POLITECNICO DI TORINO
Repository ISTITUZIONALE

One Decade of Development and Evolution of MicroRNA Target Prediction Algorithms

(Article begins on next page)

05 February 2025

Accepted Manuscript

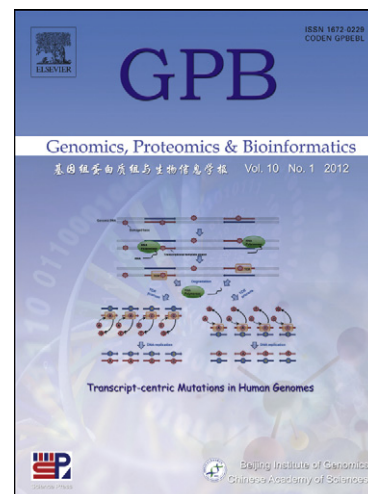One Decade of Development and Evolution of MicroRNA Target Prediction
Algorithms

Paula H. Reyes~Herrera, Elisa Ficarra

One Decade of Development and Evolution of MicroRNA Target Prediction Algorithms

Paula H. Reyes~Herrera[1*], Elisa Ficarra[2]

[1] *School of Electronic and Biomedical Engineering , Universidad Antonio Nariño, Bogotá, 110311, Colombia*

[2] *Department of Control and Computer Engineering, Politecnico di Torino, Turin, 10129 TO, Italy*

*Corresponding author.
E-mail: phreyes@gmail.com (Reyes~Herrera PH)

Submission to microRNA special issue

Running title: *Reyes~Herrera PH and Ficarra E / MicroRNA Target Prediction Status after One Decade*

Abstract

Nearly two decades have passed since the publication of the first study reporting the discovery of microRNAs (miRNAs). The key role of miRNAs in post-transcriptional gene regulation led to the performance of an increasing number of studies focusing on origins, mechanisms of action and functionality of miRNAs. In order to associate each miRNA to a specific functionality it is essential to unveil the rules that govern miRNA action. Despite the fact that there has been significant improvement exposing structural characteristics of the miRNA-mRNA interaction, the entire physical mechanism is not yet fully understood. In this respect, the development of computational algorithms for miRNA target prediction becomes increasingly important. This manuscript summarizes the research done on miRNA target prediction. It describes the experimental data currently available and used in the field and presents three lines of computational approaches for target prediction. Finally, the authors put forward a number of considerations regarding current challenges and future directions.

Keywords: MicroRNAs; MicroRNA recognition elements; Target prediction

Introduction

MicroRNAs (miRNAs) are short endogenous non-coding RNAs (ncRNAs), central actors in post-transcriptional regulation [1]. miRNAs bind the protein complex called RNA-induced silencing complex (RISC) and guide the complex toward specific sites, in particular mRNAs known as genes targets. By pairing specific sites in the mRNAs known as miRNA recognition elements (mRE), miRNAs direct post-transcriptional regulation, resulting in mRNA degradation or inhibition of protein translation.

However, the rules governing the mechanism of miRNA target regulation are not yet fully understood, making computational approaches for miRNA target prediction all the more important. In order to unveil miRNA functionality, it is critical to identify candidate targets. In fact, several computational approaches have been developed and experimental protocols have been proposed in order to improve the understanding of the mechanism.

Even though the first attempts to characterize miRNAs occurred almost twenty years ago [2], miRNAs were only reported as a significant class of small endogenous ncRNAs at the beginning of the last decade [1,3,4]. In fact, these molecules were named as miRNAs just one decade ago and since then research on miRNAs has been a focus of interest for numerous scientists worldwide, due to their powerful role in gene regulation.

As a result, research on miRNAs has flourished in the last decade. **Figure 1** shows three indicators of the research status: the number of mature miRNA sequences stored in miRBase [5], the number of publications reported in PubMed regarding miRNAs and the number of publications in PubMed with specific reference to miRNA targets. This figure not only evidences the timeline of this field, but also indicates the growing number of studies regarding miRNAs and their gene targets. It is worth noting that, according to this figure, research on gene targets of miRNAs has mainly developed after the year 2003.

Several reviews on miRNAs and their targets have been published focusing on the examination of biological principles [6,7], experimental techniques and computational prediction algorithms [8-10]. Recently, various review papers have highlighted the expansion of experimental data validating miRNA-mRNA interactions [8,9,11-14]. However, the review of miRNA target prediction algorithms is limited to the first computational algorithms which were developed based on *ab initio* strategies [7,8,10,15].

This paper attempts to present updated information, not only regarding experimental techniques, but also prediction algorithms. We summarize the current status on miRNA target prediction, pointing out the most important considerations that should be taken into account.

It is noteworthy that these considerations are addressed, not only to users of prediction tools, but also to developers. We first introduce the experimental techniques used to obtain miRNA-mRNA interactions and then present the most relevant identified characteristics of the structural interaction. Furthermore, we introduce three lines of computational algorithms for target prediction, *i.e.*, a*b initio*, machine learning and hybrid, and provide examples for each line. Finally, current challenges and future directions are discussed.

Experimental data

There is no "golden rule" regarding the technique to identify or validate miRNA-target interactions. In fact, several techniques have been used to obtain experimental data that supports miRNA-target interactions [11]. Experimental data is critical, not only to distinguish a specific interaction, but also to study features that characterize miRNA-mRNA interactions and to validate the accuracy of the computational approaches proposed. It is therefore fundamental to briefly introduce the experimental data currently available regarding miRNA-mRNA interactions. In order to understand the advantages and limitations of the data derived with each experimental technique, comments for each technique are presented in order.

Experimental techniques can be classified in two classes, depending on the type of supporting information provided: direct or indirect. In addition, the experimental data can also be categorized, depending on the resultant size of dataset: individual studies or high throughput.

The individual studies can provide direct support to validate the identified candidates. Frequently, reporter genes (such as luciferase and GFP) attached to the genes of interest were used, and the expression of reporter gene was measured before and after the introduction of miRNA to the cell [16]. Such procedure can provide direct support but fails to identify the specific mRE (particularly useful to understand the structural characteristics of the interaction). Thus in order to obtain the specific mRE, reporter genes can be attached to both the original and mutated sequences of the gene of interest. Gene expression in both samples is then measured before and after miRNA transfection [16,17]. In this way, it is possible to identify the specific site of interaction.

Moreover, the resultant experimental data size using reporter gene assays is small. Therefore a different experimental validation strategy was proposed and used in several studies [18]. In particular, expression was measured for a large number of genes through manipulating miRNA expression, either overexpression by transfection or knockdown. In the former case, a decrease in expression of target mRNAs and proteins is expected (down

regulation) with increased expression of miRNAs [18], while in the latter case, an increase in expression of target mRNAs and proteins is expected (upregulation) with the miRNA expression silenced in cells [19]. Other than that, under different biological conditions, miRNA expression varies and consequently the expression of target mRNAs and proteins varies as well [20].

Experimental techniques like microarray and PCR are commonly used to measure gene expression [18]. However, since miRNA regulates not only mRNA expression but also protein levels, in doing so, targets by inhibition of mRNA translation into protein are left out. There are a few studies that used also immunoblotting to measure protein expression [19]. Furthermore, high-throughput proteomics techniques have been proposed and are used to identify both types of targets. In particular, strategies like stable isotope labelling by/with amino acids in cell culture (SILAC) [21] and pulsed SILAC (pulsed SILAC) [22] are able to provide a high throughput dataset by using mass spectrometry (MS). Ribosome profiling is another approach for identification of both types of targets. Ribosome profiling, which is based on deep sequencing of ribosome-protected mRNA fragments [23], is a sensitive method to quantify and detect the mRNAs at the ribosome.

Nevertheless, expression-based validation strategies are indirect because the set of mRNAs/proteins with an associated microRNA induced change of expression, contains both direct targets (structural interaction) and indirect targets (the expression of the indirect target is caused by a direct target but not by the microRNA). In addition, further considerations should be taken into account depending on how the miRNA expression is manipulated. In particular, in the over-expression experiments there might be targets that, despite being affected by miRNA over-expression, do not show a high degree of down-regulation due to factors such as the saturation of the microRNA ribonucleo protein complex (miRNP) [24].

