



# New support vector machine-based method for microRNA target prediction

L. Li, Q. Gao, X. Mao and Y. Cao

Department of Bioscience & Bioengineering, South China University of Technology, Guangzhou, China

Corresponding author: Y. Cao  
E-mail: [yccao@scut.edu.cn](mailto:yccao@scut.edu.cn)

Genet. Mol. Res. 13 (2): 4165-4176 (2014)

Received January 30, 2014

Accepted June 28, 2013

Published June 9, 2014

DOI <http://dx.doi.org/10.4238/2014.June.9.3>

**ABSTRACT.** MicroRNA (miRNA) plays important roles in cell differentiation, proliferation, growth, mobility, and apoptosis. An accurate list of precise target genes is necessary in order to fully understand the importance of miRNAs in animal development and disease. Several computational methods have been proposed for miRNA target-gene identification. However, these methods still have limitations with respect to their sensitivity and accuracy. Thus, we developed a new miRNA target-prediction method based on the support vector machine (SVM) model. The model supplies information of two binding sites (primary and secondary) for a radial basis function kernel as a similarity measure for SVM features. The information is categorized based on structural, thermodynamic, and sequence conservation. Using high-confidence datasets selected from public miRNA target databases, we obtained a human miRNA target SVM classifier model with high performance and provided an efficient tool for human miRNA target gene identification. Experiments have shown that our method is a reliable tool for miRNA target-gene prediction, and a successful application of an SVM classifier. Compared with other methods, the method proposed here improves the sensitivity and accuracy of miRNA prediction. Its performance can be further

improved by providing more training examples.

**Key words:** MicroRNA; Target-prediction; Support vector machine; Machine learning

## INTRODUCTION

MicroRNA (miRNA) is a type of non-coding small RNA (~21 to 23 nucleotides), which is produced by the Dicer enzyme from a stem-loop structured RNA precursor, and is widely expressed in animal and plant cells where it functions in the form of RNA-protein complexes, termed miRISCs (Gregory, 2006). MiRNAs have been implicated in the control of development because they lead to the destruction or translational suppression of target mRNAs with homology to the miRNA. Increasing evidence suggests that miRNAs play important roles in cell differentiation, proliferation, growth, mobility, and apoptosis. They act by binding to complementary sites of the target gene to induce cleavage with near-perfect complementarity, or to repress productive translation. They also facilitate deadenylation, which leads to rapid mRNA decay (Bartel, 2004). The choice between translational inhibition and destruction is governed by the degree of mismatch between a miRNA and its target mRNA. Such complex regulating mechanisms make finding target animal mRNAs difficult when using only sequence complementarity (Rhoades et al., 2002). Nevertheless, as research in this field progresses and expands, new theories are continuously being developed, resulting in precise predictions of potential targets.

Several computational methods have been proposed for miRNA target prediction studies. These methods can be generally classified into rule-driven algorithms and data-driven algorithms. The commonly used rule-driven algorithms include miRanda, TargetScan, PITA, and RNAhybrid, among others (Lewis et al., 2003; John et al., 2004; Rehmsmeier et al., 2004; Kertesz et al., 2007), which usually operate by designing precise sequence complementarity rules and defining seed sites to scan miRNA interaction sites in the target mRNAs. The data-driven algorithms are usually based on the hidden Markov model (HMM) or the support vector machine (SVM)-learning algorithm, by establishing complex feature models to discover knowledge directly embedded in existing data. At present, the proposed data-driven methods include MirTarget2, PicTar, TargetSpy, and TargetMiner, among others (Grün et al., 2005; Wang and El Naqa, 2008; Bandyopadhyay and Mitra, 2009; Sturm et al., 2010). Some of these previous methods have shown high performance and they have been widely used; however, the exact mechanism of their inhibition of protein synthesis remains unclear. miRNA interaction sites have long been assumed to be mainly located in the 3'-untranslated region (UTR) of target genes, and only a single interaction site factor is considered in most traditional miRNA target-prediction methods. However, an increasing number of studies have indicated that many miRNA interaction sites may be found in the 5' UTR and open reading frames of target genes. Lytle et al. observed that miRNA-binding sites repress target mRNAs just as efficiently in the 5' UTR as they do in the 3' UTR, and that many genes have several target sites for either one miRNA or a few different miRNAs (Lytle et al., 2007; Hammell, 2010). Therefore, these traditional miRNA target prediction methods that are based on the single-interaction site searching theory or the scale-limited site searching theory are clearly becoming insufficient.

SVM is based on statistical learning theory and the theory of Vapnik-Chervonenkis di-

mension based on the structural risk minimization principle. According to the limited sample of information in model complexity and learning ability, the aim of SVM is to find the best compromise to obtain the best generalization ability (or generalization) (Cortes and Vapnik, 1995; Furey et al., 2000). In this paper, we propose to improve the performance of existing target-prediction methods by extending the miRNA potential interaction site searching scale to a whole target gene sequence, and by providing two-site information for training our machine-learning model. Moreover, we present an SVM-based online tool for predicting miRNA target genes.

