Prediction tools for miRNA targets: toward a better comprehension for Biologists

Aurélien Quillet, Youssef Anouar, Thierry Lecroq, Christophe Dubessy

Prediction tools for miRNA targets: toward a better comprehension for Biologists

Aurélien Quillet[1], Youssef Anouar[1], Thierry Lecroq[2], Christophe Dubessy[1*]

[1]Normandie Univ, UNIROUEN, INSERM, Laboratoire Différenciation et Communication Neuronale et Neuroendocrine, 76000 Rouen, France.
[2]Normandie Univ, UNIROUEN, UNIHAVRE, INSA Rouen, Laboratoire d'Informatique du Traitement de l'Information et des Systèmes, 76000 Rouen, France.

* Correspondence:
Dr. Christophe Dubessy
christophe.dubessy@univ-rouen.fr

Abstract

MicroRNAs (miRNAs) are small non-coding RNAs which regulate gene expression at the post-transcriptional level. Because of their wide network of interactions, miRNAs became over the past decade the focus of many studies. To streamline the amount of potential wet lab experiments, the use of miRNAs targets predictions tools is nowadays the first step undertaken. However, the predictions made are very divergent from one tool to another. This is mostly due to miRNAs complex and still not fully understood mechanism of action. Such divergences bring biologists to wonder about which tool they should use to predict miRNAs targets. To address this issue, the review highlights the main characteristics of miRNA target interaction, describes prediction models currently used, and gives some insights on predictors' performances evaluation.

Introduction

MicroRNAs (miRNAs) are small (~22 nucleotides) non-coding RNAs which act as post-transcriptional 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 shown in genome wide analysis[2,3]. According to miRbase (Release 21), the primary database of published miRNA sequences and annotation, 35 828 mature miRNA products have been identified in 223 species with 2588 of them belonging to humans[4].

miRNAs are mostly transcribed by RNA polymerase II which results in a primary miRNA (pri-miR). This pri-miR is then processed to generate a miRNA precursor (pre-miR) by DROSHA and DGCR8 complex. Afterward, Exportin 5 is responsible for the transport of the pre-miR to the cytoplasm so that it can be further processed by DICER to give the mature miRNA sequence. The biogenesis of miRNAs has been further reviewed in several publications[1,5–9].

Most importantly, each miRNAs can potentially regulate several hundreds of mRNA and one mRNA can be targeted by several miRNAs[9–12]. Because of these numerous possible interactions, miRNAs have a major effect on cellular mechanisms such as proliferation, migration, apoptosis and cell differentiation[13–15]. In 2011, Salmena et al. suggested a concept of miRNAs as mediators of a regulatory language, a way to talk between mRNAs, pseudogene transcripts and long non-coding RNAs. These transcripts have been named “competing endogenous RNAs” (ceRNAs) because their expression level depends on the action of the same miRNAs[16]. This language purpose would be to expend “the functional genetic information in the human genome” and miscommunications are expected to have a major impact in pathophysiologies[17,18]. It is true that differential expression of miRNAs has been observed in many pathologies[19] including cardiovascular[20], neurodegenerative[21], renal[22] diseases and most notably in cancers[23–25]. Therefore, increased knowledge on miRNAs action mechanisms will improve our ability to face these diseases.

Most miRNA targets are repressed at both the post-transcriptional and translational level[26]. miRNAs inhibition process requires the formation of the miRNA-induced silencing complexes (miRISCs) which is mainly composed of the Argonaute (AGO) family proteins, GW182 (glycine-tryptophan (GW) repeat-containing protein of 182 kDa) and processing bodies (P bodies)[27,28]. In most cases, the miRISC induces silencing through a combination of translational repression, deadenylation, decapping and 5′-to-3′ mRNA degradation[29,30].

Obviously, miRNAs are major actors of the epigenetic world and they have become an expanding area of study since 2001. To understand the role of miRNA in genes regulation, one needs first to focus on identifying functional miRNA targets in a pre-defined cellular and environmental context. To do so, the gold standard is to combine luciferase assay, qPCR and western-blot[31]. While a luciferase 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[31,32]. These techniques are very time consuming and allow validation of few interactions at a time. To circumvent this issue, cross-linking and immunoprecipitation approaches coupled with next generation sequencing (CLIP-seq) have been developed. They allow massive discovery of miRNA target interactions (MTIs) without the need for miRNA overexpression but the functionality of the discovered sites remains to be elucidated[31]. Even though improvements were made, many datasets generated by this type of technique contain numerous false positives due to UV crosslinking issues[33]. Experimental procedures being long and expensive, a need for in silico MTI predictions became manifest. Prediction of novel target sites is mostly achieved through building a classification or ranking model which is based on experimentally validated MTI properties (further described below). During the last decade, researchers have experimented with many different computational approaches but a consensus on how to predict MTIs has yet to be found. Currently, more than 160 target prediction tools exist (as of May 2018, from OMICtools' database)[34] which makes it even more difficult to find the one that is best designed for a particular experiment analysis. Computational predictions are plague with high false-positive/negative rates due to the small size and the binding complexity of the MTI sites[35]. Moreover, without a common method to evaluate them, it is no easy task to decide which one to test. Indeed, results lists given by each MTIs prediction algorithms for a given miRNA differ greatly in identified targets, prediction number and ranking[36]. To assist biologists in this choice, we will describe the main characteristics of a MTI as well as different up to date computational methods. The issue of algorithm performances evaluation will also be addressed.

I. Analyzable elements

Even though the mechanisms by which miRNAs act are not fully comprehended, 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.

A. Sequence features

1. Seed region

The main biological feature allowing interaction between miRNA and mRNA is defined as the "seed" region. It includes the nucleotide (nt) 2 to 8 starting from the 5' end of a 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[37–39]. Interestingly, the recognition of an adenine at the miRNA nt 1 favors miRNA-mediated protein down-regulation even when it cannot participate in a Watson-Crick interaction[40]. Seed sites are categorized in different types according to their pairing degree. The hierarchy of site efficacy is as follows: 8mer >> 7mer-m8 > 7mer-A1 >> 6mer or offset-6mer (position 3-8 match) > no site, with the 6mer differing only slightly from no site at all[2,39] (Fig. 1). Microarray experiments suggest that the majority of miRNAs' target sites are 7mer-m8 type[39]. The complexity of using the seed region in targets prediction comes from "bulges" (unpaired stretches of nucleotides located in either of the sequences) or G:U wobbles within the sequence which reduce inhibition efficiency but do not prevent it[37,40]. These sites are named "orphans" or "non-canonical" because AGO family proteins can bind to them even though there isn't a perfect seed match. They were thought to be relatively rare in mammals[2,41–43]. However, newer experimental methods tend to identify a much higher number of non-canonical sites or even sites not binding to the seed region at all (binding to the center of the miRNA or 3' end)[40,41,44–46]. A possible explanation for some of these non-canonical sites is that a "pivot-bulge" on the 6th nucleotide of the seed could enable a transitional nucleation state by stabilizing nucleation base pairing (position 2-6), allowing subsequent bulge formation and propagation of the seed interaction[42,47]. An alternative hypothesis is that non-canonical sites, since they are poorly conserved across species, may act as evolutionary intermediates between non-functional sites and canonical targets sites with a selection pressure going toward the apparition of higher affinity sites[43]. In any case, functional assays indicate a mild regulatory effect of these non-canonical sites[39,42,44]. Therefore, the usefulness of considering both fully and partially matching seed sites to improve MTIs prediction is still under discussion[48,49].

2. Compensation

While most studies consider a "canonical" site to be a full seed pairing without a bulge, miRNAs' target sites can in fact be divided into three groups: "seed" (or "seed only"), "canonical" and "3' compensatory" sites[37]. "Seed" sites, already described in the previous paragraph, have strong 5' pairing but require little or no 3' pairing. Canonical sites however

have both strong seed pairing and compensatory pairing in the 3' side of the miRNA. Finally, "3' compensatory" have weak seed pairing and strong 3' pairing[37]. It is believed that canonical sites are more effective than seed sites only. However, evaluating the effectiveness of 3' compensatory pairing is very difficult due to the number of pairing possibilities and the context dependence of this parameter[10]. Nevertheless, it has been found that additional Watson-Crick pairing of at least 4 nucleotides at position 12–17, especially from 13 to 16, enhances miRNA targeting[39]. This type of strong compensation is very rare (less than 2% of known conserved MTI) but when it occurs, then the target site is usually highly conserved across species[2].

B. Site accessibility

The complexity of miRNA-mRNA interactions makes it difficult for algorithms based on sequence matching only to be efficient. Additional parameters such as thermodynamic, 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.

1. Thermodynamic

The most basic way to consider thermodynamic is to calculate the free energy that estimate 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 feasible interaction. Since we are in the context of miRNA interaction, constraints imposed by the seed pairing must be taken into consideration. The ViennaRNA R package is the most commonly implemented to calculate the free energy of binding. It aggregates more than 20 programs/packages to solve the structure of a RNA duplex using dynamic programming[50]. Rehmsmeier et al. found that forbidding intramolecular base pairing and bulge loops seem to give a better free energy estimation[51]. They also noted that taking several nucleotides (10 and more) flanking the target site improves correlation between energy based scores and target repression[51,52]. Another possibility is to consider the hybridization energy ($\Delta\Delta G$) 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 $\Delta\Delta G$ score is then successfully correlated with the degree of miRNAs' targets repression for some interactions but not all[52].

2. Target site context

Messenger RNAs can fold into highly elaborated secondary and tertiary structures, and a perfect sequence match for a miRNA might not be structurally accessible for binding. Therefore,

context features such as 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. Of all context features, the AU content around target site is the one which favors most the interaction with a miRNA[10]. Indeed swapping a target site from an open (AU rich) UTR structure to a close one decreases site functionality[52]. A possible explanation for this is that AU-rich sequences could be recognized directly by a component of the RISC or reducing the tendency for formation of stable RNA secondary structures that could interfere with RISC binding[53]. Although there is so far a high prevalence of MTI sites found in the 3'UTR, recent papers have shown that some miRNAs can also regulate mRNAs by binding with the 5'UTR and CDS region of their targets[54,55]. Whether the site is in the 3' or 5'UTR seems to have no impact on the strength of the mRNA regulation. However, ORF target sites are not as efficient[40,56,57]. Interestingly, a recent study showed that if CDS located target sites are not as efficient to trigger mRNA degradation, they are quite potent at inhibiting translation[58]. Remarkably, some studies have shown that under different cellular conditions, miRNA–mRNA interactions with different binding sites or/and cellular localizations can increase mRNA translation[59–61]. However, the precise mechanism by which a miRNA can enhance protein synthesis has yet to be fully discovered. Thus, it is clearly important not to restrict the search to the 3'UTR for MTI predictions. Aside from the localization, the number of repetitions of a target site and their spacing on a given mRNA also affect the repressing efficiency of a miRNA[53,62]. Another important aspect to determine the functionality of an interaction, yet rarely taken into consideration, is the expression level of both miRNAs and targeted mRNAs[63]. Moreover, depending on the tissue or disease, a validated MTI can be more or less functional[64,65]. This might be due to RNA-binding proteins blocking access to the miRNA or mRNA secondary structure in that particular tissue or disease[65,66]. A screen for RNA-binding protein motifs and considering the sample tissue should therefore improve MTIs prediction.