Lately, immunoprecipitation of proteins from the RISC complex has been used to identify the mRNAs where the miRNAs bind [25]. In addition, combination of crosslinking-inmunoprecipitation and high throughput sequencing has been used to isolate the mRNAs where miRNAs and protein complex bind and to obtain sequences containing the specific site of interaction [26-27]. In particular, approaches like high-throughput sequencing of RNA isolated by crosslinking immunoprecipitation (HITS-CLIP) [26] and photoactivatable-ribonucleoside-enhanced crosslinking and immunoprecipitation (PAR-CLIP) [27] have been used to isolate, quantify, and sequence portions of mRNAs that contain the sites of miRNA-mRNA interaction. Although the sites of interaction are determined, these approaches can't identify the specific miRNA-mRNA association experimentally, which instead is estimated

by using features commonly found in experimental samples such as the seed complementary sequence.

Each of the aforementioned techniques provides an important source of information for miRNAs and genes target interaction. In particular, data with strong direct structural support is fundamental because physical interactions occur. The authors strongly advise to use data provided by direct methods to validate or train computational tools that perform the prediction based on structural characteristics.

However, indirect methods are also an important source of information. The experimental data provided contains functional targets (both direct and indirect) where the functional regulation (up or down-regulation) was induced. Evaluating the accuracy of the prediction tool (prediction of physical interaction) based on expression data does not indicate robust result since the dataset does not distinguish between indirect and direct targets. Nevertheless, a computational prediction tool can be used to distinguish direct and indirect targets in expression data sets.

Special attention should be given to the experimental data selection since indirect and direct methods perform under different assumptions. As a matter of fact, distinct target determinants between expression-based and CLIP-based data were observed in a recent study [28]. Thus, the nature of the experiment characteristics should be taken into account, in particular when the data is used to train a computational target prediction method or for validation purposes.

Databases with experimental data

The growing interest in the field has been accompanied by the continuous evolution of experimental techniques and an associated expansion of the experimental data obtained. In order to provide a common benchmark for different studies, several databases have been developed to deposit and share experimental data. In this section, the most popular databases that deposit experimental data regarding validated miRNA-mRNA interactions are presented.

The first version of Tarbase [29] was introduced in 2005 and the five previous versions of this database lacked several annotations. For example, no clear indication for records containing predicted sites (not experimentally derived) was given. However, the most recent version, Tarbase v6 [30] released in 2011, shows a number of significant improvements. In particular, several databases, such as miRecords [31], have been integrated into Tarbase v6. In addition, more details are provided regarding each interaction and it is now possible to

select data based on the experimental technique used, regulation type (up regulation, down regulation and unknown) and type of interaction (positive and negative). **Figure 2A** shows the proportions of experimental techniques used to obtain the data deposited in Tarbase v6. It is worth noting that Tarbase collects data provided by both individual studies and high-throughput studies. The largest dataset was obtained using high-throughput methods like microarrays, which alone provide an indirect type of validation as aforementioned.

miRecords [31] was released in 2008 and each record in this database was manually curated. To date, three versions of miRecords have been developed with the latest release dating back to 2010. This database contains a significant amount of miRNA-mRNA interactions, most of the interactions deposited in this database have been derived from individual studies. In fact, as shown in **Figure 2B,** the largest proportion of deposited experimental data is obtained using reporter gene assays and most of the data included in this database has a direct type of validation. Despite the fact that the latest version was released in 2010, this database is still an important resource of interactions with strong experimental support.

Two additional databases, miRtarbase [32] and starBase [33], have been recently published. miRtarbase collects miRNA-mRNA interactions and classifies miRNA-target interactions (MTIs) into four classes including functional, functional weak (indirect experimental support), non-functional and non-functional weak, depending on the power of the experimental technique used and the type of interaction (positive or negative). Classification of MTIs is particularly useful to select experimental data according to the associated support.

Finally, starBase [33] collects data provided by high-throughput CLIP-seq, such as HITS-CLIP [26] and PAR-CLIP [27]. This type of data consists of sites of interaction for the mRNA-miRNA-Argonaute complex on a transcriptome-wide scale. Although the database initially contained data from 21 CLIP-seq experiments, the number of studies using CLIP-seq experiments is likely to grow in the next few years, considering its potentiality. Specifically, 91,124 interactions for the human are currently deposited in starBase.

Features of miRNA-mRNA interactions

The identification of common characteristics for targets and specific sites with strong experimental support is fundamental in order to unveil the rules that govern miRNA-mRNA interactions. Therefore, common characteristics found in experimental data have been extracted.

In particular, characteristics associated with the duplex miRNA-site interactions are still being explored [34]. Among the duplex features, great importance has been conferred to a region in the miRNA sequence named seed [35]. A strong complementarity to the seed region was found in a significant number of experimentally-derived sites. The seed is located in the 5' section of the miRNA and different types of seeds sites were identified based on the length and complementarity (7mer-A1, 7mer-m8 and 8mer). Considering the importance of the seed region of interaction between miRNA and mRNA is commonly classified as the seed region and out-seed region. The out-seed region also plays an important role in the duplex interaction, either to reinforce the affinity (supplementary pairing) or to compensate for incomplete seed pairing (complementary pairing) [36]. In addition, within the out-seed region there was a preferential pairing from the 13th to the 16th nucleotide in the miRNA sequence [37]. Furthermore, additional characteristics regarding the duplex have been commonly found in sites of interaction [39-40]. In particular, the duplex minimum free energy associated with the interaction stability represents a determinant characteristic [39]. Frequently the conservation of the mRNA site is also an important discriminant [40].

Recently, a characteristic from the duplex was extracted from a CLIP-seq data set and investigated in a few miRNAs, such as the miR124. The feature consists of the presence of a G bulge in the 6th nucleotide of the miRNA that acts as a pivot [34]. Nevertheless, further examinations on additional datasets are needed to assess this consideration.

Not only the characteristics of the duplex delineate the interaction, characteristics associated with the environment that encloses the site in the mRNA are also significant indicators. Since the mRNA has a complex structure itself, the surrounding conditions favour/disfavour the accessibility of the miRNA to the mRE and therefore influence the interaction. A number of characteristics indicate the accessibility, such as: (1) AU content in the upstream and downstream neighbourhood and (2) AU motifs in the entire mRNA (such as AUUUA pentamer). The first characteristic promotes the miRNA access to the site [38,41], while the presence of motif signatures for RNA-binding-proteins may attenuate or enhance the regulation executed by the miRNA [42]. The presence of a GC motif downstream of the site was also extracted from sites derived experimentally [43]. In addition, features that reflect the amount of energy that must be spent in order to introduce changes from the original mRNA structure to the resulting structure after the interaction appear as important determinants [39]. In particular, the feature called ΔΔG reflects the difference in free energy between the duplex and the neighbourhood of the site initial structure [44,45]. The specific position of the site is also an important consideration. Even though the majority of the sites

have been found in the untranslated regions (UTR) of mRNAs, some mREs have also been found in the coding region [26]. However, the regulation appears to be more effective when the site of interaction is in the UTR [26]. In particular, sites in the UTR are most likely accessible if are located near the start codon or the stop codon [37]. The majority of mRE located in the UTR are found in long UTRs [46]; therefore the UTR length would seem to be another element that should be considered.

Nevertheless, none of these aforementioned characteristics are present in all the sites that have been derived experimentally [28]. One of the most commonly found characteristics is the seed, but there is a subset of sites that do not contain the seed complementarity. In fact, in the dataset derived in [26] around 27% of the identified sites did not possess seed complementary to the expressed miRNAs; such sites may bind to other miRNAs or follow different rules.

In the section Experimental Data, differences between the experimental strategies used to obtain data for miRNA-mRNA interactions were presented. Different considerations should be taken for datasets obtained from expression-based experiments and for those obtained using CLIP-seq protocols. Recently, a study [25] compared the characteristics of the interactions derived with these high-throughput strategies and noted some discrepancies. In particular, accessibility features are strongly present in data obtained with CLIP-seq techniques while duplex-related features, specifically the seed, are strong determinants for expression-based experiments. Even though there is a great overlap between the data provided by both experimental techniques, the discrepancies can be attributed to increased miRNA concentrations (overexpression) typical of expression-based experiments.

Computational algorithms

Computational algorithms for miRNA target prediction have been essential in order to identify the candidate targets and therefore the targets. Since 2003, almost one decade of development of computational miRNA target prediction algorithms has passed. Current prediction algorithms based on structural characteristics (such as the ones presented in the Features of miRNA-mRNA interactions section) can be grouped into three lines: a*b initio*, machine learning and hybrid approaches.

