Journal of Bioinformatics and Computational Biology
Vol. 8, No. 4 (2010) 763–788

DOI: 10.1142/S0219720010004896

Imperial College Press
www.icpress.co.uk

ANIMAL MICRORNA TARGET PREDICTION USING DIVERSE SEQUENCE-SPECIFIC DETERMINANTS

YUN ZHENG*

Institute of Developmental Biology
and Molecular Medicine, and School of Life Sciences
Fudan University, 220 Handan Rd., Shanghai 200433, China
zhengyun@fudan.edu.cn

WEIXIONG ZHANG*

Department of Computer Science and Engineering
Washington University in St. Louis
Campus Box 1045, St. Louis, MO 63130, USA

Department of Genetics, Washington University School of Medicine
Campus Box 8510, St. Louis, MO 63108, USA
weixiong.zhang@wustl.edu

Received 5 January 2010
Revised 6 March 2010
Accepted 7 April 2010

Many recent studies have shown that access of animal microRNAs (miRNAs) to their complementary sites in target mRNAs is determined by several sequence-specific determinants beyond the seed regions in the 5′ end of miRNAs. These factors have been related to the repressive power of miRNAs and used in some programs to predict the efficacy of miRNA complementary sites. However, these factors have not been systematically examined regarding their capacities for improving miRNA target prediction. We develop a new miRNA target prediction algorithm, called *Hitsensor*, by incorporating many sequence-specific features that determine complementarities between miRNAs and their targets, in addition to the canonical seed regions in the 5′ ends of miRNAs. We evaluate the performance of our algorithm on 720 known animal miRNA:target pairs in four species, *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster* and *Caenorhabditis elegans*. Our experimental results show that Hitsensor outperforms five popular existing algorithms, indicating that our unique scheme for quantifying the determinants of complementary sites is effective in improving the performance of a miRNA target prediction algorithm. We also examine the effectiveness of miRNA-mediated repression for the predicted targets by using a published quantitative protein expression dataset of miR-223 knockout in mouse neutrophils. Hitsensor identifies more targets than the existing algorithms, and the predicted targets of Hitsensor show comparable protein level changes to those of the existing algorithms.

Keywords: MicroRNA; sequence analysis; microRNA target prediction; sequence specific determinants.

*Corresponding authors.

1. Introduction

MicroRNAs are non-coding RNAs that regulate the expression of protein-coding genes at post-transcriptional level.[1] They function by base-paring to their target mRNAs, subsequently leading to translational repression,[1,2] mRNA cleavage[3–5] or miRNA-induced degradation.[6–8] Due to the complexity of experimental validation of miRNA targets, several computational miRNA target prediction methods have been developed, including TargetScan[9] (later updated to TargetScanS[10]), Miranda,[11,12] PicTar,[13] methods in Refs. 14 and 15, RNAHybrid,[16] rna22,[17] PITA[18] for animals, and methods in Refs. 19–21, miRU[22] for plants. Many of these methods were reviewed in Ref. 23.

Most predicted and reported complementary sites of animal miRNAs are located in the 3′ untranslated region (3′ UTR) of target mRNAs.[9–16] The imperfect complementarity between miRNAs and their targets in animals makes target prediction much harder than in plants. Many existing methods for animals[9–16] extensively make use of the seed region, which is from the 2 to 8 nucleotides from the 5′ end of a mature miRNA.

However, a substantial number of miRNA:target pairs do not have good seed regions. Brennecke *et al.*[24] found that there are mainly two types of miRNA complementary sites, 5′ dominant sites and 3′ compensatory sites. The first type constitutes most animal miRNA complementary sites.[10,13,24] For this type, 7-mer and 8-mer 5′ seed matches are sufficient to function with 3′ paring below a random noise level.[24] On the other hand, 3′ compensatory sites having insufficient 5′ seed matches, which form the second type of miRNA complementary sites, require a strong 3′ pairing in order to be functional.[24] One example is the *let-7* binding sites in *lin-41*.[25] Thus, a strong preference to the seed region by the existing methods may miss 3′ compensatory sites.

Most existing methods[9–16] also used information of evolutional conservation, which is effective for finding conserved targets. On the other hand, conservation information does not help identify species specific targets.

Many recent studies indicated that there exist other determining factors besides the seed regions. As well documented, most miRNAs start with uridine; correspondingly, their binding sites end with adenosine. Even for some miRNAs that do not begin with uridine, the position complementary to the first nucleotide of miRNA is preferentially adenosine.[26] Lewis *et al.*[10] found that seed complementary sites are often flanked by adenosines. Nielsen *et al.*[26] noticed the preference of adenosine or uridine for the site complementary to the ninth nucleotide from the 5′ end of a miRNA. They also found that an increased AU content in the 3′ of the seed region is correlated with an increased mRNA downregulation effect. Jing *et al.*[27] and Grimson *et al.*[28] further noticed that many effective sites preferentially reside within regions that are locally AU rich. As suggested in Ref. 24, 3′ compensatory sites can function because there are extensive parings in those regions. Moreover, Grimson *et al.*[28] quantified a compensatory pairing region of 12–17 nucleotides

from the 5′ end of a miRNA. In addition, Grimson *et al.*[28] also found that closely spaced sites in the 3′ UTR of a target mRNA often synergistically promote the repression of the target, and effective complementary sites often locate after the 15th nucleotide from the stop codon of the mRNA and in the first and last quarters of the 3′ UTR. All these results indicated that local AU-content, 12–17 nt pairing, closely-paced sites, site positions, along with seed pairing, are important determinants to enhance miRNA-induced repression. Furthermore, Grimson *et al.*[28] proposed linear regression models to predict the efficacy of complementary sites by combining the contribution of seed region, local AU-content, 12–17 nt region, and site positions. Their models produced quantitative scores, called Context Scores, which were correlated with the mRNA expression levels[28] and protein expression levels[30] of predicted targets. The Context Scores are also reported in the TargetScan website (http://www.targetscan.org/). A recent study by Hausser *et al.*,[31] further found that structural features are only important for miRNA-guided Argonaute binding to mRNAs, and sequence features such as the AU content of 3′ UTRs are important for mRNA degradation after investigating a set of 14 sequence and structure features.

Motivated by the observations mentioned above, we aim to systematically investigate whether these determinants are useful for improving target prediction. In comparison with the Context Scores, which emphasized the efficacy of predicted complementary sites, we focus on the performance of a prediction algorithm. In particular, we propose a novel miRNA target prediction algorithm, called Hitsensor, to exploit and combine various sequence determinants. In the Hitsensor algorithm, we introduce novel methods, which are different from those used to calculate Context Scores, to quantify the contributions from the seed region, 12–17 nt region, local AU-content, close sites and site positions. Although some existing algorithms, such as Miranda,[11] also give additional rewards to the seed region, our approach uses a new rewarding scheme to emphasize the continuously matched seed. Briefly, the Hitsensor algorithm does not use conservation information. It starts from a sequence alignment with the Smith–Waterman algorithm[29] between miRNA and its target mRNA, calculates the scores of the five determinants for each alignment site, and then adds these individual determinant contributions to the alignment score to get the total score of a miRNA complementary site. Finally, sites with total scores greater than a pre-specified threshold are outputted.

We adopt two methods to evaluate the performance of the Hitsensor algorithm, the receiver operating characteristic (ROC) curve and the signal-to-noise ratio (S2N). We use a dataset of 96 verified functional and 83 non-functional miRNA:target pairs of *Drosophila melanogaster* to quantify the contributions of individual determinants. The Hitsensor algorithm reaches an area under the ROC curve (AUC) of 0.794 and an S2N of 7.62, which are the highest among all compared algorithms, including PITA,[18] PicTar[13] and Miranda,[11] on this dataset.

We then select 541 verified functional miRNA:target pairs across four species, *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster* and *Caenorhabditis elegans*,

to validate the performance of the Hitsensor algorithm. Again, the Hitsensor algorithm makes the largest number of correct predictions, 293, among all algorithms compared. In comparison, the existing algorithms, PITA with and without flanking sequences,[18] TargetScanS,[10] PicTar[13] and Miranda,[11] have, respectively, 231, 262, 188, 138 and 123 correct predictions on these selected datasets.

Our next level of validation is to examine the strength of miRNA-mediated repression and ratio of responsiveness to miRNA upregulation or downregulation of the predicted targets. In particular, we examine the changes of the expression levels of target proteins by using a quantitative protein upregulation dataset of miR-223 knockout in mouse neutrophils.[30] Hitsensor finds more conserved targets than TargetScan and PicTar while the fold changes of target protein levels detected by Hitsensor are comparable to that from TargetScan and PicTar. The targets predicted by Hitsensor also have a larger ratio of responsiveness to miR-223 knockout than those from TargetScan and PicTar, given that all these algorithms have the same number of predicted targets. On the other hand, Hitsensor predicted more targets than TargetScan and PicTar if Hitsensor's ratio of targets responsive to miR-223 loss is comparable to the ratios of TargetScan and PicTar. Hitsensor also shows comparable performance to published results on 620 targets, with almost equal number of functional and non-functional miRNA complementary sites, prepared from the protein expression profiles in Ref. 30.

