Computational and Structural Biotechnology Journal 19 (2021) 5811–5825

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Review

Prediction methods for microRNA targets in bilaterian animals: Toward a better understanding by biologists

Aurélien Quillet [a], Youssef Anouar [a], Thierry Lecroq [b], Christophe Dubessy [a,c,*]

[a] Normandie Université, UNIROUEN, INSERM, Laboratoire Différenciation et Communication Neuronale et Neuroendocrine, 76000 Rouen, France
[b] Normandie Université, UNIROUEN, UNIHAVRE, INSA Rouen, Laboratoire d'Informatique du Traitement de l'Information et des Systèmes, 76000 Rouen, France
[c] Normandie Université, UNIROUEN, INSERM, PRIMACEN, 76000 Rouen, France

ARTICLE INFO

Article history:
Received 16 May 2021
Received in revised form 20 September 2021
Accepted 15 October 2021
Available online 18 October 2021

Keywords:
MicroRNA
Target prediction
Bioinformatics tools
Computational prediction methods
Data combination
Performance evaluation

ABSTRACT

MicroRNAs (miRNAs) are small noncoding RNAs that regulate gene expression at the posttranscriptional level. Because of their wide network of interactions, miRNAs have become the focus of many studies over the past decade, particularly in animal species. To streamline the number of potential wet lab experiments, the use of miRNA target prediction tools is currently the first step undertaken. However, the predictions made may vary considerably depending on the tool used, which is mostly due to the complex and still not fully understood mechanism of action of miRNAs. The discrepancies complicate the choice of the tool for miRNA target prediction. To provide a comprehensive view of this issue, we highlight in this review the main characteristics of miRNA-target interactions in bilaterian animals, describe the prediction models currently used, and provide some insights for the evaluation of predictor performance.

Contents

1. Introduction

MicroRNAs (miRNAs) are small (~22 nucleotides) noncoding RNAs that act as posttranscriptional regulators of gene expression for all known biological processes [1]. Indeed, between 60 and 90% of human genes are believed to be regulated by miRNAs, as revealed by genome-wide analyses [2,3]. According to miRbase (release 22), the primary database of published miRNA sequences and their annotation, a total of 48,885 mature miRNA products have been identified in 271 eukaryotic species, among which 2654 are found in humans [4]. Animal and plant miRNAs share many similarities in their biogenesis and mode of action, as revealed by

* Corresponding author at: Laboratoire Différenciation et Communication Neuronale et Neuroendocrine (DC2N), Inserm U1239, Normandie Université, Université de Rouen Normandie, Place E. Blondel, 76821 Mont-Saint-Aignan Cedex, France.
E-mail address: christophe.dubessy@univ-rouen.fr (C. Dubessy).

https://doi.org/10.1016/j.csbj.2021.10.025

biochemical and genetic studies, suggesting that they share a common ancestral origin. However, the sites of biogenesis of miRNAs, their genetic structure or the location of their genes differ between plants and animals [5–8].

In bilaterian animals, miRNAs are mostly transcribed by RNA polymerase II, which yields a primary miRNA (pri-miR). This pri-miR is then processed to generate a miRNA precursor (pre-miR) by the microprocessor complex composed of DROSHA and DGCR8 (known as Pasha in flies and nematodes). Then, exportin-5 and RAN-GTP transfer the pre-miR from the nucleus to the cytoplasm to be further processed by DICER and produce the mature-miRNA duplex sequence. The biogenesis of miRNAs in animals has been reviewed in several publications [1,9–14]. The miRNA inhibition process requires the formation of miRNA-induced silencing complexes (miRISCs), which are mainly composed of the Argonaute (AGO) family of proteins and several other proteins, such as the trinucleotide repeat containing 6 (TNRC6, known as GW182 in flies) family of proteins [14–16]. These proteins are mainly localized in cytoplasmic P-bodies, which are considered the primary sites of miRNA activity in the cytoplasm, although they can also occur in many cellular compartments, such as the nucleus, mitochondria or vesicles of the endosomal trafficking pathway [17]. In most cases, miRISC induces silencing through a combination of processes, including translational repression, deadenylation, decapping and 5′-to-3′ mRNA degradation [18,19]; however, mRNA decay is believed to be responsible for 66–90% of silencing [20,21]. Interestingly, plant miRNAs regulate their targets mainly by binding with nearly full complementarity to unique sites in the coding region. This high pairing rate mostly leads to endonucleolytic mRNA cleavage and a strong effect on a limited number of targets [6]. In contrast, miRNAs from bilaterian animals regulate transcripts via imperfect complementarity at multiple interaction sites mainly located in the 3′-UTR, which allows them to potentially regulate several hundred mRNAs, and one mRNA can be targeted by several miRNAs [13,22–25]. Because of these numerous possible interactions, miRNAs exert major effects in a variety of cellular processes, including cell proliferation, migration, apoptosis and differentiation [26–28]. Consequently, altered expression of miRNAs has been observed in many pathologies [29], including cardiovascular [30], neurodegenerative [31], and renal diseases [32], and most notably in cancers [33–35]. Therefore, improved knowledge of the mechanisms of action of miRNAs will likely impact our understanding and management of these diseases.

Because miRNAs are now considered major actors among non-coding RNAs for the regulation of gene expression, their role in this important cellular mechanism has been an expanding area of research since 2001. This is particularly challenging in bilaterian animals due to the imperfect interaction between miRNAs and their target mRNAs and the resulting large number of potential targets. To understand this role, it is essential to identify functional miRNA targets in a predefined cellular and environmental context. This goal could be achieved through the use of a combination of cell biology techniques, including gene reporter assays, quantitative PCR and western blot [36]. While a Luciferase gene reporter test can identify the direct interaction between a miRNA and its targeted mRNA region, qPCR and western blot assess the transcriptional and translational repression resulting from the interaction [36,37]. These techniques are time-consuming and allow validation of a few interactions at a time. To address this issue, cross-linking and immunoprecipitation approaches coupled with next-generation sequencing (CLIP-seq) have been developed. These techniques allow massive discovery of miRNA target interactions (MTIs) without the need for miRNA overexpression. However, the identified interactions still need additional investigations to decipher their biological meaning [36,38]. Although improvements have been made, many datasets generated by this approach contain numerous false-positives due to UV crosslinking issues [39]. Regardless of the experimental procedures, they are time-consuming and expensive; thus, in silico MTI predictions are required. Predictions of novel target sites could be achieved by building a classification or ranking model based on experimentally validated MTI properties (further described below). During the last decade, scientists have proposed many different computational approaches, although a consensus has not been reached on how to best predict MTIs. Currently, more than 192 target prediction tools have been described (as of November 2020, from OMICtools' database) [40]; therefore, it is difficult to find the best suited tool for the analysis of a particular experiment. This issue has been the subject of several reviews that discuss common prediction tools as well as the main characteristics of MTIs [41–44]. Recently, Kern et al. proposed a dedicated tool that would facilitate the choice of the most appropriate prediction tool [45]. However, computational predictions present high false-positive/negative rates due to the small size and the binding complexity of the MTI sites [46]. Moreover, without a common method to evaluate them, it is not easy to decide which one to test first. Indeed, the result lists given by each MTI prediction algorithm for a given miRNA differ greatly in the targets identified, prediction number and ranking [47]. Below, we will describe the main characteristics of MTIs in bilaterian animals as well as different up-to-date computational methods that could help biologists choose the appropriate tool, and we will provide the knowledge necessary to avoid the numerous drawbacks of these prediction tools. The issue of algorithm performance evaluation will also be addressed.

2. Analyzable elements

Although the mechanisms of action of miRNAs are not fully understood, several features of MTI have been defined through experimental work. Although each algorithm uses a different set of features, sequence complementarity, site accessibility and sequence conservation are the most commonly used.

2.1. Sequence features

2.1.1. Seed region. The main biological feature underlying the interaction between miRNA and mRNA is defined as the "seed" region, which includes nucleotides (nt) 2 to 8 starting from the 5' end of a

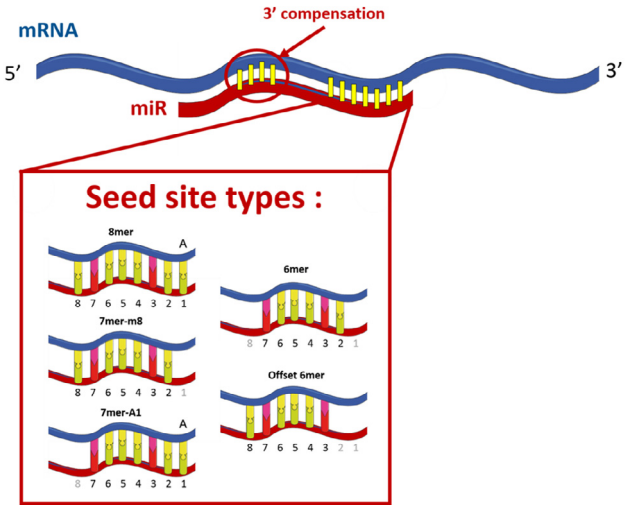

Fig. 1. microRNA seed site types. The vast majority of miRNA interactions occur through several matching possibilities of the seed region as described above. Mismatches in the seed region can still result in a functional interaction with the help of 3′ compensatory pairing.

miRNA. A perfect match with the seed region does not always induce mRNA repression, clearly indicating that this parameter alone is not sufficient to predict the interaction [48–50]. Interestingly, the recognition of an adenine at miRNA nt 1 favors miRNA-mediated protein downregulation even when it does not participate in a Watson-Crick interaction [51]. Seed sites are categorized into different types according to their pairing degree. The hierarchy of site efficacy is as follows: 8mer ≫ 7mer-m8 greater than 7mer-A1 ≫ 6mer or offset-6mer (position 3–8 match) > no site, with the 6mer differing only slightly from no site at all (Fig. 1) [2,50]. Microarray experiments suggest that the majority of miRNA target sites are 7mer-m8 type [50]. The difficulty of using the seed region in target prediction is based on the occurrence of "bulges" (unpaired stretches of nucleotides located in either one of the sequences) or G:U wobbles within the sequence that reduce (but do not prevent) inhibition efficiency [48,51]. These sites are named "orphans" or "noncanonical" because AGO proteins can bind them without a perfect seed match. They were thought to be relatively rare in mammals [2,52–54]. However, more recent experimental methods tend to identify a much higher number of noncanonical sites or even sites not binding to the seed region at all (binding to the center of the miRNA or 3' end) [51,52,55–58]. A possible explanation for some of these noncanonical sites is the existence of a "pivot bulge" on the 6th nt of the seed that could enable a transitional nucleation state by stabilizing nucleation base pairing (positions 2–6), allowing subsequent bulge formation and propagation of the seed interaction [53,59]. An alternative hypothesis is that noncanonical sites, since they are poorly conserved across species, may act as evolutionary intermediates between nonfunctional sites and canonical target sites with selection pressure going toward the appearance of higher affinity sites [54]. In any case, functional assays indicate a mild regulatory effect of these noncanonical sites [50,53,57]. Therefore, the usefulness of considering both fully and partially matching seed sites to improve MTI prediction is still a matter of debate [60,61].

2.1.2. Compensation. While most studies consider a "canonical" site to be a full seed pairing without a bulge, miRNA target sites can in fact be divided into three groups: canonical (or seed only), atypical canonical and noncanonical sites [14]. Canonical sites, which were described in the previous paragraph, have strong 5' pairing but require little or no 3' pairing. Atypical canonical sites have both strong seed pairing and supplementary pairing on the 3' side of the miRNA. Finally, noncanonical sites have weak seed pairing and strong 3' pairing. One might think that atypical canonical sites are more effective than seed sites only. However, evaluating the effectiveness of 3' supplementary pairing is very difficult due to the number of pairing possibilities and the dependence on context of this parameter [14]. Nevertheless, additional Watson-Crick pairings of at least 4 nt at positions 12–17, especially from 13 to 16, enhance miRNA targeting [50]. This type of strong compensation is very rare (less than 2% of known conserved MTIs), although when it exists, its target site is usually highly conserved across species [2].

2.2. Site accessibility. The complexity of miRNA-mRNA interactions leads to the rather weak efficiency of algorithms based on sequence matching only. Additional parameters, such as thermodynamics, UTR context or site conservation, must be considered. Site accessibility is as important as individual nucleotide matches in the seed since the action of a miRNA is mediated by a relatively large silencing complex.

