Bioinformatics, 32(9), 2016, 1316–1322
doi: 10.1093/bioinformatics/btw002
Advance Access Publication Date: 6 January 2016
Original Paper

Sequence analysis

Improving microRNA target prediction by modeling with unambiguously identified microRNA-target pairs from CLIP-ligation studies

Xiaowei Wang

Department of Radiation Oncology, Washington University School of Medicine, St Louis, MO 63108, USA

Associate Editor: Ivo Hofacker

Received on 24 June 2015; revised on 23 December 2015; accepted on 3 January 2016

Abstract

Motivation: MicroRNAs (miRNAs) are small non-coding RNAs that are extensively involved in many physiological and disease processes. One major challenge in miRNA studies is the identification of genes targeted by miRNAs. Currently, most researchers rely on computational programs to initially identify target candidates for subsequent validation. Although considerable progress has been made in recent years for computational target prediction, there is still significant room for algorithmic improvement.

Results: Here, we present an improved target prediction algorithm, which was developed by modeling high-throughput profiling data from recent CLIPL (crosslinking and immunoprecipitation followed by RNA ligation) sequencing studies. In these CLIPL-seq studies, the RNA sequences in each miRNA-target pair were covalently linked and unambiguously determined experimentally. By analyzing the CLIPL data, many known and novel features relevant to target recognition were identified and then used to build a computational model for target prediction. Comparative analysis showed that the new algorithm had improved performance over existing algorithms when applied to independent experimental data.

Availability and implementation: All the target prediction data as well as the prediction tool can be accessed at miRDB (http://mirdb.org).

Contact: xwang@radonc.wustl.edu

1 Introduction

MicroRNAs (miRNAs) are small non-coding RNAs that play important regulatory roles in many physiological and disease processes (Ambros, 2004). About 2000 human miRNAs have been reported (Kozomara and Griffiths-Jones, 2014) and collectively these miRNAs regulate the expression of thousands of gene targets at both post-transcriptional and translational levels (Baek *et al.*, 2008; Lim *et al.*, 2005; Selbach *et al.*, 2008). Reliable identification of gene targets is critical for functional characterization of miRNAs. Currently, experimental identification of miRNA targets is a time-consuming process and thus most researchers rely on computational tools to first predict a set of promising target candidates for further experimental validation. The first step in computational target prediction is to identify relevant features that are characteristic of miRNA target recognition. In recent years, significant progress has been made on this subject and multiple features that are relevant to miRNA targeting have been reported, such as target pairing to the 5′-end of the miRNA (i.e. the 'seed' sequence) and structural accessibility of the target sites for miRNA binding (Bartel, 2009; Kertesz *et al.*, 2007; Khorshid *et al.*, 2013; Liu *et al.*, 2013; Long *et al.*, 2007; Robins *et al.*, 2005; Zhao *et al.*, 2005). For algorithm development, the next step is to combine various target recognition features in a computational model for target prediction. To achieve this

important but challenging objective, high-quality experimental training data are usually required to properly weight various target features for development of prediction models. Most existing target prediction algorithms have been developed in this way by training with various high-throughput profiling data, such as with microarray profiling data (Agarwal *et al.*, 2015; Betel *et al.*, 2010; Grimson *et al.*, 2007; Wang and El Naqa, 2008) or with CLIP (crosslinking and immunoprecipitation) sequencing data (Gumienny and Zavolan, 2015; Liu *et al.*, 2013; Reczko *et al.*, 2012). However, despite steady progress in computational target prediction, available bioinformatics tools still have sub-optimal performance as revealed by subsequent validation experiments.

Currently, one major obstacle in computational algorithm development is the lack of high-quality training data from experimental studies. In recent years, multiple high-throughput studies have been performed to experimentally identify miRNA targets. For example, a major experimental strategy, CLIP has been developed to identify transcript targets associated with functional miRNA-induced silencing complex (miRNA-RISC) (Chi *et al.*, 2009; Hafner *et al.*, 2010; Zhang *et al.*, 2007). One prominent example is HITS-CLIP, which identifies short transcript sequences that are bound to the Ago protein by cross-linking the RNA strand to the protein, followed by immunoprecipitation and high-throughput RNA sequencing (Chi *et al.*, 2009). Another widely used experimental strategy for target identification is to analyze downregulated transcripts by miRNA overexpression with microarrays (Lim *et al.*, 2005; Linsley *et al.*, 2007; Wang and Wang, 2006). These high-throughput experimental datasets had previously been utilized to train multiple established algorithms for miRNA target prediction. Although highly useful, there are also major challenges associated with CLIP-seq or microarray data for target prediction modeling. One major unresolved issue is that the exact target sequence in each miRNA/target pair is unknown. As a result, computational inference, which is an error-prone process, is necessary to predict the target sites from CLIP-seq or microarray data. To address this challenge experimentally, improved CLIP strategies have been developed to directly identify miRNA-target pairs residing in the same RISC complexes (crosslinking and immunoprecipitation followed by miRNA/target ligation, termed CLIPL in our analysis). The first CLIPL-seq method, termed CLASH, was reported by Helwak and colleagues (Helwak *et al.*, 2013). Subsequently, another similar CLIPL-seq method was also reported by Grosswendt et al. (Grosswendt *et al.*, 2014). The CLIPL-seq method can unambiguously identify both a miRNA and its cognate target site in the same RISC complex, thus allowing direct characterization of miRNA target recognition features. With CLIPL-seq, thousands of miRNA-target pairs have been identified (Grosswendt *et al.*, 2014; Helwak *et al.*, 2013). In this study, public CLIPL-seq data were systematically analyzed and used to develop an improved algorithm for miRNA target prediction.

