the distance of the MRE to the closest 3′-UTR end denoted as 'distance to closest 3′-UTR end' are used. Additionally, the distance between adjacent MREs denoted as 'adjacent MRE distance', the free energy of binding as calculated with RNAhybrid (Rehmsmeier *et al*., 2004) denoted as 'free energy' and the resulting binding pattern of the 29 nt of the 3′-UTR along the MRE denoted as 'bnt1' to 'bnt29', are also used as features. All second-order interactions between all features are automatically generated and selected using F-tests.

2.3 Feature selection

To determine an optimal feature set using cross-validation, the PAR-CLIP dataset is split into three disjoint subsets, stratified for positive and negative sites. Logistic regression using the features described above is performed on each subset and a feature selection procedure minimizing the Akaike information criterion (AIC) using the stepAIC implementation in the MASS package (Venables and Ripley, 2002) for R determines an optimal set of features. For this initial set of features, the capability of each single feature to separate the complete PAR-CLIP data into sites with reads and sites without reads is tested using the Wilcoxon's test and only features with significant ($P < 0.05$) separation are retained. This feature selection procedure is performed independently for sites in CDSs and sites in 3′-UTRs (Fig. 1C). The full list of selected CDS and 3′-UTR features is provided in Supplementary Table S2.

2.4 Training and scoring

Using the identified significant features, different machine learning methods like support vector machines, neural networks, random forests and generalized linear models (GLM) (Venables and Ripley, 2002) are compared for the calculation of an MRE score. The best performance, quantified by cross-validation is obtained using GLMs. Each gene region (CDS or 3′-UTR) is represented by a separate model. The regression coefficients for all features and their significances are presented in Supplementary Table S2. The scores for all MREs identified in a region are summed into a region score (Fig. 1D).

Combining CDS and 3′-UTR targeting: for the optimal combination of the two region scores that are obtained by summation of the respective MRE scores, another generalized linear model is trained using data from the 13 different microarray experiments measuring mRNA expression changes when a miRNA is either transfected or knocked out (defined in section 2.1, Microarray data). While the PAR-CLIP data provides detailed data about miRNA target binding sites, but not about the cooperative effect of multiple target sites on a gene. Therefore, we used microarray gene expression data in order to measure the effectiveness of these sites in suppressing the expression of a gene. Genes in each dataset are sorted according to expression fold change compared with control, and the top and bottom 100 genes from each experiment are used as the true and false examples for training the generalized linear model (Fig. 1E).