2.2.1. Thermodynamic effect. The most basic approach for considering the thermodynamic effect is to calculate the free energy, which reflects the stability of the RNA binding sequences. This binding is believed to form a stable, low-energy duplex. Therefore, lower energy values indicate a more plausible interaction. Since we are in the context of miRNA interactions, constraints imposed by seed pairing must be taken into consideration. The ViennaRNA R package is the most commonly used tool to calculate the free energy of binding. It aggregates more than 20 programs/packages to solve the structure of an RNA duplex using dynamic programming [62]. Rehmsmeier *et al.* found that forbidding intramolecular base pairing and bulge loops seems to give a better free energy estimation [63]. They also noted that taking several nt (10 and more) flanking the target site improves the correlation between energy-based scores and target repression [63,64]. Another possibility is to consider the hybridization energy (ΔΔG), which is the difference between the free energy gained by the binding of the miRNA to the target, ΔG_{duplex}, and the free energy lost by unpairing the target-site nucleotides, ΔG_{open}. This ΔΔG score correlates well with the degree of miRNA target repression for some interactions but not all [64].

2.2.2. Target site context. Messenger RNAs can fold into highly elaborated secondary and tertiary structures, and a perfect miRNA sequence match might not be structurally accessible for binding. Therefore, contextual features, such as the local AU nucleotide composition, proximity to residues that can pair to miRNA nucleotides 13–16, or positioning away from the center of long UTRs, must be included in MTI prediction algorithms. Among all contextual features, the AU content around the target site favors most of the interactions with a miRNA [22]. Indeed, swapping a target site from an open (AU rich) UTR structure to a close structure decreases the site functions [64]. A possible explanation for this phenomenon is that AU-rich sequences could be recognized directly by a RISC component or may reduce the tendency to form stable RNA secondary structures that could interfere with RISC binding [65]. Although there is a high prevalence of MTI sites found in the 3′-UTR, some miRNAs can also regulate mRNAs by binding to the 5′-UTR and the coding sequence (CDS) region of their targets [66,67]. However, target sites in the open reading frame are not as efficient as the other sites [51,68,69]. Interestingly, a recent study showed that the sites in the CDS are quite potent at inhibiting translation by inducing transient ribosome stalling instead of mRNA destabilization [70,71]. Interestingly, some studies have shown that miRNA interactions with different binding sites and/or under different cellular conditions can increase mRNA translation [72–74]. However, the precise mechanism by which a miRNA can enhance protein synthesis remains to be elucidated. Thus, it is important not to restrict the search for MTI predictions to the 3′-UTR. Aside from localization, the number of repetitions of a target site and their spacing on a given mRNA also affect the inhibition efficiency of a miRNA [65,75]. Another important aspect to determine the possibility of an interaction, yet rarely taken into consideration, is the expression level of both miRNAs and targeted mRNAs [76]. Moreover, depending on the tissue or disease, a validated MTI can be more or less functional [77,78]. This might be due to RNA-binding proteins that could block access to miRNA or mRNA secondary structures in that particular tissue or disease [78,79]. Conversely, certain RNA-binding proteins, such as PUM1 and Sfpq, have been shown to promote miRNA targeting [79,80]. These RNA-binding proteins in each tissue or disease of interest must be considered to improve the predictions of MTIs.

2.3. Conservation. The level of conservation of a sequence corresponds to its presence across different species. The use of the evolutionary conservation of miRNA targets is motivated by the idea that closely related species should share common MTI sites. However, most target sites are not fully conserved over their entire length, with higher conservation often occurring in the seed region than in the other sequences of the target site. Moreover, only the percentage of 3′-pairing is generally conserved and not the nucleo-

tides themselves. Assuming that aligned sites within orthologous genes have a common origin, it was proposed to quantify site conservation in a phylogenetic tree by summing the length of all branches in which the site is present [81].

Of note, the level of conservation of a target site has to be estimated with regard to the conservation of its mRNA region and its length [2]. A stronger conservation profile has been associated with increased mRNA downregulation as assessed by microarray experiments and better MTI prediction [2,51,65,82,83]. Indeed, over 60% of human protein-coding genes have conserved targets for miRNAs, thus supporting the importance of this parameter [2]. However, since functional nonconserved MTIs exist and mediate protein translation inhibition [84], target sites cannot be filtered based on conservation criteria only. Agarwal *et al.* also observed a decrease in the performance of their predictor when considering only highly conserved sites [60]. Therefore, an ideal equilibrium must be found where conserved sites are favored as well as where nonconserved sites are also retained. Friedman *et al.* reported a high number of preferentially conserved 6mer sites [2], a surprising finding since, as mentioned above, 6mer sites typically have poor efficacy when examined experimentally [50]. A possible explanation for this result is that these sites are inactive (or less active) forms of conserved 7–8mer sites. An alternative explanation is that when binding with a 6mer, miRNA induces a function other than repressing protein output. For example, a role in mRNA subcellular localization could allow many 6mer sites to be conserved while having a poor effect on protein level inhibition [2].

3. Computational prediction methods

As mentioned in the introductive part of this review, many computational tools have been developed in the field of MTI prediction. The main objective of prediction algorithms is to select the most discriminative features within the categories of analyzable elements described above and to determine the best way to compute them to obtain the most accurate prediction.

3.1. Sequence based.

3.1.1. Heuristic scoring models. The earliest attempt to identify miRNA targets in silico was published by Stark *et al.* in 2003 [85]. The screening performed in this study was a simple two-step procedure combining sequence comparison with HMMer (alignment tool) and site accessibility using Mfold. The resulting targeted 3′-UTRs were then compared based on their conservation between *Drosophila pseudoobscura* and *Anopheles gambiae*, and they successfully validated 6 MTIs for two *Drosophila* miRNAs that they predicted using this protocol. After analyzing the characteristics of these 6 validated interactions, they described what we now know as the seed region: nucleotides 2 to 7 at the 5′ end of miRNA [85].

Following this initial report, many more studies have been performed with the aim of improving and generalizing MTI prediction. The vast majority of the described predictors utilize the seed-matching parameter since most of the reported functional MTIs have a 6mer or more. To determine this parameter, predictors either filter sequences based on a defined set of rules for seed matching [60,86] or use a score system that favors this feature [24,87,88]. However, filtering based on seed rules seems too stringent because functional MTIs can also have noncanonical seed sequences (G:U wobble or bulge). In this regard, some methods consider the binding of the first eight nucleotides as important but do not restrict it to particular seed types [89–92]. MIRZA-G (evolution of MIRZA [93]), for instance, is a recently published algorithm that allows for nonperfect seed matches if the final score for the site is above the author-defined threshold [83]. Predictors such as RNA22 [3] that do not consider seed matching at all in their predictions are rather rare. In the case of RNA22, the algorithm probes mRNA for patterns generated by comparing all known mature miRNA sequences (as of 2006) and keeps only the most similar ones. Sequence alignment results are almost always complemented with site accessibility and evolutionary inputs. Tools such as miRanda [88], RNA22 [3] and TargetScan [60,94] make use of RNA folding prediction software, such as RNAVienna [62] or Mfold [95] packages, to estimate the free energy of predicted miRNA–target duplexes and filter out the candidates above a certain threshold. Interestingly, the authors of RNAhybrid [63] used a different approach that avoids intramolecular base pairing and bulge loops, which seems to improve the estimation of the free energy [63]. In fact, some predictors, such as PicTar [24] and STarMir [96,97], use the results of RNAhybrid to filter potential target sites. As mentioned before, the authors of other predictors, such as PITA [64], prefer to consider the hybridization energy (see "Thermodynamic": II.B.1) to score miRNA–target duplex stability. Out of all the site accessibility features, the local AU content is the most implemented since it has been shown to favor MTI [60,89,90,92,98,99]. The frequency of target sites along the mRNA and the distance separating them are two other features often considered for target site context implementation [21,94,100]. The value of site conservation is frequently advocated since omitting nonconserved targets and not using this parameter drastically decrease the specificity of the method [2,90,94,101]. This parameter has been extensively analyzed by the authors of EIMMo [86], who scored MTIs based on conservation criteria only and then used Bayesian statistics to infer functionality. Therefore, EIMMo is quite efficient at predicting the mRNAs targeted by a given miRNA but not as sensitive at the target site level [102]. Feature implementation for all the algorithms cited thus far has been performed based on literature data only. To better identify the combination of features to use, the authors of miRmap decided to evaluate each of them individually before integrating them. They first screened all human transcripts for 7mer seeds and compared the performance of eleven features mentioned previously on the results from seven miRNA overexpression experiments obtained in five different studies. Based on this evaluation, they combined these features using a linear regression model, thus making it the most comprehensive MTI predictor at that time [98]. Similarly, TargetScan evaluates 26 features and eventually selects 14 to upgrade itself using a similar model in 2015 [60]. Most algorithms store the identified interactions in a publicly available database format, such as miRWalk2.0 [60,103].

3.1.2. Empirical machine learning models. The limit of rule-based methods comes from the complexity of MTIs. It is extremely difficult to take into consideration all possible aspects of these interactions. Thus, another promising direction toward better MTI prediction is data-driven (or machine learning, ML) algorithms. There are many computational models available to build such an algorithm. Unfortunately, there is no fixed rule to select one for a given problem. In general, ML methods are categorized into two groups depending on whether the output values are present in the training data (supervised learning) or not (unsupervised learning). In the field of MTI prediction, all data-driven methods use supervised learning regression (scoring system) or classifiers (categories) to differentiate functional from nonfunctional sites. The performance of each method depends on the amount and quality of the training data, the complexity of the relationship between the inputs and outputs, and the local computational restrictions (time and memory). Computational constraints depend mostly on the number of features used [104]. Since an ML approach can only be as effective as the dataset used to train it, a large high-quality dataset is therefore primordial to build an accurate model. An ideal experimental dataset would contain all types of functional MTI and as many negative experimental examples, and it would also be free

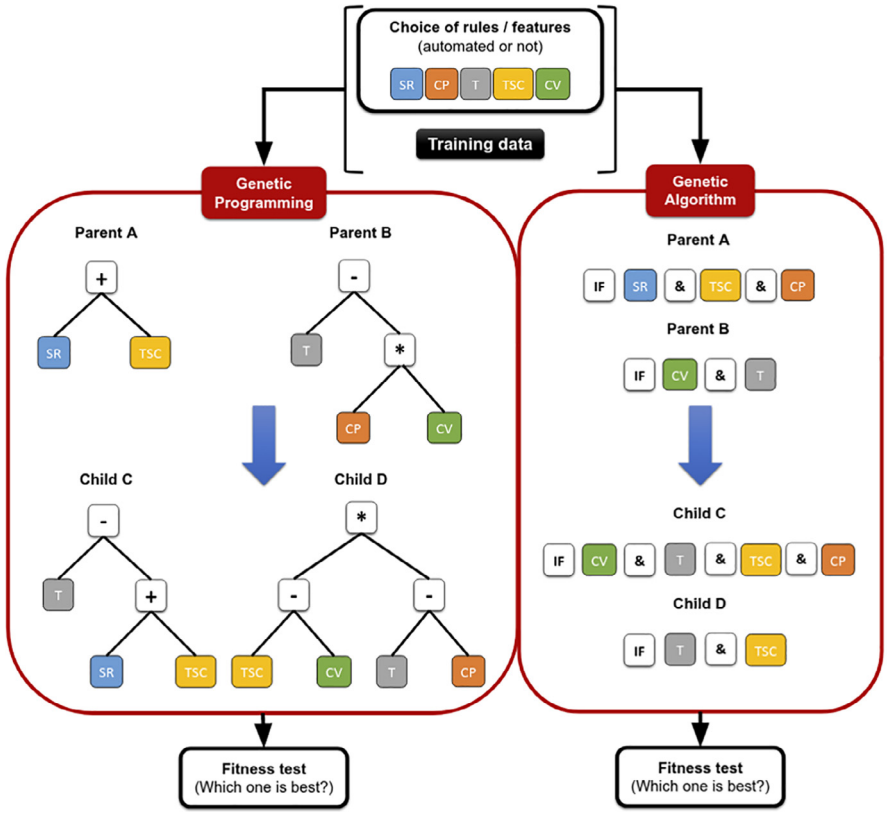

Fig. 2. Basic schematics for the genetic programming (GP) and genetic algorithm (GA). Using seed region (SR) pairing, compensation (CP) pairing, thermodynamic (T) pairing, target site context (TSC) pairing and conservation (CV) pairing on the training data, both the GP and GA will create subtree crossovers of parents A and B to form offspring C and D. A fitness test is performed for each tree (parents and offspring) to decide which one is best suited for the classification of the training data.

from any experimental biases. Since the precise mechanism of miRNA binding is not yet completely known, the aim of a data-driven algorithm is to find the best compromise of features to obtain a generalization model [105] capable of classifying an MTI in a binary fashion or according to a scoring method. Features are ranked by a metric system such as F-score (harmonic mean between precision and recall) or correlation coupled with statistics, and the top-ranked features are selected to build the algorithm. This procedure is known as feature extraction. To validate their approaches, most authors use a k-fold cross validation technique. In other words, a subset of the dataset is used for training the algorithm and the other part is used for testing it. This process is performed in general 10 times using different partitions of the original dataset, and the performance results are averaged over the rounds.