## MATERIAL AND METHODS

### Training datasets and testing datasets

To date, three comprehensive databases of experimentally supported animal miRNA targets exist worldwide (TarBase, pSILAC, and miRecords), providing comprehensive datasets to assess the features of miRNA targeting. To obtain high-quality training data to ensure accuracy of the prediction method, only experimentally verified human miRNA and target gene pairs from the three databases were selected in this study. Considering the large quantity of data, the data selected from miRecords and pSILAC were used as the positive training set, and the data selected from TarBase were used as independent testing datasets. We also obtained the miRNA detail sequence from miRBase, a searchable database of published miRNA sequences and annotations (Griffiths-Jones et al., 2006). A Perl script was used to fetch all of the target gene primary codes and 3' UTR information from the National Center for Biotechnology Information (NCBI) in real time (Larry Wall and Orwant, 2000). Another script was coded for deleting redundant datasets. Ultimately, we obtained 2778 experimentally verified miRNA:target pairs as the positive training dataset and 432 independent miRNA:target pairs for testing our SVM model.

Negative training datasets usually contribute to the specificity of an SVM classifier more significantly than do positive data (Nguyen et al., 2010). As positive miRNA:target interaction is not clearly defined, and biological experiments are not accurate with respect to negative miRNA target-gene identification, most mRNAs show no positive response to the corresponding miRNAs in experiments, and researchers will not voluntarily release this negative data. Currently, no database is available for collecting negative miRNA:target pair data. Thus, results of several miRNA overexpression microarray experiments were extracted to obtain negative training data. In such experiments, the expression level of mRNAs is inhibited when miRNA causes the destruction of target mRNAs. If miRNAs cause translational suppression of target mRNAs, their expression level remains the same. If mRNAs have an interaction site that can be targeted by some miRNAs, the expression level of the mRNA will not be significantly increased (Lim et al., 2005). Therefore, if the expression level of mRNA significantly increases in a miRNA overexpression experiment, this would indicate that the mRNA has no direct reaction site to the overexpressed miRNA.

We obtained a large amount of miRNA overexpression chip experiment data from the Gene Expression Omnibus (GEO) database of NCBI (Barrett et al., 2005). To obtain high-quality negative datasets, only data with relatively high stability ( $P$  value  $< 0.05$ ) and a large degree of upregulation (fold-change  $> 1.3$  to  $2.5$ ) were selected from a 40-chip experiment series. Ultimately, we obtained 714 negative miRNA:gene pairs (Table 1), from which 632

pairs were used as negative training datasets and the other 82 pairs were used as independent negative testing datasets.

**Table 1.** Source of negative training samples.

GEO ID	miRNA	miRNA:mRNA pairs selected
GSM156546, GSM156550	has-miR-16	59
GSM328290, GSM328287	has-miR-192	48
GSM156557, GSM156558	has-let-7c	21
GSM210900, GSM210901	has-miR-122	32
GSM210902, GSM210903	has-miR-128	6
GSM210904, GSM210905	has-miR-132	4
GSM156554, GSM156556	has-miR-20a	39
GSM156548, GSM156552	has-miR-215	86
GSM210906, GSM210907	has-miR-133a	69
GSM210908, GSM210909	has-miR-142	10
GSM210910, GSM210911	has-miR-148b	50
GSM156545, GSM156549	has-miR-15a	59
GSM156553, GSM156555	has-miR-17-5p	50
GSM187633, GSM187634, GSM187631, GSM187632	has-miR-34a	56
GSM190765, GSM190757	has-miR-34b	34
GSM190758, GSM190766	has-miR-34c	31
GSM210896, GSM210897	has-miR-7-1	2
GSM210898, GSM210899	has-miR-9	28

## Potential site searching

The traditional miRNA prediction method usually focuses on selecting interaction sites, binding free energy calculation, and evaluation of primary sequence conservation. Although these three processes are very similar, each has a different definition and arrangement (Zhang et al., 2006). For binding free energy, most of the previous methods calculate the whole section of the double-stranded target binding free energy of the miRNA:mRNA pair. However, many interaction sites are contained within the secondary structures of the mRNAs. Thus, the RNA refolding free energy should also be considered to distinguish the true interaction sites from the false complementary sites. For the evaluation of the conservation of the primary sequence, traditional methods generally adopt multiple sequence alignment (Carrillo and Lipman, 1988). Although common multiple sequence alignment methods have obviously higher computing complexity, a lighter algorithm is required to calculate the sequence conservation more efficiently, especially for the short sequences located in the interaction sites.