C. Conservation

The level of conservation of a sequence represents its presence across species. 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. There is often a higher conservation in the seed region of the target site than out of it. Moreover, it is generally only the degree of 3' pairing that is conserved but not the nucleotide identity. Assuming that aligned sites within orthologous genes have a single origin,

it was proposed to quantify site conservation in a phylogenetic tree by summing all branches length for which the site is present[67].

Of note, the level of conservation of a target site has to be done with regard of the conservation of its mRNA region and length[2]. Stronger conservation profile has been associated with increased mRNA down-regulation using microarray experiments and better MTI prediction[2,37,40,53,68,69]. Indeed, over 60% of human protein-coding genes are conserved targets of miRNAs, supporting the importance of this parameter[2]. However, since functional non-conserved MTIs exist and mediate protein translation inhibition[70], sites cannot be filtered based on conservation criteria. Moreover, Agarwal et al. also found a decrease of performance of their predictor when considering only highly conserved sites [48]. Therefore, an ideal equilibrium needs to be found where conserved sites are favored and non-conserved sites are retained. Friedman et al. found 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[39]. A possible explanation for this result is that these sites are inactive (or less active) decay products of conserved 7–8mer sites. An alternative possibility is that when binding with a 6mer, a 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].

II. Computational prediction methods

As mentioned in the Introduction chapter, many computational tools have been developed in the field of MTIs prediction. The main objective of prediction algorithms is to select the most discriminative features, within the categories of analyzable elements described above, and to find how to compute them to get a better prediction accuracy.

A. Sequence based

1. Heuristic scoring models

The earliest attempt to identify in silico miRNA targets was published by Stark et al. in 2003[71]. Their screening was a simple two steps procedure combining sequence comparison with HMMer (alignment tool) and site accessibility using Mfold. The resulting targeted 3'UTR were then compared based on their conservation between *Drosophila pseudoobscura* and *Anopheles gambiae*. Using this protocol, they successfully validated 6 MTIs for 2 Drosophila miRNAs. After analyzing the characteristics of these 6 validated interactions, they started to describe what we now know as the seed region: the first eight nucleotide at the 5'end of the miRNA[71].

Following this publication, many more attempts have been made to improve and generalize MTI prediction. The vast majority of predictors utilize the seed-matching parameter since most of the reported functional MTI have a 6mer or more. To do so, predictors either filter sequences based on a defined set of rules for seed-matching[48,72] or use a score system favoring this feature[12,73,74]. However, filtering based on seed rules seems too stringent because functional MTIs can also have non-canonical seed (G:U wobble or bulge). In this regard, some methods consider binding of the first eight nucleotides as important but do not restrict it to particular seed types[75–78]. MIRZA-G (evolution of MIRZA[79]) for instance, is a recently published algorithm that allows non-perfect seed matches if the final score for the site is above the author defined threshold[69]. Rare are the predictors, like RNA22[3], that do not consider seed-matching at all in its prediction. Instead, it searches the mRNA for patterns generated by comparing all known mature miRNA sequences (as of 2006) and keeps only the significantly similar ones. Sequence alignment results are almost always complemented with site accessibility and evolutionary inputs. Tools such as miRanda[74], RNA22[3], and TargetScan[80] make use of RNA folding prediction software, like RNAVienna[50] or Mfold[81] packages, to estimate free energy of predicted miRNA–target duplexes and filter out candidates above a certain threshold. Interestingly, authors of RNAhybrid[51] used a different approach preventing intramolecular base pairing and bulge loops which seems to improve the estimation of the free energy[51]. This led predictors such as PicTar[12] and STarMir[82,83] to filter potential target sites based on results from RNAhybrid. As mentioned before, other predictors, such as PITA[52], 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 one since it has been shown to favor MTI[48,75,76,78,84,85]. The frequency of target sites along the mRNA and the distance separating them are two other features often considered for target site context implementation[80,86,87]. The value of site conservation is quite argued since omitting non-conserved targets is irrelevant and not using this parameter at all decreases drastically the specificity of the method[2,76,80,88]. This has been widely studied by the authors of EIMMo[72] who score MTIs based on conservation criteria only and then use Bayesian statistics to infer functionality. This makes EIMMo quite efficient at predicting the mRNAs targeted by a given miRNA but not as sensitive at the duplex level[89]. Features implementation for all algorithms cited so far has been done based on literature knowledge only. To better identify what combination of features to use, miRmap's authors decided to evaluate each feature individually before integrating them. They first screened all human transcripts for 7mer seeds and compared the performance of eleven features mentioned previously on results from seven miRNA

overexpression experiments coming from five studies. Based on this evaluation, they combined these features using a linear regression model, making it the most comprehensive MTI predictor of the time[84]. Likewise, TargetScan evaluated 26 features and eventually selected 14 of them to upgrade itself with a similar model in 2015[48]. Most algorithms end-up storing all resulting interactions in a publicly available database format like miRWalk2.0[48,90].

2. Empirical machine learning models

The limit of rule-based method comes from the complexity of MTIs. It is extremely difficult for a human being to take into consideration all possible aspects of these interactions. Thus, another promising direction toward better MTI prediction is a data driven (or machine learning, ML) algorithm. There are many computational models available to build such an algorithm. Unfortunately, there is no fixed rule as to which one to select 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 classifier (categories) to differentiate functional from non-functional 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, as well as the local computational restrictions (time and memory). Computational constraints depend mostly on the number of features used[91]. Since a 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 while, of course, being free 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 features compromise to obtain a generalization model[92] able to classify a MTI in a binary fashion or according to a scoring method. Features are ranked by a metric system like F-score (harmonic mean between precision and recall) or correlation coupled with statistics and the top-ranked ones are selected to build the algorithm. This procedure is known as features extraction. To validate their approaches, most authors use a 10-fold cross validation technique. In other words, a subset of the dataset is used for training the algorithm and the other part for testing it. This is done 10 times using different partitions of the original dataset and the performance results are averaged over the rounds.

a) Genetic programming

Genetic programming is a ML method which generates functions (represented as trees) using the different rules or features implemented in order to best describe a positive interaction[93,94] (Fig. 2). One of the first ML model developed with this method was TargetBoost in 2005[94]. This model is one of the rare type 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 nucleotides as negative examples. These motifs are then weighted with a boosting algorithm which 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[93]. 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. No features extraction done in this model nor did they apply a conservation or site density filter. The data from 3 miRNAs were used to train the model and tested it on the data of another miRNA using the "leave one out" method. TargetBoost compared itself to RNAhybrid and another algorithm named nucleus and was either as good as each of them or more performant depending on the dataset used for testing[95]. 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 C4.5 decision tree instead of boosting[96]. The output of C4.5 algorithm results in several rule sets to take as inputs for the genetic algorithm. To begin, their algorithm was trained and tested on a small dataset, taken from TarBase database (version 3.0) and containing 48 positive and 16 negative examples[97–99]. They obtained a 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.[100]) containing 113 positive and 312 negative examples and therefore, they obtained 97% accuracy. Authors relate 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 making it more likely for this algorithm to perform well.

b) Probabilistic based classifier

A commonly used method is to model the relationship between features and the output categories using probabilities with a Naïve 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). A MTI is then classified based on the product of all features probabilities[91] (Fig. 3).

NBmiRTar[101] is an example of such a probabilistic machine learning method. Using both 'seed' and 'out-seed' features, they applied the NB classifier on predictions from miRanda taking its scoring and free energy calculation as filters. They used the same dataset of 3,000 random 30 nt strings for negative examples as 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 non-conserved MTIs, authors did not use sequence conservation in their model, which has the inconvenience of generating 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 would need to be tested on more than 13 positives targets. Also using a Bayesian probabilistic method, GenMiR3[102] (an evolution of GenMiR++[103]) considers the hybridization energy, target site conservation (PhastCons algorithm[104]) and context information (5 sequence features) to establish a prior probability for the target site to be functional. They tested the performance of each features using multiple linear regression models and cross-validation. Hybridization energy seems to be the feature which enhanced the most 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, even though they restricted their training data to colorectal cancer MTIs, CRCmiRTar[105] authors compared different ML approaches (NB, SVM, Random forest (RF), Artificial Neural Network (ANN)) and found NB classifier to be the most sensitive and specific method. This algorithm also showed to be more efficient when compared to other tools on an independent colorectal cancer-specific test dataset. The tissue origin of the samples seems therefore to be a parameter that should be included in MTI predictions.

Another probabilistic model in use for MTIs prediction is the Random Forest (RF) classifier. Each tree of the forest is a predictor which 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 and the majority defines the predicted class for this example[106] (Fig. 4). The mechanism used to grow the trees allows to easily estimate the most important set of features and is also easily interpretable. An example of such model is RFMirTarget[107]. They used the dataset published by Bandyopadhyay and Mitra[85] containing 289 experimentally validated functional pairs and 289 "systematically identified tissue-specific negative examples" to train a RF classifier. Since no site alignment is given in this dataset, they used miRanda to define potential MTI sites sequences and alignments. After testing, their model proved to be more efficient on

their training set than other types of machine learning (support vector machines and NB based) and was able to identify more positive targets than TargetSpy 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 has been developed[64]. 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 instance is labeled positive, then the bag is labeled as functional. Since authors of this algorithm deem 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, they did not make a comparison with RFMirTarget which is the closest related method to MBSTAR. Recently, authors of TarPmir decided to use CLASH (crosslinking, ligation, and sequencing of miRNA-RNA hybrids) data, a new high-throughput experimental method to identify MTIs, to train a RF-based model for MTI predictions[108]. The advantage of CLASH compare 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[44]. Since no other CLASH dataset was available at the time, the performances of this method has been tested on three independent PAR-CLIP datasets. Validated MTIs were identified using DIANA-TarBase (v7.0)[97]. Even though TarPmir came out better than three other commonly used algorithms, it still only achieved 55% recall and 19% precision, leaving much space for improvement. However, since CLASH data includes many "non-seed" MTIs, TarPmir can better predict most sites of this type.