3.1.2.1. Genetic programming. Genetic programming is an ML method that generates functions (represented as trees) using the different rules or features implemented to best describe a positive interaction [106,107] (Fig. 2). One of the first ML models developed with this method was TargetBoost in 2005 [107]. This model is one of the rare types of algorithms that does not use the seed matching criteria to predict MTIs. Instead, TargetBoost creates sequence motifs from a set of 36 experimentally validated MTIs (from the literature) and 3,000 random strings of 30 nt as negative examples. These motifs are then weighted with a boosting algorithm that eventually returns a score indicating the probability of interaction. Boosting algorithms combine a set of simple rules (or features) by assigning to each one of them a weight. The idea is to form a single model with better performance than each rule taken individually [106]. The final score is calculated by summing the number of true and false-positive/negative hits and the relative weights given by the algorithm for each sequence. Feature extraction is not performed, and conservation or site density filters are not applied in this model. The data from 3 miRNAs were used to train the model, which was tested on the data of another miRNA using the "leave one out" method. Compared with RNAhybrid and another algorithm named nucleus, TargetBoost was either as good as or more performant depending on the dataset used for testing [107]. To improve the performance of this type of model, a recent study by Rabiee-Ghahfarrokhi *et al.* used a genetic algorithm (Fig. 2) in combination with a C4.5 decision tree instead of boosting [108]. The output of the C4.5 algorithm results in several rule sets that can be taken as inputs for the genetic algorithm. First, their algorithm was trained and tested on a small dataset taken from the TarBase database (version 3.0) and containing 48 positive and 16 negative examples [109–111]. They obtained 94% accuracy using a 10-fold cross-validation method for testing. This performance was confirmed by training and testing the model on a different dataset (taken from Ahmadi *et al.* [112]) containing 113 positive and 312 negative examples and therefore showed 97% accuracy. The authors related the high performances of their method to the set of rules used as inputs. However, in both cases, the training and testing datasets were not independent, thus increasing the likelihood that this algorithm will perform well.

3.1.2.2. Probabilistic based classifier. A commonly used method is to model the relationship between the features and the output categories using probabilities with a naive Bayes (NB) classifier. In other words, this model computes the probability that a feature belongs to a certain class (in our case, positive or negative). An MTI is then classified based on the product of the probabilities of all features (Fig. 3) [104]. NBmiRTar is an example of such a probabilistic machine learning method [113]. Using both 'seed' and 'out-seed' features, the NB classifier was applied to predictions

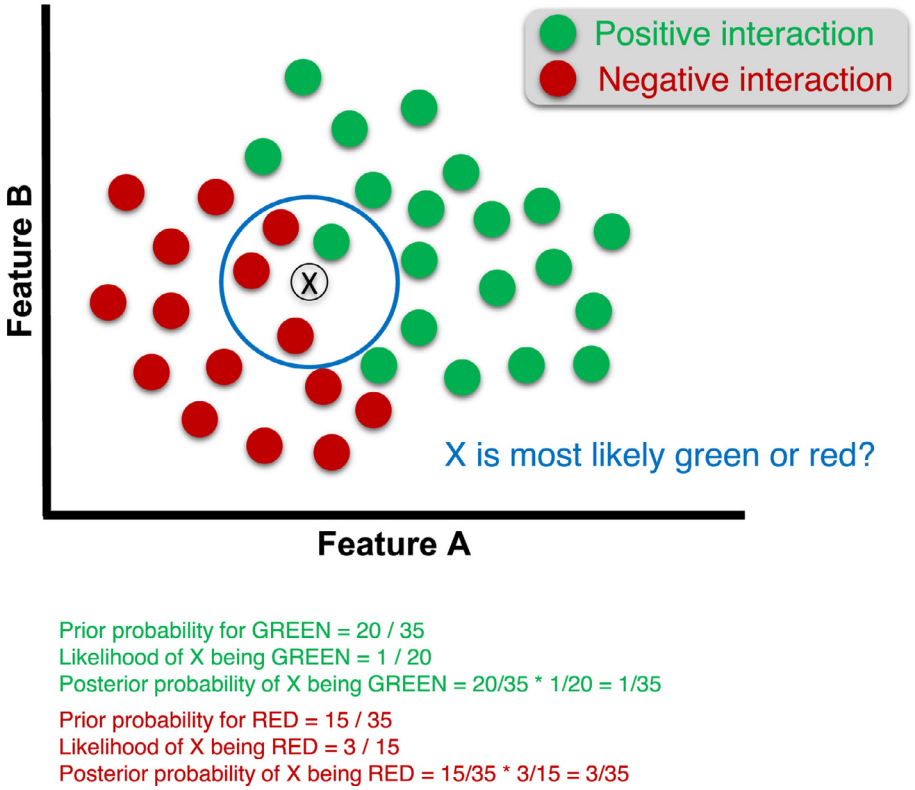

Fig. 3. Naive Bayes classification. The probability that a given interaction is positive or negative is calculated for multiple sets of features. The final decision of the algorithm is the product of all probabilities.

from miRanda, with its scoring and free energy calculation taken as filters. Moreover, the same dataset of 3000 random 30 nt strings were used as negative examples for the TargetBoost method. Interestingly, the two most important features in this model discriminate seed pairing mismatches ("number of bulges in the seed" and "number of bulges in the seed with length 1"). To avoid excluding nonconserved MTIs, the authors did not use sequence conservation in their model, which generates a large number of MTIs. Nevertheless, they claim to be able to reduce this number of MTIs while retaining most of the positive targets (10 out of 13) by using a high score threshold. However, the consistency of this model needs to be tested on more than 13 positive targets. Additionally, using a Bayesian probabilistic method, GenMiR3 [114] (an evolution of GenMiR++ [115]) considers the hybridization energy, target site conservation (PhastCons algorithm [116]) and context information (5 sequence features) to establish a prior probability for the target site to be functional. The authors tested the performance of each feature using multiple linear regression models and cross-validation and found that hybridization energy had the greatest enhancing effect on the predictive power of this model. Expression data for miRNAs and mRNAs were also used to compute a final (or posterior) probability for the site to be functional. Unfortunately, no performance evaluation is available for GenMiR3. Interestingly, although the training data was restricted to colorectal cancer MTIs, the CRCmiRTar authors compared different ML approaches (NB, SVM, random forest (RF), artificial neural network (ANN)) and found that the NB classifier was the most sensitive and specific method [117]. This algorithm also proved to be more efficient than other tools on an independent colorectal cancer-specific test dataset. The tissue origin of the samples therefore seems to be a parameter that should be included in MTI predictions.

Another probabilistic model used for MTI predictions is the random forest (RF) classifier. Each tree of the forest is a predictor that depends on the values and order of a randomly selected subset of features. When an unlabeled example is given to the algorithm, each tree votes, with the majority defining the predicted class for this example (Fig. 4) [118]. The mechanism used to grow the trees allows us to easily estimate the most important set of features and is also easily interpretable. An example of such a model is RFMirTarget [119]. The authors used the dataset published by Bandyopadhyay and Mitra [99] that contains 289 experimentally validated functional pairs and 289 "systematically identified tissue-specific negative examples" to train an RF classifier. Since no site alignment was given in this dataset, they used miRanda to define potential MTI site sequences and alignments. After testing, their model proved to be more efficient on their training set than other types of machine learning methods (support vector machines and NB-based) and was able to identify more positive targets than TargetSpy [120] and miRanda while generating a higher false-positive rate. Using the same training dataset, a multiple instance learning random forest classifier (MIL-RF) called MBSTAR was developed [77]. This model considers potential binding sites as instances and miRNA-mRNA pairs as bags. Thus, a bag can contain several instances. If at least one of the instances is labeled positive, then the bag is labeled as functional. Since the authors of this algorithm deem the secondary structure of the target to be more important than site hybridization, the top features used by MBSTAR are nucleotide patterns in the flanking areas of the potential site and are not seed-related. MBSTAR achieves an accuracy of 78% on a large independent dataset (2nd best is miRanda with 58%). Unfortunately, the authors did not perform a comparison with RFMirTarget, which is the closest related method to MBSTAR. Recently, the authors of TarPmir used data from CLASH

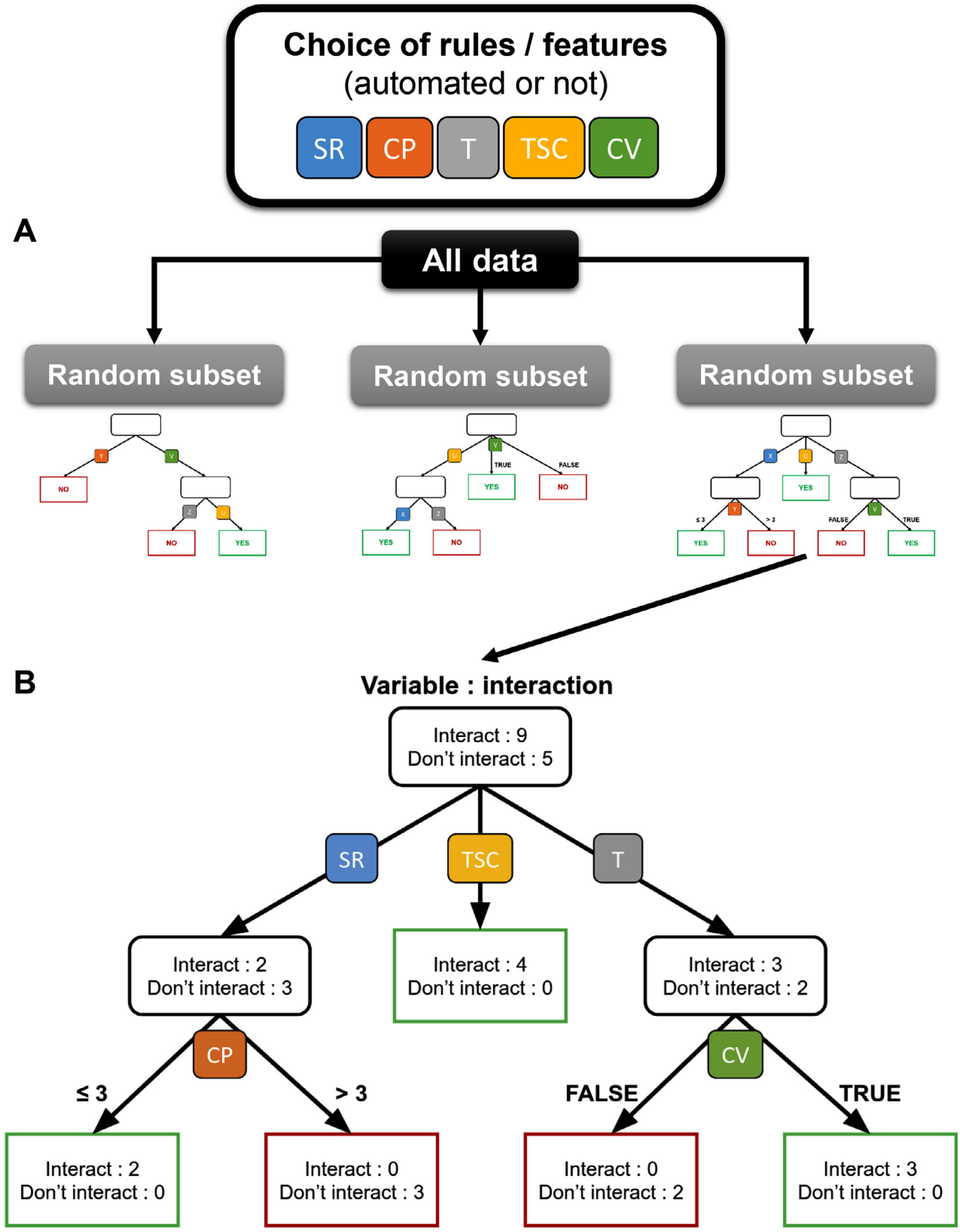

Fig. 4. Random forest (RF) classifier. A) All data are randomly sorted into subsets to generate several trees using a predefined set of rules to optimize the split. In this example, seed region (SR) pairing, compensation (CP) pairing, thermodynamic (T) pairing, target site context (TSC) pairing and conservation (CV) pairing are used. B) This specific tree considers an interaction to occur if it possesses all necessary parameters to fall in one of the green leaves. The RF algorithm returns the prediction made by the majority of the trees. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

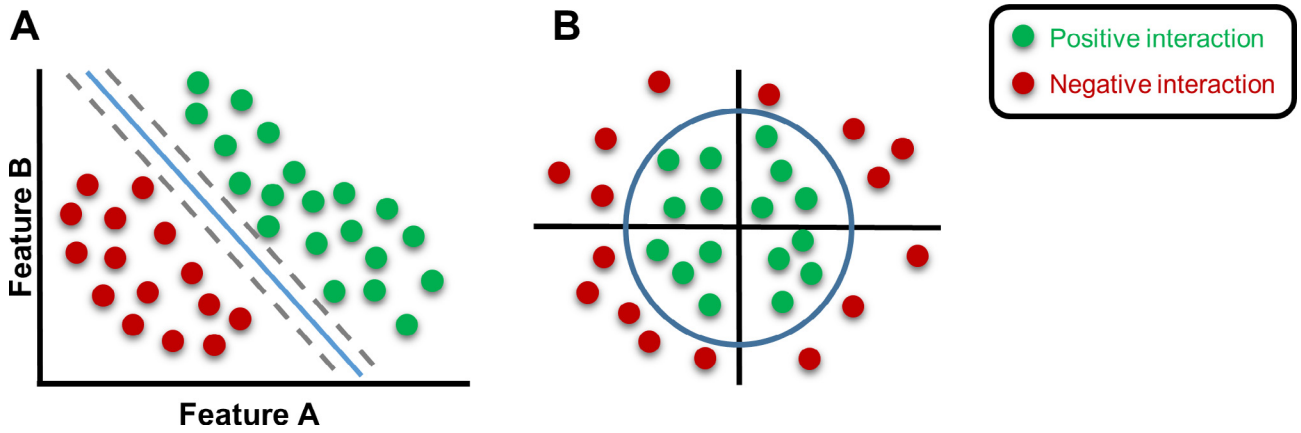

Fig. 5. Nonlinear support vector machine (SVM). A) SVM constructs hyperplanes (gray dotted lines) in a multidimensional space (as many as the number of features being used) that separates cases of different class labels. B) Biological data are rarely separable by straight lines, and a transformation is often used to obtain a nonlinear separation model.