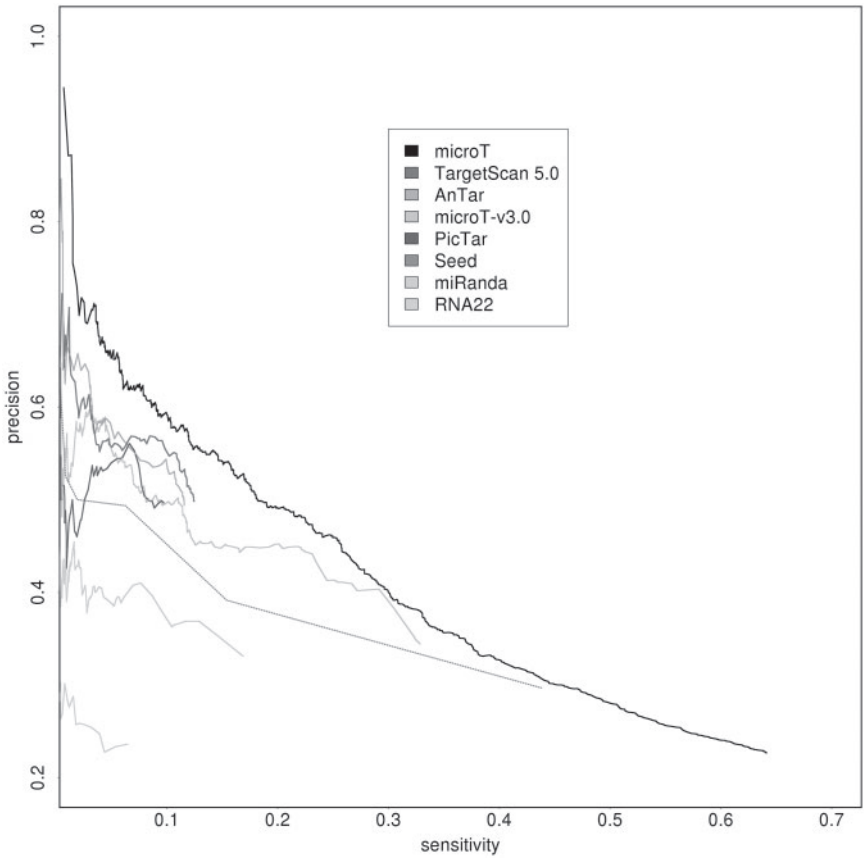

Fig. 2. A precision receiver operating curve (pROC) analysis of the predictions for different target prediction methods (defined in Section 2). Using a decreasing score cutoff, the prediction sensitivity and precision for each method is tested on the dataset from Selbach *et al*. (2008). The dashed line shows the performance of the seed measure that counts the number of miRNA seed matches in the 3′-UTR of a gene. A distinct increase of both sensitivity and precision for microT-CDS can be observed.

3 RESULTS

3.1 Addition of target sites in coding regions enables a more sensitive target prediction

The developed algorithm is tested on a large independent test dataset provided by Selbach *et al*. (2008). This set provides the experimentally supported targets for five miRNAs identified through a high-throughput method (Section 2.1.3). Approximately half of the 2447 genes, that are considered as targets of these miRNAs, do not carry a single miRNA seed (nucleotides 2–7 from the 5′-end of the miRNA) match in their 3′-UTR sequences and are thus not recognized by existing miRNA target prediction programs. The combined (CDS and 3′-UTR) model presented here increases the sensitivity in this dataset from 52% to 65% in comparison to the 3′-UTR-only region model, keeping the specificity at the same level of 32%. This corresponds for the particular dataset of five miRNAs to 293 additional correctly predicted targets (see Supplementary Fig. S1).

To test the significance of the additional CDS model, the predicted results are compared with a partly random predictor, where for each miRNA, the scores of the two models are shuffled by combining the 3′-UTR score of each gene with a randomly selected CDS score from a target gene of the same miRNA. The performance of this randomized predictor is significantly lower than the combined model (Supplementary Fig. S1), demonstrating a significant and synergistic contribution of targeting in the CDS.

The combined model is also compared with other widely used miRNA target prediction programs such as TargetScan 5.0 (Friedman *et al*., 2009), PicTar (Lall *et al*., 2006), RNA22 (Miranda *et al*., 2006), miRanda (John *et al*., 2004), DIANA-microT-v3.0 (Maragkakis *et al*., 2009a; b), AnTar (Wen *et al*., 2011) and a seed measure, whose prediction score is defined through the number of miRNA seed matches on the 3′-UTR of genes. The latter has been shown in a comparison of (Alexiou *et al*., 2009) to be more sensitive than many other published prediction programs at that time (Section 2). The sensitivity and precision of all of these programs is measured at different prediction score cutoffs, yielding precision recall curves shown in Figure 2. The DIANA-microT-CDS program exhibits the highest sensitivity at any level of specificity in comparison with the other six programs. Interestingly, a high increase in sensitivity is observed at lower specificity values, outperforming also the seed measure.

The validity of using a specific prediction model for the additional CDS sites is verified in a comparison with predictions of TargetScan 5.0 that also uses sites in the coding region. Obtaining the scores for TargetScan 5.0 using the sequence covering both the CDS and the

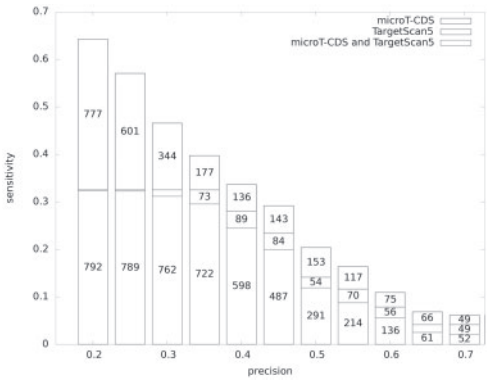

Fig. 3. Analysis of the overlap of different target prediction methods. The number of correctly predicted targets is shown for targets predicted only by DIANA-microT-CDS, predicted only by TargetScan 5.0 and predicted by both programs, respectively. The comparison evaluates the 2447 known targets in the (Selbach *et al*., 2008) dataset at specific score thresholds corresponding to different prediction precision levels.