c) *Support vector machines*

Support Vector Machines (SVMs) are machine learning algorithms made 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 end-up 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[91] (Fig. 5). Most SVMs used for MTI prediction are non-linear and based on a similarity function, called a kernel, between pairs of samples (miRNA:mRNA)[75,77,78,85,109]. MiTarget was one of the first algorithm to implement a SVM to predict MTI and showed equal performances than popular predictors such as miRanda, TargetScan or RNAhybrid[78]. Interestingly, SVMicrO implemented

two SVMs, one for site and one for UTR-related features[75]. Naturally, the most important features of the site-SVM are seed based but surprisingly, the 3' context region conservation 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 not using it at all and some showing it as an important parameter or not[46,75,78,109,110]. As for the UTR-SVM of SVmicrO, predictions result mainly from the number of positive sites in the UTR (the more the better) and the score of each of these sites (the higher the better) as well as the length of the UTR. SVMicrO showed overall better performance than Pictar, miRanda, mirTarget, TargetScan and PITA. Another SVM approach, MiREE, proposed a hybrid solution by combining genetic programming for the miRNA duplex characteristics (sequence homology and thermodynamic) and a non-linear SVM for the context features[77]. Just like SVMicrO, its most important features are seed related. This method obtained a 95% accuracy on human MTI predictions which is higher (2nd best is miTarget with >60%) than the other compared methods in this review. Surprisingly, Aviskar, a recently published predictor used a linear SVM model because it has the advantage of being directly interpretable from the weights of each feature and for its ease of implementation[46]. However, as mention above, this type of machine learning is expected to perform poorly due to the complexity of MTIs. As a result, even though Aviskar obtained a 98% recall on human MTI, it showed poor accuracy with 30% of all predicted targets being misclassified. Interestingly, Li et al. proposed to improve the performance of miRNA target prediction by searching a second MTI on the whole mRNA sequence after finding one in the 3'UTR[111]. Thus, they trained a SVM on a two sites search dataset of validated MTIs from miRecords and pSILAC (quantitative proteomics) experiment. When tested on an independent dataset, it showed higher performance than other commonly used methods (PicTar, MirTarget2, miRanda, PITA, TargetSpy, TargetMiner, and TargetScan). Trying to improve both the prediction model and the training dataset, Lu et Leslie created chimiRic, a two SVMs model based on CLASH and AGO-CLIP sequencing data[112]. One SVM uses both data types for duplex prediction and the other one serves for AGO sites discrimination (true or not). This strategy has the advantage of training on a large dataset of interacting miR-target duplexes but without any guaranties on their functionality. Nevertheless, it shows superior performances to MIRZA, MirTarget, TargetScan, miRanda and Diana-microT-CDS.

d) Artificial neural networks

Artificial Neural Networks (ANNs, also called neural networks) systems have been developed in the same idea than interconnected neurons in the brain. 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, to eventually return a prediction. 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 neurons ordering can change to best suit the training data[113] (Fig. 6). One of the first MTI prediction method using an ANN was MTar[114]. Unlike most of the algorithms of the time who heavily focus on seed region matching, MTar aimed to efficiently identify MTIs no matter the type of interaction. It first calculates a complementarity score to decide in which of these 3 categories (determined from Betel et al., 2010) the site falls: 5'seed-only, 5'dominant and 3'canonical. 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% less targets for each miRNA compared to conventional methods with a 94.5% sensitivity and 90.5% specificity. Using a very similar model to the one of MTar, HomoTarget uses a pattern recognition neural network (PRNN) coupled with a principle component analysis (PCA) for features selection[100]. It contains 16 input nodes, 14 neurons in the hidden layer and 2 units in the output layer. Unlike MTar, HomoTarget is focusing on the seed region to predict MTIs since it filters sequences based on standard seed rules. HomoTarget was trained on a 425 examples dataset and showed a 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.

e) Training datasets

As mentioned above, a good training dataset needs to have a high amount of high quality examples. The training dataset truly is a critical aspect of all machine learning methods. A difficult challenge in creating a 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[12,78,80,94,101,115]. TargetMiner's authors (who later also created MBSTAR) especially emphasized this issue[85]. 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, in a given tissue, a miRNA and its potential targeted mRNA were both overexpressed, then this pair was retained as a negative example. Using this method, 289 negative MTI were generated. A subset of negative examples were then confirmed on a separate pSILAC dataset[26]. To complete the

dataset, 289 experimentally validated positive sites were retrieved from miRecords and TarBase[97,98,116]. Using an independent dataset (187 positive and 59 negative pairs), TargetMiner showed a 74% accuracy when NBmiRTar and MirTarget2 only had 51% and 46% respectively (lower than in their original publications) clearly showing the importance of the testing dataset on the 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 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[46,109,110]. The development of high throughput methods brought the tendency to include the most amount of examples regardless of the lack of functional testing. Being used by many predictors, several datasets marked the history of MTIs prediction such as the one published 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)[26,41,117,118]. As mentioned in introduction, miRNA targets are not necessarily repressed at the mRNA level, making microarray data not sufficient to fully encompass the action of a miRNA. Use of complementary proteomics data is strongly suggested in this case. Moreover, under-expressed mRNA/protein levels measured by high throughput experiments can be due to secondary effects of miRNA regulation[119]. Recently, some predictors were trained on CLASH experiments which identifies both AGO binding miRNA and target sites on a transcriptome-wide scale. However, one needs to be careful with CLASH data as the specificity of the ligation and the exhaustivity of the captured MTIs are questionable[33,112]. At the moment, as difficult and expensive as it might be to acquire the data, combining all these technologies (CLIP-seq, CLASH, microarray and pSILAC) seems the best solution to be able to rely on large training datasets.

3. Popular prediction tools

When published, most if not all prediction algorithms compare themselves to miRanda, Diana-microT-CDS and/or TargetScan. This is because biologists have mostly been using these 3 heuristic scoring methods to identify MTIs before going for wet-lab experiments. Their popularity is mostly due to their age, frequent updates and a strong adaptation ability to new advances in MTIs prediction.

In the direct foot-steps of the Stark method, 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 assigning values for each nucleotide pairing with higher scores for seed matching[74]. Site conservation is also included in the tested features and 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 done by a support vector regression algorithm (mirSVR) based on mRNA expression change[76,120]. They trained mirSVR on a set of nine microRNA transfection experiments performed on HeLa cells from Grimson et al[39]. The score resulting from mirSVR is intended to estimate the efficiency of miRNA regulation on a given target site and not the probability of regulating this site. With this model as well, authors found that the most important features are related to the seed region. The mirSVR upgrade showed significant better performances than the previous version of miRanda and seems slightly above TargetScan[76].

Diana-microT is an algorithm published in 2004 that first searches for the miRNA-recognition elements (MREs), which include Watson-Crick pairing identification and minimum binding energy calculation using 38 nucleotides window, in the 3'UTR of a mRNA. A second parameter takes into account the miRNA-associated proteins complex which impacts both pairing between the miRNA and its target and site accessibility[121]. In 2009, microT was updated to filter MREs that do not have at least a 7mer in the seed region. Authors also decided to integrate 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 results with a set of mock miRNAs. An enrichment analysis is also done with all potential MREs for a given miRNA using KEGG pathways database. Results are highlighted in the significantly identified pathways[122]. In 2012, the algorithm was renamed DIANA-microT-CDS because numerous studies had shown that the coding region of a mRNA can be targeted by a miRNA with measurable effect on its degradation. Therefore, microT now screens 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 nucleotides window on the 3'-UTR or CDS. The prediction method scores differently the 3'UTR and CDS region and then combined these scores to create the final estimation for the whole mRNA[123]. This last update showed better performance than miRanda and TargetScan at the time of the publication (2012).

Released as a freely available web-tool in 2003 by Bartel's group, TargetScan first used conservation of miRNAs and mRNA UTR as a filter and then seed matching (length and

frequency), 3' compensation and folding free energy as prediction features[80,124]. The algorithm progressively evolved (last version: v7.0, 2015) to take into consideration all analyzable elements of MTI previously described[2,39,48,124–126]. 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[125]. The resulting models were collectively called the context++ model. When multiple sites are present, individual context++ scores are summed to rank that predicted 3'UTR. Over the years, site conservation became one of the features of TargetScan instead of being used as filter. With a relatively weak contribution to the context++ score, non-conserved targets can even make it to the top predictions. After thoroughly analyzing CLIP datasets, TargetScan authors concluded that "non-canonical sites might exist but have not yet been characterized to the point that they can be used for miRNA target prediction" and they therefore did not include these sites into their predictions[48]. 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 found no better performances compare to linear regression model[48]. This result is consistent with a similar test done by Vejnar et al. in 2012[84]. The most recent version for TargetScan (2015) showed better performance than 15 other predictors (miRanda and microT included) when tested on the dataset from Linsley et al[118]. With 8 publications describing its content and updates, TargetScan is so far the most widely used MTI prediction tool by the scientific community (3180 citations from web of science core collection as of Mai 2018)[89,127,128].

B. Data combination

Due to the small overlap of results (5-70%) between all previously cited methods[129], researchers often combine all results from different prediction tools to strengthen the likelihood of studying true positive MTIs. Several strategies to combine MTI predictions have been proposed.

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 popular tools using statistics and/or mRNA/protein expression data[90,130–133]. It is with such an intersection strategy that Kuhn et al. validated the interaction of the human angiotensin II type 1 receptor (hAT1R) with miR-155 leading them to suggest to cross results between at least two MTI predictors before going for experimental investigations[134]. 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[35]. Therefore, intersecting results do not increase the probability of retaining true positives. Moreover, approaches based on intersection of predictions may lead to decreased sensitivity because of possibly omitting valid interactions as shown by Sethupathy et al[135]. This is supported by Oliveira et al.[136] who showed that the union of the results from several prediction tools was more efficient than the intersection. However, when ranking of 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 purpose). Nevertheless, these databases have the advantage of giving a wide panel of predictions for a given miRNA with an edge for miRWalk which has been recently updated. However, most users have not enough understanding of MTI predictions to decide which database to take or remove from the union and intersections strategies to be efficient.

2. Ensemble methods

Because of the limits of the intersection strategy, others have used the union with a rescoring method to better rank MTIs according the likelihood of being true. It was first investigated by DeConde et al. in 2006 with an algorithm that combines ranked lists of miRNA targets from five microarray studies and re-rank the targets using a statistical test proposed by Tusher et al[137]. Performances of this method compared to other tools was not evaluated. While this was done on experimental data, other methods have used aggregation strategies on predicted MTIs from several popular tools. It is the case of MiRror-Suite which gathered predicted and/or validated MTIs from 18 databases making it possible to analyze about 40 000 genes and 2500 miRNAs[138]. The aggregation strategy consisted in creating a set of potential targets using a several filters (species, miR family, cell line, number of databases etc.) and then calculating the probability of a MTI to be functional based on a hypergeometric test. However, its ranking performances were not compared to 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 clearly out-performed aggregated methods[139]. The good performances of similar combination approaches were also confirmed with a model that aggregates 9 predictive algorithms[140]. Others, like BCmicrO and ComiR, have used more complex strategies for the combination step with a NB classifier for BCmicrO and a SVM for ComiR[141,142]. Interestingly, ComiR takes into consideration inputted miRNAs expression levels in its rescoring methods. Of note, ComiR was especially designed to predict the targets of a set of miRNAs and to consider combinatory interactions. As expected, all aggregation methods

were able to outperform, in term of MTI ranking, each aggregated database taken individually. This was also confirmed with the aggregation method miRabel (soon to be published) using a very large dataset (982 411 common interactions). MiRabel uses a statistic R package (RobustRankAgreg) to rescore each MTI from their ranks in 3 databases (miRanda, PITA and SVMicrO). 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)[143].

III. Performances evaluation

Since the prediction tools are designed for biologists, the ease of use should be a criterion in the overall performances. These tools usually come in 3 different platform usage: web-service, downloadable programs or R/python packages. The first kind is the most used because of its user-friendly aspect. However, ease of use being generally inversely proportional to flexibility, it also offers the least amount of freedom in sequence analysis[127].