2 Methods

2.1 Data retrieval and processing

Two public CLIPL-seq datasets were included in our analysis, including one from the Helwak study (CLASH data) (Helwak *et al.*, 2013) and the other from the Grosswendt study (Grosswendt *et al.*, 2014). Raw RNA-seq data from the CLASH study were downloaded from the NCBI GEO database (Barrett *et al.*, 2013). In addition, the list of curated miRNA/target pairs was downloaded from the journal's website (Helwak *et al.*, 2013). In the CLASH study, HEK293 cells were first crosslinked by UV irradiation. Crosslinked RISC complexes containing both miRNAs and target transcripts were then immunoprecipitated. As a key step in the protocol, RNA–RNA ligation was performed to covalently ligate the miRNA and its target transcript in the same RISC complex. The identity of both the miRNAs and the ligated target transcripts was determined by sequencing of the chimeric RNA strands. Similar to the CLASH analysis, the Grosswendt dataset contains a list of ligated pairs of miRNAs and their cognate target transcripts captured from the same RISC complexes. In the Grosswendt study, ligated miRNA-target pairs were identified from *Caenorhabditis elegans*, using an experimental protocol similar to that described in the Helwak study. Ligated miRNA-target pairs were also identified in mammalian cells from multiple public studies (mainly in human HEK293 cells and mouse T cells) (Grosswendt *et al.*, 2014). The list of human miRNA-target pairs identified from the Grosswendt study was downloaded from the journal's website (Grosswendt *et al.*, 2014) and included in this study.

For RNA-seq raw data, the sequence reads were aligned to the transcriptome with BLAT (Kent, 2002). miRNA sequences were downloaded from miRBase (Kozomara and Griffiths-Jones, 2014), and all other transcript sequences as well as gene mapping index files were downloaded from the NCBI ftp site (NCBI Resource Coordinators, 2015). The 3′-UTR (untranslated region) sequences were parsed with BioPerl (http://www.bioperl.org) based on GenBank annotations. Orthologous gene relationships were computed using the NCBI HomoloGene database (NCBI Resource Coordinators, 2015). Predicted miRNA targets by public computational algorithms were retrieved from the respective public websites [TargetScan 7.0 (Agarwal *et al.*, 2015), http://targetscan.org; DIANA-MicroT (Reczko *et al.*, 2012), http://diana.cslab.ece.ntua.gr; miRanda-mirSVR (Betel *et al.*, 2010), http://microrna.org; RNA22 (Miranda *et al.*, 2006), https://cm.jefferson.edu/rna22/]. The target transcript IDs from all these algorithms were mapped to NCBI Gene IDs for subsequent comparative analysis.

Public high-throughput profiling data were used for evaluation of target prediction algorithms. The microarray data reported by Hafner *et al.* (2010) were analyzed to evaluate the impact of miRNAs on regulation of target RNA expression. In this microarray study, 25 miRNAs were suppressed in HEK293 cells by antisense oligonucleotide inhibitors, and the impact on the transcriptome was determined with Affymetrix Human U133Plus2 chips. Microarray raw data from the Hafner study were downloaded from the NCBI GEO database, and then normalized and log2-transformed using the Bioconductor RMA method (http://www.bioconductor.org). Array signals from transcripts of the same gene were combined and averaged. Genes with undetectable expression in HEK293 cells were excluded from further analysis. The fold change of gene expression due to miRNA inhibition was calculated by comparing to the negative control arrays.