The first algorithms proposed are in the a*b initio* line. These algorithms perform the prediction based on the structural features extracted from data with experimental support [47-55]. They are based on computational models that do not use the experimental data directly. Machine learning (ML) approaches, on the other hand, use computational algorithms that rely

directly on experimental data to train a classifier [57-68]. In this way, the classifier is able to identify a candidate target site based on similarity to the experimental training set. Machine learning algorithms started to appear when the number of interactions with experimental support increased significantly.

Each line has an associated pitfall. For a*b initio* algorithms it is the high number of false positives [26] and for machine learning approaches it is the reduced number of negative interactions with experimental support (negative interactions are often not published and not recorded in databases). The set of predictions generated by a*b initio* algorithms contains a notable number of false positives. In order to overcome this problem, *ab initio* algorithms use several restrictions to retain candidates that have a high probability of being targets and filter out false positives. However with filtering, several true positives may also be discarded.

Machine learning approaches identify the probable candidates (positives) from the unlikely candidates based on the experimental data that represents positive and negative interactions. However, negative experimentally-identified interactions are usually discarded and therefore the currently available negative set (negative interactions with experimental support) is quite poor compared to the positive set.

The drawbacks of a*b initio* and machine learning algorithms, have led to the development of hybrid algorithms with characteristics from both lines incorporated. These hybrid algorithms integrate merits from each line in order to meet the current challenges of prediction algorithms.

The most popular *ab initio*, machine learning and hybrid algorithms are briefly discussed below. In addition, a summarizing table of the algorithms with several considerations can be found in the supplementary material (Table S1).

Ab initio algorithms

- miRanda [47,48] uses a weighted dynamic programming algorithm to obtain the candidate sequences. This algorithm uses a score to rank the predictions that consists of a weighted sum based on matches, mismatches and G:U wobbles. Initially, miRanda [47] used features such as seed complementarity and duplex free energy; the most recent version also takes into account a conservation measure based on the PhastCons conservation score. The algorithm and the set of target predictions are available online (http://www.microrna.org).

- TargetScan [35,37,49]: this algorithm requires the seed complementary at least for 6 nt and considers the different seed types that have been defined, with a certain hierarchy (6mer offset <6mer<7mer-A1<7mer-m8<8mer) [36]. Moreover, TargetScan ranks the sites

using a context score based on seed complementarity, conservation and AU content in the site vicinity. In the recent release of the latest version of TargetScan[50], a number of additional determinants have been integrated while retaining the previous considerations. In particular, a multiple linear regression trained on 74 filtered datasets was used to integrate determinants such as seed-pairing stability (SPS) and target-site abundance (TA). TargetScan is available online (http://www.targetscan.org/).

- Pictar [51]: this algorithm has strict requirements regarding the seed and also takes into account the overall duplex stability based on the minimum free energy. Once the sites are aligned, the targets are ranked based on a score derived using a hidden Markov model that considers the site conservation. Predictions obtained with Pictar are available online (http://pictar.mdc-berlin.de/).

- RNA22 [52] is a pattern-based discovery strategy to identify the candidate targets. First, a Markov chain is used for pattern discovery, but only the most statistically significant patterns are retained to identify target islands (areas where many statistically significant patterns map). Consequently, the target islands are paired with miRNAs. The target islands that represent candidate binding sites for miRNAs are selected based on user-imposed parameters (minimum number of base pairs, maximum number of unpaired bases and maximum allowed free energy). RNA22 is available online (http://cbcsrv.watson.ibm.com/rna22.html).

- RNAHybrid [53] is an algorithm that finds the minimum free energy not only for short sequences (miRNA-mRE) as most of the previously-reported algorithms, but also for the entire miRNA-mRNA. The user can impose several restrictions, such as the number of unpaired bases and free energy allowed, to reduce the set of resulting predictions. RNAHybrid is available online (http://bibiserv.techfak.uni-bielefeld.de/rnahybrid/) .

- PITA [44] is a proposal that considers not only the specific duplex interaction information, but also takes into account the accessibility to the site in the mRNA. Accessibility is considered as the difference between the minimum free energy of the entire complex and the energy that originally had a short region of the mRNA near the site, $\Delta\Delta G$. The user can impose different restrictions to reduce the resultant set of candidates (minimum seed size, G:U bobbles and unpaired bases). PITA is available online (http://genie.weizmann.ac.il/pubs/mir07/).

- EIMMO [54] is an algorithm that scores the sites based on the conservation score and uses a Bayesian method. It infers the phylogenetic distribution for the functional sites. To

characterize miRNA function, associations between the predicted targets and biochemical pathways (KEGG) are searched. EIMMO is available online (http://www.mirz.unibas.ch/ElMMo2/).

- • DIANA [55]: this algorithm measures the goodness of an interaction based on its specific characteristics. Each gene is weighted taking into consideration conserved as well as non-conserved sites. Moreover, a signal to noise ratio (SNR) is obtained for each interaction to estimate the number of false positives. DIANA is available online (http://diana.cslab.ece.ntua.gr/microT/) .

An interesting performance comparison of the nine most popular *ab initio* algorithms was carried out [8] on the dataset obtained in [22]. The results evidence the usage of strict restrictions by these algorithms to reduce the number of false positives. In particular, algorithms such as PicTar, TargetScan and Diana, have significantly-compromised sensitivity (~10%) in order to achieve a remarkable precision (~50%). However, the test used an expression-based dataset in which a set of miRNAs was overexpressed. As previously mentioned, when evaluating the performance on an indirect dataset, further considerations, such as the presence of indirect targets in both the positive and negative sets, and the effects caused by the overexpression like the miRNP saturation, must be taken into account.

It is also important to highlight a test performed in [56] using four *ab initio* algorithms (PITA, TargetScan, PicTar and miRanda) on CLIP-seq datasets obtained from Starbase. The intention of the test was to present the coverage obtained by these algorithms on CLIP-seq data. miRanda demonstrated the best sensitivity overall (66%), while TargetScan and PicTar were characterized by the lowest sensitivities (20%). This result validates (1) differences in the target determinants provided by expression-based and CLIP-based data and (2) strict restrictions imposed by *ab initio* algorithms like PicTar and TargetScan.

Machine learning and hybrid algorithms

Machine learning approaches appeared later than *ab initio* approaches. Nevertheless, the importance of these methods has grown since the data with experimental support started to grow significantly. Representatives from this line are briefly described as follows. Since machine learning algorithms strongly rely on experimental data, we also specify the size of the respective training dataset.

- • TargetBoost [57] consists of a boosting algorithm that assigns weights to sequence patterns of 30 nucleotides. The negative dataset used for training consists of 300 randomly-

generated sequences, and the positive data set consists of 36 interactions with experimental support. The set of predictions obtained with the algorithm for the *C. elegans* can be found online, but the algorithm itself is not currently available.

- miTarget [58] is an algorithm that uses a support vector machine (SVM) with an radial basis function (RBF) as kernel, to predict the candidate targets. It is based on structural, thermodynamic and positional features. The negative set used for training consists of 83 interactions with experimental support plus 163 negative interactions inferred from experimental data. The positive dataset consists of 152 interactions with experimental support. miTarget is available online (http://cbit.snu.ac.kr/~miTarget/introduction.html).

- Ensemble Algorithm [59], a post-processing step for miRanda, consists of 10 SVM (polynomial kernels). The prediction is based on features from the miRNA-mRE interactions, and features from the mRNA targets. The negative and positive datasets used for training consist of 16 and 48 experimentally-verified interactions, respectively.

- NBmiRTar [60] consists of a post-processing step to miRanda. First a filter based on the folding energy is applied, followed by a filter based on the score obtained by miRanda and score obtained by a Naïve Bayes classifier. The prediction is based on structural features from the miRNA-mRE duplex features and observed sequence features. The negative dataset was composed of 38 negative interactions with experimental support and 133,316 generated target sites for artificial miRNA sequences, while the positive dataset consists of 225 interactions with experimental support.

- miRTarget2 [61] uses an SVM classifier to obtain set of predictions. The features used include characteristics from the miRNA-mRE duplex and from the mRNA. The positive and negative datasets for training consists of 1,017 negative interactions and 454 positive interactions with experimental support, respectively. The predicted interactions are available online (http://mirdb.org).

- miRTif [62] starts from the combination of the sets predicted by miRanda, PicTar and Targetscan. It then uses an SVM classifier (RBF kernel) based on features from the miRNA-mRE interaction. The positive and negative datasets contain 195 and 21 interactions with experimental support, respectively. In addition, the negative dataset contains 17 interactions inferred from experimental data. miRTif is available online (http://mirtif.bii.a-star.edu.sg/).