(crosslinking, ligation, and sequencing of miRNA-RNA hybrids), a high-throughput experimental method for identifying MTIs, to train an RF-based model for MTI predictions [121]. The advantage of CLASH compared to CLIP-seq experiments is that it provides both the miRNA and the corresponding target sequences. The training dataset was published by Helwak *et al.* in 2013 and contains 18,534 MTIs for 399 miRNAs [57]. Since no other CLASH datasets were available at the time, the performances of this method were tested on three independent PAR-CLIP datasets. Validated MTIs were identified using DIANA-TarBase (v7.0) [109]. Although TarPmir scored better than three other commonly used algorithms, it only achieved 55% recall and 19% precision, leaving much space for improvement. However, since CLASH data include many "non-seed" MTIs, TarPmir can better predict most sites of this type.

3.1.2.3. Support vector machines. Support vector machines (SVMs) are machine learning algorithms generated to identify the best hyperplanes (linear separation between positive and negative data) while maximizing the margin of error. The training data points that are on the margin hyperplanes are called "support vectors". In the field of biology, however, it is impossible to separate all training data points by a straight line. Thus, some will be located within the margin or on the wrong side of the hyperplane. SVMs are then formulated to soften the impact of these points or use more support vectors. SVMs often use a nonlinear curve to create a decision boundary between data points (Fig. 5) [104]. Most SVMs used for MTI prediction are nonlinear and based on a similarity function called a kernel between pairs of samples (miRNA:mRNA) [89,91,92,99,122]. MiTarget was one of the first algorithms to implement an SVM to predict MTI and showed equal performances to popular predictors, such as miRanda, TargetScan or RNAhybrid [92]. Interestingly, SVMicrO implemented two SVMs, one for the site and one for UTR-related features [89]. Naturally, the most important feature of the site-SVM is seed-based, although conservation of the 3' context region of the interaction was the 2nd best ranked feature. The debate over the use of conservation criteria has been quite active in the field of SVM, with some researchers not using it at all and others showing that it is an important parameter [56,89,92,120,122]. For the UTR-SVM of SVmicrO, predictions result mainly from the number of positive sites in the UTR (the greater the better) and the score of each of these sites (the higher the better) as well as the length of the UTR. At the time of writing this review, SVMicrO showed overall better performance than Pictar, miRanda, mirTarget, TargetScan and PITA; however, this tool has never been updated and is no longer maintained. In the SVM approach MiREE, a hybrid solution is proposed by combining genetic programming for miRNA duplex characteristics (sequence homology and thermodynamics) and a nonlinear SVM for context features [91]. Similar to SVMicrO, its most important features are seed-related. This method obtained a 95% accuracy on human MTI predictions, which is higher than the other methods compared in this review (2nd best is miTarget at greater than 60%). Surprisingly, the Avishkar predictor used a linear SVM model because it has the advantage of being directly interpretable from the weights of each feature and easy to implement [56]. However, as mentioned above, this type of machine learning is expected to perform poorly due to the complexity of MTIs. As a result, even though Avishkar obtained a 98% recall on human MTI, the method showed poor accuracy, with 30% of all predicted targets being misclassified. Interestingly, Li *et al.* proposed improving the performance of miRNA target prediction by searching a second MTI on the whole mRNA sequence after finding one in the 3'-UTR [123]. Thus, they trained an SVM on a two-site search dataset of validated MTIs from miRecords and pSILAC (quantitative proteomics) experiments. When tested on an independent dataset, it showed higher performance than other commonly used methods (PicTar, MirTarget2, miRanda, PITA, TargetSpy, TargetMiner, and TargetScan). To improve both the prediction model and the training dataset, Lu et Leslie created chimiRic, a two-SVM model based on CLASH and AGO-CLIP sequencing data [124]. One SVM uses both data types for duplex prediction, and the other serves for AGO site discrimination (true or not). This strategy has the advantage of training on a large dataset of interacting miR-target duplexes but does not guarantee their functionality. Nevertheless, it shows a superior performance to MIRZA, MirTarget, TargetScan, miRanda and DianamicroT-CDS.

3.1.2.4. Artificial neural networks. Artificial neural networks (ANNs, also called neural networks) have been developed using interconnected neurons in the brain as a model. Features are used as input nodes in this model to feed the "neurons" or working units of the algorithm, which then create new combinations (hidden layers) of these inputs following principles such as fuzzy logic, genetic algorithm or Bayesian statistics, and a prediction is eventually returned. Weight factors are assigned to each neuron to modulate its impact on the predicted result. The model is computed to be adaptive so that weight factors and neuron ordering can change to best suit the training data [125] (Fig. 6). One of the first MTI prediction methods using an ANN was MTar [126]. Unlike most of the current algorithms, which heavily focus on seed region matching, MTar aimed to efficiently identify MTIs regardless of the type of interaction. It first calculates a complementarity score to deter-

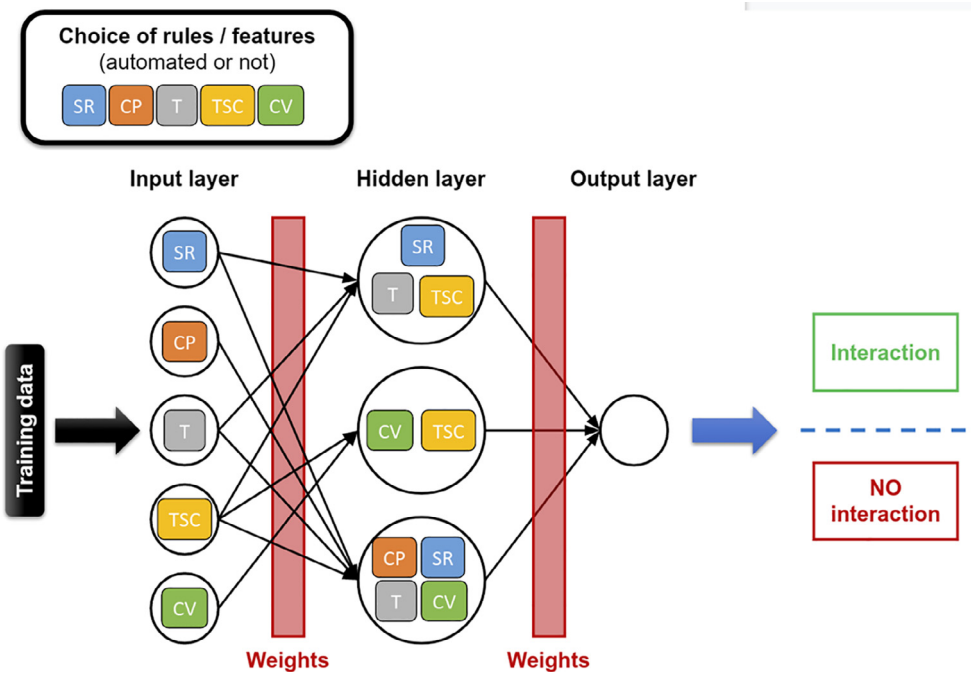

Fig. 6. Neural Network. Selected features (here, seed region (SR) pairing, compensation (CP) pairing, thermodynamic (T) pairing, target site context (TSC) pairing and conservation (CV) pairing) are used as input signals in this feedforward partially connected neural network example. Each node decides what to send to the next node following various principles, such as fuzzy logic, genetic algorithms or Bayesian statistics. Weight factors are applied to each edge. Eventually, an output layer will combine all results in one or several nodes (one in this example), thus allowing the classifier to make a decision. The model can change the weights and node ordering to best classify the training data.

mine the category into which the site falls: 5' seed-only, 5' dominant and 3' canonical (determined from Betel *et al.* [90]). Three different ANNs were trained depending on the site category. They contain 16 input nodes, 9 neurons in the hidden layer and 1 unit in the output layer. This method produces more than 90% fewer targets for each miRNA compared to conventional methods with 94.5% sensitivity and 90.5% specificity. Using a very similar model to that of MTar, HomoTarget uses a pattern recognition neural network (PRNN) coupled with principal component analysis (PCA) for feature selection [112]. It contains 16 input nodes, 14 neurons in the hidden layer and 2 units in the output layer. Unlike MTar, HomoTarget focuses on the seed region to predict MTIs since it filters sequences based on standard seed rules. HomoTarget was trained on a dataset of 425 examples and showed 99% specificity using cross-validation. These two algorithms quickly achieved high performance values due to the limited number of duplexes in their training and testing datasets. It would be interesting to test them on independent and larger datasets.

3.1.2.5. Training datasets. As mentioned above, a good training dataset needs to have a high number of high-quality examples. The training dataset is a critical aspect of all machine learning methods. A major challenge in creating an MTI dataset is to generate real negative examples. The strategy of creating random nucleotide sequences of varying lengths was tried for a few models but was then quickly disregarded because such sequences often interact with miRNAs, as shown in the signal-to-noise ratio experiments of previous studies [24,92,94,107,113,127]. TargetMiner's authors (who later also created MBSTAR) emphasized this issue [99]. Instead of generating random sequences as negative MTIs, they crossed the predictions of other algorithms (miRanda, TargetScanS, PicTar and DIANA-micro-T) with microarray experiments. If a miRNA and its potential targeted mRNA were both overexpressed in a given tissue, then this pair was retained as a negative example. Using this method, 289 negative MTI were generated. A subset of negative examples was then confirmed on a separate pSILAC dataset [128]. To complete the dataset, 289 experimentally validated positive sites were retrieved from miRecords and TarBase [109,110,129]. Using an independent dataset (187 positive and 59 negative pairs), TargetMiner showed 74% accuracy when NBmiRTar and MirTarget2 only had 51% and 46%, respectively, (lower than reported in their original publications), clearly showing the importance of the testing dataset in performance evaluation. Furthermore, they showed that TargetMiner performs better when trained with their negative dataset than with an artificially generated negative set. They confirmed this finding by obtaining similar results with the model of NBmiRTar when repeating the experiment. While validated interactions are most often taken from miRecords or TarBase, some predictors, such as MirTarget2, TargetSpy and Avishkar, were directly trained with positive interactions inferred from microarray or CLIP-seq experiments [56,120,122]. The development of high-throughput methods fostered the tendency to include the largest number of interactions regardless of functional testing. Several datasets used by many predictors marked the history of MTI prediction methods, such as those reported by Linsley *et al.* in 2007 (microarray), Selbach *et al.* in 2008 (pSILAC), Chi *et al.* in 2009 (HITS-CLIP) or Hafner *et al.* in 2010 (PAR-CLIP) [53,128,130,131]. As mentioned in the introduction, miRNAs do not necessarily reduce mRNA levels; thus, microarray data insufficient to fully reflect the action of a miRNA. The use of complementary proteomics data is recommended in this case. Moreover, underexpressed mRNA/protein levels measured by high-throughput experiments can be due to indirect effects of miRNA action [132]. Recently, some predictors were trained on CLASH experiments, which identified both AGO-binding miRNAs and target sites on a transcriptome-wide scale. However, some caution must be taken with CLASH data because several issues related to the specificity of the ligation and the functionality and exhaustivity of the captured MTIs remain unsolved [39,124]. At present, as difficult and expensive as it might be to acquire the

data, combining all these technologies (CLIP-seq, CLASH, microarray and pSILAC) seems to represent the best solution for the use of large training datasets.