2. Materials and Methods

2.1. *Datasets*

As summarized in Table 1, we extracted 720 *experimentally verified* miRNA:target pairs for four species from Ref. 18, the TarBase[32] and Ref. 33. Kertesz *et al.*[18] summarized a dataset with 190 *Drosophila melanogaster* miRNA:target pairs, 102 functional and 88 non-functional. Because the target genes of 6 and 5 pairs from 102 functional and 88 non-functional sets, respectively, have no 3′ UTR in the FlyBase (http://flybase.bio.indiana.edu/), we only use the remaining 96 functional and 83 non-functional pairs, i.e. dme96P and dme83N in Table 1, which are used as the training dataset to find optimal quantifications of the five determinants.

In addition, the TarBase contains another 16 functional miRNA:target pairs of *Drosophila* not in dme96P, which form dme16P in Table 1. The cel, hsa and mmu datasets are for worm *Caenorhabditis elegans*, human *Homo sapiens* and mouse *Mus musculus* and downloaded from the TarBase. After removing some miRNA:target pairs of worm, human and mouse in the TarBase because either their miRNA or target sequences are not available, we have 14, 440 and 49 pairs in cel, hsa and mmu datasets. The unc-hsa dataset in Table 1 consists of 22 of the 23 non-conserved human miRNA:target pairs in Ref. 33, because we did not find 3′ UTR for 1 of the 23 pairs in Ref. 33. The detailed list of these 720 miRNA:target pairs are given in Supplementary Table S1.

Table 1. The miRNA:target pairs used in training and testing. The training and testing parts are reported in literature. The last four datasets are prepared from the protein expression profiles in Ref. 30. The details are given in the main text.

	No.	Functionality	Reference
training			
dme96P + 83N	179	96 func. and 83 non-func.	18
testing			
dme16P	16	functional	TarBase[32]
cel	14	functional	TarBase[32]
hsa	440	functional	TarBase[32]
mmu	49	functional	TarBase[32]
unc-hsa	22	functional	33
subtotal	541		
Total	720		
from protein expression profiles			
miR-1Trans	168	83 func./85 non-func.	30
miR-124Trans	107	56 func./51 non-func.	30
miR-181Trans	149	76 func./73 non-func.	30
miR-223KO	196	100 func./96 non-func.	30
Total	620		

Because there are no or very limited negative datasets for studying miRNA and mRNA interaction, we prepare 4 additional datasets from the protein expression profiles in Ref. 30. By using a similar approach in Ref. 31, we choose the top 25% downregulated (or upregulated for miR-223 knockout sample) genes that have at least one 8-mer or 7-mer seed match in their 3′ UTRs as positive samples. As negative samples, we take the 25% least-changing genes with seed matches in their 3′ UTRs, i.e. those genes whose log2 protein expression fold changes are closest to 0 when comparing the samples with transfected miRNAs (or knock-out of miR-223) to the control samples. We then have 168, 107, 149 and 196 targets from, respectively, the miR-1 transfection, miR-124 transfection, miR-181 transfection and miR-223 knock-out protein expression datasets in Ref. 30, which are listed as miR-1Trans, miR-124Trans, miR-181Trans and miR-223KO, respectively, in Table 1 with their details shown in the second to fifth sheet of Supplementary Table S1.

The sequences of miRNAs studied were downloaded from the miRBase (release 14).[34] The sequences of mRNA targets were from NCBI RefSeq database (ftp://ftp.ncbi.nih.gov/refseq/H_sapiens/H_sapiens/RNA for hsa, ftp://ftp.ncbi.nih.gov/refseq/M_musculus/Contigs/RNA for mmu) and NCBI CoreNucleotide database and the FlyBase for dme96P, dme84N, dme16P and cel.

2.2. *Algorithms compared*

We will compare Hitsensor with five benchmark methods, i.e. PITA with (PITAf) and without (PITAn) flanking sequences,[18] TargetScanS,[10] PicTar[13] and Miranda.[11] The features used by the algorithms compared are summarized in Table 2 and discussed in detail in the following. All these algorithms make use

Table 2. The features used by the six algorithms compared. "opt." means optional. ΔG_{open} is energy cost of unpairing the 3′ UTR of target. ΔG_{duplex} is the free energy of miRNA:target duplex.

	HITS	MIRA	PITAn	PITAf	TSS	PicTar
Seed	√	√	√	√	√	√
12–17 nt	√			√	√	
seed flank	√			√	√	
close site	√	√	√	√	√	√
site position	opt.					
ΔG_{open}			√	√		
ΔG_{duplex}		√	√	√	√	√
conservation					√	√

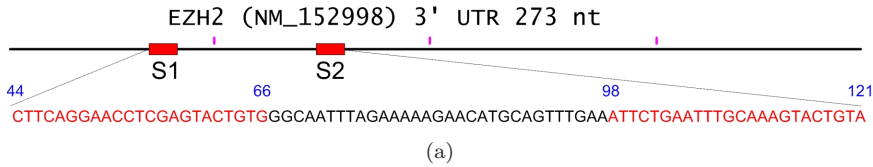
of the seed region, although in different ways. Hitsensor and Miranda give additional rewards to Watson–Crick pairs in seed regions with different schemes (to be discussed in the next section). PITAn, PITAf and TargetScanS directly find perfect seed regions.[18,10] PicTar prefers perfect seed matches but also allows imperfect seed matches.[13,33] Hitsensor and TargetScanS explicitly use the 12–17 nt region. Hitsensor, PITAf and TargetScanS employ the flanking regions of seeds.[18,33] Hitsensor uses close site determinants and optionally uses site position determinants, while other methods take the site number into consideration. PITA, i.e. both PITAf and PITAn, is the only algorithm that considers the free energy of 3′ UTR before miRNA binding (ΔG_{open}) by employing the energy gain after and before a miRNA binds its target, i.e. $\Delta G_{duplex} - \Delta G_{open}$.[18] Miranda, TargetScanS and PicTar compute the free energy of miRNA:target duplex, ΔG_{duplex}, with different methods.[33] Finally, conservation information is used by TargetScanS, PicTar, and optionally by Miranda[33]; therefore, these algorithms are conservation based.

The results of TargetScanS were downloaded from the TargetScan website (http://www.targetscan.org/), for both conserved and nonconserved miRNA families. The results of PicTar were downloaded from annotation databases of dm2, hg17, mm7 and ce2 of the UCSC genome browser (http://hgdownload.cse.ucsc.edu/). The results of PITA were downloaded from (http://genie.weizmann.ac.il/pubs/mir07/mir07_data.html) for targets catalog with and without 3-nt upstream and 15-nt downstream flanking sequences. We used a local version of the Miranda algorithm (version 1.9), available at the Miranda website (http://www.microrna.org/miranda_new.html), to obtain its results.

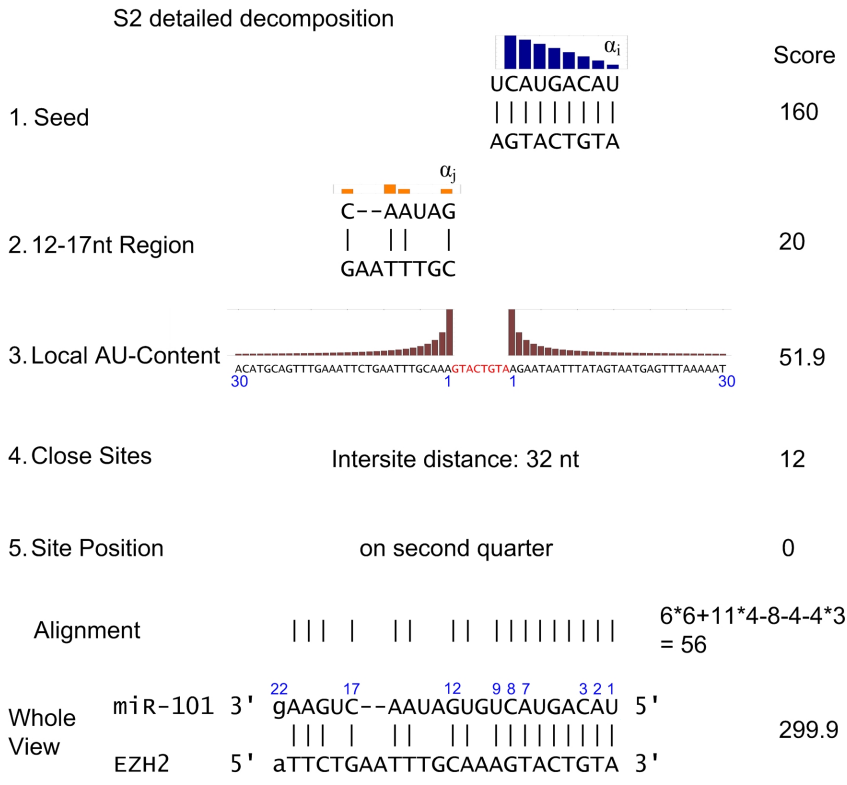
Because some verified complementary sites, such as miR-431 complementary sites on RTL1/Rtl1,[35] are located in coding regions of targets, we applied Hitsensor and Miranda (without conservation information) separately to 3′ UTRs and coding sequences (CDS) to examine miRNA complementary sites in CDS.