- TargetMiner [63] first selects a set of sites based on the seed complementarity. It then uses an SVM classifier (RBF kernel) based on mRNA and miRNA-mRE duplex features. The positive dataset is composed of 476 positive interactions and the negative data set contains 59

experimental interactions plus 289 inferred negative interactions. TargetMiner is available online (http://www.isical.ac.in/~bioinfo_miu/targetminer20.html).

- mTar [64] first selects 3 classes of sites (5' seed only, 5' dominant and 3' canonical). It then uses an artificial neural network to classify targets and non-targets based on features from the miRNA-mRNA interaction. The dataset used for training contains 340 positive miRNA-mRNA interactions and 400 negative ones. mTar is available online (http://www.rgcb.res.in/downloads/Mtar.rar).

- TargetSpy [65] generates candidate zones, which are merged and a ranking of the zones is performed. This algorithm uses an automatic feature selection based on compositional, structural and base pairing features. The positive and negative dataset contains 3872 positive and 4540 negative instances, respectively, derived using the HITS-CLIP [9] protocol. TargetSpy is available online (http://www.targetspy.org/).

- mirSVR, an algorithm proposed by the developers of miRanda, is a hybrid approach that uses miRanda followed by a support vector regression (SVR). In practice, miRanda is used to obtain set of predictions, which are then ranked using a machine learning approach called mirSVR [66]. mirSVR is trained based on expression changes caused by miRNA overexpression, to obtain a score for each prediction that represents an empirical probability of down regulation. Sets of predictions provided by this algorithm are available online (http://www.microrna.org).

- miRror [67] is a tool based on the notion of miRNA combinatorial mode of action. The algorithm combines several *ab initio* predictors into a unified platform by incorporating a statistical measure. miRror is available online (http://www.proto.cs.huji.ac.il/mirror/search.php).

- miREE [68] is a hybrid algorithm composed of two parts. The first part uses a genetic algorithm to generate a set of sequences that represent an optimal diverse population for binding sites. These binding sites are then mapped to mRNAs and classified as targets and non-targets using SVM (RBF Kernel), a machine learning technique. The positive and negative datasets contain 324 and 351 interactions, respectively. This algorithm is available online (http://didattica-online.polito.it/eda/miREE/).

The negative dataset is one of the principal limitations for machine learning approaches, since negative interactions are not commonly published. Several strategies have been used to overcome this drawback. In particular, the negative dataset has been expanded by using 1) random sequences, 2) predictions for artificial miRNA sequences and 3) sequences of genes

that were not regulated by the miRNAs, from which the negative interaction sites were not extracted. For the first two solutions, the generated interactions do not have experimental support. Therefore, the third alternative is strongly recommended due to the presence of experimental support.

A recent comparison of the performance for five machine learning and hybrid algorithms was carried out in [68]. The comparison test was performed on a direct dataset in which each record had strong experimental support and was obtained from public databases. This comparison evidences that most of the algorithms are characterized by an unbalanced performance (sensitivity and specificity). In particular, the machine learning prediction algorithms were designed to reduce the number of false positives (principal limitation of *ab initio* proposals), which has significant impact on the sensitivity and the overall accuracy.

Further considerations

There has been a significant advance on development of microRNA target prediction algorithms. However, there is still room for further improvement. In particular, it is worth to integrate data derived using the lately-developed protocols (CLIP-seq), which show remarkable potential. A few approaches [65,68] used CLIP-seq datasets to train machine learning algorithms and for validation purposes. Nevertheless, the nature of the data (high-throughput and direct validation) will most likely lead to unveil additional rules and take prediction algorithms to a different stage.

Moreover, expression-based data is also an important source of information. In fact, the use of expression-based data together with computational prediction algorithms shows significant potential, if with the right assumptions. When integrating expression data with prediction tools, it is important to associate different rules with the mechanisms of regulation [50] (mRNA abundance and mRNA stability) and distinguish between direct and indirect targets. In addition, by using expression-based data, it is possible to correlate the structural interaction with the degree of regulation exerted by the miRNA-mRNA interactions [66].

Furthermore, the interest in miRNA target prediction tools is not limited only to obtaining a set of candidate targets. Day by day, it is becoming more important to understand the miRNA functionality (functional correlation between the multiple targets both direct and indirect). In order to unveil the miRNA functionality, it is important to integrate set of predicted candidates with a different source of information and the degree of miRNA-exerted regulation in expression-based experiments is a feasible alternative. Moreover, information

available in databases that contain pathways with experimental support and Gene Ontology is helpful.

Another important direction in the field is to understand the joint action mechanisms involved in transcriptional and post-transcriptional regulation, cooperative action of transcription factors and miRNAs, and also the collective action of miRNAs and RNA-binding proteins.

Finally, exploring the interactions of miRNAs not only with coding RNA but also with the entire transcriptome might be helpful in order to understand not only the miRNA functionality but also the role of other RNAs such as long ncRNAs.

Conclusion

Significant progress has been made in computational algorithms for miRNA target prediction during the last decade. In particular, this evolution has been influenced by the development of experimental protocols that expanded the datasets available with experimental support. Albeit the *ab initio* algorithms were first proposed, with the expansion of experimental data, the use of machine learning and hybrid proposals is very promising.

Candidate target genes for prediction algorithms built under structural assumptions should be validated with experimental data with strong structural support (direct validation). The use of expression data to validate the prediction results for algorithms based on structural assumptions is susceptible to misunderstandings.

Competing interest

The authors developed one of the computational algorithms [68] but avoided any preference in this manuscript.

References

[1] Bartel D. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 2004; 116: 281-97.

[2] Lee RC, Feinbaum RL, Ambros V. The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell 1993; 75: 843-54.

[3] Lee RC, Ambros V. An extensive class of small RNAs in *Caenorhabditis elegans*. Science 2001; 294: 862-4.

[4] Lagos-QuintanaM, Rauhut R, Lendeckel W, Tuschl T. Identication of novel genes coding for small expressed RNAs. Science 2001; 294: 853-8.

[5] Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ. miRBase: tools for microRNA genomics. Nucleic Acids Res 2008; 36: D154-8.

[6] Cai Y, Yu X, Hu S, Yu J. A brief review on the mechanisms of miRNA regulation. Genomics Proteomics Bioinformatics 2009; 7: 147-54.

[7] Mendes ND, Freitas AT, Sagot MF. Current tools for the identification of miRNA genes and their targets. Nucleic Acids Res 2009; 37: 2419-33.

[8] Alexiou P, Maragkakis M, Papadopoulos GL, Reczko M, Hatzigeorgiou AG. Lost in translation: an assessment and perspective for computational microRNA target identification. Bioinformatics 2009; 25:3049–55.

[9] Witkos TM, Koscianska E, Krzyzosiak WJ. Practical aspects of microRNA target prediction. Curr Mol Med 2011; 11: 93-109.

[10] Min H, Yoon S. Got target? Computational methods for microRNA target prediction and their extension. Exp Mol Med 2010; 42: 233-44.

[11] Lund AH. Experimental identification of microRNA targets. Gene 2010; 451: 1-5.

[12] Jin H, Tuo W, Lian H, Liu Q, Zhu X.Q, Gao H. Strategies to identify microRNA targets: new advances. N Biotechnol 2010; 27: 734-8

[13] Thomson DW, Bracken CP, Goodall GJ. Experimental strategies for microRNA target identification. Nucleic Acids Res 2011; 39: 6845-53.

[14] Thomas M, Lieberman J, Lal A. Desperately seeking microRNA targets. Nat Struct Mol Biol 2010; 17:1169–74.

[15] Watanabe Y, Tomita M, Kanai A. Computational methods for microRNA target prediction. Methods Enzymol 2007; 427: 65–86.

[16] Kiriakidou M, Nelson PT, Kouranov A, Fitziev P, Bouyioukos C, Mourelatos Z, et al. A combined computational-experimental approach predicts human microRNA targets. Genes Dev 2004; 18:1165-78.

[17] Didiano D, Hobert O. Molecular architecture of a miRNA-regulated 3' UTR. RNA 2008; 14: 1297-317.

[18] Lim LP, Lau NC, Garrett-Engele P, Grimson A, Schelter JM, Castle J, et al. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. Nature 2005; 433: 769-73.