It is common to consider state of the art tools which harbor a greater correlation between their predictions and protein or RNA downregulation[144]. However, that would be the case if the downregulation is directly due to the miRNA transfection which is far from certain in high throughput experiments. 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[145]. It plots the sensitivity or True Positive Rate (TPR) against specificity or False Positive Rate (FPR) with TPR = TP/(TP+FN) while FPR = FP/(FP+TN). An MTI is considered to be a True Positives (TP) if it has been predicted and 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 it is sadly not the same for tested but not validated interactions. Therefore, in the case of MTI prediction, a non-negligible part of FPs and TNs are mislabeled creating biases in ROC analysis [140]. 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)[146]. An alternative is to plot the cumulated precision versus the normalized scores (sorted in descending order)[140]. Both methods have the advantage of not taking TN in

consideration which minimized the number of mislabeled MTI in the analysis. The problem is not completely solved however since these methods are still depending on FP. The use of both ROC and PR analysis is thus recommended for complete performance evaluation of a MTI prediction tool.

Unfortunately, not all reviews or published algorithms use the same type of measurement to evaluate performances which makes publications nearly impossible to be compared. A common mistake, which tends to disappear nowadays, is to use the training dataset to evaluate prediction performances. Indeed, using several datasets for truly evaluating predictors' performances 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[89,133,147]. Using all measurements addressed above and more, Fan et Kurgan[89] compared 7 target predictors with 4 testing datasets. Even though TargetScan and miRmap looked the strongest in this review, there was no consistent best predictor across all possible measurements. Of note, TargetScan performs systematically well across the vast majority of studies comparing MTIs prediction algorithms, closely followed by Diana-microT-CDS and miRanda-mirSVR.

Despite the increasing enthusiasm for the field of MTI prediction, much improvements remain to come. MTI prediction is a complex challenge and overcoming it will necessarily reside in a closer concertation between multidisciplinary teams. Nevertheless, there is no doubt that future studies on MTI prediction will eventually bring us a greater ability to quickly identify major contributors to the epigenetic network and therefore a better understanding of human physiology.

IV. Conclusion

All prediction algorithms use a combination of sequence, site accessibility and conservation features to identify potential MTIs. However, since the mechanisms of miRNAs action are not yet fully understood, predictors still have a high false positive rate. To improve accuracy, different computational methods have been tested. None so far have shown consistently better performances. Surprisingly, empirical methods do not seem to perform better than heuristic methods suggesting that actual training datasets do not efficiently capture all possible MTI examples. There is clearly a great need for standardizing methods to compare algorithms. Overall, 3 predictors, TargetScan, miRanda and Diana-microT seem to perform well across benchmarking reviews. Until better algorithms come to be developed, ensemble methods seem

to be the most efficient strategies to get an integrated vision of the target predictions for a given miRNA. Ultimately, efficient MTI prediction will reduce the time and resources spent validating miRNA targets and therefore increase the speed at which molecular biologists elucidate the role of miRNAs in healthy and pathological conditions.

V. References

1. Bartel, D. P. MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. *Cell* **116,** 281–297 (2004).
2. Friedman, R. C., Farh, K. K.-H., Burge, C. B. & Bartel, D. P. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.* **19,** 92 (2009).
3. Miranda, K. C. *et al.* A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. *Cell* **126,** 1203–1217 (2006).
4. Kozomara, A. & Griffiths-Jones, S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* **42,** D68-73 (2014).
5. Catalanotto, C., Cogoni, C. & Zardo, G. MicroRNA in Control of Gene Expression: An Overview of Nuclear Functions. *Int. J. Mol. Sci.* **17,** (2016).
6. Lin, S. & Gregory, R. I. MicroRNA biogenesis pathways in cancer. *Nat. Rev. Cancer* **15,** 321–333 (2015).
7. Ha, M. & Kim, V. N. Regulation of microRNA biogenesis. *Nat. Rev. Mol. Cell Biol.* **15,** 509–524 (2014).
8. Oliveto, S., Mancino, M., Manfrini, N. & Biffo, S. Role of microRNAs in translation regulation and cancer. *World J. Biol. Chem.* **8,** 45–56 (2017).
9. Karbiener, M., Glantschnig, C. & Scheideler, M. Hunting the Needle in the Haystack: A Guide to Obtain Biologically Meaningful MicroRNA Targets. *Int. J. Mol. Sci.* **15,** 20266–20289 (2014).
10. Bartel, D. P. MicroRNA Target Recognition and Regulatory Functions. *Cell* **136,** 215–233 (2009).

11. Hamzeiy, H., Allmer, J. & Yousef, M. Computational methods for microRNA target prediction. *Methods Mol. Biol. Clifton NJ* **1107,** 207–221 (2014).
12. Krek, A. *et al.* Combinatorial microRNA target predictions. *Nat. Genet.* **37,** 495–500 (2005).
13. Dror, S. *et al.* Melanoma miRNA trafficking controls tumour primary niche formation. *Nat. Cell Biol.* **18,** 1006–1017 (2016).
14. Li, Y. *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. Biokhimiia* **82,** 474–482 (2017).
15. 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.* doi:10.1016/j.biopha.2015.07.037
16. Li, Y. *et al.* Systematic review of computational methods for identifying miRNA-mediated RNA-RNA crosstalk. *Brief. Bioinform.* (2017). doi:10.1093/bib/bbx137
17. Salmena, L., Poliseno, L., Tay, Y., Kats, L. & Pandolfi, P. P. A ceRNA hypothesis: the Rosetta stone of a hidden RNA language? *Cell* **146,** 353–358 (2011).
18. Su, X. *et al.* microRNAs and ceRNAs: RNA networks in pathogenesis of cancer. *Chin. J. Cancer Res.* **25,** 235–239 (2013).
19. Maqbool, R. & Hussain, M. U. MicroRNAs and human diseases: diagnostic and therapeutic potential. *Cell Tissue Res.* (2014). doi:10.1007/s00441-013-1787-3
20. Bronze-da-Rocha, E. MicroRNAs Expression Profiles in Cardiovascular Diseases. *BioMed Res. Int.* **2014,** (2014).
21. Basak, I., Patil, K. S., Alves, G., Larsen, J. P. & Møller, S. G. microRNAs as neuroregulators, biomarkers and therapeutic agents in neurodegenerative diseases. *Cell. Mol. Life Sci. CMLS* **73,** 811–827 (2016).

22. Szeto, C.-C. & Li, P. K.-T. MicroRNAs in IgA nephropathy. *Nat. Rev. Nephrol.* **10,** 249–256 (2014).

23. Di Leva, G., Garofalo, M. & Croce, C. M. microRNAs in cancer. *Annu. Rev. Pathol.* **9,** 287–314 (2014).

24. Oom, A. L., Humphries, B. A. & Yang, C. MicroRNAs: Novel Players in Cancer Diagnosis and Therapies. *BioMed Res. Int.* **2014,** (2014).

25. Cheng, Q., Yi, B., Wang, A. & Jiang, X. Exploring and exploiting the fundamental role of microRNAs in tumor pathogenesis. *OncoTargets Ther.* **6,** 1675–1684 (2013).

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

27. Parker, R. & Sheth, U. P bodies and the control of mRNA translation and degradation. *Mol. Cell* **25,** 635–646 (2007).

28. 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* **11,** 1640–1647 (2005).

29. Jonas, S. & Izaurralde, E. Towards a molecular understanding of microRNA-mediated gene silencing. *Nat. Rev. Genet.* **16,** 421–433 (2015).

30. Fabian, M. R., Sonenberg, N. & Filipowicz, W. Regulation of mRNA translation and stability by microRNAs. *Annu. Rev. Biochem.* **79,** 351–379 (2010).

31. Chou, C.-H. *et al.* miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. *Nucleic Acids Res.* **44,** D239–D247 (2016).

32. Campos-Melo, D., Droppelmann, C. A., Volkening, K. & Strong, M. J. Comprehensive Luciferase-Based Reporter Gene Assay Reveals Previously Masked Up-Regulatory Effects of miRNAs. *Int. J. Mol. Sci.* **15,** 15592–15602 (2014).

33. Broughton, J. P. & Pasquinelli, A. E. Identifying Argonaute binding sites in Caenorhabditis elegans using iCLIP. *Methods San Diego Calif* **63,** 119–125 (2013).

34. Henry, V. J., Bandrowski, A. E., Pepin, A.-S., Gonzalez, B. J. & Desfeux, A. OMICtools: an informative directory for multi-omic data analysis. *Database J. Biol. Databases Curation* **2014,** (2014).

35. Ritchie, W., Flamant, S. & Rasko, J. E. J. Predicting microRNA targets and functions: traps for the unwary. *Nat. Methods* **6,** 397–398 (2009).

36. Sedaghat, N., Fathy, M., Modarressi, M. H. & Shojaie, A. Combination of Supervised and Unsupervised Approaches for miRNA Target Prediction. *IEEE/ACM Trans. Comput. Biol. Bioinform.* (2017). doi:10.1109/TCBB.2017.2727042

37. Brennecke, J., Stark, A., Russell, R. B. & Cohen, S. M. Principles of MicroRNA–Target Recognition. *PLoS Biol.* **3,** (2005).

38. Didiano, D. & Hobert, O. Perfect seed pairing is not a generally reliable predictor for miRNA-target interactions. *Nat. Struct. Mol. Biol.* **13,** 849–851 (2006).

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

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

41. Chi, S. W., Zang, J. B., Mele, A. & Darnell, R. B. Ago HITS-CLIP decodes miRNA-mRNA interaction maps. *Nature* **460,** 479–486 (2009).

42. Chi, S. W., Hannon, G. J. & Darnell, R. B. An alternative mode of microRNA target recognition. *Nat. Struct. Mol. Biol.* **19,** 321–327 (2012).

43. Seok, H., Ham, J., Jang, E.-S. & Chi, S. W. MicroRNA Target Recognition: Insights from Transcriptome-Wide Non-Canonical Interactions. *Mol. Cells* **39,** 375–381 (2016).

44. Helwak, A., Kudla, G., Dudnakova, T. & Tollervey, D. Mapping the Human miRNA Interactome by CLASH Reveals Frequent Noncanonical Binding. *Cell* **153,** 654–665 (2013).

45. Moore, M. J. *et al.* miRNA–target chimeras reveal miRNA 3′-end pairing as a major determinant of Argonaute target specificity. *Nat. Commun.* **6,** (2015).

46. Ghoshal, A., Shankar, R., Bagchi, S., Grama, A. & Chaterji, S. MicroRNA target prediction using thermodynamic and sequence curves. *BMC Genomics* **16,** (2015).

47. Stefani, G. & Slack, F. J. A 'pivotal' new rule for microRNA-mRNA interactions. *Nat. Struct. Mol. Biol.* **19,** 265–266 (2012).

48. Agarwal, V., Bell, G. W., Nam, J.-W. & Bartel, D. P. Predicting effective microRNA target sites in mammalian mRNAs. *eLife* **4,** (2015).

49. Friedersdorf, M. B. & Keene, J. D. Advancing the functional utility of PAR-CLIP by quantifying background binding to mRNAs and lncRNAs. *Genome Biol.* **15,** R2 (2014).

50. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms Mol. Biol. AMB* **6,** 26 (2011).

51. Rehmsmeier, M., STEFFEN, P., HÖCHSMANN, M. & GIEGERICH, R. Fast and effective prediction of microRNA/target duplexes. *RNA* **10,** 1507–1517 (2004).

52. Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U. & Segal, E. The role of site accessibility in microRNA target recognition. *Nat. Genet.* **39,** 1278–1284 (2007).

53. Nielsen, C. B. *et al.* Determinants of targeting by endogenous and exogenous microRNAs and siRNAs. *RNA* **13,** 1894–1910 (2007).

54. Moretti, F., Thermann, R. & Hentze, M. W. Mechanism of translational regulation by miR-2 from sites in the 5′ untranslated region or the open reading frame. *RNA* **16,** 2493–2502 (2010).

55. Qu, H. *et al.* microRNA-558 facilitates the expression of hypoxia-inducible factor 2 alpha through binding to 5′-untranslated region in neuroblastoma. *Oncotarget* **7,** 40657–40673 (2016).

56. Gu, S., Jin, L., Zhang, F., Sarnow, P. & Kay, M. A. The biological basis for microRNA target restriction to the 3' untranslated region in mammalian mRNAs. *Nat. Struct. Mol. Biol.* **16,** 144–150 (2009).

57. Lytle, J. R., Yario, T. A. & Steitz, J. A. 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.* **104,** 9667–9672 (2007).

58. Hausser, J., Syed, A. P., Bilen, B. & Zavolan, M. Analysis of CDS-located miRNA target sites suggests that they can effectively inhibit translation. *Genome Res.* **23,** 604–615 (2013).

59. Niepmann, M. Activation of hepatitis C virus translation by a liver-specific microRNA. *Cell Cycle Georget. Tex* **8,** 1473–1477 (2009).

60. Ørom, U. A., Nielsen, F. C. & Lund, A. H. MicroRNA-10a Binds the 5′UTR of Ribosomal Protein mRNAs and Enhances Their Translation. *Mol. Cell* **30,** 460–471 (2008).

61. Ni, W.-J. & Leng, X.-M. Dynamic miRNA–mRNA paradigms: New faces of miRNAs. *Biochem. Biophys. Rep.* **4,** 337–341 (2015).

62. Sætrom, P. *et al.* Distance constraints between microRNA target sites dictate efficacy and cooperativity. *Nucleic Acids Res.* **35,** 2333–2342 (2007).

63. Shu, J. *et al.* Dose-dependent differential mRNA target selection and regulation by let-7a-7f and miR-17-92 cluster microRNAs. *RNA Biol.* **9,** 1275–1287 (2012).

64. Bandyopadhyay, S., Ghosh, D., Mitra, R. & Zhao, Z. MBSTAR: multiple instance learning for predicting specific functional binding sites in microRNA targets. *Sci. Rep.* **5,** 8004 (2015).

65. Erhard, F. *et al.* Widespread context dependency of microRNA-mediated regulation. *Genome Res.* **24,** 906–919 (2014).

66. Ciafrè, S. A. & Galardi, S. microRNAs and RNA-binding proteins. *RNA Biol.* **10,** 934–942 (2013).

67. Kheradpour, P., Stark, A., Roy, S. & Kellis, M. Reliable prediction of regulator targets using 12 Drosophila genomes. *Genome Res.* **17,** 1919–1931 (2007).

68. Farh, K. K.-H. *et al.* The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. *Science* **310,** 1817–1821 (2005).

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

70. Stark, A., Brennecke, J., Bushati, N., Russell, R. B. & Cohen, S. M. Animal MicroRNAs Confer Robustness to Gene Expression and Have a Significant Impact on 3′UTR Evolution. *Cell* **123,** 1133–1146 (2005).

71. Stark, A., Brennecke, J., Russell, R. B. & Cohen, S. M. Identification of Drosophila MicroRNA Targets. *PLoS Biol.* **1,** (2003).

72. Gaidatzis, D., van Nimwegen, E., Hausser, J. & Zavolan, M. Inference of miRNA targets using evolutionary conservation and pathway analysis. *BMC Bioinformatics* **8,** 69 (2007).

73. Burgler, C. & Macdonald, P. M. Prediction and verification of microRNA targets by MovingTargets, a highly adaptable prediction method. *BMC Genomics* **6,** 88 (2005).

74. Enright, A. J. *et al.* MicroRNA targets in Drosophila. *Genome Biol.* **5,** R1 (2003).

75. Liu, H., Yue, D., Chen, Y., Gao, S.-J. & Huang, Y. Improving performance of mammalian microRNA target prediction. *BMC Bioinformatics* **11,** 476 (2010).

76. 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.* **11,** R90 (2010).

77. Reyes-Herrera, P. H., Ficarra, E., Acquaviva, A. & Macii, E. miREE: miRNA recognition elements ensemble. *BMC Bioinformatics* **12,** 454 (2011).

78. Kim, S.-K., Nam, J.-W., Rhee, J.-K., Lee, W.-J. & Zhang, B.-T. miTarget: microRNA target gene prediction using a support vector machine. *BMC Bioinformatics* **7,** 411 (2006).

79. Khorshid, M., Hausser, J., Zavolan, M. & van Nimwegen, E. A biophysical miRNA-mRNA interaction model infers canonical and noncanonical targets. *Nat. Methods* **10,** 253–255 (2013).

80. Lewis, B. P., Shih, I., Jones-Rhoades, M. W., Bartel, D. P. & Burge, C. B. Prediction of Mammalian MicroRNA Targets. *Cell* **115,** 787–798 (2003).

81. Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31,** 3406–3415 (2003).

82. Rennie, W. *et al.* STarMir: a web server for prediction of microRNA binding sites. *Nucleic Acids Res.* **42,** W114–W118 (2014).

83. Kanoria, S. *et al.* STarMir Tools for Prediction of microRNA binding sites. *Methods Mol. Biol. Clifton NJ* **1490,** 73–82 (2016).

84. Vejnar, C. E. & Zdobnov, E. M. miRmap: Comprehensive prediction of microRNA target repression strength. *Nucleic Acids Res.* **40,** 11673–11683 (2012).

85. Bandyopadhyay, S. & Mitra, R. TargetMiner: microRNA target prediction with systematic identification of tissue-specific negative examples. *Bioinforma. Oxf. Engl.* **25,** 2625–2631 (2009).

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

87. Guo, H., Ingolia, N. T., Weissman, J. S. & Bartel, D. P. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* **466,** 835–840 (2010).

88. Marín, R. M., Šulc, M. & Vaníček, J. Searching the coding region for microRNA targets. *RNA* **19,** 467–474 (2013).

89. Fan, X. & Kurgan, L. Comprehensive overview and assessment of computational prediction of microRNA targets in animals. *Brief. Bioinform.* bbu044 (2014). doi:10.1093/bib/bbu044

90. Dweep, H. & Gretz, N. miRWalk2.0: a comprehensive atlas of microRNA-target interactions. *Nat. Methods* **12,** 697 (2015).

91. Baştanlar, Y. & Ozuysal, M. Introduction to machine learning. *Methods Mol. Biol. Clifton NJ* **1107,** 105–128 (2014).

92. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20,** 273–297 (1995).

93. Saetrom, P. Predicting the efficacy of short oligonucleotides in antisense and RNAi experiments with boosted genetic programming. *Bioinforma. Oxf. Engl.* **20,** 3055–3063 (2004).

94. SÆTROM, O., SNØVE, O. & SÆTROM, P. Weighted sequence motifs as an improved seeding step in microRNA target prediction algorithms. *RNA* **11,** 995–1003 (2005).

95. Rajewsky, N. & Socci, N. D. Computational identification of microRNA targets. *Dev. Biol.* **267,** 529–535 (2004).

96. Rabiee-Ghahfarrokhi, B., Rafiei, F., Niknafs, A. A. & Zamani, B. Prediction of microRNA target genes using an efficient genetic algorithm-based decision tree. *FEBS Open Bio* **5,** 877–884 (2015).

97. Vlachos, I. S. *et al.* DIANA-TarBase v7.0: indexing more than half a million experimentally supported miRNA:mRNA interactions. *Nucleic Acids Res.* **43,** D153–D159 (2015).

98. Sethupathy, P., Corda, B. & Hatzigeorgiou, A. G. TarBase: A comprehensive database of experimentally supported animal microRNA targets. *RNA N. Y. N* **12,** 192–197 (2006).

99. Yan, X. *et al.* Improving the prediction of human microRNA target genes by using ensemble algorithm. *FEBS Lett.* **581,** 1587–1593 (2007).

100. Ahmadi, H. *et al.* HomoTarget: A new algorithm for prediction of microRNA targets in Homo sapiens. *Genomics* **101,** 94–100 (2013).

101. Yousef, M., Jung, S., Kossenkov, A. V., Showe, L. C. & Showe, M. K. Naïve Bayes for microRNA target predictions--machine learning for microRNA targets. *Bioinforma. Oxf. Engl.* **23,** 2987–2992 (2007).

102. Huang, J. C., Frey, B. J. & Morris, Q. D. Comparing sequence and expression for predicting microRNA targets using GenMiR3. *Pac. Symp. Biocomput. Pac. Symp. Biocomput.* 52–63 (2008).

103. Huang, J. C. *et al.* Using expression profiling data to identify human microRNA targets. *Nat. Methods* **4,** 1045–1049 (2007).

104. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15,** 1034–1050 (2005).

105. Amirkhah, R. *et al.* Naïve Bayes classifier predicts functional microRNA target interactions in colorectal cancer. *Mol. Biosyst.* **11,** 2126–2134 (2015).

106. Breiman, L. Random Forests. *Mach. Learn.* **45,** 5–32 (2001).

107. Mendoza, M. R. *et al.* RFMirTarget: Predicting Human MicroRNA Target Genes with a Random Forest Classifier. *PLoS ONE* **8,** (2013).

108. Ding, J., Li, X. & Hu, H. TarPmiR: a new approach for microRNA target site prediction. *Bioinformatics* btw318 (2016). doi:10.1093/bioinformatics/btw318

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

110. Sturm, M., Hackenberg, M., Langenberger, D. & Frishman, D. TargetSpy: a supervised machine learning approach for microRNA target prediction. *BMC Bioinformatics* **11,** 292 (2010).

111. Li, L., Gao, Q., Mao, X. & Cao, Y. New support vector machine-based method for microRNA target prediction. *Genet. Mol. Res. GMR* **13,** 4165–4176 (2014).

112. Lu, Y. & Leslie, C. S. Learning to Predict miRNA-mRNA Interactions from AGO CLIP Sequencing and CLASH Data. *PLoS Comput. Biol.* **12,** (2016).

113. Churpek, M. M. *et al.* Multicenter Comparison of Machine Learning Methods and Conventional Regression for Predicting Clinical Deterioration on the Wards. *Crit. Care Med.* **44,** 368–374 (2016).

114. Chandra, V., Girijadevi, R., Nair, A. S., Pillai, S. S. & Pillai, R. M. MTar: a computational microRNA target prediction architecture for human transcriptome. *BMC Bioinformatics* **11,** S2 (2010).