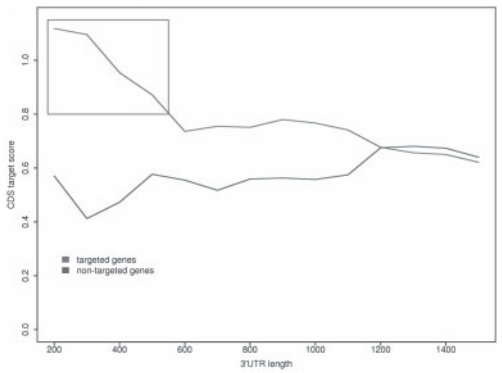

Fig. 4. Preferential occurrence of MREs in the CDS for short 3′-UTRs. Comparing the sum of the predicted site scores in coding sequence (CDS score) against various 3′-UTR sizes of targeted (green line) and non-targeted (blue line) genes on an independent test set reveals a significantly higher number of sites in CDS for genes with 3′-UTR lengths shorter than 500 nt (red box, $P < 0.05$, Wilcoxon's test).

3′-UTR as input, predictions with >10% lower precision compared with DIANA-microT-CDS are obtained (Supplementary Fig. S2).

In order to scrutinize the improvement of DIANA-microT-CDS to the top-performing program TargetScan 5.0, the overlap between the targets predicted by DIANA-microT-CDS and TargetScan on the Selbach *et al*. dataset is measured ranging from 50 to 70%, depending on the precision level. This indicates that a large fraction of novel targets, as also shown in Figure 3, are predicted only by DIANA-microT-CDS. Particularly, at lower precision levels the number of correct predictions is almost doubled using DIANA-microT-CDS.

The performance of DIANA-microT-CDS program in the detection of CDS target sites is also evaluated on the high-throughput sequencing of RNA isolated by cross-linking immunoprecipitation (HITS-CLIP) dataset of Chi *et al*. (2009). In this dataset, the Argonaute-mRNA binding sites corresponding to mouse miRNA targets are measured and used here. Of the top 20 expressed miRNAs in this experiment, seven are not in the set of miRNAs used for the development of our algorithm and are used here as an independent test set. Out of the genes targeted by these microRNAs, genes having HITS-CLIP clusters only in the CDS and not in the 3′-UTR are collected, resulting in 1210 CDS target sites. DIANA-microT-CDS is capable of predicting the location of 286 of these sites correctly. In order to estimate if this could also happen by chance, the locations of the predicted sites is randomized 100 times. The randomized model is able to locate only 10.3 out of the 1210 real binding sites, leading to an estimated ratio of true over randomly predicted sites >27.

A test of the DIANA-microT-CDS algorithm on the five individual cases of experimentally verified CDS targeting mentioned in the introduction, recalls three positive cases (for the genes: Dnmt3b, Dicer and HNF4a). This is in agreement with our estimated sensitivity and is currently the only available computational prediction for this type of sites. The contribution of target sites located in the CDS is further verified in additional tests on the microarray experiments measuring the effect of over- or underexpression of six miRNAs not contained in the training set used for constructing the MRE predictors (mir-98, miR-124, miR-335, miR-122a, miR-132, miR-142). Comparing our algorithm when using only target sites in the 3′-UTR with the algorithm using all sites on this data, the sensitivity of detecting verified targeted genes when using the same score cut-off is increased from 42.7 to 46.8% by more than 4%, while the false positive predictions and the precision of the predictions remains at the same level (see also Supplementary Fig. S3). This corresponds to 25 correctly predicted additional targets in the CDS in this set of 600 verified targets.

3.2 Genes with shorter 3′-UTR have significantly more targets in coding regions

To gain more insight into the mechanism underlying CDS targeting, the relations between CDS and 3′-UTR targeting is investigated in the dataset of Selbach *et al*. Comparing the CDS target scores with the 3′- UTR length of the same target gene, it is found that genes with 3′-UTRs <500 nt have a significantly higher CDS target score (Wilcoxon's test, $P < 0.05$). The red region in Figure 4 indicates all 3′-UTR lengths with significantly higher CDS sores, indicating likely targeting in the CDS. Such preference could not be observed for the group of genes that are measured as not targeted by miRNAs in the same proteomics experiment.