[19] Ji R, Cheng Y, Yue J, Yang J, Liu X, Chen H, et al. MicroRNA expression signature and antisense-mediated depletion reveal an essential role of MicroRNA in vascular neointimal lesion formation. Circ Res 2007; 100: 1579–88.

[20] Pandey P, Brors B, Srivastava PK, Bott A, Boehn SNE, Groene HJ, et al. Microarray-based approach identifies microRNAs and their target functional patterns in polycystic kidney disease. BMC Genomics 2008;9:624

[21] Baek D, Villén J, Shin C, Camargo FD, Gygi SP, Bartel DP. The impact of microRNAs on protein output. Nature 2008; 455: 64-71.

[22] Selbach M, Schwanhausser B, Thierfelder N, Fang Z, Khanin R, Rajewsky N. Widespread changes in protein synthesis induced by microRNAs. Nature 2008; 455: 58-63.

[23] Guo H, Ingolia NT, Weissman JS, Bartel DP. Mammalian microRNAs predominantly act to decrease target mRNA levels. Nature 2010; 466: 835-40.

[24] Khan AA, Betel D, Miller ML, Sander C, Leslie CS, Marks DS. Transfection of small RNAs globally perturbs gene regulation by endogenous microRNAs. Nat Biotechnol 2009; 27: 549-55.

[25] Tan LPP, Seinen E, Duns G, de Jong D, Sibon OC, Poppema S, et al. A high throughput experimental approach to identify miRNA targets in human cells. Nucleic Acids Res 2009; 37:e137.

[26] Chi SW, Zang JB, Mele A, Darnell RB. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. Nature 2009; 460: 479-86.

[27] Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, et al. Transcriptome-wide identication of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell 2010; 141: 129-41.

[28] Wen J, Parker BJ, Jacobsen A, Krogh A. MicroRNA transfection and AGO-bound CLIP-seq data sets reveal distinct determinants of miRNA action. RNA 2011; 17: 820-34.

[29] Papadopoulos GL, Reczko M, Simossis VA, Sethupathy P, Hatzigeorgiou AG. The database of experimentally supported targets: a functional update of TarBase. Nucleic Acids Res 2009: 37: D155-8.

[30] Vergoulis T, Vlachos IS, Alexiou P, Georgakilas G, Maragkakis M, Reczko M, et al. TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support. Nucleic Acids Res 2012; 40: D222-9.

[31] Xiao F, Zuo Z, Cai G, Kang S, Gao X, Li T. miRecords: an integrated resource for microRNA-target interactions. Nucleic Acids Res 2009; 37: D105-10.

[32] Hsu SD, Lin FM, Wu WY, Liang C, Huang WC, Chan WL, et al. miRTarBase: a database curates experimentally validated microRNA-target interactions. Nucleic Acids Res 2011; 39: D163-9.

[33] Yang JH, Li JH, Shao P, Zhou H, Chen YQ, Qu LH. starBase: a database for exploring microRNA-mRNA interaction maps from Argonaute CLIP-Seq and Degradome-Seq data. Nucleic Acids Res 2011; 39: D202-9.

[34] Stefani G, Slack FJ. A 'pivotal' new rule for microRNA-mRNA interactions. Nat Struct Mol Biol 2012; 19: 265-6.

[35] Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell 2005; 120:15-20.

[36] Bartel DP. MicroRNAs: target recognition and regulatory functions. Cell 2009; 136: 215-33.

[37] Grimson A, Farh KK, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Mol Cell 2007; 27: 91-105.

[38] Barreau C, Paillard L, Osborne BH. AU-rich elements and associated factors: are there unifying principles? Nucleic Acids Res 2005; 33: 7138-50.

[39] Muckstein U, Tafer H, Hackermüller J, Bernhart SH, Stadler PF,Hofacker IL. Thermodynamics of RNA–RNA binding. Bioinformatics 2006;22:1177–82.

[40] Altuvia Y, Landgraf P, Lithwick G, Elefant N, Pfeffer S, Aravin A, et al. Clustering and conservation patterns of human microRNAs. Nucleic Acids Res 2005; 33: 2697–706.

[41] Enright AJ. MicroRNA Research. Fundamentals, Reviews and Perspectives. The 2008 Collection Booklet. Denmark: Exiqon, Inc.

[42] Jacobsen A, Wen J, Marks DS, Krogh A. Signatures of RNA binding proteins globally coupled to effective microRNA target sites. Genome Res 2010; 20: 1010-9.

[43] Schmidt T, Mewes HW, Stümplen V. A novel putative miRNA target eEnhancer signal. PLoS One 2009; 4: e6473.

[44] Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E. The role of site accessibility in microRNA target recognition. Nat Genet 2007; 39: 1278-84.

[45] Zhao Y, Samal E, Srivastava D. Serum response factor regulates a muscle-specific microRNA that targets Hand2 during cardiogenesis. Nature 2005; 436: 214-20

[46] Sandberg R, Neilson JR, Sarma A, Sharp PA, Burge CB. Proliferating Cells Express mRNAs with Shortened 3' Untranslated Regions and Fewer MicroRNA Target Sites. Science 2008; 320:1643–7

[47] John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS. Human microRNA targets. PLoS Biol 2004; 2: e363

[48] Betel D, Wilson M, Gabow A, Marks DS, Sander C. The microRNA.org resource: targets and expression. Nucleic Acids Res 2008; 36: D149- 53.

[49] Friedman RC, Farh KK, Burge CB, Bartel DP. Most mammalian mRNAs are conserved targets of microRNAs. Genome Res 2009; 19: 92-105.

[50] Garcia DM, Baek D, Shin C, Bell GW, Grimson A, Bartel DP. Weak seed pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. Nat Struct Mol Biol 2011; 18: 1139-46.

[51] Lall S, Grün D, Krek A, Chen K,Wang YLL, Dewey CNN, et al. A genome-wide map of conserved microRNA targets in *C. elegans*. Curr Biol 2006; 16: 460- 71.

[52] Miranda KC, Huynh T, Tay Y, Ang YSS, Tam WLL, Thomson AM, et al. A pattern-based method for the identification of microRNA binding sites and their corresponding heteroduplexes. Cell 2006; 126: 1203-17.

[53] Kruger J, Rehmsmeier M. RNAhybrid: microRNA target prediction easy, fast and flexible. Nucleic Acids Res 2006; 34: W451-4.

[54] Gaidatzis D, van Nimwegen E, Hausser J, Zavolan M. Inference of miRNA targets using evolutionary conservation and pathway analysis. BMC Bioinformatics 2007; 8: 69.

[55] Maragkakis M, Alexiou P, Papadopoulos G, Reczko M, Dalamagas T, Giannopoulos G, et al. Accurate microRNA target prediction correlates with protein repression levels. BMC Bioinformatics 2009; 10: 295.

[56] Kumar A, Wong AK, Tizard ML, Moore RJ, Lefèvre C. miRNA_Targets: A database for miRNA target predictions in coding and non-coding regions of mRNAs. Genomics 2012; doi: 10.1016/j.ygeno.2012.08.006.

[57] Saetrom O, Snǿve O, Saetrom P. Weighted sequence motifs as an improved seeding step in microRNA target prediction algorithms. RNA 2005; 11: 995-1003.

[58] Kim SK, Nam JW, Rhee JK, Lee WJ, Zhang BT. miTarget: microRNA target-gene prediction using a support vector machine. BMC Bioinformatics 2006; 7: 411.

[59] Yan X, Chao T, Tu K, Zhang Y, Xie L, Gong Y, et al. Improving the prediction of human microRNA target genes by using ensemble algorithm. FEBS Lett 2007; 581: 1587-93.

[60] Yousef M, Jung S, Kossenkov AV, Showe LC, Showe MK. Naïve Bayes for microRNA target predictions-machine learning for microRNA targets. Bioinformatics 2007; 23: 2987-92.

[61] Wang X, El Naqa IM. Prediction of both conserved and nonconserved microRNA targets in animals. Bioinformatics 2008; 24: 325-32.

[62] Yang Y, Wang YP, Li KB. MiRTif: a support vector machine-based microRNA target interaction filter. BMC Bioinformatics 2008; 9: S4.

[63] Bandyopadhyay S, Mitra R. TargetMiner: microRNA target prediction with systematic identication of tissue-specific negative examples. Bioinformatics 2009; 25: 2625-31.

[64] Chandra V, Girijadevi R, Nair AS, Pillai SS, Pillai RM. MTar: a computational microRNA target prediction architecture for human transcriptome. BMC Bioinformatics 2010; 11: S2.