2.3. *Sequence-specific determinants*

We use the example in Fig. 1 to show how to use a set of parameters, called reward bases, to quantify the five sequence-specific determinants of miRNA complementary



(a)



(b)

Fig. 1. A schematic view of sequence-specific determinants that affect hsa-miR-101 binding to the 3′ UTR of EZH2 (NM_000609). (a) The two predicted binding sites of hsa-miR-101, red boxes, in 3′ UTR of EZH2 that is represented by the black solid line. The quarter points of the 3′ UTR are indicated by the pink points above the 3′ UTR. (b) Detailed decomposition of different determinants for site S2. With the values indicated with the bars, α_i and α_j above the seed and 12–17 nt region are the numbers of continuous matches at that position that are defined in Eq. (1) and Eq. (2), respectively. For the local AU-content determinant, the weights of the position are represented by the heights of the bar above the nucleotides. The reward base for seed (R), 12–17 nt region (U), local AU-content (B), close sites (D) and site position (Q) determinant are 4, 4, 8, 12 and 12, respectively. The AUScore is calculated with Eq. (3). Based on the rules described in Materials and Methods, the close site score and site position score are 12 and 0, respectively. The alignment score is calculated by rewarding 6×6 to 6 GC pairs, rewarding 11×4 to 11 AU pairs, penalizing $-8 - 4$ (-8 for one gap opening and -4 for one gap extension) to two gaps, and penalizing -4×3 to 4 mismatches, i.e. $36 + 44 - 12 - 12 = 56$.

sites, i.e. seed region, 12–17 nt region, local AU-content, close sites and site positions. Different values can be given to the reward bases to adjust the contributions of different determinants. In our implementation, we have assigned optimal default values, 8, 4, 44, 12 and 0 to reward bases of the seed region, 12–17 nt region, local AU-content, close sites and site-position determinant, respectively. We will discuss how to obtain these values of the reward bases in the Results section.

2.3.1. *Seed determinant*

Continuously matched seed regions are critical for repressing target mRNA or inducing target mRNA degradation.[10,24,28] To capture the importance of continuous matches in seed regions, we design a new score scheme that rewards functional, continuously matched seed regions with larger scores than discontinuously matched counterparts, which often occur by chance. Formally, we give a reward to the seed region based on Eq. (1):

$$\text{SeedScore} = R \times \sum_{i=1}^{8} (\alpha_i - \beta_i \times 2), \tag{1}$$

where R is the reward base of seed determinant, α_i is the number of continuous Watson–Crick matches from the 5′ end of a miRNA and is reset to 0 when a mismatch or a G:U pair occurs, and β_i is the number of continuous mismatches or G:U pairs from the 5′ end of a miRNA and is renewed to 0 when a Watson–Crick pair appears. α_i and β_i in Eq. (1) serve as a reward to continuous matches and a penalty to mismatches and G:U pairs, respectively.

In addition, because 8-mer perfect seeds are more effective to repress targets than 7-mer ones,[26,28] we also adopt the following empirical rules: if there is a continuously paired 8-mer seed, an additional reward of $3R$ will be given; if there exists a continuously paired 7-mer seed with a G:U pair or mismatch at the first nucleotide, an additional reward of $2.5R$ and $2R$ will be added; if there are at least seven continuously paired nucleotides and the ninth nucleotide is a Watson–Crick pair, an additional reward of R will be given. Finally, if there are totally more than two mismatches or G:U pairs, we give an additional penalty of $R \times (n_m + n_{G:U})$, where n_m and $n_{G:U}$ are the number of mismatches and the number of G:U pairs from the first to the eighth nucleotide of the miRNA, respectively.

For example, in site 2 (S2) of Fig. 1, hsa-miR-101 is continuously paired to 3′ UTR of EZH2 from the first to ninth nucleotide. Thus, this site receives a seed score of 160, i.e. $(1 + 2 + \cdots + 8) \times 4 = 144$ based on Eq. (1), plus 12 for a continuous 8-mer pair and 4 for a paired ninth nucleotide. As another example, if there was a mismatch at the fifth nucleotide of S2, then α_5 to α_8 would become 0 to 3 [see Supplemental Fig. S1(a)]. Therefore, the seed score would be $(1 + \cdots + 4) \times 4 - 8 + (1 + 2 + 3) \times 4 = 56$, which is 104 less than a continuously matched 8-mer seed, where -8 is the penalty to a mismatch at position 4. In contrast, if the reward

is determined by the number of Watson–Crick pairs, as used by Miranda[11] [see Fig. S1(b)], the difference between the two cases is only $4 \times 8 - (4 \times 7 - 4) = 8$.

2.3.2. *12–17 nt region determinant*

The continuously matched 12–17 nt region is important and compensatory to imperfect seed region,[24] and enhances miRNA binding.[28] Therefore, similar to the *SeedScore* in Eq. (1), we reward the 12–17 nt region with Eq. (2).

$$\text{TwelveSeventeenScore} = U \times \sum_{j=1}^{6}(\alpha_j - \beta_j \times 2), \tag{2}$$

where U is the reward base for the 12–17 nt region determinant, α_j and β_j have the same values as α_i and β_i in Eq. (1) except starting from 12 nt of a miRNA. Similar to the seed region, we also give an additional penalty of $U \times (n_m + n_{G:U})$ where n_m and $n_{G:U}$ are the number of mismatches and the number of G:U pairs from 12 to 17 nt, respectively, if there are more than two mismatches or G:U pairs in the 12–17 nt region.

A complementary site with sufficient matches in the seed region can function with little support from the pairing from the 3′ end of the miRNA.[24] Therefore, if there exists at least one basic 6-mer (2–7 nt) seed match, we will not give a penalty to the 12–17 nt region, i.e. $penalty = 0$. On the contrary, if a complementary site does not contain a 6-mer seed match and 12 to 17 nt form a 6-mer continuous Watson–Crick match, we will give an additional reward of $6U$ to 12–17 nt determinant, and set the *SeedScore* in Eq. (1) to 0 if it is negative.

For the example in Fig. 1, site S2 has an 8-mer matched seed, thus *penalty* to 12–17 nt region is zero. There are in total four Watson–Crick paired nucleotides with two of them continuously matched, thus the total reward is $4+4+8+4 = 20$.

2.3.3. *Local AU-content determinant*

We calculate the score of local AU-content with Eq. (3).

$$\text{AUScore} = \left(\sum_{i=1}^{30}\frac{1}{i} \times IsAU_{\text{up}}(i) + \sum_{j=1}^{30}\frac{1}{j} \times IsAU_{\text{down}}(j)\right) \times B, \tag{3}$$

where $IsAU_{\text{up}}(i)$, a variable indicating whether a position on a mRNA beginning from the opposite of 9 nt of the miRNA is A or U(T), will be 1 if the nucleotide at position i is A or U(T), or 0 otherwise; $IsAU_{\text{down}}(j)$, similar to $IsAU_{\text{up}}(i)$, indicates whether a position on a mRNA beginning from the −1 nt nucleotide opposite to the miRNA is A or U(T), and will be 1 if the nucleotide at position j is A or U(T), or 0 otherwise. Because local AU preference normally appears with continuous seed match,[10,28] we allocate different reward base B and $0.25B$ to sites with and without perfectly matched 6-mer seeds (2–7 nt) to further differentiate functional sites with

perfect seeds to those with imperfect seeds normally due to random chance. Because AU preference immediately beside seed region is important and decreases fast when the distance from the seed increases,[10,28] the weights of these A and U, $1/i$ and $1/j$, are decreasing when the distance between them and seed, i and j, increases. As shown in Fig. 1, the weights of local A and U around seed are reflected by the height of the bars above the corresponding nucleotides. Thus, the sum operations in Eq. (3) capture the effects of A and U in the flanking region of the seed. For the example in Fig. 1, because the site has a matched 8-mer seed, B is 8, and the score of local AU-content is 51.9, following Eq. (3).

2.3.4. *Close sites determinant*

If a miRNA has more than one complementary sites on a target, these sites may synergistically repress the target, when they have an intersite distance between 19 to 34 nt.[28] Thus, we first find all sites with at least a 6-mer matched seed or a total score from other determinants greater than that of an 8-mer matched seed plus 8 additional paired nucleotides, and then calculate the distances between these sites. If the distance between two sites is within 19 to 34 nt, we give a close site score of D. In Fig. 1, sites S2 and S1 have a close site score of $D = 12$, because S2 and S1 are 32 nt apart and S1 has a 7-mer matched seed.

2.3.5. *Position determinant*

We give a position score of Q if a complementary site is located in the first or last quarter of a 3′ UTR, and an additional reward of $0.5Q$ if the 3′ UTR is longer than 1300 nt. This is because complementary sites in the first and last quarters of 3′ UTRs longer than 1300 nt are more effective.[28] However, if a complementary site is located within the first 15 nt of the first quarter of a 3′ UTR, we will not give reward to it, because such a site is weaker than those in other regions of the 3′ UTR.[28] The position determinant is only applicable to the miRNA complementary sites in 3′ UTRs of target mRNAs. For the example in Fig. 1, no position score is given to site S2, which is located in the second quarter of the 3′ UTR.