Complementary matching between miRNAs and their binding sites is always used as a significant basis for identifying miRNA targets. Undoubtedly, definition of complementary seed sites is very important in any miRNA target-prediction method. It also plays a constructive role in previous rule-driven prediction methods. Many methods improve specificity by limiting the complete complementary pairing in the seed region. Lewis used a signal to noise ratio (SNR) analysis method to demonstrate that the complete complementary pairing of 5' terminal bases has special contributions for improving SNR prediction (Lewis et al., 2005). Thus, we devised the following search rules to improve the accuracy of miRNA target site searching. We also used the RNAhybrid program to search for the sequence block with the lowest free energy when paired complementarily with specified miRNA.

- 1) The region between nucleotides 2 and 8 at the 5' end of the miRNA is defined as the

seed region, and the number of consecutive bases should not be less than five in the seed region.

2) The number of mismatched bases should be less than four in the whole interaction site.

3) The number of consecutive mismatched bases should not be more than three in the whole interaction site.

4) The free energy value of the double-strand complementarily paired RNA should be lower than -30.

Unlike many other methods, we searched two potential interaction sites from the target mRNA. We searched for the first matching site only in the 3' UTR region of the target mRNA, from which we obtained the first lowest free energy value interaction site. Then, we blocked the site from the whole target mRNA sequence and searched for another matching site in the rest of the area. This time, the searching scope was no longer limited to the 3' UTR region. Finally, we obtained two potential interaction sites, the primary sequence code, details of complementary pairing, the lowest energy value, and the P value. All of this information was returned to the main program.

To reduce the number of omitted miRNA targets to its minimum, our program will also search for two lower free energy sites, and will automatically return their information to the main program if no complementary site confined to the searching rules listed above is found in the whole mRNA sequence.

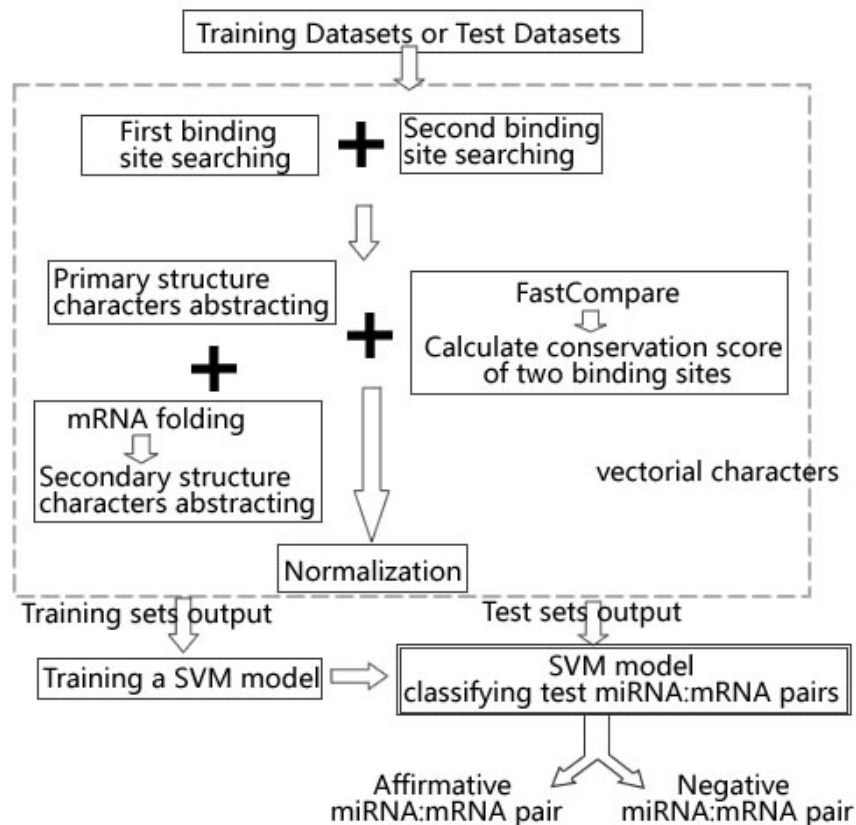
## Vectors and scheme

During SVM training and classification, only numerical features can be input as feature vectors. However, most of the site information returned to the main program contained RNA sequence or complementary pairing features. Therefore, we needed to map the sequence information in numerical form, which is the definition of feature vectors. The purpose of defining feature vectors is to design high-quality vectors that can reflect the information contained in the sample as comprehensively as possible, to allow the learning machine to obtain maximum knowledge, and to verify that the mathematical model fits the sample characteristics (Fung and Mangasarian, 2004). No fixed mode or unchangeable method exists for defining feature vectors, which requires an algorithm design with specific considerations for different practical problems. We defined 133 feature vectors that fall into three classes: primary structural features, secondary structural features, and sequence conservation features. Specifically, we defined 1) 84 primary structural feature vectors: single-base number, double-base number, triple-base number, and GC ratio; 2) two interaction site feature vectors: the lowest free energy and P value; 3) 46 secondary structure feature vectors: paired base number to total base number ratio, unpaired base number to total base number ratio, refolding free energy, number of stems of various sizes, number of loops of various sizes, total number of stems and loops, and sizes of all kinds of stems and loops; and 4) one site sequence conservative score feature vector.