[65] Sturm M, Hackenberg M, Langenberger D, Frishman D. TargetSpy: a supervised machine learning approach for microRNA target prediction. BMC Bioinformatics 2010; 11: 292.

[66] Betel D, Koppal A, Agius P, Sander C, Leslie C. Comprehensive modelling of microRNA targets predicts functional non-conserved and non-canonical sites. Genome Biol 2010; 11: R90.

[67] Friedman Y, Naamati G, Linial M. MiRror: a combinatorial analysis web tool for ensembles of microRNAs and their targets. Bioinformatics 2010; 26: 1920-1.

[68] Herrera PR, Ficarra E, Acquaviva A, Macii E. miREE: miRNA recognition elements ensemble. BMC Bioinformatics 2011; 12: 454.

Figure legends

Figure 1 MicroRNA research timeline

Shown in the graph is the number of microRNA mature sequences deposited in miRbase (white), PubMed-reported publications on microRNAs (gray) and microRNA targets (black), respectively.

Figure 2 Experimental techniques used to obtain data deposited in public databases

Experimental techniques used to derive human microRNA-mRNA interactions in Tarbase v6–2011 (**A**), miRecords v3–2010 (**B**) and miRtarbase (**C**), respectively, were shown as pie chart. **D.** Distribution of miRNA-target interaction (MTI) types found in miRtarbase.

Supplementary material

Table S1 Summary for microRNA-target prediction algorithms
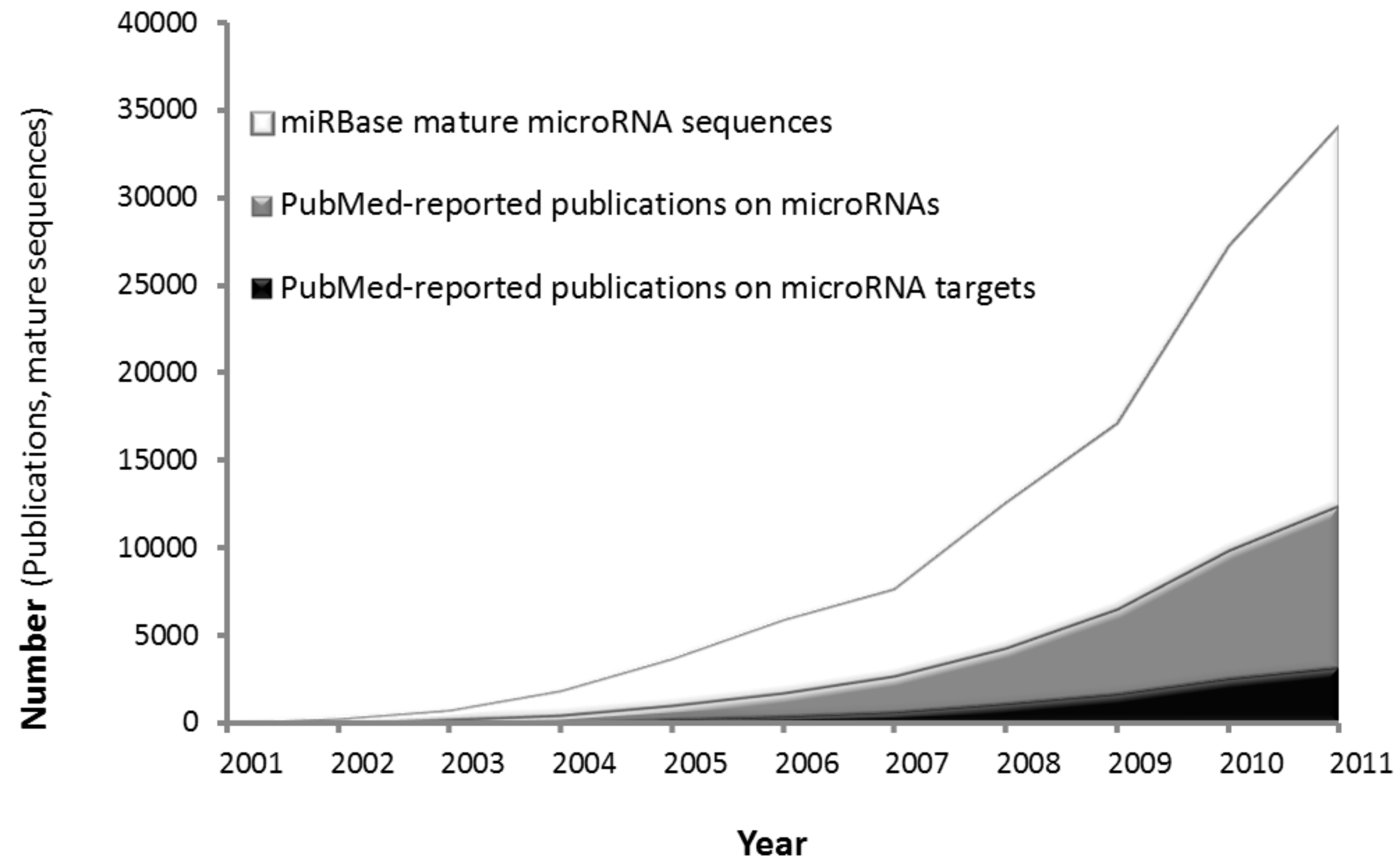
Figure 1