2.4. *The Hitsensor algorithm*

Hitsensor first uses a modified Smith–Waterman (SW) algorithm[29] to find regions with sufficient matches between miRNAs and their targets. Instead of performing alignments with matched nucleotides, e.g. A-A and C-C, Hitsensor finds complementary nucleotides, i.e. G-C, A-U and G-U "wobble" pairs that have rewards of $+6, +4$ and $+2$, respectively, in alignment. The affine gap penalty, i.e. the penalty increasing linearly with the length of gap after the initial gap opening penalty, is used for gap opening (-8) and gap extension (-4). The algorithm gives a penalty of -3 to known mismatch nucleotides and a penalty of -1 to mismatches to unspecified nucleotides (i.e. "N") in mRNAs. The algorithm will first recursively search for

miRNA complementary sites on the whole target mRNA sequence. If a site has a positive alignment score, the algorithm will keep it for further analysis.

After obtaining a list of sites, Hitsensor will continue to evaluate the sequence-specific determinants for all sites and set the scores for the determinants. The final score of a complementary site is then the sum of the scores of all determinants and alignment score from the Smith–Waterman algorithm. For example, the final score for S2 in Fig. 1 is 299.9, which is the sum of the scores of different determinants and the alignment score. If the final score of a given pair is greater than a user-specified threshold, Hitsensor will output this site. Finally, the max score of all sites for a given miRNA:target pair is used as the representative score of the pair to reflect the best possible binding of the pair. This information is useful because even though many miRNA:target pairs carry a single complementary site,[33] a large number of them have multiple complementary sites. And when multiple sites exist, the most accessible site should be more likely to be bound than the other sites since a site with a larger final score should be more accessible than one with a smaller final score.

In some extreme cases, we found that some miRNA:pairs with perfect seed matches, such as dme-miR-79 versus bap, have optimal SW alignments with imperfectly matched seed regions. Consequently, these sites will have low final scores based on our score scheme. To correct this drawback due to application of the SW alignment, Hitsensor will check whether the target has regions that perfectly match 2–7 nt of the miRNA, if it fails to find complementary sites after evaluating all determinants. If such regions are found, Hitsensor will cut the flanking sequences, upstream 29 nt and downstream 1 nt, of these regions, re-evaluate the determinants and output these sites if they satisfy the specified threshold.

In addition, we also implemented the reward of an adenosine (A) opposite the first nucleotide of the miRNA, because recognition of A from first nucleotide of miRNA favors miRNA-mediated protein downregulation.[30]

Hitsensor does not use conservation in prediction. If a miRNA family in two species targets the same homolog genes in two species, we defined the miRNA:target pair as conserved. Then, conserved and nonconserved miRNA:target pairs were chosen to compare the performance of Hitsensor on these cases.

We have implemented the Hitsensor algorithm with the Java programming language. The software package and documents are available upon request.

2.5. *Evaluation methods*

2.5.1. *The receiver operating characteristic (ROC) curve*

The ROC curve shows the sensitivity versus false positive ratios (fpr, i.e. 1 - specificity) under different score thresholds. The area under the curve (AUC) measures the ability of the algorithm to correctly classify functional and non-functional miRNA:target pairs. On an ROC curve, the point nearest to the upper left corner

provides the optimal algorithm setting, where the algorithm reaches the optimal balance between sensitivity and specificity (i.e. 1 -fpr).

2.5.2. *Signal-to-noise ratio*

The signal-to-noise (S2N) ratio is often used to evaluate the performance of target prediction algorithms.[9,13] We use the scores of verified functional miRNA:target pairs as the scores of positive samples and the scores of verified non-functional miRNAs as values of negative samples to generate the signal-to-noise ratio.

3. Results

3.1. *Improved performance by incorporating diverse sequence-specific determinants*

3.1.1. *Examining effects of different determinants*

To find optimal quantifications of determinants, we exclusively changed the reward base for one of the five determinants, i.e. seed region (R), 12–17 nt region (U), local AU-content (B), close sites (D) and site position (Q), from 0 to 20, and obtained the ROC curves of the training datasets (dme96P + dme88N). In other words, when we change one of the reward bases, we fix the values of other reward bases. The results are listed in Supplementary Figures S2(a) to S2(e), respectively. The AUC and S2N against various values of the five reward bases are given in Figs. S3(a) and S3(b). As shown in Figs. S2(a) and S2(b), the algorithm had a very different performance, and reached best AUC and S2N when $R = 8$ and $U = 0$ in Figs. S3(a) and S3(b). But after reviewing Fig. S2(b), we found that the algorithm had the optimal tradeoff between sensitivity and specificity when $U = 4$. The increasing reward base of local AU-content, B, had a beneficial effect on the AUC and S2N of the algorithm, although less significantly than R and U [Figs. S3(a) and S3(b)]. After testing various B values, we found that AUC reached the maximal value when B was around 40 [Fig. S3(c)]. When $B = 44$, the Hitsensor algorithm had its best tradeoff between sensitivity and specificity [Fig. S2(f)]. Various reward bases of close site determinant, D, had little effect on the performance of the algorithm [Figs. S2(d) and S3(a),(b)]. We also found that increasing Q, the reward base of position determinant, could decrease AUC and S2N values [Figures S2(e) and S3(a),(b)]. Therefore, we applied $R = 8$, $U = 4$, $B = 44$, $D = 12$ and $Q = 0$ to both the training and testing datasets. The obtained ROC curve, AUC and S2N of Hitsensor on training datasets, as well as those from other algorithms, are given in Fig. 2, while the number of positive predictions, i.e. samples predicted as functional miRNA:target pairs, for all datasets at the optimal settings of the compared algorithms are listed in Table 3. The optimal thresholds of the compared algorithms are obtained with their ROC curves, as discussed in Methods. The complete lists of Hitsensor predictions when using 3′ UTRs and CDS are given in Tables S2 and S3, respectively.

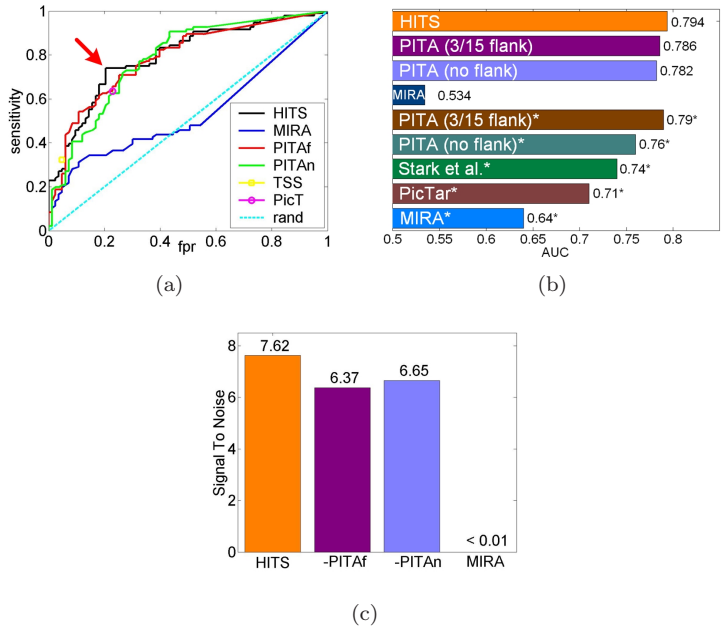


Fig. 2. The comparisons of different algorithms. (a) The ROC curve, (b) AUC and (c) S2N of the compared algorithms for the training dataset (dme96P + dme83N). HITS, MIRA, PITAf, PITAn, TSS and PicT stand for the Hitsensor, Miranda, PITA with flanking sequences, PITA without flanking sequences, TargetScanS, and PicTar algorithm, respectively. In part (a), the results obtained by a random scoring of the targets are shown by a dashed line. The point pointed by the red arrow was the best tradeoff between sensitivity and specificity reached by the Hitsensor algorithm. *In part (b), these results are obtained from Ref. 18 on 190 pairs.

Table 3. The number of positive predictions of the compared algorithms. The *subtotal* row lists the total number of correct predictions on all testing datasets. The last row shows the threshold scores to obtain the results. Algorithm names are the same as those in Fig. 2. The best prediction performances, i.e. the largest numbers for datasets with functional pairs and the smallest number for dme83N with non-functional pairs, are shown in bold face.

	3′UTR						CDS		3′UTR+CDS	
	HITS	MIRA	PITAn	PITAf	TSS	PicTar	HITS	MIRA	HITS	MIRA
dme96P	**72**	40	69	69	31	61	13	21	77	48
dme83N	17	25	22	22	**4**	19	15	25	32	38
dme16P	**11**	1	5	2	2	10	0	0	11	1
cel	6	4	8	**9**	4	4	1	2	7	6
hsa	**237**	96	226	202	151	117	50	54	268	132
mmu	30	21	17	11	**31**	7	14	16	39	30
unc-hsa	**9**	1	6	7	0[a]	0[a]	3	5	10	6
subtotal	**293**	123	262	231	188	138	68	77	335	175
threshold	*472*	*139*	*−6.8*	*−2.2*	NA	NA	*472*	*139*	*472*	*139*

[a]Results from Ref. 33.

We also tried different rewards to adenosine (A) opposite the first nucleotide of miRNA. The results showed that increasing reward of first A resulted in a marginal increase of AUC and S2N values (not shown), with a maximal AUC value when reward to first A is around 52.