The robustness of this observation is tested by randomly combining the CDS scores with the 3′-UTR scores. In only 553 out of 10 000 randomizations, a significantly higher CDS score is tested for the 3′-UTR shorter than 500 nt is detected ($P < 0.05$, Wilcoxon's test). Similarly, when analyzing the miRNA target genes as observed from 13 microarray experiments (Section 2), the genes identified as targeted only in the CDS are observed to have significantly shorter 3′-UTR sequences than genes targeted only on the 3′-UTR ($P < 10^{-13}$, Wilcoxon's test). These findings suggest that evolutionary pressure might enforce the presence of additional sites on the CDS in cases where there is restricted space on the 3′-UTR.

4 DISCUSSION

High-throughput proteomics experiments that measure changes for thousands of genes both on the mRNA and the protein level reveal that approximately half of the genes whose expression is increased/decreased after miRNA transfection/knockout do not carry

a single corresponding miRNA seed match in their 3′-UTR sequence (Baek *et al*., 2008; Selbach *et al*., 2008). The program introduced here enables the recognition of 12% of these downregulated genes as additional targets of miRNAs, having their targets in coding regions. A list of all genes predicted to be targeted only in the CDS is contained in Supplementary Table S3. This list is predicted with an expected precision of 50% and contains on average 64 such genes per miRNA.

The analysis of the recent data for miRNA-associated protein immunoprecipitation and the subsequent RNA sequencing results in a program that uses several features that are different from other programs. Generally, evolutionary conservation is a strong indication for MRE functionality (Friedman *et al*., 2009; Kiriakidou *et al*., 2004; Lewis *et al*., 2003). However, the coding sequences of genes usually have a significantly higher background conservation level than 3′-UTR sequences due to their underlying amino acid content. Therefore, a specific feature for conservation of MREs in coding regions is incorporated here, exploiting the conservation of synonymous codons.

A feature analysis for MREs in 3′-UTRs reveals a number of novel significant features, such as the requirement for increased accessibility in the mRNA secondary structure at the start of an MRE. In several cases, the synergistic effect of two features is more informative than the two features used independently. For example, the higher mRNA AU content in the region surrounding an MRE (Grimson *et al*., 2007) when combined with the free energy of the binding complex ($P < 10^{-15}$, Wald test) gains higher significance than any of these features alone. Interestingly, this gain suffices to eliminate the AU content as an independent feature. The analysis reveals also that functional MREs in the CDS preferentially require a stronger binding than MREs in the 3′-UTR. MREs in coding regions require a perfect binding along the miRNA seed region and mismatches disrupt their functionality, which was also found by Fang and Rajewsky (2011).

As the only resource to provide target predictions accounting also for target sites in the CDS, and moreover specifying the predicted binding locations of all sites, the results of DIANA-microT-CDS are available through the DIANA web server (Maragkakis *et al*., 2009b) at www.microrna.gr/microT-CDS.

ACKNOWLEDGEMENTS

We thank N. Koziris of the National Technical University of Athens for providing their computational cluster to conduct experiments and supporting all DIANA-microT-CDS web.

Funding: Project 09 SYN - 13 -1055 ‘MIKRORNA’ by the ESPA program of the Greek General Secretariat for Research and Technology.

Conflict of Interest: none declared.

REFERENCES

Abdelmohsen, K. *et al*. (2010) miR-519 suppresses tumor growth by reducing HuR levels. *Cell cycle*, **9**, 1354–1359.

Alexiou, P. *et al*. (2009) Lost in translation: an assessment and perspective for computational microRNA target identification. *Bioinformatics*, **25**, 3049–3055.

Baek, D. *et al*. (2008) The impact of microRNAs on protein output. *Nature*, **455**, 64–71.

Bartel, D.P. (2009) MicroRNAs: target recognition and regulatory functions. *Cell*, **136**, 215–233.

Chi, S.W. *et al*. (2009) Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature*, **460**, 479–486.

Ding, Y. *et al*. (2004) Sfold web server for statistical folding and rational design of nucleic acids. *Nucleic Acids Res.*, **32**, W135–W141.