3.1.3. Commonly used prediction tools. Most if not all prediction algorithms are usually compared to miRanda, Diana-microT-CDS and/or TargetScan because these three heuristic scoring methods have generally be used by biologists to identify MTIs prior to wet-lab experiments. Their popularity is due to their long history, frequent updates and strong adaptation ability to new advances in MTI prediction.

In the direct foot-step method proposed by Stark [85], miRanda (2003) was developed to further identify MTIs in animals. MiRanda uses the ViennaRNA package to calculate the thermodynamic folding energy of interaction and a scoring matrix, and it assigns values for each nucleotide pairing, with the higher scores used for seed matching [88]. Site conservation is also included in the tested features, and the results are ranked according to the conservation score. From 2004 to 2010, miRanda was upgraded to integrate target site context (global, local and at the duplex level), with a final scoring performed by a support vector regression algorithm (mirSVR) based on mRNA expression change [90,133]. The authors trained mirSVR on a set of nine microRNA transfection experiments performed in HeLa cells by Grimson *et al.* [50]. The score resulting from mirSVR is intended to estimate the efficiency of miRNA action on a given target site and not the probability of regulating this site. With this model, the authors found that the most important features are related to the seed region. The upgrading of mirSVR showed significantly better performances than the previous version of miRanda and seems slightly above TargetScan [90].

Diana-microT is an algorithm published in 2004 that first searches for miRNA-recognition elements (MREs) in the 3′-UTR of a mRNA, including Watson-Crick pairing identification and minimum binding energy calculation using a 38-nt window. A second parameter takes into account the miRNA-associated protein complex, which impacts both the pairing between the miRNA and its target and the site accessibility [134]. In 2009, microT was updated to filter MREs that do not have at least a 7mer in the seed region. The authors also decided to integrate the conservation profiles of MREs using 27 species. Eventually, each considered 3′-UTR is ranked by the weighted sum of the scores of all its identified MREs, and a precision score is calculated by comparing the results with a set of mock miRNAs. An enrichment analysis was also performed with all potential MREs for a given miRNA using the KEGG pathway database. The results are highlighted in the significant pathways that were identified [135]. In 2012, the algorithm was renamed DIANA-microT-CDS because numerous studies have shown that the mRNA coding region can be targeted by a miRNA with a measurable effect on its degradation. Therefore, microT is now used to screen for MREs in this mRNA region, and associated conservation scores are also calculated. Moreover, a dynamic programming algorithm identifies the optimal alignment for the miRNA extended seed sequence (nucleotides 1–9 from the 5′-end of the miRNA) with a 9-nucleotide window on the 3′-UTR or CDS. The prediction method scores the 3′-UTR and CDS region differently and then combines these scores to create the final estimation for the whole mRNA [136]. This last update showed better performance than miRanda and TargetScan at the time of its publication in 2012.

Released as a freely available web tool in 2003 by Bartel's group, TargetScan first used conservation of miRNAs and mRNA UTRs as a filter and then seed matching (length and frequency), 3′ compensation and folding free energy as prediction features [94,137]. The algorithm progressively evolved (last version: v7.2, 2018) to take into consideration all analyzable elements of MTIs that were previously described [2,50,60,137–139]. TargetScan broke down these elements into 14 features using multiple linear regression models (one for each of the four common seed types, off-set 6mer included) trained on microarray datasets published by Garcia *et al.* in 2011 [138]. The resulting models were collectively called the context++ model. When multiple sites are present, individual context++ scores are summed to rank the predicted 3′-UTR. Over time, site conservation has become one of the features of TargetScan and is no longer used as a filter. With a relatively weak contribution to the context++ score, nonconserved targets can even represent the top prediction. After thoroughly analyzing CLIP datasets, the TargetScan authors concluded that "noncanonical sites might exist but have not yet been characterized to the point that they can be used for miRNA target prediction"; therefore, they did not include these sites in their predictions [60]. They also evaluated the use of other more complex types of regression (e.g., linear regression models with interaction terms, lasso/elastic net-regularized regression, multivariate adaptive regression splines, random forest, boosted regression trees, and iterative Bayesian model averaging) but did not find any improvement compared to that of linear regression models [60]. This result is consistent with a similar test performed by Vejnar *et al.* in 2012 [98]. The version of TargetScan described in 2015 showed better performance than 15 other predictors (miRanda and microT included) when tested on the dataset from Linsley *et al.* [131]. With 8 publications describing its content and updates, TargetScan is currently the most widely used MTI prediction tool by the scientific community (more than 1700 citations from PubMed as of November 2020) [102,140,141].

3.2. Data combination

Due to the moderate overlap of the results (5–70%) between all previously cited methods [142], investigators often combine the predictions of different tools to obtain mainly true positive MTIs. Several strategies to combine MTI predictions have been proposed as described below.

3.2.1. Union and intersection. Assuming that an interaction predicted by more than one algorithm is more likely to be functional, databases such as miRWalk, miRSystem or miRGator store and compare results predicted by several tools using statistics and/or mRNA/protein expression data [103,143–146]. Using such an intersection strategy, Kuhn *et al.* validated the interaction of the human angiotensin II type-1 receptor (hAT1R) with hsa-miR-155 and suggested based on their findings to cross results between at least two MTI predictors before undertaking experimental investigations [147]. Ritchie *et al.*, however, demonstrated that targets resulting from the intersection of two lists of predictions are not more likely to be present in the intersection of two other lists [46]. Therefore, intersecting results do not increase the probability of retaining true positives. Moreover, approaches based on the intersection of predictions may lead to decreased sensitivity by possibly omitting valid interactions, as shown by Sethupathy *et al.* [148]. In fact, Oliveira *et al.* [149] showed that union of the results obtained by several prediction tools was more efficient than their intersection. However, when ranking MTIs is required, this method should not be used since it increases the rate of false positives and therefore decreases the specificity of the predictions, which is the most important aspect for ranking purposes. Nevertheless, these databases have the advantage of giving a wide panel of predictions for a given miRNA, with an edge observed for miRWalk, which has been recently updated [103]. Overall, most users do not have enough understanding of MTI predictions to decide which database to take or remove from the union and intersection strategies to be efficient.

3.2.2. Ensemble methods. Because of the limits of the intersection strategy, the union with a rescoring method has been used to

better rank MTIs according to the likelihood of being true. This strategy was first explored by DeConde *et al.* in 2006 using an algorithm that combines ranked lists of miRNA targets from five microarray studies and a reranking of the targets using a statistical test proposed by Tusher *et al.* [150,151]. The performance of this method compared to other tools was not evaluated. Although this work was performed using experimental data, other methods have used aggregation strategies on MTIs predicted by several commonly used tools. For example, MiRror-Suite gathered predicted and/or validated MTIs from 18 databases and allowed for the analysis of approximately 40,000 genes and 2,500 miRNAs [152]. The aggregation strategy consisted of creating a set of potential targets using several filters (species, miRNA family, cell line, number of databases, etc.) and then calculating the probability of an MTI being functional based on a hypergeometric test. However, the ranking performances of this algorithm were not compared to that of other methods. Alternative strategies were tested, such as ExprTarget, which used a multivariate logistic regression model to combine the scores of 3 databases (miRanda, PicTar and TargetScan) and which clearly outperformed other methods based on aggregation [153]. The good performance of similar combination approaches was also confirmed with a model that aggregated 9 predictive algorithms [154]. Others, such as BCmicrO and ComiR, have used more complex strategies for the combination step with an NB classifier for BCmicrO and an SVM for ComiR [155,156]. Interestingly, ComiR takes into consideration the expression levels of miRNAs in its rescoring methods. Of note, ComiR was designed to specifically predict the targets of a set of miRNAs and to consider combinatory interactions. As expected, all aggregation methods were able to outperform in terms of MTI ranking and each aggregated database was considered individually. This finding was also confirmed with our aggregation method named miRabel using a very large dataset (982,411 common interactions) [157]. MiRabel uses a statistic R package (RobustRankAgreg) to rescore MTIs according to their ranks in 4 databases (miRanda, PITA, SVMicrO and Target). This recently published method showed better or equal ranking specificity when compared to other (not aggregated) popular prediction tools. The biological relevance of combined miRNA target predictions from multiple prediction algorithms can also be enhanced by prioritizing results based on functional ranking (inferred from Gene Ontology and enrichment analysis) [158].

4. Performance evaluation

Since prediction tools are designed for biologists, the ease of use should be a criterion in the overall performance. These tools are usually presented in three different platform types: web service, downloadable programs or R/python packages. The first type is the most commonly used because of its user-friendly features; however, ease of use is generally inversely related to flexibility; thus, this tool offers the least degree of freedom in sequence analysis [140].

Programs that exhibit a greater correlation between their predictions and protein or RNA downregulation are commonly considered state-of-the-art tools [41]; however, this would not be the case if the downregulation was directly due to miRNA transfection, which is far from being certain in high-throughput experiments because the experimental conditions can induce false positives.

A more interesting and widely used evaluation method is the area under the receiver operating characteristic (ROC) curve (AUC), which is now well recognized for its capacity to evaluate the performance of classifiers [159]. It plots the sensitivity or true positive rate (TPR) against specificity or false-positive rate (FPR) with TPR = TP/(TP + FN) and FPR = FP/(FP + TN). An MTI is considered to be a true positive (TP) if it has been predicted and experimentally validated, a true negative (TN) if it has been neither predicted nor validated, a false-positive (FP) if it has been predicted and not validated, and a false negative (FN) if validated but not predicted. TPs are readily available through several databases, but this is unfortunately not the case for tested but not validated interactions. Therefore, in the field of MTI prediction, a nonnegligible part of FPs and TNs are mislabeled, thus creating biases in ROC analyses [154]. To complement the ROC analysis, the precision (TP/(TP + FP)) can be plotted versus the recall (same as TPR), and the AUC can also be used for classifier performance evaluation (PR analysis) [160]. An alternative is to plot the cumulated precision versus the normalized scores (sorted in descending order) [154]. Both methods have the advantage of not taking TN into consideration, which minimizes the number of mislabeled MTIs in the analysis. The problem is not completely solved, however, since the accuracy of these methods still depends on the included FP. The use of both ROC and PR analysis is thus recommended for complete performance evaluation of an MTI prediction tool. Unfortunately, not all published algorithms use the same type of parameters to evaluate the performance, which makes comparisons almost impossible. A common pitfall that has been increasingly avoided is to use the training dataset to evaluate prediction performances. Indeed, using several datasets to truly evaluate the performances of predictors is crucial. To address this issue, several independent reviews have already benchmarked some of the previously presented tools, with some predictors being in all benchmarking papers [102,146,161]. Using all the measurements mentioned above and additional measurements, Fan and Kurgan [102] compared 7 target predictors with 4 testing datasets. Although TargetScan and miRmap appeared to be the strongest in this report, a consistent best predictor was not observed across all the possible measurements. Of note, TargetScan performs systematically well in the vast majority of studies comparing MTI prediction algorithms, and it is closely followed by Diana-microT-CDS and miRanda-mirSVR.

These prediction tools are often misused because they do not predict the biological functionality of the interaction between miRNA and mRNA. Indeed, it is unlikely that each predicted miRNA target is sufficiently dose-sensitive to be functionally regulated by miRNAs. Moreover, several studies have shown that some miRNA target prediction software programs are contaminated by high false-positive rates, although this information is rarely emphasized [162,163]. Thus, some mRNAs can efficiently titrate miRNAs, which may contribute to the conservation of miRNA binding sites for ineffectively repressed targets. Another possible explanation would be that phylogenetically conserved interaction sites are conserved for reasons independent of their interaction with miRNAs, which would lead to the overconservation of seed sequences and thus to an increase in the false-positive rate. A better understanding of MTI prediction will likely improve the performance of these bioinformatics tools.

5. Summary and outlook

All miRNA target prediction algorithms use a combination of the sequence, site accessibility and conservation features to identify potential MTIs. However, since the mechanisms of miRNA action are not yet fully understood, predictors still have a high false-positive rate. To improve the accuracy of these tools, different computational methods have been tested. However, none thus far has shown a systematically higher performance regardless of the parameters considered. Surprisingly, empirical methods do not seem to perform better than heuristic methods, suggesting that current training datasets do not efficiently capture all possible MTIs. Additionally, standardization methods are required to com-

pare the algorithms. MTI prediction is challenging, and overcoming the difficulties will require closer coordination between multidisciplinary teams. Overall, 3 predictors, TargetScan, miRanda and Diana-microT, perform well, as reported in benchmarking reviews [102,146,161]. Until better algorithms are developed, ensemble methods seem to be the most efficient strategies to obtain an integrated vision of target predictions for a given miRNA. Ultimately, efficient MTI prediction will reduce the time and resources spent validating miRNA targets and therefore increase the ability of molecular biologists to elucidate the role of miRNAs and their targets under physiological and pathological conditions.