115. Rodriguez, A., Griffiths-Jones, S., Ashurst, J. L. & Bradley, A. Identification of mammalian microRNA host genes and transcription units. *Genome Res.* **14,** 1902–1910 (2004).

116. Xiao, F. *et al.* miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Res.* **37,** D105-110 (2009).

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

118. Linsley, P. S. *et al.* Transcripts Targeted by the MicroRNA-16 Family Cooperatively Regulate Cell Cycle Progression. *Mol. Cell. Biol.* **27,** 2240–2252 (2007).

119. Zhang, H.-M. *et al.* Transcription factor and microRNA co-regulatory loops: important regulatory motifs in biological processes and diseases. *Brief. Bioinform.* **16,** 45–58 (2015).

120. John, B. *et al.* Human MicroRNA Targets. *PLoS Biol.* **2,** (2004).

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

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

123. Reczko, M., Maragkakis, M., Alexiou, P., Grosse, I. & Hatzigeorgiou, A. G. Functional microRNA targets in protein coding sequences. *Bioinforma. Oxf. Engl.* **28,** 771–776 (2012).

124. Lewis, B. P., Burge, C. B. & Bartel, D. P. Conserved Seed Pairing, Often Flanked by Adenosines, Indicates that Thousands of Human Genes are MicroRNA Targets. *Cell* **120,** 15–20 (2005).

125. Garcia, D. M. *et al.* Weak Seed-Pairing Stability and High Target-Site Abundance Decrease the Proficiency of lsy-6 and Other miRNAs. *Nat. Struct. Mol. Biol.* **18,** 1139–1146 (2011).

126. Nam, J.-W. *et al.* Global analyses of the effect of different cellular contexts on microRNA targeting. *Mol. Cell* **53,** 1031–1043 (2014).

127. Riffo-Campos, Á. L., Riquelme, I. & Brebi-Mieville, P. Tools for Sequence-Based miRNA Target Prediction: What to Choose? *Int. J. Mol. Sci.* **17,** (2016).

128. Ekimler, S. & Sahin, K. Computational Methods for MicroRNA Target Prediction. *Genes* **5,** 671–683 (2014).

129. Hammell, M. Computational methods to identify miRNA targets. *Semin. Cell Dev. Biol.* **21,** 738–744 (2010).

130. 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.* **44,** 839–847 (2011).

131. Lu, T.-P. *et al.* miRSystem: An Integrated System for Characterizing Enriched Functions and Pathways of MicroRNA Targets. *PLoS ONE* **7,** e42390 (2012).

132. Nam, S., Kim, B., Shin, S. & Lee, S. miRGator: an integrated system for functional annotation of microRNAs. *Nucleic Acids Res.* **36,** D159–D164 (2008).

133. Roberts, J. T. & Borchert, G. M. Computational Prediction of MicroRNA Target Genes, Target Prediction Databases, and Web Resources. *Methods Mol. Biol. Clifton NJ* **1617,** 109–122 (2017).

134. Kuhn, D. E. *et al.* Experimental Validation of miRNA Targets. *Methods San Diego Calif* **44,** 47–54 (2008).

135. Sethupathy, P., Megraw, M. & Hatzigeorgiou, A. G. A guide through present computational approaches for the identification of mammalian microRNA targets. *Nat. Methods* **3,** 881–886 (2006).

136. Oliveira, A. C. *et al.* Combining Results from Distinct MicroRNA Target Prediction Tools Enhances the Performance of Analyses. *Front. Genet.* **8,** (2017).

137. DeConde, R. P. *et al.* Combining results of microarray experiments: a rank aggregation approach. *Stat. Appl. Genet. Mol. Biol.* **5,** Article15 (2006).

138. Friedman, Y., Karsenty, S. & Linial, M. miRror-Suite: decoding coordinated regulation by microRNAs. *Database J. Biol. Databases Curation* **2014,** (2014).

139. Gamazon, E. R. *et al.* ExprTarget: An Integrative Approach to Predicting Human MicroRNA Targets. *PLoS ONE* **5,** (2010).

140. Tabas-Madrid, D. *et al.* Improving miRNA-mRNA interaction predictions. *BMC Genomics* **15 Suppl 10,** S2 (2014).

141. Coronnello, C. & Benos, P. V. ComiR: combinatorial microRNA target prediction tool. *Nucleic Acids Res.* **41,** W159–W164 (2013).

142. Yue, D., Guo, M., Chen, Y. & Huang, Y. A Bayesian decision fusion approach for microRNA target prediction. *BMC Genomics* **13,** S13 (2012).

143. Li, J. *et al.* Functional combination strategy for prioritization of human miRNA target. *Gene* **533,** 132–141 (2014).

144. Oulas, A. *et al.* Prediction of miRNA targets. *Methods Mol. Biol. Clifton NJ* **1269,** 207–229 (2015).

145. Zweig, M. H. & Campbell, G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin. Chem.* **39,** 561–577 (1993).

146. Jesse Davis & Mark Goadrich. The Relationship Between Precision-Recall and ROC Curves. *Proceeding* (2006).

147. Bradley, T. & Moxon, S. An Assessment of the Next Generation of Animal miRNA Target Prediction Algorithms. *Methods Mol. Biol. Clifton NJ* **1580,** 175–191 (2017).

Figures legends

Figure 1: microRNA seed site types

The vast majority of miRNA interactions happens 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.

Figure 2: Basic schematics for Genetic Programming (GP) and Genetic Algorithm (GA)

Using training data, both GP and GA will create subtree crossover of parents A and B to form offspring C and D. A fitness test is done for each tree (parents and offspring) to decide which one is best suited for the classification of the training data.

Figure 3: Naïve 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.

Figure 4: Random Forest (RF) classifier

A) All data are being subsetted randomly to generate several trees using a predefined set of rules to optimize the split. B) This specific tree considers an interaction to happen if this one possesses all necessary parameters to fall in one of the green leafs. The RF algorithm returns the prediction made by the majority of the trees.

Figure 5: Nonlinear Support Vector Machine (SVM)

A) SVM constructs hyperplanes (grey dotted lines) in a multidimensional space that separates cases of different class labels. B) Biological data being rarely separable by straight lines, a transformation is often used to get a nonlinear separation model.

Figure 6: Neural Network

Selected features are used as input signals in this feedforward partially connected neural network example. Each node decides what to send to the next one following principles such as fuzzy logic, genetic algorithm 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) allowing the classifier to make a decision. The model can change the weights and nodes ordering in order to best classify the training data.

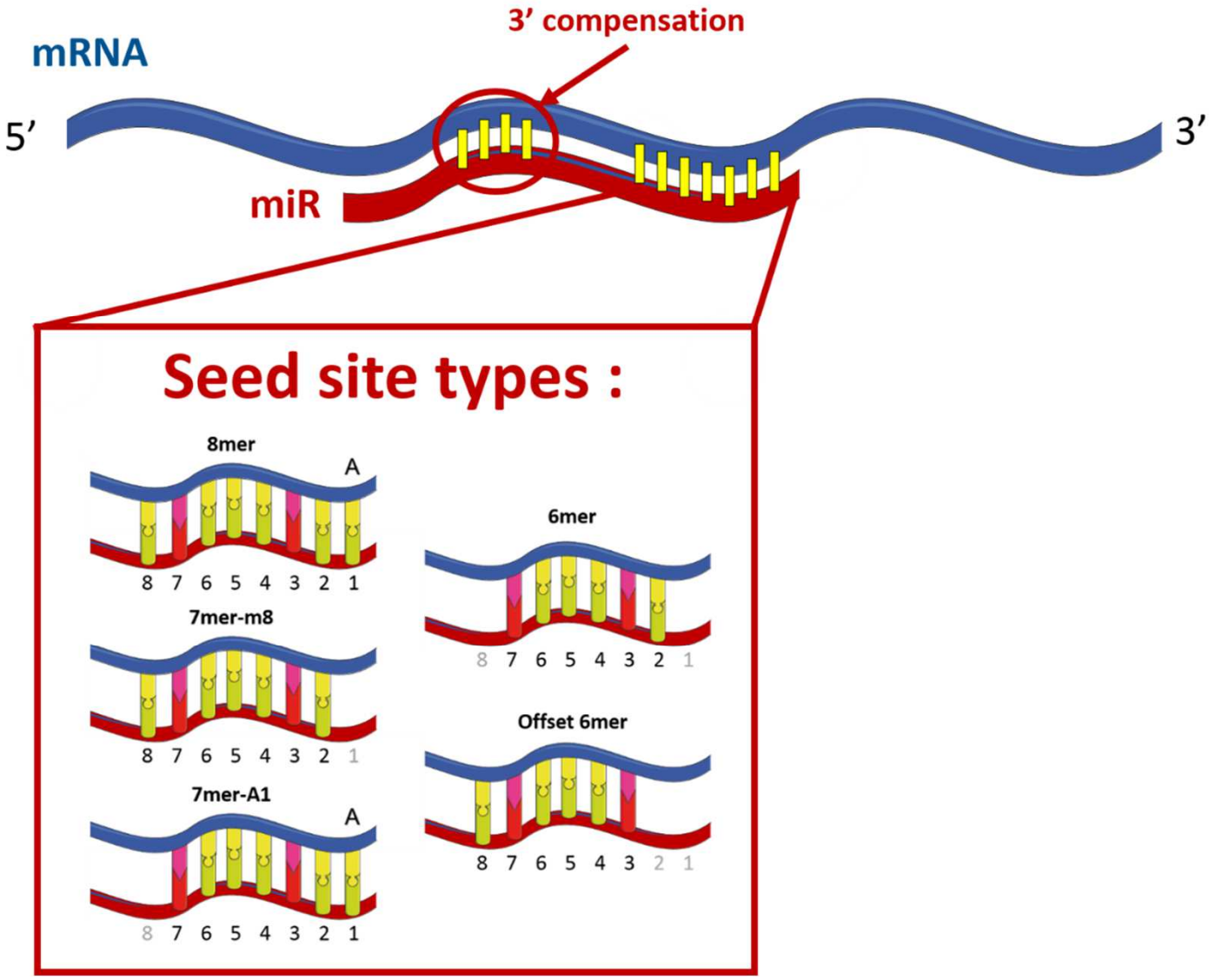

3’ compensation
mRNA
5’
3’
miR
Seed site types :
8mer
A
8 7 6 5 4 3 2 1
6mer
8 7 6 5 4 3 2 1
7mer-m8
8 7 6 5 4 3 2 1
Offset 6mer
8 7 6 5 4 3 2 1
7mer-A1
A
8 7 6 5 4 3 2 1