2.2 Computational tools and data analysis

RNA secondary structures were calculated with RNAfold (Hofacker, 2003). The predicted structures were analyzed to determine whether individual nucleotides were base-paired or exposed. Statistical computing was performed with the R package (http://www.r-project.org/). Statistical significance (*P*-value) for the training features was calculated with Student's *t*-test or χ^2 test. LIBSVM was used to build miRNA target prediction models with support vector machines (SVMs) (http://www.csie.ntu.edu.tw/~cjlin/libsvm/). For SVM analysis, a radial basis function (RBF) was used for kernel transformation. Optimization of the RBF kernel parameters was done

with grid search and cross-validation according to the recommended protocol by LIBSVM. Recursive feature selection was done with Weka (http://www.cs.waikato.ac.nz/ml/weka/) to evaluate the independent contribution of each training feature in the SVM modeling. All predicted targets and the newly developed prediction tool are available at miRDB (http://mirdb.org).

3 Results

3.1 Identification of miRNA targets from CLIPL-seq data

The CLIPL method represents a major improvement over conventional CLIP methods, as both the miRNAs and their cognate target transcripts from the same RISC complexes can be simultaneously identified. In this way, the one-to-one relationship can be unambiguously established for a miRNA and its cognate target transcript in the same RISC complex. Thus, CLIPL-seq data have provided an unprecedented opportunity to identify features that are characteristic of miRNA target recognition. Two CLIPL datasets (Grosswendt *et al.*, 2014; Helwak *et al.*, 2013) were analyzed and combined in this study (described in 'Methods' section).

From CLIPL studies, a large number of identified miRNA targets were of non-canonical type, i.e. without involving perfect pairing to the miRNA seed region. However, based on recent analysis of CLIPL-seq data, non-canonical targeting usually does not lead to significant target downregulation at either the RNA or protein level (Wang, 2014). In support of this view, although there are exceptional cases indicating the functional relevance of non-canonical targets, almost all validated miRNA targets reported in literature are of the canonical type. Thus, in this study, we decided to focus on canonical targets, as most functional miRNA studies are centered on regulation of target expression (more details presented later in 'Discussion' section). Specifically, all identified miRNA/target pairs from CLIPL studies were screened for perfect pairing between a miRNA seed (positions 2–8) and its target-binding site. The CLASH and Grosswendt datasets contain 705 and 1443 canonical miRNA/target pairs, respectively. These 2148 pairs of miRNA/targets were combined and used as positive training data for further target feature analysis.

Negative training data were also compiled from the CLIPL-seq studies. Specifically, a set of potential non-target sequences were selected based on the following criteria: (i) The non-target site does not overlap with any sequence tags identified in the CLIPL experiment, including both ligated miRNA/mRNA chimeras and free unligated mRNA tags (which constituted the vast majority of all RNA-seq reads); (ii) the target site is from a transcript with detectable expression level in the cells as revealed by microarrays; and (iii) the target site pairs perfectly to the seed sequence of a miRNA, which is randomly selected from a list of all expressed miRNAs in the cells. By implementing these screening criteria, 4000 non-target sequences were randomly selected as negative controls for target feature analysis.

3.2 Characterizing target recognition features

Target and non-target sequences were directly compared to identify features that are characteristics of miRNA target recognition.

3.2.1 Patterns of nucleotide usage in the target site

Local nucleotide composition surrounding the seed-binding sites was compared between the target sites and non-target sites. In general, the target sites had significantly lower GC content compared with the non-target sites (P = 1.3E-123 with Student's t-test). All four mono-nucleotide counts were significantly different in the target sites compared with the non-target sites, with G being the most underrepresented (15 and 24% in targets and non-targets, respectively, P = 9.6E-129; Table 1). Interestingly, there was only a slight bias against C in the target sites (21 versus 23%, P = 3.9E-7), indicating that low overall GC content was most likely a reflection of strong bias against G in the target-binding sites.

Among all 16 dinucleotide counts of the target sites, GG was most underrepresented and UA/AU/UU was most overrepresented (Table 1). Positions 2–8 in the target sites are seed-binding site, and, as shown in Figure 1A, the most significantly different bases in target sites (as compared with non-target sites) were all located adjacent to the seed-binding site. Among all the bases in the target site, the one immediately upstream of the seed-binding site (position 9) was most significant (P = 7.7E-111). Among all target sites, 57% had U at position 9; in contrast, U was present at the same position in only 27% of the non-target sites (enrichment ratio = 2.1, Table 2). Due to the strong preference for U at position 9 in the target site, all three other bases were significantly depleted. Of note, there was no preference for A, and thus the enrichment of U at position 9 cannot be explained by the general requirement of low GC content in the target sites. Previous studies have demonstrated that there was a strong preference for A at position 1 (Bartel, 2009). Consistently, A was found to be significantly enriched at position 1 (49% in target sites versus 24% in non-target sites, P = 1.7E-85). Besides positions 1 and 9, significant differences in base usage were observed at multiple other positions, especially at positions 10–13 (Table 2, Fig. 1A). Of note, G was significantly depleted at these positions. In contrast, the usage of C was not different at most of these positions.