Duursma, A.M. *et al*. (2008) miR-148 targets human DNMT3b protein coding region. *RNA*, **14**, 872–877.

Elcheva, I. *et al*. (2009) CRD-BP protects the coding region of betaTrCP1 mRNA from miR-183-mediated degradation. *Mol Cell.*, **35**, 240–246.

Fang, Z. and Rajewsky, N. (2011) The impact of miRNA target sites in coding sequences and in 3′UTRs. *PLoS One*, **6**, e18067.

Forman, J.J. and Coller, H.A. (2010) The code within the code: microRNAs target coding regions. *Cell cycle*, **9**, 1533–1541.

Forman, J.J. *et al*. (2008) A search for conserved sequences in coding regions reveals that the let-7 microRNA targets Dicer within its coding sequence. *Proc. Natl Acad. Sci. USA*, **105**, 14879–14884.

Friedman, R.C. *et al*. (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.*, **19**, 92–105.

Gennarino, V.A. *et al*. (2009) MicroRNA target prediction by expression analysis of host genes. *Genome Res.*, **19**, 481–490.

Grimson, A. *et al.*. (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Cell.* **27**, 91–105.

Hafner, M. *et al*. (2010) Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, **141**, 129–141.

John, B. *et al*. (2004) Human MicroRNA targets. *PLoS Biol.*, **2**, e363.

Kiriakidou, M. *et al*. (2004) A combined computational-experimental approach predicts human microRNA targets. *Genes Dev.*, **18**, 1165–1178.

Lall, S. *et al*. (2006) A genome-wide map of conserved microRNA targets in C. elegans. *Curr. Biol.*, **16**, 460–471.

Lewis, B.P. *et al*. (2003) Prediction of mammalian microRNA targets. *Cell*, **115**, 787–798.

Linsley, P.S. *et al*. (2007) Transcripts targeted by the microRNA-16 family cooperatively regulate cell cycle progression. *Mol. Cell Biol.*, **27**, 2240–2252.

Maragkakis, M. *et al*. (2009a) Accurate microRNA target prediction correlates with protein repression levels. *BMC Bioinformatics*, **10**, 295.

Maragkakis, M. *et al*. (2009b) DIANA-microT web server: elucidating microRNA functions through target prediction. *Nucleic Acids Res.*, **37**, W273–W276.

Miranda, K.C. *et al*. (2006) A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. *Cell*, **126**, 1203–1217.

Papadopoulos, G.L. *et al*. (2009) The database of experimentally supported targets: a functional update of TarBase. *Nucleic Acids Res.*, **37**, D155–D158.

Rehmsmeier, M. *et al*. (2004) Fast and effective prediction of microRNA/target duplexes. *RNA*, **10**, 1507–1517.

Schnall-Levin, M. *et al*. (2011) Unusually effective microRNA targeting within repeat-rich coding regions of mammalian mRNAs. *Genome Res.*, **21**, 1395–1403.

Schnall-Levin, M. *et al*. (2010) Conserved microRNA targeting in Drosophila is as widespread in coding regions as in 3′UTRs. *Proc. Natl Acad. Sci. USA*, **107**, 15751–15756.

Selbach, M. *et al*. (2008) Widespread changes in protein synthesis induced by microRNAs. *Nature*, **455**, 58–63.

Takagi, S. *et al*. (2010) MicroRNAs regulate human hepatocyte nuclear factor 4alpha, modulating the expression of metabolic enzymes and cell cycle. *J. Biol. Chem.*, **285**, 4415–4422.

Tay, Y. *et al*. (2008) MicroRNAs to Nanog, Oct4 and Sox2 coding regions modulate embryonic stem cell differentiation. *Nature*, **455**, 1124–1128.

Venables, W. and Ripley, B. (2002) *Modern Applied Statistics with S.* Springer, New York.

Wang, W.-X. *et al*. (2011) Individual microRNAs (miRNAs) display distinct mRNA targeting ‘rules’. *RNA Biol.*, **7**, 373–380.

Wang, X. and Wang, X. (2006) Systematic identification of microRNA functions by combining target prediction and expression profiling. *Nucleic Acids Res.*, **34**, 1646–1652.

Wen, J. *et al*. (2011) MicroRNA transfection and AGO-bound CLIP-seq datasets reveal distinct determinants of miRNA action. *RNA*, **17**, 820–834.