3.1.2. *miRNA complementary sites in* 3′ *UTRs and CDS*

Although most verified animal miRNA complementary sites are located in 3′ UTRs of targets,[9–16] some mammalian coding genes also have miRNA complementary sites in their CDS.[35,36] As shown in Table 3, both Hitsensor and Miranda predicted more miRNA complementary sites in 3′ UTRs than in CDS. For instance, Hitsensor predicted 237 sites in 3′ UTRs while only 50 sites in CDS for hsa. It is important to note that we found that some miRNAs can have complementary sites in both 3′ UTRs and CDS. We found that among the 50 miRNAs that have complementary sites in CDS of human genes (on hsa dataset), 19 also have complementary sites in 3′ UTRs (Table S4). The regulatory effects of these 50 miRNAs on CDS can be well explained by the microarray gene expression profiles of the targets (see Table S4) in Ref. 7. This suggests that these miRNA sites in CDS might play roles in the regulation of the targets. Furthermore, many miRNA complementary sites in CDS of RTL1/Rtl1, such as those of miR-136 and miR-341, have been directly verified with 5′ RACE.[35]

We also find that the miRNA complementary sites of two miRNA:target pairs, hsa-miR-125b:DDX19B/mmu-miR-125b-5p:Ddx19b and miR-431:RTL1/Rtl1, in CDS are conserved between human and mouse (see Table S3). It is interesting to point out that miR-125b:DDX19B was listed as a non-conserved pair in Ref. 33 because there were no conserved complementary sites in 3′ UTRs. However, our findings suggest that the regulatory relation of miR-125b and DDX19B is conserved between human and mouse through miR-125b complementary sites in CDS of DDX19B. As to be shown in Fig. 3(a), the conservation of miR-431 complementary sites in CDS of RTL1/Rtl1 have been verified in Ref. 35. In addition, a recent study also demonstrated that miR-148 targets coding region of human DNMT3b, which is conserved in mammals.[36]

These findings suggest that the 3′ UTRs of animals have evolved to accommodate most miRNA complementary sites, meanwhile coding regions still maintain a small portion of miRNA complementary sites.

3.2. *Comparisons with the existing methods*

Hitsensor achieved the best overall performance in all algorithms compared on both training and testing datasets; the results are shown in Fig. 2 and Table 3. On the training datasets, Hitsensor reached a sensitivity of 75% (72/96) and a specificity of 79.5% (1−17/83), which are 3% and 6% higher than those of PITA, respectively. As shown in Table 3, PITA had the best performance among all existing algorithms. This was also shown in Fig. 2(a), where the closest point of all ROC curves to

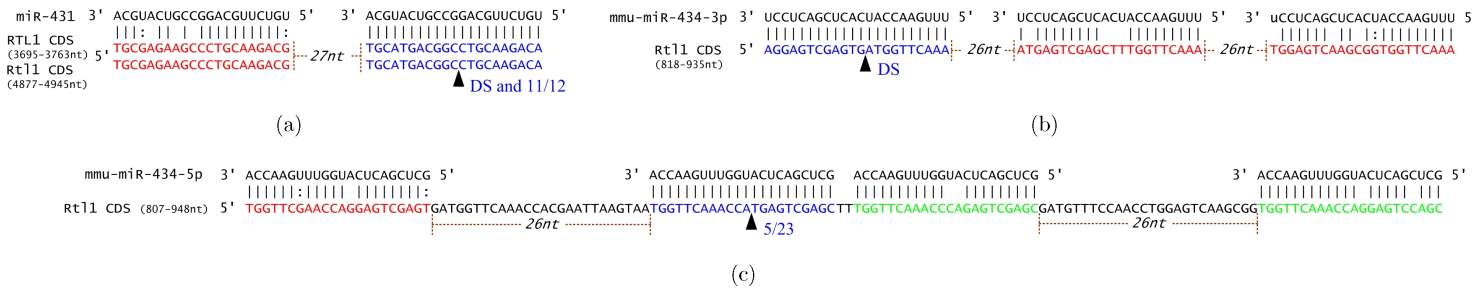


Fig. 3. Predicted putative synergistic miRNA binding sites on RTL1 (of *Homo sapiens*)/Rtl1 (of *Mus musculus*). Blue sites were reported in Ref. 35. Red and green sites are putative synergistic complementary sites predicted by the Hitsensor algorithm. Black arrows indicate cleavage sites reported in Ref. 35, which were identified by RLM 5′ RACE either by direct sequencing of the PCR products (DS) or by sequencing of individual cloned products. The numbers indicate the fraction of clones that identify the blue cleavage site.[35] (a) Conserved miR-331 complementary sites on RTL1/Rtl1. (b) mmu-miR-434-3p complementary sites on Rtl1. (c) mmu-miR-434-5p complementary sites on Rtl1.

the upper-left corner is on the ROC curve of Hitsensor. We attribute this to the 12–17 nt determinant used by Hitsensor. As discussed earlier, Hitsensor could reach optimal tradeoff between sensitivity and specificity when the reward base of 12–17 nt region, U, was 4 [see Fig. S2(b)]. Meanwhile, other algorithms compared did not use information from 12–17 nt region, as shown in Table 2. If taking CDS of targets into account, Hitsensor could have a sensitivity of 80.2% and specificity of 70% on the training datasets (see Table 3).

On the testing sets, Hitsensor had an overall sensitivity of 54.2% (293/541), again the highest among all compared algorithms. When compared with the best sensitivity of the existing algorithms (from PITAn), Hitsensor had an improvement of 5.8%. Hitsensor found another 42 pairs, 7.8%, on all test datasets if both 3′ UTRs and CDS were considered, as shown in Table 3. On individual datasets, Hitsensor performed the best in four out of the seven datasets, shown in bold fonts in Table 3. On dme83N, Hitsensor produced 17 false positive predictions, which was only larger than that of TargetScanS. However, the sensitivities of TargetScanS were much lower than Hitsensor, except for the mmu dataset.

Hitsensor reached an AUC value of 0.794 that is slightly higher than those of PITA, with and without flanking sequences, and much higher than that of Miranda [Fig. 2(b)]. As reported in Ref. 18, PITAf had an AUC of 0.79 on 190 samples, which were higher than those from the method in Ref. 15, PicTar[13] and Miranda[11] [see Fig. 2(b)]. PITA had a similar performance on our slightly reduced datasets to that reported by Ref. 18, which suggests that it is meaningful to compare our results with those methods in Ref. 18 [starred methods in Fig. 2(b)]. Again, Hitsensor had a higher AUC value than those methods in Ref. 18 [see Fig. 2(b)]. Miranda performed better on the 190 samples in Ref. 18 than on our training data with 179 samples, which might have resulted from different versions of Miranda and/or different methods to calculate miRNA:target scores. Hitsensor also had higher S2N values when compared with PITA and Miranda, as shown in Fig. 2(c). Wang and Naqa[37] also used the AUC to evaluate their method. Their models reached AUC values of 0.79 and 0.77 with and without the conservation information, respectively.[37] Hitsensor obtained a slightly better AUC value than that of Wang and Naqa's method[37] even though Hitsensor did not use mRNA expression information.

As shown in Table 3, PITA performed well by using free energy of target 3′ UTRs and miRNA:target duplex (Table 2). In contrast, Hitsensor achieved an overall better performance than PITA without employing the thermodynamical information used by PITA, which is computationally expensive. Because all algorithms used seed information, we attribute this improvement to two unique features that Hitsensor used, the 12–17 nt region and the local AU-content (Table 2). As discussed early, 12–17 nt region is effective to improve the tradeoff between sensitivity and specificity [Fig. S2(b)]. The reward to local AU-content determinant improved the AUC of Hitsensor [Fig. S3(c)]. In addition, the score of local AU-content is computationally cheaper to compute than the free energy of 3′ UTR and miRNA:target duplex used by PITA.

Table 4. The overlapped predictions in 3′ UTRs of different algorithms on the dme96P (below upper-left to lower-right diagonal) and hsa datasets (above upper-left to lower-right diagonal). The value in each cell means the overlapped predictions of the two algorithms from the row and column of the cell. The last row and column list the total number of commonly predicted pairs with other algorithms for the algorithm in this column and row on dme96P and hsa datasets, respectively. Algorithm names are the same as those in Fig. 2.

hsa dme96P	HITS	MIRA	PITAn	PITAf	TSS	PicT	Total
HITS	—	95	171	147	129	101	643
MIRA	19	—	75	62	42	35	309
PITAn	55	37	—	167	120	93	626
PITAf	57	37	62	—	100	74	550
TSS	28	12	24	26	—	109	500
PicT	56	30	47	49	28	—	412
Total	215	135	225	231	118	210	—

We also compared the overlapped predictions of different algorithms for the dme96P and hsa datasets, and the results are shown in Table 4. For a given algorithm, the total number of overlapped predictions showed capability of this algorithm to find predictions from other algorithms compared. We thus listed the total number of overlapped predictions in the last column (for hsa dataset) and last row (for dme96P dataset). For instance, Hitsensor respectively had 643 and 215 total common predictions for hsa and dme96P with the other five algorithms compared. As shown in Table 4, Hitsensor made a much larger number of common predictions than Miranda, PITAf, TargetScanS and Pictar for the hsa dataset. For the dme96P dataset, Hitsensor, PITAn, PITAf and PicTar made comparable number of total common predictions, and the total common predictions of Miranda and TargetScanS were much smaller than the other four algorithms compared. These indicate that Hitsensor could successfully find major parts of correct positive predictions produced by other algorithms. For example, Hitsensor found 171 out of the 226 (75.7%) predictions of the hsa dataset from PITAn.