Funding

This work was supported by the Ligue Contre le Cancer de Normandie, Conseil Régional de Normandie, Institut National de la Santé et de la Recherche Médicale (UMRS1239), University of Rouen Normandy and the European Community. Europe has become involved in regional development through the ERDF program. The funding bodies were not involved in the design of the study, the writing of the review or the decision to submit for publication.

CRediT authorship contribution statement

Aurélien Quillet: Conceptualization, Investigation, Writing – original draft. **Youssef Anouar:** Writing – review & editing, Funding acquisition. **Thierry Lecroq:** Investigation, Writing – review & editing. **Christophe Dubessy:** Conceptualization, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 2004;116(2):281–97. https://doi.org/10.1016/S0092-8674(04)00045-5.
[2] Friedman RC, Farh KKH, Burge CB, Bartel DP. Most mammalian mRNAs are conserved targets of microRNAs. Genome Res 2009;19:92–105. https://doi.org/10.1101/gr.082701.108.
[3] Miranda KC, Huynh T, Tay Y, Ang Y-S, Tam W-L, Thomson AM, et al. A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. Cell 2006;126(6):1203–17. https://doi.org/10.1016/j.cell.2006.07.031.
[4] Kozomara A, Griffiths-Jones S. MiRBase: annotating high confidence microRNAs using deep sequencing data. Nucl Acids Res 2014;42(D1): D68–73. https://doi.org/10.1093/nar/gkt1181.
[5] Voinnet O. Origin, biogenesis, and activity of plant microRNAs. Cell 2009;136 (4):669–87. https://doi.org/10.1016/j.cell.2009.01.046.
[6] Moran Y, Agron M, Praher D, Technau U. The evolutionary origin of plant and animal microRNAs. Nat Ecol Evol 2017;1(3). https://doi.org/10.1038/s41559-016-0027.
[7] Yu Y, Jia T, Chen X. The 'how' and 'where' of plant microRNAs. New Phytol 2017;216:1002–17. https://doi.org/10.1111/nph.14834.
[8] Bråte J, Neumann RS, Fromm B, Haraldsen AAB, Tarver JE, Suga H, et al. Unicellular origin of the animal microRNA machinery. Curr Biol 2018;28 (20):3288–3295.e5. https://doi.org/10.1016/j.cub.2018.08.018.
[9] Catalanotto C, Cogoni C, Zardo G. MicroRNA in control of gene expression: an overview of nuclear functions. Int J Mol Sci 2016;17(10):1712. https://doi.org/10.3390/ijms17101712.
[10] Lin S, Gregory RI. MicroRNA biogenesis pathways in cancer. Nat Rev Cancer 2015;15(6):321–33. https://doi.org/10.1038/nrc3932.
[11] Ha M, Kim VN. Regulation of microRNA biogenesis. Nat Rev Mol Cell Biol 2014;15(8):509–24. https://doi.org/10.1038/nrm3838.
[12] Oliveto S, Mancino M, Manfrini N, Biffo S. Role of microRNAs in translation regulation and cancer. World J Biol Chem 2017;8(1):45. https://doi.org/10.4331/wjbc.v8.i1.45.
[13] Karbiener M, Glantschnig C, Scheideler M. Hunting the needle in the haystack: a guide to obtain biologically meaningful microRNA targets. Int J Mol Sci 2014;15:20266–89. https://doi.org/10.3390/ijms151120266.
[14] Bartel DP. Metazoan microRNAs. Cell 2018;173(1):20–51. https://doi.org/10.1016/j.cell.2018.03.006.
[15] Parker R, Sheth U. P Bodies and the control of mRNA translation and degradation. Mol Cell 2007;25(5):635–46. https://doi.org/10.1016/j.molcel.2007.02.011.
[16] Rehwinkel J, Behm-Ansmant I, Gatfield D, Izaurralde E. A crucial role for GW182 and the DCP1:DCP2 decapping complex in miRNA-mediated gene silencing. RNA 2005;11:1640–7. https://doi.org/10.1261/rna.2191905.
[17] Trabucchi M, Mategot R. Subcellular heterogeneity of the microRNA machinery. Trends Genet 2019;35(1):15–28. https://doi.org/10.1016/j.tig.2018.10.006.
[18] Jonas S, Izaurralde E. Towards a molecular understanding of microRNA-mediated gene silencing. Nat Rev Genet 2015;16(7):421–33. https://doi.org/10.1038/nrg3965.
[19] Fabian MR, Sonenberg N, Filipowicz W. Regulation of mRNA translation and stability by microRNAs. Annu Rev Biochem 2010;79(1):351–79. https://doi.org/10.1146/annurev-biochem-060308-103103.
[20] Eichhorn S, Guo H, McGeary S, Rodriguez-Mias R, Shin C, Baek D, et al. MRNA destabilization is the dominant effect of mammalian microRNAs by the time substantial repression ensues. Mol Cell 2014;56(1):104–15. https://doi.org/10.1016/j.molcel.2014.08.028.
[21] Guo H, Ingolia NT, Weissman JS, Bartel DP. Mammalian microRNAs predominantly act to decrease target mRNA levels. Nature 2010;466 (7308):835–40. https://doi.org/10.1038/nature09267.
[22] Bartel DP. MicroRNAs: target recognition and regulatory functions. Cell 2009;136(2):215–33. https://doi.org/10.1016/j.cell.2009.01.002.
[23] Hamzeiy H, Allmer J, Yousef M. Computational methods for microRNA target prediction. Methods Mol Biol 2014;1107:207–21. https://doi.org/10.1007/978-1-62703-748-8_12.
[24] Krek A, Grün D, Poy MN, Wolf R, Rosenberg L, Epstein EJ, et al. Combinatorial microRNA target predictions. Nat Genet 2005;37(5):495–500. https://doi.org/10.1038/ng1536.
[25] Gebert LFR, MacRae IJ. Regulation of microRNA function in animals. Nat Rev Mol Cell Biol 2019;20(1):21–37. https://doi.org/10.1038/s41580-018-0045-7.
[26] Dror S, Sander L, Schwartz H, Sheinboim D, Barzilai A, Dishon Y, et al. Melanoma miRNA trafficking controls tumour primary niche formation. Nat Cell Biol 2016;18(9):1006–17. https://doi.org/10.1038/ncb3399.
[27] Li Y, Shan Z, Liu C, Yang D, Wu J, Men C, et al. MicroRNA-294 promotes cellular proliferation and motility through the PI3K/AKT and JAK/STAT pathways by upregulation of NRAS in bladder cancer. Biochem 2017;82 (4):474–82. https://doi.org/10.1134/S0006297917040095.
[28] Xia H, Long J, Zhang R, Yang X, Ma Z. MiR-32 contributed to cell proliferation of human breast cancer cells by suppressing of PHLPP2 expression. Biomed Pharmacother 2015;75:105–10. https://doi.org/10.1016/j.biopha.2015.07.037.
[29] Maqbool R, Hussain MU. MicroRNAs and human diseases: diagnostic and therapeutic potential. Cell Tissue Res 2014;358(1):1–15. https://doi.org/10.1007/s00441-013-1787-3.
[30] Bronze-da-Rocha E. MicroRNAs expression profiles in cardiovascular diseases. Biomed Res Int 2014;2014:1–23. https://doi.org/10.1155/2014/985408.
[31] Basak I, Patil KS, Alves G, Larsen JP, Møller SG. MicroRNAs as neuroregulators, biomarkers and therapeutic agents in neurodegenerative diseases. Cell Mol Life Sci 2016;73(4):811–27. https://doi.org/10.1007/s00018-015-2093-x.
[32] Szeto C-C, Li P-T. MicroRNAs in IgA nephropathy. Nat Rev Nephrol 2014;10 (5):249–56. https://doi.org/10.1038/nrneph.2014.50.
[33] Di Leva G, Garofalo M, Croce CM. MicroRNAs in cancer. Annu Rev Pathol Mech Dis 2014;9(1):287–314. https://doi.org/10.1146/pathmechdis.2014.9.issue-110.1146/annurev-pathol-012513-104715.
[34] Oom AL, Humphries BA, Yang C. MicroRNAs: Novel players in cancer diagnosis and therapies. Biomed Res Int 2014;2014:1–13. https://doi.org/10.1155/2014/959461.
[35] Cheng Q, Yi B, Wang A, Jiang X. Exploring and exploiting the fundamental role of microRNAs in tumor pathogenesis. Onco Targets Ther 2013;6:1675–84. https://doi.org/10.2147/OTT.S52730.
[36] Chou C-H, Chang N-W, Shrestha S, Hsu S-D, Lin Y-L, Lee W-H, et al. miRTarBase 2016: Updates to the experimentally validated miRNA-target interactions database. Nucl Acids Res 2016;44(D1):D239–47. https://doi.org/10.1093/nar/gkv1258.
[37] Campos-Melo D, Droppelmann CA, Volkening K, Strong MJ. Comprehensive luciferase-based reporter gene assay reveals previously masked up-regulatory effects of miRNAs. Int J Mol Sci 2014;15:15592–602. https://doi.org/10.3390/ijms150915592.
[38] Bottini S, Pratella D, Grandjean V, Repetto E, Trabucchi M. Recent computational developments on CLIP-seq data analysis and microRNA targeting implications. Brief Bioinform 2017;19:1290–301. https://doi.org/10.1093/bib/bbx063.
[39] Broughton JP, Pasquinelli AE. Identifying argonaute binding sites in caenorhabditis elegans using iCLIP. Methods 2013;63(2):119–25. https://doi.org/10.1016/j.ymeth.2013.03.033.
[40] Henry VJ, Bandrowski AE, Pepin AS, Gonzalez BJ, Desfeux A. OMICtools: an informative directory for multi-omic data analysis. Database (Oxford) 2014;2014. doi: 10.1093/database/bau069.
[41] Oulas A, Karathanasis N, Louloupi A, Pavlopoulos GA, Poirazi P, Kalantidis K, et al. Prediction of miRNA targets. Methods Mol Biol 2015;1269:207–29. https://doi.org/10.1007/978-1-4939-2291-8_13.