3.2.2 Structural accessibility of the target site

Multiple studies have shown that structural accessibility plays an important role in miRNA target recognition (Kertesz *et al.*, 2007; Long *et al.*, 2007; Robins *et al.*, 2005; Zhao *et al.*, 2005). With the CLIPL training data, it is now possible to precisely determine the accessibility of individual nucleotides in the target sites. Specifically, local secondary structures of the target sites were calculated with RNAfold (Hofacker, 2003). In general, the target sites were significantly more accessible compared with the non-target sites, as measured by the free energy value (ΔG) of the folding structures (P =

Table 1. Significant nucleotide counts of the target sites[a]

Nucleotides	Enrichment ratio	P-value
G	61%	9.6E-129
U	131%	4.5E-88
A	110%	1.5E-10
C	91%	3.9E-07
AU	164%	3.3E-71
UA	168%	7.0E-63
AG	56%	1.1E-58
UU	162%	3.4E-55
GG	45%	8.9E-54
GC	59%	1.2E-42
GA	62%	3.4E-36
UG	71%	9.3E-33
AC	129%	1.2E-14
CG	67%	3.9E-07
CC	81%	1.4E-06
AA	118%	8.6E-06
GU	85%	1.2E-05

[a]The enrichment ratio is defined by: (average count per target site)/(average count per non-target site). The P-values were calculated with Student's t-test.

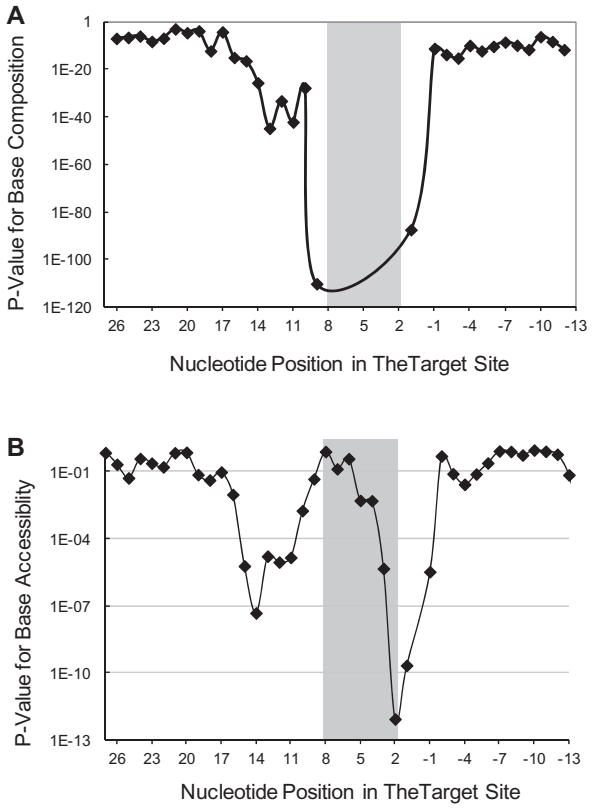

Fig. 1. Position-specific sequence and structural features for the target sites. Nucleotides at positions 2–8 (highlighted with a shaded box) in the target site are the miRNA seed-binding region. Statistical significance for each nucleotide position was calculated by comparing the target sites with non-target sites for base usage or accessibility. **(A)** Differences in base usage at individual positions between the target site and non-target site. P-values were calculated with χ^2 test. **(B)** Differences in structural accessibility at individual positions between the target site and non-target site. Local secondary structure was calculated with RNAfold, and the fractions of unstructured bases at individual positions of the target and non-target sites were compared with Student's *t*-test

Table 2. Difference in base composition between the target sites and non-target sites[a]

Base position	A	C	G	U	A or U	*P*-value
1	2.03	0.62	0.31	1.06	**1.51**	4.4E-88
9	0.79	0.43	0.54	**2.06**	1.43	7.7E-111
10	1.36	0.79	0.59	1.21	**1.28**	2.0E-28
11	1.12	0.98	**0.38**	1.45	1.30	8.3E-43
12	1.14	1.14	**0.43**	1.22	1.18	6.2E-34
13	1.19	1.09	**0.37**	1.31	1.25	1.9E-45

[a]The seed-binding site is designated as positions 2–8. Each value in the table represents base enrichment ratio for each position in the target sites as compared with non-target sites. The most significant enrichment ratio at each position was highlighted in bold, with the *P*-value presented in the table.