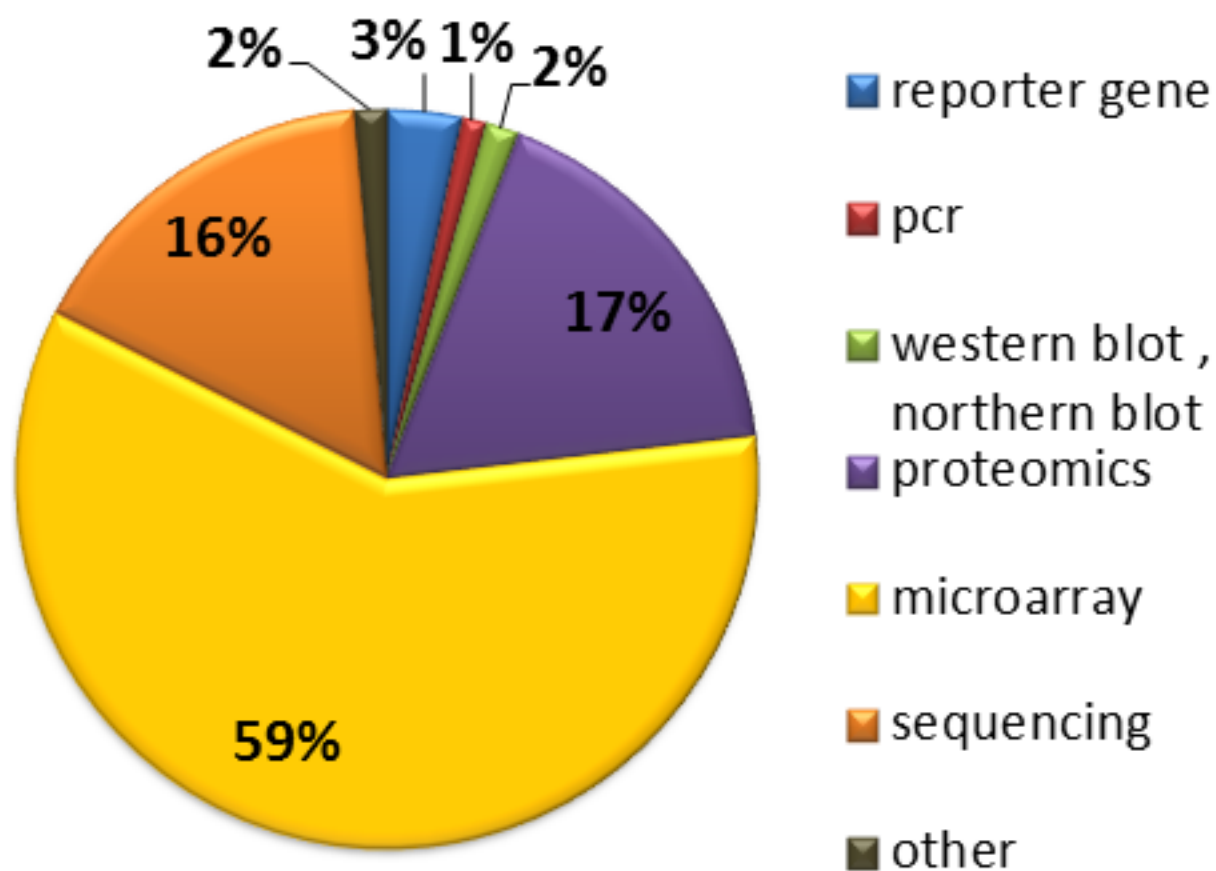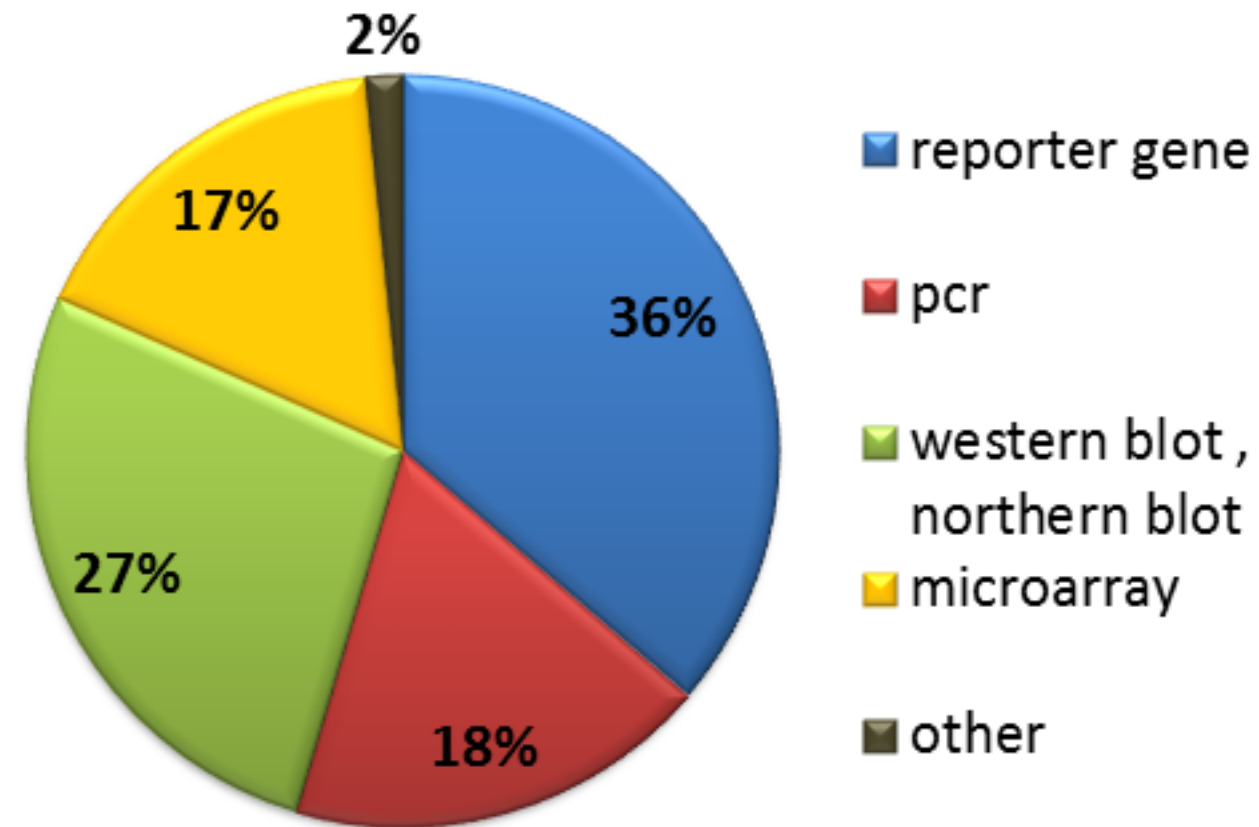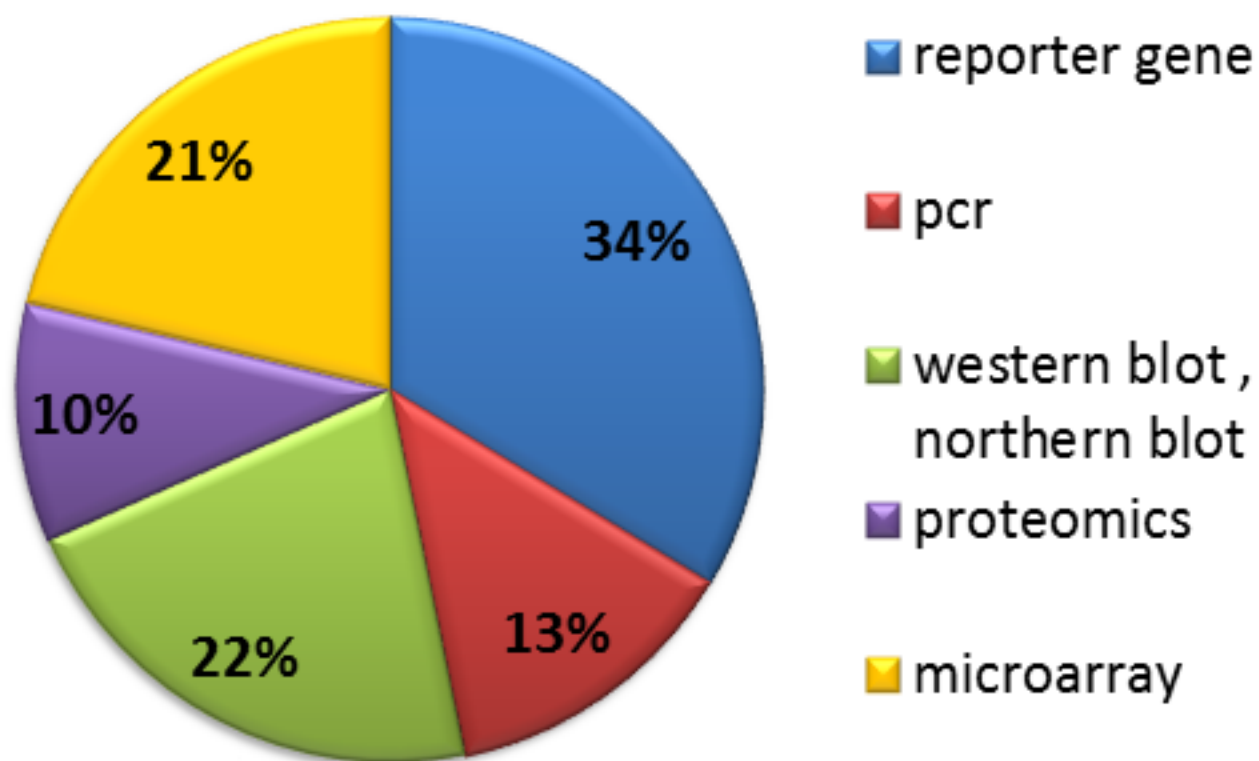# microRNA Research timeline

Figure 2

A Tarbase v6 (2011)

- reporter gene
- pcr
- western blot, northern blot
- proteomics
- microarray
- sequencing
- other

2% 3% 1% 2%
16%
17%
59%

B miRecords v3 (2010)

- reporter gene
- pcr
- western blot, northern blot
- microarray
- other

2%
17%
36%
27%
18%

C miRtarbase

- reporter gene
- pcr
- western blot, northern blot
- proteomics
- microarray

21%
34%
10%
22%
13%

D MTI in miRtarbase

- Functional MTI
- Functional MTI (Weak)
- Non-Functional MTI
- Non-Functional MTI (Weak)