3.3. *Synergistic complementary sites*

It has been observed that miRNAs can act synergistically in post-transcriptional regulation.[28,38] This has also been observed in our results, listed in Supplemental Table S5. We found that 12 miRNA:target pairs, which span over 11 miRNAs and 10 targets, in Table S5, have putative synergistic complementary sites of the same miRNA in the selected datasets.

We analyzed the complementary sites on RTL1 (of *Homo sapiens*)/Rtl1 (of *Mus musculus*) in Fig. 3, where Hitsensor predicted a total of 6 new synergistic complementary sites (red and green sites), in addition to the 3 blue sites reported in Ref. 35. The Hitsensor algorithm predicted two conserved synergistic miR-431

complementary sites on RTL1/Rtl1, as shown in Fig. 3(a). Davis *et al.*[35] reported that 11 out of 12 clones correspond to the blue site. This suggests that at least some of the clones might be produced by the newly found red site in Fig. 3(a). Figure 3(b) shows that, in addition to the site reported in Ref. 35, Hitsensor predicted two more complementary sites of mmu-miR-434-5p. Davis *et al.*[35] reported that only 5 out of 23 clones were shown to be the cleavage product of mmu-miR-434-5p at the position pointed by the arrow in Fig. 3(c), and other clones were supposed to be random Rtl1 degradation products.[35] However, the predicted synergistic mmu-miR-434-5p complementary sites in Fig. 3(c) suggest that the remaining clones are very likely to be cleavage products from the newly predicted red mmu-miR-434-5p sites. Furthermore, Hitsensor also predicted another pair of synergistic mmu-miR-434-5p sites, i.e. the green sites in Fig. 3(c), which are 2 nt downstream of the blue site. They only have two mismatched nucleotides and have an intersite distance of 26 nt. They might also produce some of the remaining clones detected by Davis *et al.*[35]

Figure 3 shows that there exist two levels of cooperative miRNA-induced repression on Rtl1. First, at least three miRNAs, mmu-miR-431, mmu-miR-434-3p and mmu-miR-434-5p cooperatively repress Rtl1 by binding to their respective complementary sites on Rtl1. Furthermore, Davis *et al.*[35] reported that mmu-miR-127, mmu-miR-136, mmu-miR-433-3p and mmu-miR-433-5p are also involved in repressing Rtl1. Second, several copies of mmu-miR-431, mmu-miR-434-3p and mmu-miR-434-5p may bind to their respective synergistic complementary sites and collaboratively repress Rtl1.

3.4. *Performance on protein expression data*

Two methods can be used to examine the strength of miRNA-mediated repression and ratio of responsiveness to miRNA upregulation or downregulation of the predicted targets, either to transfect a specific miRNA to a sample[28,30] or to delete a specific miRNA from a sample[30] before measuring the changes of the mRNA and protein levels in the sample. As argued by Baek *et al.*,[30] responsive proteins may not necessarily be the direct targets of transfected miRNAs. We thus used the protein upregulation dataset of miR-223 knockout in mouse neutrophils[30] to examine the miRNA-mediated repression of predicted targets. This dataset provides both quantitative information of proteomic changes of about 3,800 coding genes in wild-type and miR-223-deficient mice. The latest prediction results of TargetScan,[9,10] Miranda,[11,12] PicTar,[13] PITA,[18] and miRBase Targets[34] were downloaded from their corresponding websites, and then mapped to the protein change dataset of the 3,800 genes. Duplicate miRNA:target pairs were removed before calculating the mean protein fold changes. To test the effect of adenosine (A) opposite the first nucleotide of miRNA, we tested two sets of parameters for Hitsensor, one without rewards to the first A and the other with a reward of 52 to the first A. Accordingly, the threshold score of Hitsensor for the latter case was increased to 518 chosen with

the ROC curve of the dme96P and dme83N datasets in Table 1 by choosing the point nearest to the upper left corner (as discussed in Evaluation Methods).

We first compared the mean protein fold changes of conserved miR-223 targets predicted by the algorithms analyzed; the results are shown in Fig. 4(a). As mentioned in Materials and Methods, we selected targets that have homologous genes in human for the comparisons with the other algorithms. The version of Hitsensor that rewards the first adenosine (A) [Hitsensor (A) in Fig. 4] performed slightly better than Hitsensor, TargetScan and Pictar. Moreover, Hitsensor also predicted more targets than TargetScan and PicTar. As in Fig. 4(b), only the predictions from TargetScan (ordered by Total Context Score[28]) showed a significant difference in its top and bottom third ($P < 0.01$, Mann-Whitney U-test), although generally for all the algorithms that we tested the mean fold changes of the top third are greater than those in the bottom third. For non-conserved targets, Hitsensor also predicted targets with greater mean fold changes than those from PITA and targets with 7–8 mer sites, shown in Fig. 4(c) (left part). The top third predictions from PITA were also significantly different from its bottom third predictions ($P < 0.01$, Mann-Whitney U-test), as shown in Fig. 4(c) (right part).

We next examined the cumulative distribution of protein changes; the results are shown in Fig. 4(d). The cumulative distributions of Hitsensor with rewards to the first adenosine and TargetScan are the two best cases among the distributions of all algorithms compared. Because the maximal cumulative difference from the no-site distribution, black curve in Fig. 4(d), gives an estimation of percentage of targets responsive to miR-223 deletion,[30] we then compared the algorithms in two ways. First, given the same number of predictions, we compared the maximal cumulative differences of different algorithms. Second, given the same (or comparable) maximal cumulative difference, we compared the number of predictions from different algorithms. Thus, we show in Fig. 4(e) the maximal cumulative differences from the no-site distribution when the number of predictions is fixed at 41, which is the number of predictions from TargetScan. Hitsensor (with or without reward to first A) had a slightly better maximal cumulative difference than TargetScan and PicTar. Next, we compared the number of predictions when different algorithms had comparable maximal cumulative differences with respect to the no-site distribution. As shown in Fig. 4(f), Hitsensor predicted more targets than TargetScan and PicTar, with an increase of 49% and 97%, respectively.

Then, we examined the overlapped predictions between Hitsensor and TargetScan and between Hitsensor and PicTar. Hitsensor identified 27 of the 41 predictions of TargetScan and 21 of the 31 predictions of PicTar, respectively.

Finally, we compared the results of Hitsensor on miR-1Trans, miR-124Trans, miR-181Trans and miR-223KO with the methods reported in Ref. 31. Hitsensor achieved AUC values of 0.56, 0.58, 0.55 and 0.64 on miR-1Trans, miR-124Trans, miR-181Trans and miR-223KO, respectively (lines 1 to 4 in Table 5). Hitsensor had an AUC value of 0.58 on the combined dataset of these four datasets, as shown on line 5 in Table 5. The optimal score of combined dataset is 507, which is slightly