2.8E-119 with Student's *t*-test). At individual nucleotide level, multiple bases surrounding the 3′-end of the seed-binding site (position 2) were significantly more exposed in target sites compared with non-target sites (Fig. 1B). In particular, the accessibility of the base at position 2 was the most significant ($P = 8.8E-13$). Thus, free exposure of the 3′-end of the seed-binding site is a strong feature for miRNA target recognition. Interestingly, there was no difference in base accessibility surrounding the 5′-end of the seed-binding site (positions 6–9). This implies that it is possible that individual nucleotides involved in miRNA seed binding play distinct roles in target recognition, with the 3′-end of the seed matching site first recognized by the miRNA, followed by successive unwinding and seed pairing for the entire seed-binding region.

3.2.3 Seed pairing with the target site

As reported previously, the seed-binding sequence of a target site is often conserved across multiple species (Bartel, 2009; Wang and El Naqa, 2008). The level of seed site conservation was evaluated in five species, including human, mouse, rat, dog and chicken. The count of conserved species for each site was used to evaluate seed site conservation. Consistent with previous reports, seed site conservation was very significantly different between target and non-target sites ($P = 7.3E-248$ with χ^2 test). The seed conservation feature has been included as a principle requirement in most existing target prediction tools. Although useful for the elimination of a large number of false positives, the stringent requirement of sequence conservation in the mean time also excludes many bona fide target sites that are not evolutionarily conserved. To account for this in our prediction algorithm, sequence conservation was used as a contributing selection feature, but not as a requirement. In this way, both conserved and non-conserved target sites could be identified.

Previous studies have demonstrated that canonical 7-mer miRNA seed, occupying positions 2–8, is the most important seed type (Bartel, 2009). Besides the 7-mer seed, other types of seed may also play important roles in target recognition. For example, a perfect terminal base match will convert a 7-mer seed matching site into an 8-mer seed matching site (binding to positions 1–8 in the miRNA). Among all the target sites, 33% had a terminal base match (i.e. 8-mer seed-binding sites), in contrast to 25% of the non-target sites ($P = 6.3E-10$ with χ^2 test). Furthermore, recent analysis of the CLASH data indicates that thermodynamic stability of the seed/target binding is a significant determinant of target recognitions patterns (Wang, 2014). To this end, the stability of base pairing between miRNA seed- and target-binding site was calculated with the nearest neighbor method using thermodynamic parameters for RNA–RNA base pairing (Xia *et al.*, 1998). Compared with non-target sites, miRNA seed/target binding was generally less stable as indicated by free energy calculation ($P = 8.7E-7$ with Student's *t*-test). The potential role of non-seed-based target binding was also investigated. In this analysis, all pentamers from the non-seed region of a miRNA were screened against the 5′-end of the target-binding site to identify any sequence match. In this way, significant enrichment of pentamer-matching motifs were discovered in the target sites ($P = 2.0E-15$ with χ^2 test).

3.2.4 Location of the target site

The activity of a target-binding site is related to its location in the 3′-UTR (Gaidatzis *et al.*, 2007; Grimson *et al.*, 2007). A target site buried in the middle of a long UTR will make the site less accessible to miRNA binding. In our analysis, the target and non-target sites were compared to determine any difference in their UTR location. Overall, target sites were preferentially located in shorter 3′-UTRs as compared with non-target sites ($P = 2.2E-20$ with Student's *t*-test for the evaluation of UTR length difference between the two groups). Consistent with previous studies, the CLIPL analysis confirmed that target sites tend to reside toward either end of the 3′-UTR; specially, target sites were more likely to be found within 200 nucleotides from either end of 3′-UTR as compared with

non-target sites ($P = 1.6E\text{-}27$ with χ^2 test). Conversely, non-target sites were more likely to be found to be at least 800 bases away from either end of the 3′-UTR ($P = 2.9E\text{-}30$ with χ^2 test).

3.3 Target prediction model and genome-wide target prediction

The above described structural and sequence features were modeled in a SVM framework for algorithm development. Recursive feature elimination (RFE) was performed to identify the most important independent features for target prediction. In this RFE analysis, all the sequence features were analyzed collectively in an SVM framework. The least predictive feature was first identified by SVM and subsequently eliminated. The remaining features were then evaluated again collectively to identify the next least predictive feature for elimination. The process was repeated with one feature (the least predictive one from the collection of all remaining features) eliminated from each cycle until only one feature was left. SVM-RFE is especially useful to determine the independent contribution of each feature for model performance. Fifty top ranking sequence features selected by SVM-RFE are listed in Table 3, and were used for building target prediction models. A scoring system was developed to compute prediction scores for all genes in the genome. Most genes have a single predicted target-binding site, while some genes have multiple predicted sites. In both cases, the target prediction score was computed for each gene as the following:

$$S = 100^* \left(1 - \prod_{i=1}^{n} P_i\right),$$

where n represents the number of predicted target sites and P_i represents the statistical significance for each predicted site as estimated by the SVM model. The scores for single-site genes were calculated using the same equation with $n = 1$. These scores, ranging from 0 to 100, were used to rank the relative significance of the predicted targets. A gene candidate with a score of over 50 was predicted to be a target. The final prediction model, which we named MirTarget, was then used for genome-wide prediction of miRNA targets in human, mouse, rat, dog or chicken. The predicted targets are presented in an online database, miRDB (http://mirdb.org). Detailed statistics of genome-wide target prediction are presented in the Statistics page of miRDB. Of note, among all predicted targets, 23% were non-conserved targets. Beside the 3′-UTR, predicted unconventional target sites in the coding region or 5′-UTR are also presented in the Custom Prediction page of miRDB. Furthermore, the prediction tool, MirTarget can be accessed via the miRDB web server interface.