Figure 1

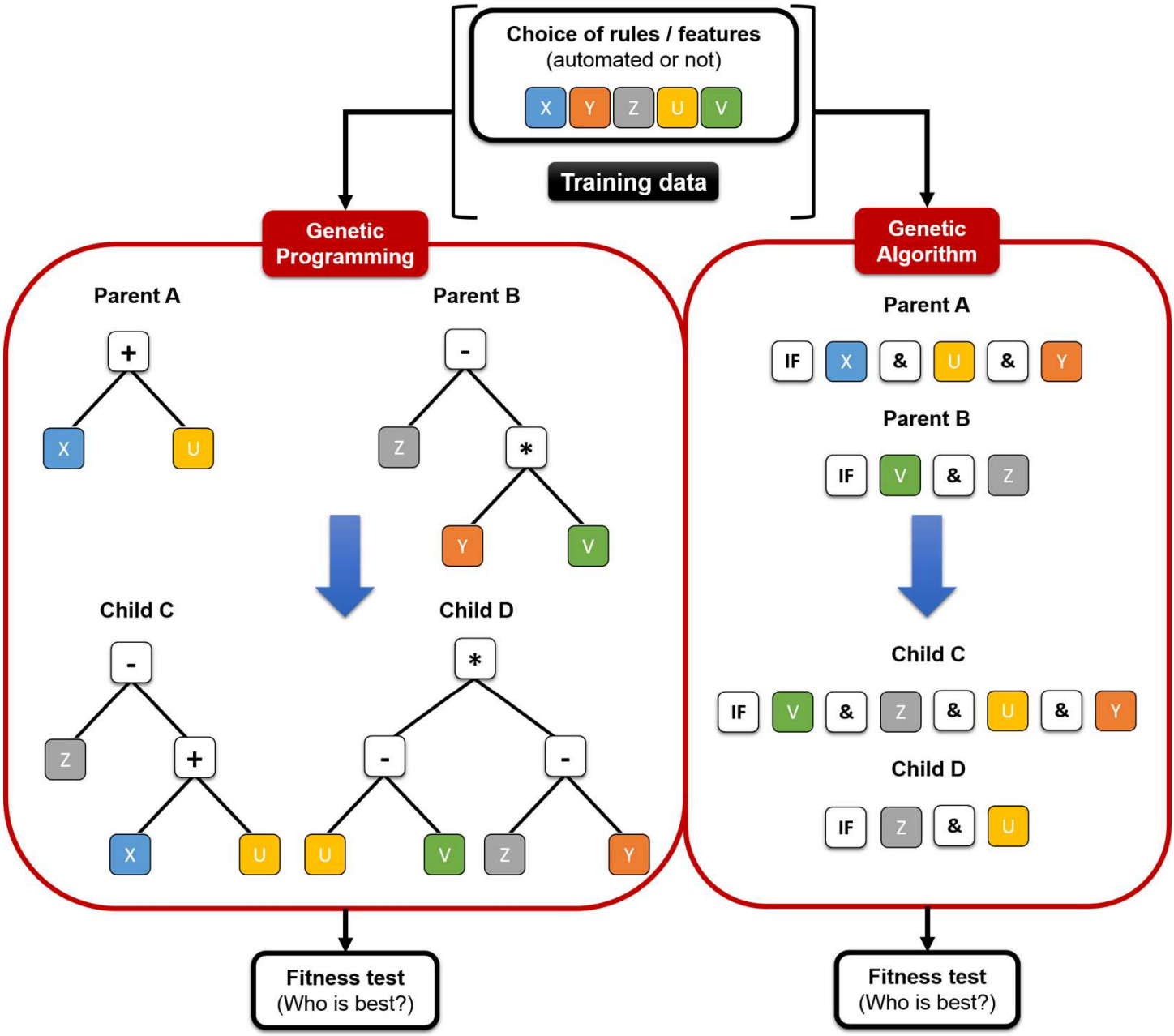
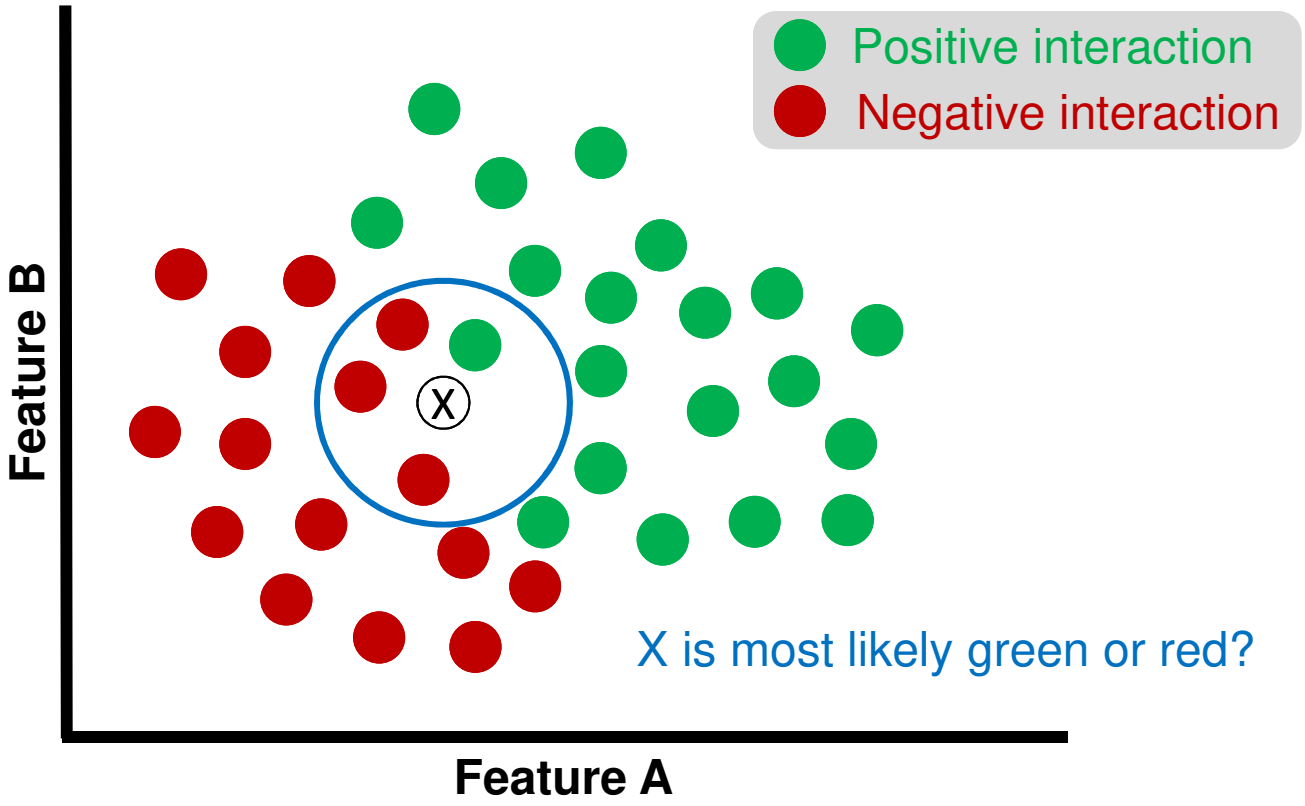

Choice of rules / features
(automated or not)
X Y Z U V
Training data
Genetic Programming
Parent A
Parent B
Child C
Child D
Fitness test
(Who is best?)
Genetic Algorithm
Parent A
IF X & U & Y
Parent B
IF V & Z
Child C
IF V & Z & U & Y
Child D
IF Z & U
Fitness test
(Who is best?)

Figure 2



Prior probability for GREEN = 20 / 35
Likelihood of X being GREEN = 1 / 20
Posterior probability of X being GREEN = 20/35 * 1/20 = 1/35

Prior probability for RED = 15 / 35
Likelihood of X being RED = 3 / 15
Posterior probability of X being RED = 15/35 * 3/15 = 3/35

Figure 3

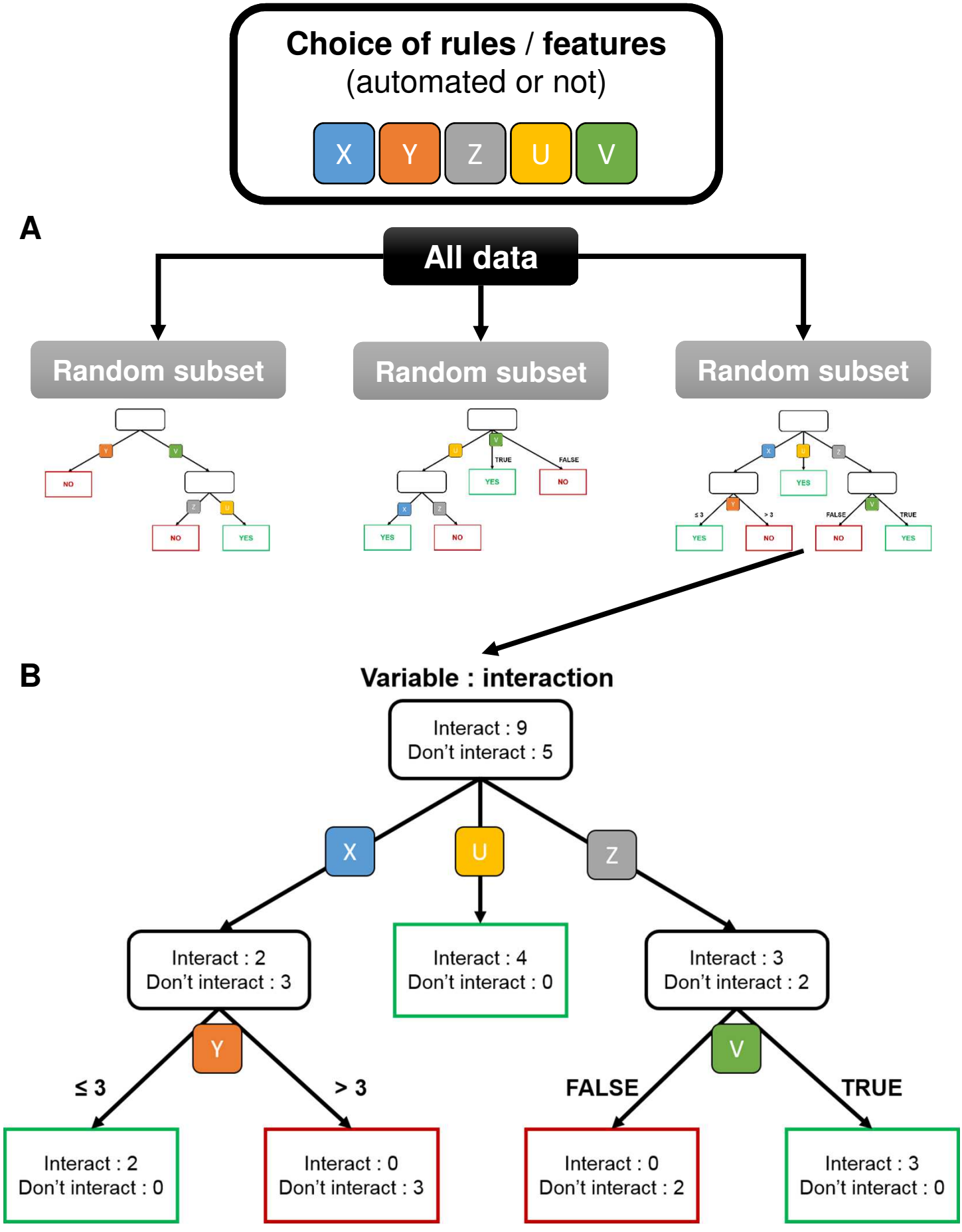
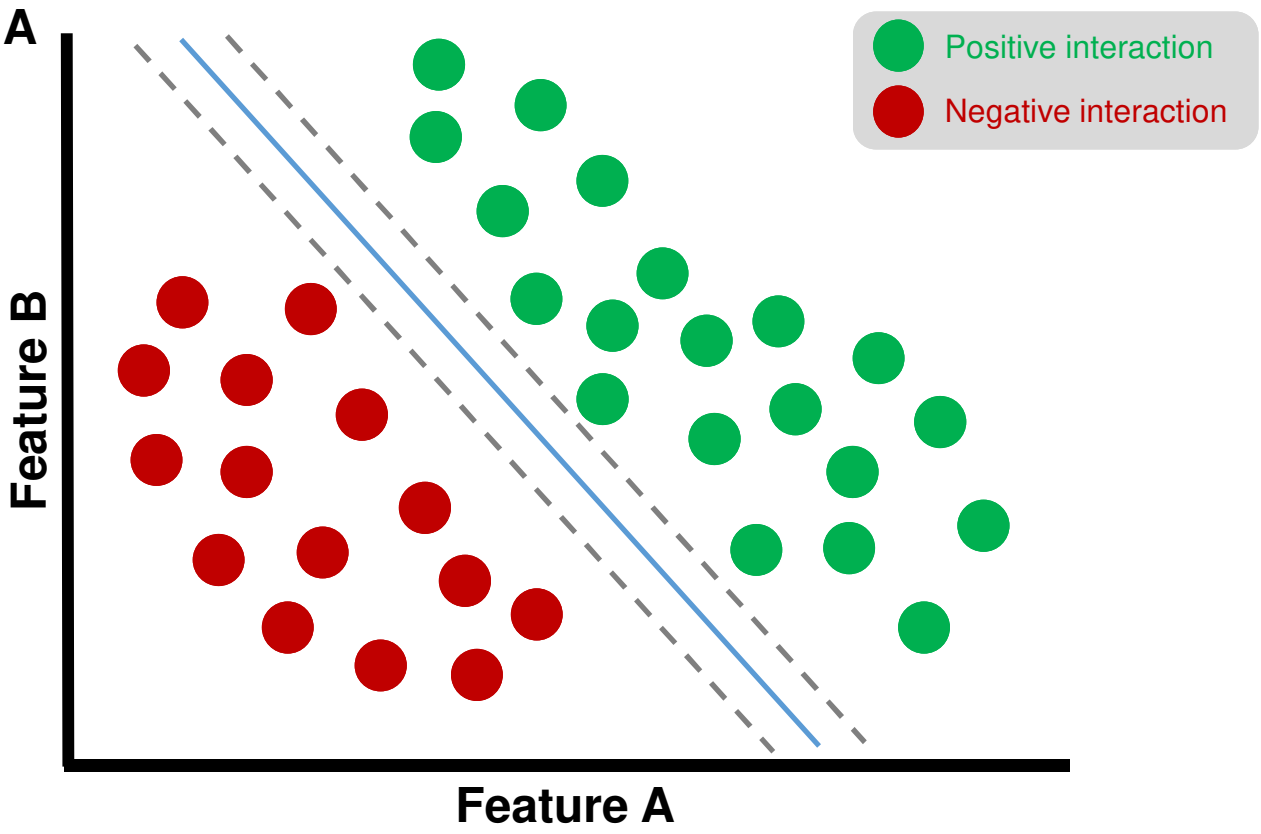