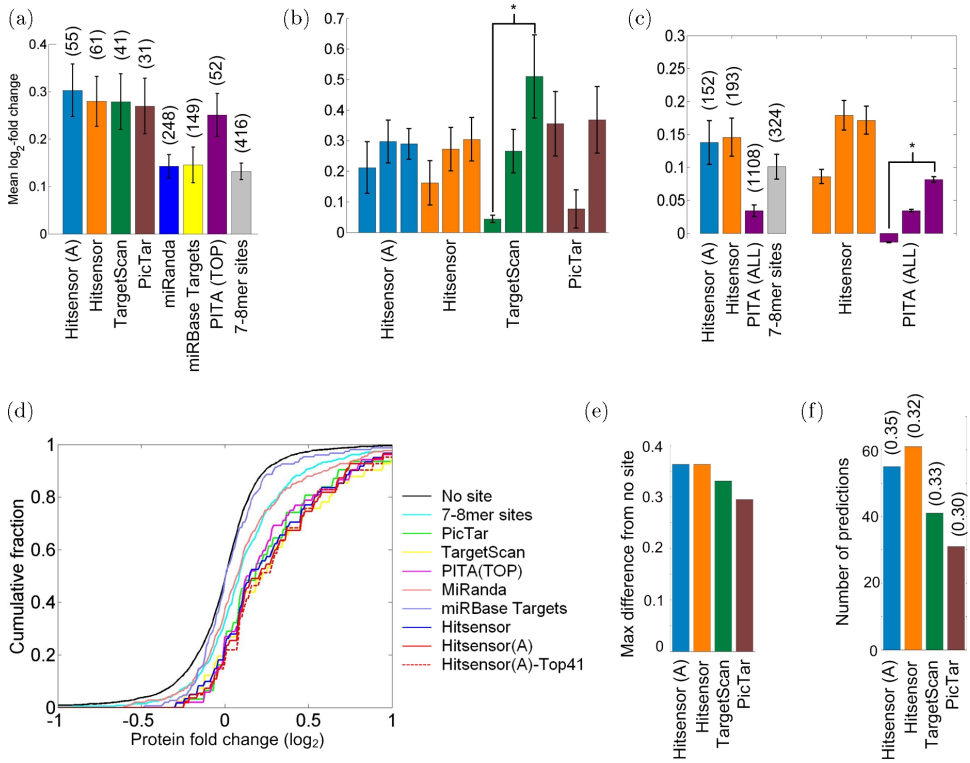


Fig. 4. The correspondence between computationally predicted targets and their changes of protein expression levels in miR-223 knockout mice. (a) Mean protein expression level change (upregulation) of the predicted targets from algorithms that used conservation information, where error bars are ± standard error. Only conserved miRNA:target pairs are used for Hitsensor. The 7–8 mer sites are targets with 7–8 mer complementary sites in their 3′ UTRs. The numbers in parenthesis are the numbers of predicted targets. (b) The relationship between protein upregulation and scores given by different prediction algorithms. The protein fold changes of predicted targets are divided into three equal-size bins according to their scores given by different algorithms. The three bars from left to right correspond to the bottom, central, and top third. Statistically significant differences between the bottom and top third are indicated (asterisk, $P < 0.01$, Mann-Whitney U-test). One predicted target Slc29a1, with a fold change of 2.28, of Hitsensor is regarded as an outlier ($P < 10^{-5}$, by assuming a normal distribution for the changes of protein expression levels of targets predicted by Hitsensor) and removed before drawing figure of Hitsensor. (c) Mean protein fold upregulation of predicted targets that did not use conservation information, with legend the same as in (a) and (b). (d) Cumulative distribution of protein upregulation of predicted targets of miR-223. Plotted curves show the fraction of proteins that change at least to the degree indicated on the x axis. The no site curve shows the distribution of genes without 6–8 mer miR-223 sites in their 3′ UTRs. (e) The maximal cumulative difference between predicted targets and the no-site distribution when the numbers of predictions are comparable. Only the top 41, the number of predictions of TargetScan, predictions of Hitsensor are used. To correct for bumpiness in the cumulative distributions, we calculated the bumpiness values for different sample sizes as described in Ref. 28, i.e. 0.062 for 55, 0.058 for 61, 0.079 for 41, and 0.070 for 31 predictions, respectively. The plotted values are their original maximal differences minus the bumpiness values corresponding to their sample sizes. (f) The numbers of predictions when the maximal cumulative differences are comparable. The numbers in parenthesis are maximal cumulative differences after subtracting their corresponding bumpiness values.

Table 5. The performance of Hitsensor on miR-1Trans, miR-124Trans, miR-181Trans, and miR-223KO datasets.

	Data	AUC	Optimal score	S2N
1	miR-1Trans	0.56	535	0.81
2	miR-124Trans	0.58	536	0.73
3	miR-181Trans	0.55	487	1.22
4	miR-223KO	0.64	479	3.29
5	Overall	0.58	507	3.22

greater than 472 obtained on the dme96P and dme83N datasets in Table 1. In comparison, Hausser *et al.*[31] achieved AUC values of 0.57, 0.60 and 0.59 with generalized linear models (GLM) trained on the datasets which were prepared with mRNA expression, protein expression and both, respectively. TargetScanS reached the best AUC of 0.65 and PITAf had an AUC of 0.57.[31]

4. Discussion

We have studied the effects of different sequence-specific determinants on predicting miRNA target complementary sites and developed a new miRNA target prediction algorithm which we called Hitsensor. The Hitsensor algorithm has a superior performance over five benchmark miRNA target prediction methods that we compared on an extensive collection of experimentally validated datasets. We attribute the performance of Hitsensor to three major aspects.

First, we used various determinants proposed in previous studies, especially in but not limited to Refs. 28, 26, 30, 31, 24 in our methods. We also included a new scheme to quantify the conventional seed region used by the other algorithms. As discussed in Methods, our quantification method to seed region, as well as 12–17 nt region, has given much higher rewards to continuously matched seed regions than discontinuously matched counterparts, which might be produced by random chance. This has helped to distinguish functional miRNA:target pairs to randomly paired non-functional miRNA and mRNA. Another important factor contributing to the success of Hitsensor is local AU content around seed region. Functional miRNA complementary sites are often located in AU-rich regions in 3′ UTRs of targets.[27,28] Appropriate reward to local AU content has helped to improve the AUC and optimal sensitivity versus specificity of Hitsensor, as shown in Figs. S3(c) and S2(f).

Second, Hitsensor could predict species specific miRNA:target relations. In comparison, TargetScanS, PicTar and Miranda used conservation information in their prediction, which let them miss some species specific miRNA:target pairs, as shown by their predictions on the unc-hsa dataset in Table 3.

Finally, as shown in Table 3, we found that 13.5% (13/96) and 12.6% (68/541) of functional pairs of training and testing datasets, respectively, have predicted complementary sites in CDS of targets. Some of the predicted complementary sites

in CDS had been verified in Ref. 35. As shown in Results, Hitsensor had better sensitivities when taking CDS into account. These results suggest that CDS of targets contain substantial percentage of miRNA complementary sites and should not be ignored when performing target prediction for animal miRNAs, although most miRNA complementary sites are located in 3′ UTRs, as shown in Table 3 and in Refs. 9–16. Hitsensor made 15 positive predictions for the dme83N dataset when using CDS. Further experiments are necessary to verify whether these sites function or not, because only 3′ UTRs of the targets were tested with reporter gene assays (see Ref. 18 and references therein).

As shown in Fig. 2 and Table 3, PITA also performed well on the selected datasets. These suggest that difference between free energy of miRNA:target duplex and energy cost of unpairing the 3′ UTR of target used by PITA is useful information in predicting animal miRNA targets. Hitsensor does not use the folding energy of miRNA:target duplex and 3′ UTRs. However, there is a relationship between local AU-content and energy cost of unpairing the 3′ UTR, because high local AU-content around seeds reduces the energy costs to make seeds accessible for miRNAs loaded in the RNA-induced silencing complex (RISC, see Ref. 1). These imply that the seed and its flanking region are two critical factors that affect the performance of target prediction algorithms. Hitsensor performed better than PITA except for the cel dataset (see Table 3) because it used additional information from 12–17 nt determinant.

The results on the protein level change dataset suggested that Hitsensor predicted more conserved targets than TargetScan and PicTar. Meanwhile, the conserved targets predicted by Hitsensor had an average miRNA-mediated repressive strength comparable to that from TargetScan. Generally, a higher Hitsensor score of a miRNA:target pair indicated a stronger miRNA-mediated repression. With a comparable number of predictions, Hitsensor had a better maximal cumulative difference than TargetScan and PicTar. Furthermore, when the maximal cumulative differences are comparable, Hitsensor predicted more targets than TargetScan and PicTar. Generally, when the threshold for Hitsensor was increased, it predicted less targets, while its maximal cumulative difference was increased. A high threshold can be used to find a small set of reliable and responsive targets; likewise, a low threshold can be adopted to find a relatively large set of targets with an acceptable maximal cumulative difference from the no-site distribution. We included in supplementary materials tables of scores versus maximal cumulative differences calculated from the protein and mRNA datasets of miR-223 knockout in Ref. 30 to help choose Hitsensor thresholds.

Due to the lack of negative samples, we used protein expression profiles to prepare negative and positive miRNA targets as those having the least and greatest protein level changes, similar to those used in Ref. 31. Although the GLM models in Ref. 31 consist of as many as 14 non-redundant features, the AUC values of these models are not significantly better than that of Hitsensor. This suggests that the sequence-specific features have their limitations in predicting the miRNA-induced protein expression changes, or that there may be a better strategy of using these

features. Another possibility is that the genes with the least protein fold-changes may carry true miRNA target sites, since the miRNA-mediated repression of protein expression is often mild.[39] In addition, another study found that targets of miR-124 showed larger changes at the mRNA level than the protein level, with average decreases of 35% and 12%, respectively.[40] In light of these previous observations, we examined the miR-124Trans data, and found that 22 of the 51 non-functional targets have negative log2 protein fold-changes, with the smallest value of −0.13, and the log2 protein fold change of a 12% decrease is −0.18. These results further suggest that the genes that have the least protein changes are not necessarily the non-functional targets. Hausser *et al.*[31] also suggested that these protein expression profiles were preliminary for studying the effects of different determinants.

On the other hand, conserved seeds lead to greater repression at the protein level than non-conserved seeds.[39] TargetScanS performed better than other methods, probably because it used both conservation and transcriptional expression profiles with transfected miRNAs.[31]

The design of the Hitsensor is simple. The scores for individual determinants are calculated separately after aligning miRNAs to their potential target sequences with the Smith–Waterman algorithm. Therefore, if new knowledge of miRNA target recognition emerges, it can be added to the existing scores as additional determinants.

As a final note, target expression data are needed to build regression models to calculate the Context Scores. However, mRNA and/or protein expression data on tissues with a transfected miRNA or a deleted miRNA may not be easily available. This makes it difficult to calculate Context Scores if different miRNAs have different effects on the expression levels of their targets. In contrast, Hitsensor only uses the sequences of miRNAs and their targets, which makes Hitsensor applicable to a broader arena than methods based on Context Scores.