The performance of MirTarget was evaluated with receiver operating characteristic (ROC) curve analysis. CLIPL-seq training data were analyzed to determine the sensitivity and specificity of the prediction model. To reduce potential overtraining risk, 10-fold cross-validation was performed. Specifically, the dataset was randomly divided into 10 sample groups of equal size. For each iteration, samples from one group were removed and an SVM model was trained using samples from the remaining nine groups. The removed samples were then used for independent model testing. The process was repeated until all the sample groups had been used independently for model testing. For each iteration, a slightly different SVM model was generated for testing. In the end, the prediction results from all ten SVM models were added together and presented in Figure 2. Besides MirTarget, four established algorithms (TargetScan, DIANA-MicroT, miRanda-mirSVR and RNA22) were also evaluated in this ROC analysis. As shown in Figure 2, MirTarget had the best performance among all the algorithms, with an area under the curve of 0.86.

Table 3. List of all target recognition features that were used for SVM target prediction modeling[a]

Feature name	RFE rank
Seed site conservation	1
Target site location in UTR	2
GC content of target site	3
UG count	4
AG count	5
UTR length	6
Free energy of seed sequence binding	7
Pentamer motif match	8
GC content of 3′ end of target site	9
GC content of upstream nucleotides of target site	10
Position 1 A	11
Position 9 U	12
Position 1 G	13
CG count	14
Position 13 G	15
UA count	16
Position 11 G	17
AC count	18
A count	19
Position 9 C	20
Position 12 G	21
Target site <200 n.t. to UTR end	22
Position 10 AU	23
AA count	24
Position 1 AU	25
GG count	26
Position 16 G	27
GA count	28
Position 14 G	29
Position 9 A	30
Position 13 C	31
Position 9 AU	32
Position 13 AU	33
C count	34
CU count	35
AU count	36
CC count	37
Position 12 C	38
Position 11 AU	39
Position 0 G	40
Position 11 C	41
Position 15 G	42
GU count	43
Position 16 U	44
UC count	45
CA count	46
Position 16 C	47
G count	48
Target site > 800 n.t. to UTR end	49
Position 11 U	50

[a]Top-ranking target recognition features were identified by SVM-RFE. The seed-binding site is designated as positions 2–8. These features were summarized by analysis of target sites in 3′-UTR sequences.

3.4 Algorithm evaluation with independent experimental data

One common concern in computational data modeling is that the model may work well on the training data, but not as well on

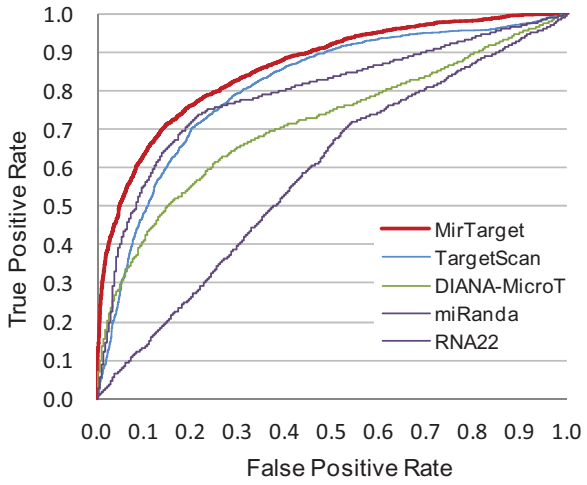

Fig. 2. ROC analysis of algorithm performance. ROC curves were constructed to evaluate MirTarget and four other target prediction algorithms using CLIPL-seq data. Ten-fold cross-validation was performed for MirTarget prediction to reduce potential over-training risk