Choice of rules / features
(automated or not)
X
Y
Z
U
V
A
All data
Random subset
Random subset
Random subset
NO
NO
YES
TRUE
FALSE
YES
NO
YES
NO
YES
≤ 3
> 3
FALSE
TRUE
YES
NO
NO
YES
B
Variable : interaction
Interact : 9
Don't interact : 5
X
U
Z
Interact : 2
Don't interact : 3
Interact : 4
Don't interact : 0
Interact : 3
Don't interact : 2
Y
V
≤ 3
> 3
FALSE
TRUE
Interact : 2
Don't interact : 0
Interact : 0
Don't interact : 3
Interact : 0
Don't interact : 2
Interact : 3
Don't interact : 0

Figure 4


A
Positive interaction
Negative interaction
Feature B
Feature A

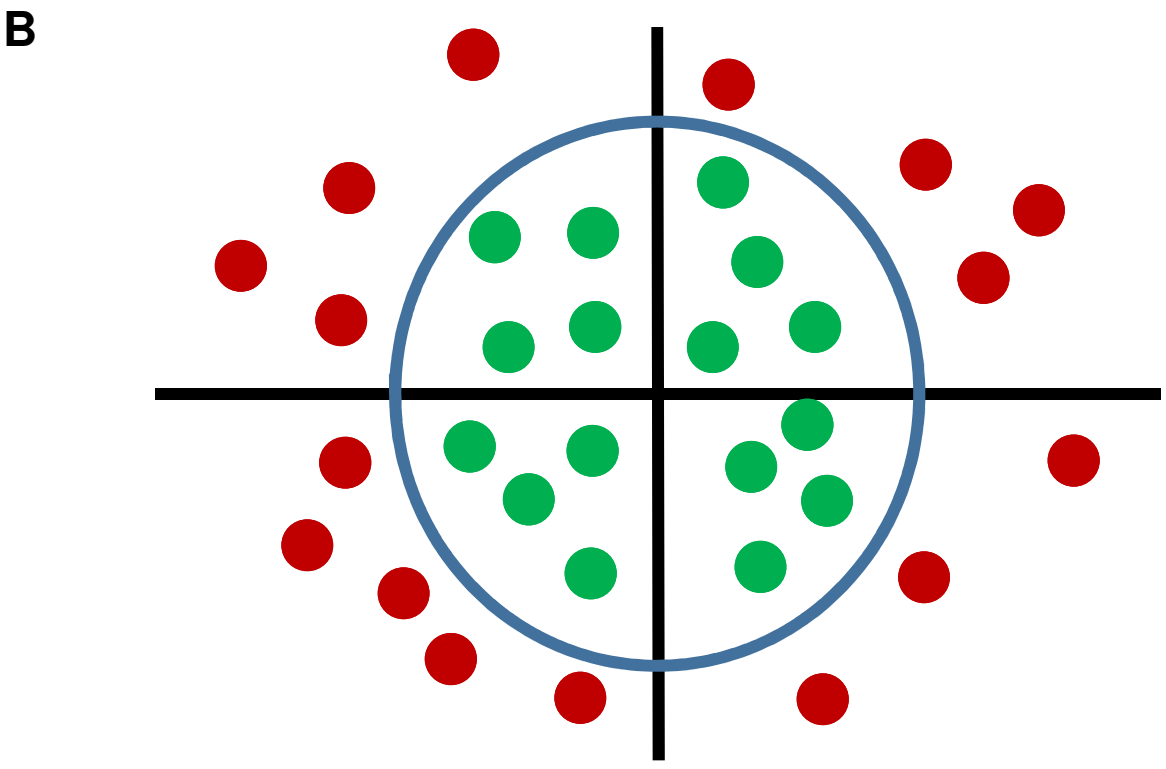

B

Figure 5

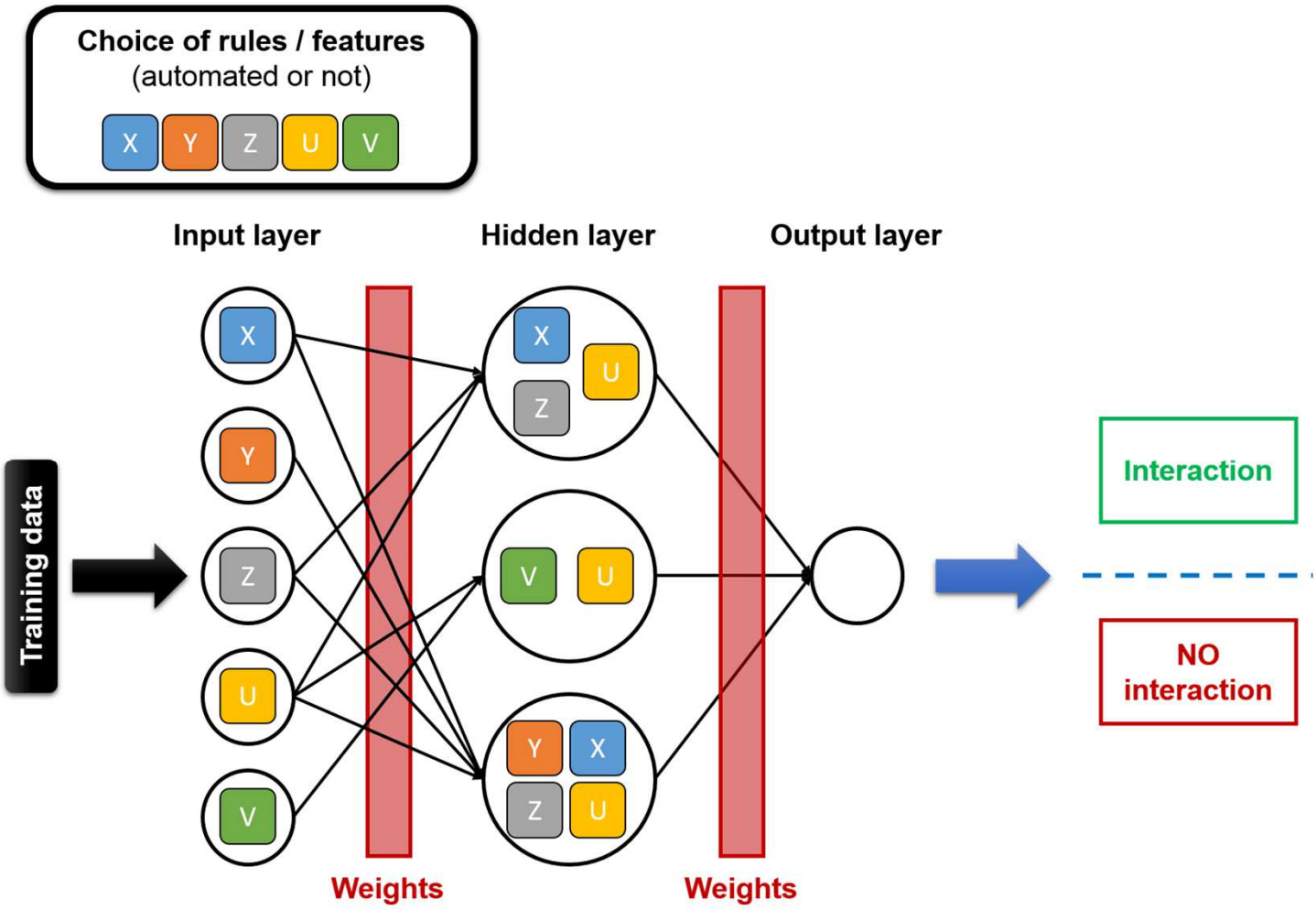

Choice of rules / features
(automated or not)
X
Y
Z
U
V
Input layer
Hidden layer
Output layer
Training data
Weights
Weights
Interaction
NO interaction

Figure 6