1%
9%
50%
40%
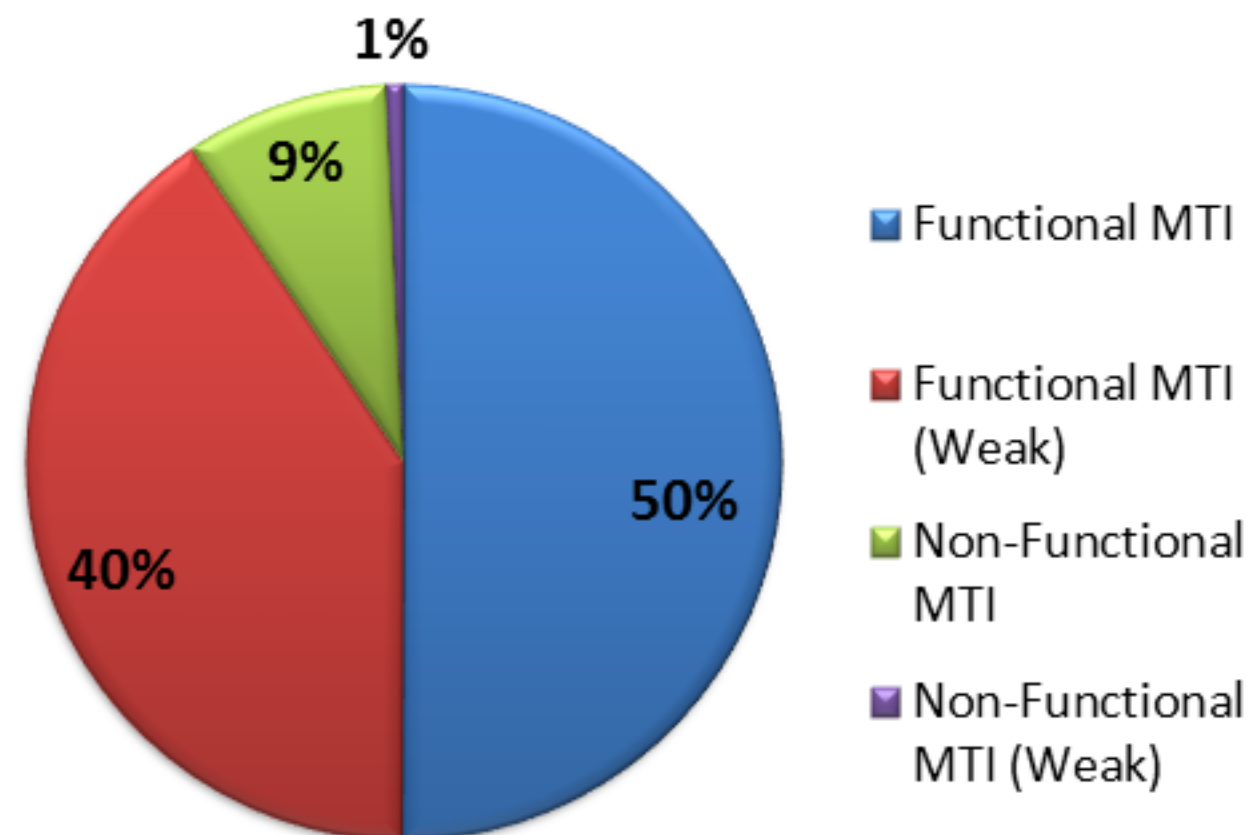
Table S1

Table S1 **Summary for microRNA-target prediction algorithms**

Approach	Algorithm name	Considerations	Website	Reference
Ab initio	miRanda	FEATURES: It strongly considers features like seed complementarity and duplex minimum free energy. The most recent version uses a conservation measure (PhastCons).	http://www.microrna.org	[47] [48]
	TargetScan	FEATURES: It strongly considers seed complementarity, noting a hierarchy. It ranks the predictions using a context score (seed complementarity, conservation and AU content). In the recent release, seed-pairing stability and target-site abundance were integrated.	http://www.targetscan.org/	[35] [37] [49] [50]
	PicTar	FEATURES: It has strict requirements regarding seed complementarity, duplex minimum free energy and conservation. NOTE: Predictions are available but the algorithm is not available	http://pictar.mdc-berlin.de/	[51]
	RNA22	FEATURES: It identifies the most statistically-significant patterns in order to recognize the candidate binding sites. NOTE: The user can add restrictions (minimum number of base pairs, maximum number of unpaired bases and maximum allowed free energy)	http://cbcsrv.watson.ibm.com/rna22.html	[52]
	RNAHybrid	FEATURES: This algorithm finds the minimum free energy for the entire microRNA-mRNA duplex. NOTE: The user can add restrictions such as the number unpaired bases and minimum free energy allowed.	http://bibiserv.techfak.uni-bielefeld.de/rnahybrid/	[53]
	PITA	FEATURES: This is the first algorithm that takes into account accessibility to the site in the mRNA ($\Delta\Delta G$). NOTE: The user can impose different restrictions (minimum seed size,	http://genie.weizmann.ac.il/pubs/mir07/	[44]

G:U bobbles and unpaired bases)

	EIMMO	FEATURES: It uses a conservation score derived from a phylogenetic distribution for the functional sites.	http://www.mirz.unibas.ch/ElMMo2/	[54]
	DIANA	FEATURES: This algorithm correlates structural score (sum of site scores considering duplex minimum free energy and conservation) with protein fold change.	http://diana.cslab.ece.ntua.gr/microT/	[55]
Machine learning	TargetBoost	This is the first machine learning algorithm ever developed for microRNA target prediction. ALGORITHM: Consists of a boosting algorithm that assigns weights to sequence patterns. NEGATIVE SET: Randomly generated sequences. NOTE: Problems with site availability. Predictions are available but the algorithm is not available	http://www.interagon.com/demo/	[57]
	miTarget	ALGORITHM: Consists of an SVM Classifier (RBF kernel). FEATURES: It uses structural characteristics of mREs. NEGATIVE SET: Uses inferred data from mRNAs where the interaction site was deleted and as result no repression was observed.	http://cbit.snu.ac.kr/~miTarget/introduction.html	[58]
	Ensemble Algorithm	ALGORITHM: Consists of the ensemble of 10 SVM (polynomial kernels). FEATURES: Characteristics from the microRNA-mRE interaction and features from the mRNA target. NEGATIVE SET: Uses only 16 samples with experimental support. NOTE: Problems with site availability.	http://www.biosino.org/~kanghu/mRTP/mRTP.html	[59]
	NbmiRTar	ALGORITHM: It consists of naïve Bayes classifier to filter predictions. FEATURES: It uses structural characteristics of mREs.	http://wotan.wistar.upenn.edu/NBmiRTar/	[60]

		NEGATIVE SET: Uses generated sites for artificial microRNA sequences.		
		NOTE: Problems with site availability.		
	miRTarget2	ALGORITHM: Consists of an SVM classifier.	Predictions available at: http://mirdb.org	[61]
		FEATURES: Characteristics from the mRE and from the mRNA.		
		NEGATIVE SET: Uses 1017 negative interactions with experimental support.		
	miRTif	ALGORITHM: It consists of an SVM classifier (RBF kernel).	http://mirtif.bii.a-star.edu.sg/	[62]
		FEATURES: It uses features from the microRNA-mRE interaction.		
		NEGATIVE SET: Contains 17 inferred interactions from experimental data.		
	miRror	FEATURES: This algorithm is based on the notion of microRNA combinatorial mode of action.	http://www.proto.cs.huji.ac.il/mirror/search.php	[67]
		This algorithm uses expression-based experimental data.		
Hybrid	TargetMiner	ALGORITHM: The first part selects the target sites based on the seed complementarity. The second part uses an SVM classifier (RBF kernel).	http://www.isical.ac.in/~bioinfo_miu/targetminer20.html	[63]
		FEATURES: Characteristics from the mRNA and from the microRNA-mRE duplex.		
		NEGATIVE SET: Consists of 59 experimental samples plus 289 inferred negatives from expression-based experimental data.		
	mTar	ALGORITHM: The first part selects 3 classes of sites (5' seed only, 5' dominant and 3' canonical). The second part, uses an artificial neural network to classify.	http://www.rgcb.res.in/downloads/Mtar.rar	[64]
		FEATURES: This algorithm uses features from the microRNA-mRE interaction.		
		NEGATIVE SET: Contains 400 samples with experimental support		

TargetSpy	ALGORITHM: This algorithm uses generated candidate zones. It uses a multiboost classifier.	http://www.targetspy.org/	[65]
	FEATURES: Compositional, structural and base-pairing features.		
	NEGATIVE SET: Consists of 4540 sites derived using CLIP-seq protocols.		
mirSVR	ALGORITHM: The first part uses miRanda (same developers). The second part uses an SVR trained on expression-based experimental data.	http://www.microrna.org	[66]
	NOTE: Predictions are available but the site is not available.		
miREE	ALGORITHM: The first part uses a genetic algorithm to generate optimal candidate sites. The second part uses an SVM (RBF Kernel).	http://didattica-online.polito.it/eda/miREE/	[68]
	FEATURES: Characteristics from the mRNA:mRE interaction and accessibility features.		
	NEGATIVE SET: Contains 46 negative records and 305 inferred negatives from data with experimental support.		

Note: Only algorithms that developed both parts (*ab initio* part and machine learning part) were considered as hybrid approaches.