Acknowledgments

The research was supported in part by grant 10ZR1403000 of STCSM and a start-up grant of Fudan University to YZ and two NSF grants (IIS-0535257 and DBI-0743797), two NIH grants (AR058681 and AI057160) and a grant from the Alzheimer's Association to WZ. We thank Ian Kniseley for his help in proofreading the manuscript.

References

1. Bartel DP, MicroRNAs: Genomics, biogenesis, mechanism, and function, *Cell* **116**:281–297, 2004.
2. Esquela-Kerscher A, Slack FJ, Oncomirs — microRNAs with a role in cancer, *Nat Rev Cancer* **6**:259–269, 2006.
3. Llave C, Xie Z, Kasschau KD, Carrington JC, Cleavage of scarecrow-like mRNA targets directed by a class of Arabidopsis miRNA, *Science* **297**(5589):2053–2056, 2002.
4. Tang G, Reinhart BJ, Bartel DP, Zamore PD, A biochemical framework for RNA silencing in plants, *Genes Dev* **17**(1):49–63, 2003.

5. Yekta S, Shih I-H, Bartel DP, MicroRNA-directed cleavage of HOXB8 mRNA, *Science* **304**(5670):594–596, 2004.
6. Bagga S, Bracht J, Hunter S, Massirer K, Holtz J, Eachus R, Pasquinelli AE, Regulation by let-7 and lin-4 miRNAs results in target mRNA degradation, *Cell* **122**:553–563, 2005.
7. Lim LP, Lau NC, Garrett-Engele P, Grimson A, Schelter, JM, Castle J, Bartel DP, Linsley PS, Johnson JM, Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs, *Nature* **433**:769–773, 2005.
8. Wu L, Fan J, Belasco JG, MicroRNAs direct rapid deadenylation of mRNA, *Proc Natl Acad Sci USA* **103**(11):4034–4039, 2006.
9. Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP, Burge, CB, Prediction of mammalian microRNA targets, *Cell* **115**(7):787–798, 2003.
10. Lewis BP, Burge CB, Bartel DP, Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets, *Cell* **120**:15–20, 2005.
11. Enright A, John B, Gaul U, Tuschl T, Sander C, Marks D, microRNA target detection, *Genome Biol* **5**:R1, 2003.
12. John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS, Human microRNA targets, *PLoS Biol* **2**(11):e363, 2004.
13. Krek A, Grün D, Poy MN, Wolf R, Rosenberg L, Epstein EJ, Macmenamin P, daPiedade Id, Gunsalus KC, Stoffel M, Rajewsky N, Combinatorial microRNA target predictions, *Nat Genet* **37**(5):495–500, 2005.
14. Rajewsky N, Socci N, Computational identification of microRNA targets, *Genome Biol* **5**(2):P5, 2004.
15. Stark A, Brennecke J, Bushati N, Russell RBB, Cohen SMM, Animal microRNAs confer robustness to gene expression and have a significant impact on 3'utr evolution, *Cell* **123**(6):1133–1146, 2005.
16. Rehmsmeier M, Steffen P, Hochsmann M, Giegerich R. Fast and effective prediction of microRNA/target duplexes, *RNA* **10**(10):1507–1517, 2004.
17. Miranda KC, Huynh T, Tay Y, Ang Y-S, Tam W-L, Thomson AM, Lim B, Rigoutsos I, A pattern-based method for the identification of microRNA binding sites and their corresponding heteroduplexes, *Cell* **126**(6):1203–1217, 2006.
18. Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E, The role of site accessibility in microRNA target recognition, *Nat Genet* **39**(10):1278–1284, 2007.
19. Rhoades MW, Reinhart BJ, Lim LP, Burge CB, Bartel B, Bartel DP, Prediction of plant microRNA targets, *Cell* **110**(4):513–520, 2002.
20. Wang XJ, Reyes JL, Chua NH, Gaasterland T, Prediction and identification of arabidopsis thaliana microRNAs and their mRNA targets, *Genome Biol* **5**(9):R65, 2004.
21. Jones-Rhoades MW, Bartel DP, Computational identification of plant microRNAs and their targets, including a stress-induced miRNA, *Mol Cell* **14**(6):787–799, 2004.
22. Zhang Y, miRU: An automated plant miRNA target prediction server, *Nucleic Acids Res* **33**(suppl_2):W701–704, 2005.
23. Rajewsky N, microRNA target predictions in animals, *Nat Genet* **38 Suppl 1**(6s):2006.
24. Brennecke J, Stark A, Russell RBB, Cohen SMM, Principles of microRNA-target recognition, *PLoS Biol* **3**(3):2005.
25. Vella MC, Choi E-Y, Lin S-Y, Reinert K, Slack FJ, The C. elegans microRNA let-7 binds to imperfect let-7 complementary sites from the lin-41 3'UTR, *Genes Dev* **18**(2):132–137, 2004.
26. Nielsen CB, Shomron N, Sandberg R, Hornstein E, Kitzman J, Burge CB, Determinants of targeting by endogenous and exogenous microRNAs and siRNAs, *RNA* **13**(11):1894–1910, 2007.

27. Jing Q, Huang S, Guth S, Zarubin T, Motoyama A, Chen J, Padova FD, Lin S, Gram H, Han J, Involvement of microRNA in AU-rich element-mediated mRNA instability, *Cell* **120**(5):623–634, 2005.
28. Grimson A, Farh KK-H, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP, MicroRNA targeting specificity in mammals: Determinants beyond seed pairing, *Mol Cell* **27**(1):91–105, 2007.
29. Smith TF, Waterman MS, Identification of common molecular subsequences, *Journal of Molecular Biology* **147**:195–197, 1981.
30. Baek D, Villen J, Shin C, Camargo FD, Gygi SP, Bartel DP, The impact of microRNAs on protein output, *Nature* **455**(7209):64–71, 2008.
31. Hausser J, Landthaler M, Jaskiewicz L, Gaidatzis D, Zavolan M, Relative contribution of sequence and structure features to the mRNA binding of Argonaute/EIF2C-miRNA complexes and the degradation of miRNA targets, *Genome Res* **19**(11):2009–2020, 2009.
32. Sethupathy P, Corda B, Hatzigeorgiou AG, TarBase: A comprehensive database of experimentally supported animal microRNA targets, *RNA* **12**(2):192–197, 2006.
33. Sethupathy P, Megraw M, Hatzigeorgiou AG, A guide through present computational approaches for the identification of mammalian microRNA targets, *Nat Methods* **3**(11):881–886, 2006.
34. Griffiths-Jones S, Grocock RJ, vanDongen S, Bateman A, Enright AJ, miRBase: microRNA sequences, targets and gene nomenclature, *Nucleic Acids Res* **34**(suppl_1):D140–144, 2006.
35. Davis E, Caiment F, Tordoir X, Cavaillé J, Ferguson-Smith A, Cockett N, Georges M, Charlier C, RNAi-mediated allelic trans-interaction at the imprinted rtl1/peg11 locus, *Curr Biol* **15**(8):743–749, 2005.
36. Duursma AM, Kedde M, Schrier M, leSage C, Agami R, miR-148 targets human DNMT3b protein coding region, *RNA* **14**(5):872–877, 2008.
37. Wang X, El Naqa IM, Prediction of both conserved and nonconserved microRNA targets in animals, *Bioinformatics* **24**(3):325–332, 2008.
38. Doench JG, Sharp PA, Specificity of microRNA target selection in translational repression, *Genes Dev* **18**(5):504–511, 2004.
39. Selbach M, Schwanhausser B, Thierfelder N, Fang Z, Khanin R, Rajewsky N, Widespread changes in protein synthesis induced by microRNAs, *Nature* **455**(7209):58–63, 2008.
40. Hendrickson DG, Hogan DJ, McCullough HL, Myers JW, Herschlag D, Ferrell JE, Brown PO, Concordant regulation of translation and mRNA abundance for hundreds of targets of a human microRNA, *PLoS Biol* **7**(11):e1000238+, 2009.

Yun Zheng is an associate professor of Institute of Developmental Biology and Molecular Medicine, and School of Life Sciences, Fudan University, Shanghai, China. He received his B.Eng. in Manufacturing Engineering from Beijing University of Aeronautics and Astronautics, and his Ph.D. in Computer Science from Nanyang Technological University, Singapore. Dr. Zheng was a research fellow of National University of Singapore from 2005 to 2006, and a post-doctorial research associate of Washington

University in St. Louis from 2006 to 2009. Dr. Zheng is interested in computational methods related to microRNA-guided gene regulations, machine learning, and theoretical foundations of computational learning.

Weixiong Zhang is a full professor of Computer Science and of Genetics at Washington University in St. Louis, Missouri, USA. He received his B.S. and M.S. in Computer Engineering from Tsinghua University, Beijing, China, and his M.S. and Ph.D. in Computer Science from University of California at Los Angeles (UCLA). Professor Zhang's research interests include computational biology and genomics, artifical intelligence, data mining, and combinatorial optimization. He is currently associate editors of *PLoS Computational Biology*, *Journal of Alzheimer's Disease*, *International Journal of Artifical Intelligence*, and *AI Communications — The European Journal on Artificial Intelligence*; he also serves on the editorial board of *Journal of Artificial Intelligence Research*.