[42] Shukla V, Varghese VK, Kabekkodu SP, Mallya S, Satyamoorthy K. A compilation of Web-based research tools for miRNA analysis. Brief Funct Genomics 2017;16:249–73. https://doi.org/10.1093/bfgp/elw042.
[43] Akhtar MM, Micolucci L, Islam MS, Olivieri F, Procopio AD. A practical guide to miRNA target prediction. Methods Mol Biol 2019;1970:1–13. https://doi.org/10.1007/978-1-4939-9207-2_1.
[44] Monga I, Kumar M. Computational resources for prediction and analysis of functional miRNA and their targetome. Methods Mol Biol 2019;1912:215–50. https://doi.org/10.1007/978-1-4939-8982-9_9.
[45] Kern F, Backes C, Hirsch P, Fehlmann T, Hart M, Meese E, et al. What's the target: Understanding two decades of in silico microRNA-target prediction. Brief Bioinform 2020;21:1999–2010. doi: 10.1093/bib/bbz111.
[46] Ritchie W, Flamant S, Rasko JEJ. Predicting microRNA targets and functions: traps for the unwary. Nat Methods 2009;6(6):397–8. https://doi.org/10.1038/nmeth0609-397.
[47] Sedaghat N, Fathy M, Modarressi MH, Shojaie A. Combining supervised and unsupervised learning for improved mirna target prediction. IEEE/ACM Trans Comput Biol Bioinforma 2018;15:1594–604. https://doi.org/10.1109/TCBB.2017.2727042.
[48] Brennecke J, Stark A, Russell RB, Cohen SM. Principles of microRNA-target recognition. PLoS Biol 2005;3(3):e85. https://doi.org/10.1371/journal.pbio.0030085.
[49] Didiano D, Hobert O. Perfect seed pairing is not a generally reliable predictor for miRNA-target interactions. Nat Struct Mol Biol 2006;13(9):849–51. https://doi.org/10.1038/nsmb1138.
[50] Grimson A, Farh K-H, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Mol Cell 2007;27(1):91–105. https://doi.org/10.1016/j.molcel.2007.06.017.
[51] Baek D, Villén J, Shin C, Camargo FD, Gygi SP, Bartel DP. The impact of microRNAs on protein output. Nature 2008;455(7209):64–71. https://doi.org/10.1038/nature07242.
[52] Chi SW, Zang JB, Mele A, Darnell RB. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. Nature 2009;460(7254):479–86. https://doi.org/10.1038/nature08170.
[53] Chi SW, Hannon GJ, Darnell RB. An alternative mode of microRNA target recognition. Nat Struct Mol Biol 2012;19(3):321–7. https://doi.org/10.1038/nsmb.2230.
[54] Seok H, Ham J, Jang ES, Chi SW. MicroRNA target recognition: Insights from transcriptome-wide non-canonical interactions. Mol Cells 2016;39:375–81. https://doi.org/10.14348/molcells.2016.0013.
[55] Bottini S, Hamouda-Tekaya N, Tanasa B, Zaragosi LE, Grandjean V, Repetto E, et al. From benchmarking HITS-CLIP peak detection programs to a new method for identification of miRNA-binding sites from Ago2-CLIP data. Nucl Acids Res 2017;45:. https://doi.org/10.1093/nar/gkx007e71.
[56] Ghoshal A, Shankar R, Bagchi S, Grama A, Chaterji S. MicroRNA target prediction using thermodynamic and sequence curves. BMC Genomics 2015;16:999. https://doi.org/10.1186/s12864-015-1933-2.
[57] Helwak A, Kudla G, Dudnakova T, Tollervey D. Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. Cell 2013;153(3):654–65. https://doi.org/10.1016/j.cell.2013.03.043.
[58] Moore MJ, Scheel TKH, Luna JM, Park CY, Fak JJ, Nishiuchi E, et al. MiRNA-target chimeras reveal miRNA 3′-end pairing as a major determinant of Argonaute target specificity. Nat Commun 2015;6(1). https://doi.org/10.1038/ncomms9864.
[59] Stefani G, Slack FJ. A "pivotal" new rule for microRNA-mRNA interactions. Nat Struct Mol Biol 2012;19(3):265–6. https://doi.org/10.1038/nsmb.2256.
[60] Agarwal V, Bell GW, Nam JW, Bartel DP. Predicting effective microRNA target sites in mammalian mRNAs. Elife 2015;4. doi: 10.7554/eLife.05005.
[61] Friedersdorf MB, Keene JD. Advancing the functional utility of PAR-CLIP by quantifying background binding to mRNAs and lncRNAs. Genome Biol 2014;15(1):R2. https://doi.org/10.1186/gb-2014-15-1-r2.
[62] Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, et al. ViennaRNA Package 2.0. Algorithms Mol Biol 2011;6:26. doi: 10.1186/1748-7188-6-26.
[63] Rehmsmeier M, Steffen P, Höchsmann M, Giegerich R. Fast and effective prediction of microRNA/target duplexes. RNA 2004;10:1507–17. https://doi.org/10.1261/rna.5248604.
[64] Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E. The role of site accessibility in microRNA target recognition. Nat Genet 2007;39(10):1278–84. https://doi.org/10.1038/ng2135.
[65] Nielsen CB, Shomron N, Sandberg R, Hornstein E, Kitzman J, Burge CB. Determinants of targeting by endogenous and exogenous microRNAs and siRNAs. RNA 2007;13:1894–910. https://doi.org/10.1261/rna.768207.
[66] Moretti F, Thermann R, Hentze MW. Mechanism of translational regulation by miR-2 from sites in the 5′ untranslated region or the open reading frame. RNA 2010;16(12):2493–502. https://doi.org/10.1261/rna.2384610.
[67] Qu H, Zheng L, Song H, Jiao W, Li D, Fang E, et al. microRNA-558 facilitates the expression of hypoxia-inducible factor 2 alpha through binding to 5'-untranslated region in neuroblastoma. Oncotarget 2016;7(26):40657–73.
[68] Gu S, Jin L, Zhang F, Sarnow P, Kay MA. Biological basis for restriction of microRNA targets to the 3′ untranslated region in mammalian mRNAs. Nat Struct Mol Biol 2009;16(2):144–50. https://doi.org/10.1038/nsmb.1552.
[69] Lytle JR, Yario TA, Steitz JA. Target mRNAs are repressed as efficiently by microRNA-binding sites in the 5′ UTR as in the 3′ UTR. Proc Natl Acad Sci U S A 2007;104(23):9667–72. https://doi.org/10.1073/pnas.0703820104.
[70] Hausser J, Syed AP, Bilen B, Zavolan M. Analysis of CDS-located miRNA target sites suggests that they can effectively inhibit translation. Genome Res 2013;23(4):604–15. https://doi.org/10.1101/gr.139758.112.
[71] Zhang K, Zhang X, Cai Z, Zhou J, Cao R, Zhao Ya, et al. A novel class of microRNA-recognition elements that function only within open reading frames. Nat Struct Mol Biol 2018;25(11):1019–27. https://doi.org/10.1038/s41594-018-0136-3.
[72] Niepmann M. Activation of hepatitis C virus translation by a liver-specific microRNA. Cell Cycle 2009;8(10):1473–7. https://doi.org/10.4161/cc.8.10.8349.
[73] Ørom UA, Nielsen FC, Lund AH. MicroRNA-10a binds the 5′UTR of ribosomal protein mRNAs and enhances their translation. Mol Cell 2008;30(4):460–71. https://doi.org/10.1016/j.molcel.2008.05.001.
[74] Ni WJ, Leng XM. Dynamic miRNA-mRNA paradigms: new faces of miRNAs. Biochem Biophys Reports 2015;4:337–41. https://doi.org/10.1016/j.bbrep.2015.10.011.
[75] Sætrom P, Heale BSE, Snøve O, Aagaard L, Alluin J, Rossi JJ. Distance constraints between microRNA target sites dictate efficacy and cooperativity. Nucl Acids Res 2007;35:2333–42. doi: 10.1093/nar/gkm133.
[76] Shu J, Xia Z, Li L, Liang ET, Slipek N, Shen D, et al. Dose-dependent differential mRNA target selection and regulation by let-7a-7f and miR-17-92 cluster microRNAs. RNA Biol 2012;9(10):1275–87. https://doi.org/10.4161/rna.21998.
[77] Bandyopadhyay S, Ghosh D, Mitra R, Zhao Z. MBSTAR: Multiple instance learning for predicting specific functional binding sites in microRNA targets. Sci Rep 2015;5:8004. https://doi.org/10.1038/srep08004.
[78] Erhard F, Haas J, Lieber D, Malterer G, Jaskiewicz L, Zavolan M, et al. Widespread context dependency of microRNA-mediated regulation. Genome Res 2014;24(6):906–19. https://doi.org/10.1101/gr.166702.113.
[79] Ciafrè SA, Galardi S. microRNAs and RNA-binding proteins: a complex network of interactions and reciprocal regulations in cancer. RNA Biol 2013;10(6):934–42. https://doi.org/10.4161/rna.24641.
[80] Bottini S, Hamouda-Tekaya N, Mategot R, Zaragosi L-E, Audebert S, Pisano S, et al. Post-transcriptional gene silencing mediated by microRNAs is controlled by nucleoplasmic Sfpq. Nat Commun 2017;8(1). https://doi.org/10.1038/s41467-017-01126-x.
[81] Kheradpour P, Stark A, Roy S, Kellis M. Reliable prediction of regulator targets using 12 Drosophila genomes. Genome Res 2007;17(12):1919–31. https://doi.org/10.1101/gr.7090407.
[82] Farh K-H, Grimson A, Jan C, Lewis BP, Johnston WK, Lim LP, et al. Biochemistry: the widespread impact of mammalian microRNAs on mRNA repression and evolution. Science (80-) 2005;310(5755):1817–21. https://doi.org/10.1126/science:1121158.
[83] Gumienny R, Zavolan M. Accurate transcriptome-wide prediction of microRNA targets and small interfering RNA off-targets with MIRZA-G. Nucl Acids Res 2015;43:1380–91. https://doi.org/10.1093/nar/gkv050.
[84] Stark A, Brennecke J, Bushati N, Russell RB, Cohen SM. Animal microRNAs confer robustness to gene expression and have a significant impact on 3′UTR evolution. Cell 2005;123(6):1133–46. https://doi.org/10.1016/j.cell.2005.11.023.
[85] Stark A, Brennecke J, Russell RB, Cohen SM. Identification of Drosophila microRNA targets. PLoS Biol 2003;1(3):e60. https://doi.org/10.1371/journal.pbio.0000060.
[86] Gaidatzis D, van Nimwegen E, Hausser J, Zavolan M. Inference of miRNA targets using evolutionary conservation and pathway analysis. BMC Bioinf 2007;8:69. https://doi.org/10.1186/1471-2105-8-69.
[87] Burgler C, Macdonald PM. Prediction and verification of microRNA targets by moving targets, a highly adaptable prediction method. BMC Genomics 2005;6:88. https://doi.org/10.1186/1471-2164-6-88.
[88] Enright AJ, John B, Gaul U, Tuschl T, Sander C, Marks DS. MicroRNA targets in Drosophila. Genome Biol 2003;5:R1. https://doi.org/10.1186/gb-2003-5-1-r1.
[89] Liu H, Yue D, Chen Y, Gao SJ, Huang Y. Improving performance of mammalian microRNA target prediction. BMC Bioinf 2010;11:476. https://doi.org/10.1186/1471-2105-11-476.
[90] Betel D, Koppal A, Agius P, Sander C, Leslie C. Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. Genome Biol 2010;11(8):R90. https://doi.org/10.1186/gb-2010-11-8-r90.
[91] Reyes-Herrera PH, Ficarra E, Acquaviva A, Macii E. MiREE: MiRNA recognition elements ensemble. BMC Bioinf 2011;12:454. https://doi.org/10.1186/1471-2105-12-454.
[92] Kim SK, Nam JW, Rhee JK, Lee WJ, Zhang BT. miTarget: MicroRNA target gene prediction using a support vector machine. BMC Bioinf 2006;7:411. https://doi.org/10.1186/1471-2105-7-411.
[93] Khorshid M, Hausser J, Zavolan M, van Nimwegen E. A biophysical miRNA-mRNA interaction model infers canonical and noncanonical targets. Nat Methods 2013;10(3):253–5. https://doi.org/10.1038/nmeth.2341.
[94] Lewis BP, Shih I-H, Jones-Rhoades MW, Bartel DP, Burge CB. Prediction of mammalian MicroRNA targets. Cell 2003;115(7):787–98. https://doi.org/10.1016/S0092-8674(03)01018-3.
[95] Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. Nucl Acids Res 2003;31(13):3406–15. https://doi.org/10.1093/nar/gkg595.
[96] Rennie W, Liu C, Carmack CS, Wolenc A, Kanoria S, Lu J, et al. STarMir: A web server for prediction of microRNA binding sites. Nucl Acids Res 2014;42:W114-8. doi: 10.1093/nar/gku376.