independent unseen data. Thus, the best way to evaluate the performance of MirTarget would be to apply it to independent testing data. In this study, the general applicability of MirTarget was evaluated with independent microarray data (Hafner *et al.*, 2010). In this public study, 25 miRNAs were concurrently inhibited and the impact on target RNA expression was evaluated with microarrays. In this analysis, predicted targets by different algorithms were evaluated in the context of gene expression changes. First, we examined cumulative target distribution in relation to transcriptional expression changes. As shown in Figure 3A, the targets predicted by MirTarget were most upregulated as compared with the targets predicted by other algorithms. The average transcriptional expression change from top-ranking targets (300 targets on average per miRNA with the highest prediction scores from each algorithm) was also evaluated. As shown in Figure 3B, targets predicted by MirTarget had on average the highest fold change in upregulation. Further, all target prediction scores from each algorithm were correlated to gene expression changes by Pearson correlation analysis. As shown in Figure 3C, MirTarget prediction scores had the highest correlation to gene expression upregulation as compared with other algorithms. True miRNA targets were expected to be upregulated due to miRNA inhibition. Thus, the microarray profiling data suggested that MirTarget was most effective at identifying true miRNA targets.

4 Discussion

One major obstacle in computational target prediction is the lack of reliable experimental data to guide computational data modeling. To map the target sites for individual miRNAs more precisely, CLIPL methods have been developed recently to link a miRNA to its target site in the same RISC complex. Thus, CLIPL-seq profiling data have provided an unprecedented opportunity to study miRNA target recognition features in a more precise manner as compared with conventional approaches (Breda *et al.*, 2015). In particular, by analyzing individual nucleotide positions in the target sites, many novel position-specific sequence and structural features have been identified (presented in Fig. 1 and Table 2). Further, CLIPL analysis also confirmed certain well-characterized target features such as

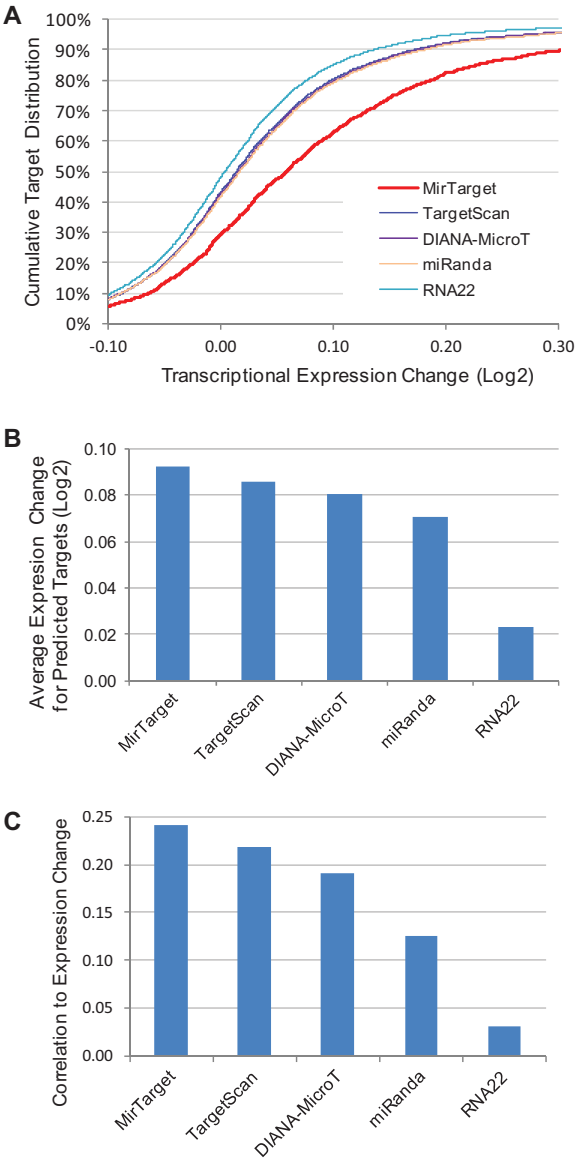

Fig. 3. Evaluation of target prediction algorithms with miRNA inhibition profiling data. Microarray profiling data were analyzed to identify target upregulation resulting from simultaneous inhibition of 25 miRNAs (Hafner *et al.*, 2010). Gene targets predicted by each algorithm were analyzed to identify the fraction of predicted targets that were upregulated as revealed by microarrays. **(A)** The curves in the plot represent cumulative distributions of all predicted targets by individual algorithms that were upregulated due to miRNA inhibition. **(B)** Average expression changes for top-ranking targets predicted by individual algorithms. For each algorithm, the average expression change of 300 gene targets on average per miRNA with the highest prediction scores was presented. **(C)** Correlation between gene expression changes and target prediction scores computed by each algorithm, as represented by the absolute value of Pearson correlation coefficient

cross-species conservation of seed-binding sites and the relevance of target site location within the transcript. By combining both known and newly identified target recognition features, an improved target prediction model, MirTarget was developed by training with CLIPL-seq data. The robust performance of MirTarget has also been validated with independent experimental profiling data.