[97] Kanoria S, Rennie W, Liu C, Carmack CS, Lu J, Ding Y. STarMir tools for prediction of microRNA binding sites. Methods Mol Biol 2016;1490:73–82. https://doi.org/10.1007/978-1-4939-6433-8_6.
[98] Vejnar CE, Zdobnov EM. MiRmap: Comprehensive prediction of microRNA target repression strength. Nucl Acids Res 2012;40:11673–83. https://doi.org/10.1093/nar/gks901.
[99] Bandyopadhyay S, Mitra R. TargetMiner: MicroRNA target prediction with systematic identification of tissue-specific negative examples. Bioinformatics 2009;25:2625–31. https://doi.org/10.1093/bioinformatics/btp503.
[100] Maragkakis M, Alexiou P, Papadopoulos GL, Reczko M, Dalamagas T, Giannopoulos G, et al. Accurate microRNA target prediction correlates with protein repression levels. BMC Bioinf 2009;10(1). https://doi.org/10.1186/1471-2105-10-295.
[101] Marin RM, Sulc M, Vanicek J. Searching the coding region for microRNA targets. RNA 2013;19(4):467–74. https://doi.org/10.1261/rna.035634.112.
[102] Fan X, Kurgan L. Comprehensive overview and assessment of computational prediction of microRNA targets in animals. Brief Bioinform 2015;16(5):780–94. https://doi.org/10.1093/bib/bbu044.
[103] Dweep H, Gretz N. MiRWalk2.0: A comprehensive atlas of microRNA-target interactions. Nat Methods 2015;12:697. doi: 10.1038/nmeth.3485.
[104] Baştanlar Y, Özuysal M. Introduction to machine learning. Methods Mol Biol 2014;1107:105–28. https://doi.org/10.1007/978-1-62703-748-8_7.
[105] Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20(3):273–97. https://doi.org/10.1007/BF00994018.
[106] Saetrom P. Predicting the efficacy of short oligonucleotides in antisense and RNAi experiments with boosted genetic programming. Bioinformatics 2004;20(17):3055–63. https://doi.org/10.1093/bioinformatics/bth364.
[107] Sætrom O, Snøve O, Sætrom P. Weighted sequence motifs as an improved seeding step in microRNA target prediction algorithms. RNA 2005;11:995–1003. https://doi.org/10.1261/rna.7290705.
[108] Rabiee-Ghahfarrokhi B, Rafiei F, Niknafs AA, Zamani B. Prediction of microRNA target genes using an efficient genetic algorithm-based decision tree. FEBS Open Bio 2015;5:877–84. https://doi.org/10.1016/j.fob.2015.10.003.
[109] Vlachos IS, Paraskevopoulou MD, Karagkouni D, Georgakilas G, Vergoulis T, Kanellos I, et al. DIANA-TarBase v7.0: Indexing more than half a million experimentally supported miRNA:mRNA interactions. Nucleic Acids Res 2015;43:D153–9. doi: 10.1093/nar/gku1215.
[110] Sethupathy P, Corda B, Hatzigeorgiou AG. TarBase: A comprehensive database of experimentally supported animal microRNA targets. RNA 2006;12:192–7. https://doi.org/10.1261/rna.2239606.
[111] 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. doi: 10.1016/j.febslet.2007.03.022.
[112] Ahmadi H, Ahmadi A, Azimzadeh-Jamalkandi S, Shoorehdeli MA, Salehzadeh-Yazdi A, Bidkhori G, et al. HomoTarget: a new algorithm for prediction of microRNA targets in Homo sapiens. Genomics 2013;101(2):94–100. https://doi.org/10.1016/j.ygeno.2012.11.005.
[113] 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. https://doi.org/10.1093/bioinformatics/btm484.
[114] Huang JC, Frey BJ, Morris QD. COmparing sequence and expression for predicting microRNA targets using GENMiR3. Pacific Symp Biocomput 2008, PSB 2008 2008:52–63. doi: 10.1142/9789812776136_0007.
[115] Huang JC, Babak T, Corson TW, Chua G, Khan S, Gallie BL, et al. Using expression profiling data to identify human microRNA targets. Nat Methods 2007;4(12):1045–9. https://doi.org/10.1038/nmeth1130.
[116] Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 2005;15:1034–50. https://doi.org/10.1101/gr.3715005.
[117] Amirkhah R, Farazmand A, Gupta SK, Ahmadi H, Wolkenhauer O, Schmitz U. Naïve Bayes classifier predicts functional microRNA target interactions in colorectal cancer. Mol Biosyst 2015;11(8):2126–34. https://doi.org/10.1039/C5MB00245A.
[118] Breiman L. Random forests. Mach Learn 2001;45:5–32. https://doi.org/10.1023/A:1010933404324.
[119] Mendoza MR, da Fonseca GC, Loss-Morais G, Alves R, Margis R, Bazzan ALC, et al. RFMirTarget: predicting human microRNA target genes with a random forest classifier. PLoS ONE 2013;8(7):e70153. https://doi.org/10.1371/journal.pone.0070153.
[120] Sturm M, Hackenberg M, Langenberger D, Frishman D. TargetSpy: a supervised machine learning approach for microRNA target prediction. BMC Bioinf 2010;11:292. https://doi.org/10.1186/1471-2105-11-292.
[121] Ding J, Li X, Hu H. TarPmiR: A new approach for microRNA target site prediction. Bioinformatics 2016;32(18):2768–75. https://doi.org/10.1093/bioinformatics/btw318.
[122] Wang X, El Naqa IM. Prediction of both conserved and nonconserved microRNA targets in animals. Bioinformatics 2008;24:325–32. https://doi.org/10.1093/bioinformatics/btm595.
[123] Li L, Gao Q, Mao X, Cao Y. New support vector machine-based method for microRNA target prediction. Genet Mol Res 2014;13:4165–76. https://doi.org/10.4238/2014.June.9.3.
[124] Lu Y, Leslie CS, Chen K. Learning to predict miRNA-mRNA interactions from AGO CLIP sequencing and CLASH data. PLoS Comput Biol 2016;12(7):e1005026. https://doi.org/10.1371/journal.pcbi.1005026.
[125] Churpek MM, Yuen TC, Winslow C, Meltzer DO, Kattan MW, Edelson DP. Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. Crit Care Med 2016;44(2):368–74. https://doi.org/10.1097/CCM.0000000000001571.
[126] Chandra V, Girijadevi R, Nair AS, Pillai SS, Pillai RM. MTar: a computational microRNA target prediction architecture for human transcriptome. BMC Bioinf 2010;11:S2. https://doi.org/10.1186/1471-2105-11-S1-S2.
[127] Rodriguez A, Griffiths-Jones S, Ashurst JL, Bradley A. Identification of mammalian microRNA host genes and transcription units. Genome Res 2004;14:1902–10. https://doi.org/10.1101/gr.2722704.
[128] Selbach M, Schwanhäusser B, Thierfelder N, Fang Z, Khanin R, Rajewsky N. Widespread changes in protein synthesis induced by microRNAs. Nature 2008;455(7209):58–63. https://doi.org/10.1038/nature07228.
[129] Xiao F, Zuo Z, Cai G, Kang S, Gao X, Li T. miRecords: an integrated resource for microRNA-target interactions. Nucl Acids Res 2009;37(Database):D105–10. https://doi.org/10.1093/nar/gkn851.
[130] Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, et al. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell 2010;141(1):129–41. https://doi.org/10.1016/j.cell.2010.03.009.
[131] Linsley PS, Schelter J, Burchard J, Kibukawa M, Martin MM, Bartz SR, et al. Transcripts targeted by the microRNA-16 family cooperatively regulate cell cycle progression. Mol Cell Biol 2007;27(6):2240–52. https://doi.org/10.1128/MCB.02005-06.
[132] Zhang H-M, Kuang S, Xiong X, Gao T, Liu C, Guo A-Y. Transcription factor and microRNA co-regulatory loops: Important regulatory motifs in biological processes and diseases. Brief Bioinform 2015;16(1):45–58. https://doi.org/10.1093/bib/bbt085.
[133] John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS, et al. Human microRNA targets. PLoS Biol 2004;2(11):e363. https://doi.org/10.1371/journal.pbio.
[134] 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. https://doi.org/10.1101/gad.1184704.
[135] Maragkakis M, Reczko M, Simossis VA, Alexiou P, Papadopoulos GL, Dalamagas T, et al. DIANA-microT web server: elucidating microRNA functions through target prediction. Nucl Acids Res 2009;37:W273–6. https://doi.org/10.1093/nar/gkp292.
[136] Reczko M, Maragkakis M, Alexiou P, Grosse I, Hatzigeorgiou AG. Functional microRNA targets in protein coding sequences. Bioinformatics 2012;28:771–6. https://doi.org/10.1093/bioinformatics/bts043.
[137] 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(1):15–20. https://doi.org/10.1016/j.cell.2004.12.035.
[138] 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(10):1139–46. https://doi.org/10.1038/nsmb.2115.
[139] Nam J-W, Rissland O, Koppstein D, Abreu-Goodger C, Jan C, Agarwal V, et al. Global analyses of the effect of different cellular contexts on microRNA targeting. Mol Cell 2014;53(6):1031–43. https://doi.org/10.1016/j.molcel.2014.02.013.
[140] Riffo-Campos Á, Riquelme I, Brebi-Mieville P. Tools for sequence-based miRNA target prediction: what to choose? Int J Mol Sci 2016;17(12):1987. https://doi.org/10.3390/ijms17121987.
[141] Ekimler S, Sahin K. Computational methods for microRNA target prediction. Genes (Basel) 2014;5:671–83. https://doi.org/10.3390/genes5030671.
[142] Hammell M. Computational methods to identify miRNA targets. Semin Cell Dev Biol 2010;21(7):738–44. https://doi.org/10.1016/j.semcdb.2010.01.004.
[143] Dweep H, Sticht C, Pandey P, Gretz N. MiRWalk – database: prediction of possible miRNA binding sites by "walking" the genes of three genomes. J Biomed Inform 2011;44(5):839–47. https://doi.org/10.1016/j.jbi.2011.05.002.
[144] Lu T-P, Lee C-Y, Tsai M-H, Chiu Y-C, Hsiao CK, Lai L-C, et al. MiRSystem: an integrated system for characterizing enriched functions and pathways of microRNA targets. PLoS ONE 2012;7(8):e42390. https://doi.org/10.1371/journal.pone.0042390.
[145] Nam S, Kim B, Shin S, Lee S. miRGator: an integrated system for functional annotation of microRNAs. Nucl Acids Res 2008;36:D159–64. https://doi.org/10.1093/nar/gkm829.
[146] Roberts JT, Borchert GM. Computational prediction of microRNA target genes, target prediction databases, and web resources. Methods Mol Biol 2017;1617:109–22. https://doi.org/10.1007/978-1-4939-7046-9_8.
[147] Kuhn DE, Martin MM, Feldman DS, Terry AV, Nuovo GJ, Elton TS. Experimental validation of miRNA targets. Methods 2008;44(1):47–54. https://doi.org/10.1016/j.ymeth.2007.09.005.
[148] Sethupathy P, Megraw M, Hatzigeorgiou AG. A guide through present computational approaches for the identification of mammalian microRNA targets. Nat Methods 2006;3(11):881–6. https://doi.org/10.1038/nmeth954.
[149] Oliveira AC, Bovolenta LA, Nachtigall PG, Herkenhoff ME, Lemke N, Pinhal D. Combining results from distinct microRNA target prediction tools enhances

the performance of analyses. Front Genet 2017;8:59. https://doi.org/10.3389/fgene.2017.00059.
[150] DeConde RP, Hawley S, Falcon S, Clegg N, Knudsen B, Etzioni R. Combining results of microarray experiments: a rank aggregation approach Article15. Stat Appl Genet Mol Biol 2006:5. https://doi.org/10.2202/1544-6115.1204.
[151] Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 2001;98 (9):5116–21. https://doi.org/10.1073/pnas.091062498.
[152] Friedman Y, Karsenty S, Linial M. MiRror-Suite: Decoding coordinated regulation by microRNAs. Database 2014;2014. doi: 10.1093/database/bau043.
[153] Gamazon ER, Im H-K, Duan S, Lussier YA, Cox NJ, Dolan ME, et al. ExprTarget: an integrative approach to predicting human microRNA targets. PLoS ONE 2010;5(10):e13534. https://doi.org/10.1371/journal.pone.0013534.
[154] Tabas-Madrid D, Muniategui A, Sánchez-Caballero I, Martínez-Herrera DJ, Sorzano COS, Rubio A, et al. Improving miRNA-mRNA interaction predictions. BMC Genomics 2014;15(S10). https://doi.org/10.1186/1471-2164-15-S10-S2.
[155] Coronnello C, Benos PV. ComiR: combinatorial microRNA target prediction tool. Nucl Acids Res 2013;41:W159–64. https://doi.org/10.1093/nar/gkt379.
[156] Yue D, Guo M, Chen Y, Huang Y. A Bayesian decision fusion approach for microRNA target prediction. BMC Genomics 2012;13(Suppl 8):S13. https://doi.org/10.1186/1471-2164-13-s8-s13.
[157] Quillet A, Saad C, Ferry G, Anouar Y, Vergne N, Lecroq T, et al. Improving bioinformatics prediction of microRNA targets by ranks aggregation. Front Genet 2020;10. https://doi.org/10.3389/fgene.2019.01330.
[158] Li J, Zhang Y, Wang Y, Zhang C, Wang Q, Shi X, et al. Functional combination strategy for prioritization of human miRNA target. Gene 2014;533 (1):132–41. https://doi.org/10.1016/j.gene.2013.09.106.
[159] Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem 1993;39:561–77. https://doi.org/10.1093/clinchem/39.4.561.
[160] Davis J, Goadrich M. The relationship between precision-recall and ROC curves. ACM Int Conf Proceeding Ser 2006;148:233–40. https://doi.org/10.1145/1143844.1143874.
[161] Bradley T, Moxon S. An assessment of the next generation of animal miRNA target prediction algorithms. Methods Mol Biol 2017;1580:175–91. https://doi.org/10.1007/978-1-4939-6866-4_13.
[162] Pinzón N, Li B, Martinez L, Sergeeva A, Presumey J, Apparailly F, et al. MicroRNA target prediction programs predict many false positives. Genome Res 2017;27(2):234–45. https://doi.org/10.1101/gr.205146.116.
[163] Fridrich A, Hazan Y, Moran Y. Too many false targets for MicroRNAs: challenges and pitfalls in prediction of miRNA targets and their gene ontology in model and non-model organisms. BioEssays 2019;41(4):1800169. https://doi.org/10.1002/bies.v41.410.1002/bies.201800169.