Both canonical and non-canonical target sites have been identified in CLIPL-seq studies. However, our recent CLIPL-seq analysis

indicates that non-canonical targeting usually does not lead to target downregulation (Wang, 2014). As such, non-canonical targeting represents a set of interesting but poorly understood miRNA/target interactions that are not directly related to regulation of target expression. In contrast, canonical targeting involving perfect seed pairing usually leads to significant target downregulation. This raised an interesting question as to how an effective miRNA target should be defined. Non-canonical binding between a miRNA and a transcript could be functionally important, but may not involve target downregulation. One possibility is that non-canonical miRNA binding to a transcript may be related to the competing endogenous RNAs hypothesis as a way to sequester miRNAs from regulating the expression of canonical targets (Salmena *et al*., 2011). In this study, the main focus is to identify miRNA target sites, which leads to significant downregulation of target expression upon binding by the miRNA. Thus, only canonical target sites are considered by the prediction model. However, we would like to emphasize that prediction of non-canonical-binding sites may also be of practical importance, especially when functional impacts of such non-canonical interactions are better characterized in the future.

Funding

This work was supported by grant (R01GM089784) from the National Institutes of Health.

Conflict of Interest: none declared.

References

Agarwal,V. *et al*. (2015) Predicting effective microRNA target sites in mammalian mRNAs. *Elife*, **4**, e05005.

Ambros,V. (2004) The functions of animal microRNAs. *Nature*, **431**, 350–355.

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

Barrett,T. *et al*. (2013) NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res*., **41**, D991–D995.

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

Betel,D. *et al*. (2010) Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biol*., **11**, R90.

Breda,J. *et al*. (2015) Quantifying the strength of miRNA-target interactions, *Methods*, **85**, 90–99.

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

Gaidatzis,D. *et al*. (2007) Inference of miRNA targets using evolutionary conservation and pathway analysis. *BMC Bioinformatics*, **8**, 69.

Grimson,A. *et al*. (2007) MicroRNA Targeting Specificity in Mammals: Determinants beyond Seed Pairing. *Mol Cell*, **27**, 91–105.

Grosswendt,S. *et al*. (2014) Unambiguous identification of miRNA:target site interactions by different types of ligation reactions. *Mol. Cell*, **54**, 1042–1054.

Gumienny,R. and Zavolan,M. (2015) Accurate transcriptome-wide prediction of microRNA targets and small interfering RNA off-targets with MIRZA-G. *Nucleic Acids Res*., **43**, 1380–1391.

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

Helwak,A. *et al*. (2013) Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell*, **153**, 654–665.

Hofacker,I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res*., **31**, 3429–3431.

Kent,W.J. (2002) BLAT–the BLAST-like alignment tool. *Genome Res*., **12**, 656–664.

Kertesz,M. *et al*. (2007) The role of site accessibility in microRNA target recognition. *Nat. Genet*., **39**, 1278–1284.

Khorshid,M. *et al*. (2013) A biophysical miRNA-mRNA interaction model infers canonical and noncanonical targets. *Nat. Methods*, **10**, 253–255.

Kozomara,A. and Griffiths-Jones,S. (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res*., **42**, D68–D73.

Lim,L.P. *et al*. (2005) Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature*, **433**, 769–773.

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

Liu,C. *et al*. (2013) CLIP-based prediction of mammalian microRNA binding sites. *Nucleic Acids Res*., **41**, e138.

Long,D. *et al*. (2007) Potent effect of target structure on microRNA function. *Nat. Struct. Mol. Biol*., **14**, 287–294.

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.

NCBI Resource Coordinators. (2015) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*., **43**, D6–D17.

Reczko,M. *et al*. (2012) Functional microRNA targets in protein coding sequences. *Bioinformatics*, **28**, 771–776.

Robins,H. *et al*. (2005) Incorporating structure to predict microRNA targets. *Proc. Natl. Acad. Sci. USA*, **102**, 4006–4009.

Salmena,L. *et al*. (2011) A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell*, **146**, 353–358.

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

Wang,X. (2014) Composition of seed sequence is a major determinant of microRNA targeting patterns. *Bioinformatics*, **30**, 1377–1383.

Wang,X. and El Naqa,I.M. (2008) Prediction of both conserved and nonconserved microRNA targets in animals. *Bioinformatics*, **24**, 325–332.

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

Xia,T. *et al*. (1998) Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry*, **37**, 14719–14735.

Zhang,L. *et al*. (2007) Systematic identification of *C. elegans* miRISC proteins, miRNAs, and mRNA targets by their interactions with GW182 proteins AIN-1 and AIN-2. *Mol. Cell*, **28**, 598–613.

Zhao,Y. *et al*. (2005) Serum response factor regulates a muscle-specific microRNA that targets Hand2 during cardiogenesis. *Nature*, **436**, 214–220.