Secondary structure features are important because they imitate the shape and mechanism of miRNA:site pairing. As most of the lengths of the returned sequences were less than 40 bases, precisely predicting the secondary structure of such a short mRNA region is very difficult. Thus, we obtained 40 base sequences located at the 5' and 3' ends of the returned site from the origin mRNA. The returned sequence was called the core block, and the other 80 base sequences that were fetched later were called the non-core block. Then, we obtained an mRNA sequence with approximately 120 bases and used the RNAfold program to predict its

secondary structure (Hofacker et al., 1994). After processing the secondary structure feature vector extraction, we obtained 46 secondary structure features, including paired base number to total base number ratio, unpaired base number to total base number ratio, refolding free energy, number of stems of various sizes, number of loops of various sizes, total numbers of stems and loops, and sizes of all kinds of stems and loops.

High levels of genetic conservation in the miRNA binding sequence were revealed with several related studies. This result also fits the hypothesis that functional sequences show greater degrees of conservation than other regions. FastCompare (Elemento and Tavazoie, 2007) is a single and efficient algorithm for searching global conservative adjustment factors between two species. It aims to shorten the length of the base elements, which renders the use of multiple sequence alignment to evaluate sequence conservation unnecessary (Elemento and Tavazoie, 2005). Owing to its flexibility and practicability, the FastCompare algorithm was here adopted to calculate the conservative scores. We used it to calculate the binding site located at the 4mer factor scores. The sum of all of the 4mers' conservative scores was regarded as the total score of the whole sequence. Then, we obtained the conservative eigenvectors and returned them to the main program. The main modeling flow chart is shown in Figure 1.



**Figure 1.** Scheme of the SVM modeling and testing flow.



## Modeling and optimal parameters

An SVM model was trained with 2778 positive and 703 negative miRNA:gene pair eigenvector datasets. The LibSVM library was used to train and run the SVM. Five main types of commonly used SVM kernels exist: linear kernel, polynomial kernel, radial basis function (RBF) kernel, Gaussian RBF Kernel, and sigmoid kernel. Owing to its ability to solve linear inseparability problems, easy personal computing, and few-parameter characters (Amari and Wu, 1999), the RBF kernel has been widely applied in solving numerous bioinformatics classification problems. Thus, we chose the C-SVC and RBF kernels for our two-category classifier. To achieve the best SVM classifier results, the RBF kernel parameter  $\gamma$  and the slack penalty coefficient  $C$  need to be calibrated. We used grid-based search tools provided by the LibSVM library to select the optimal parameters  $C$  and  $\gamma$  from the training datasets. Finally, the optimal parameters  $C = 4$  and  $\gamma = 0.0156$  were obtained, and the search results are shown in Figure 2.

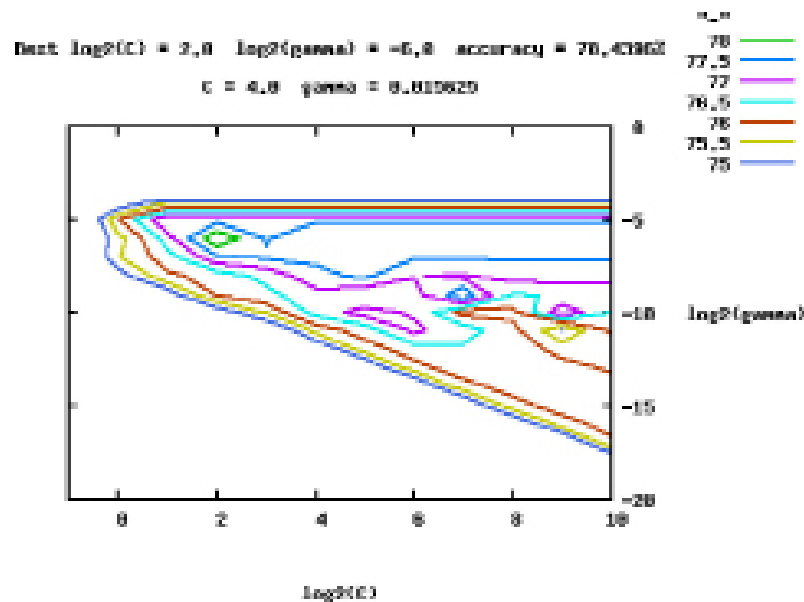


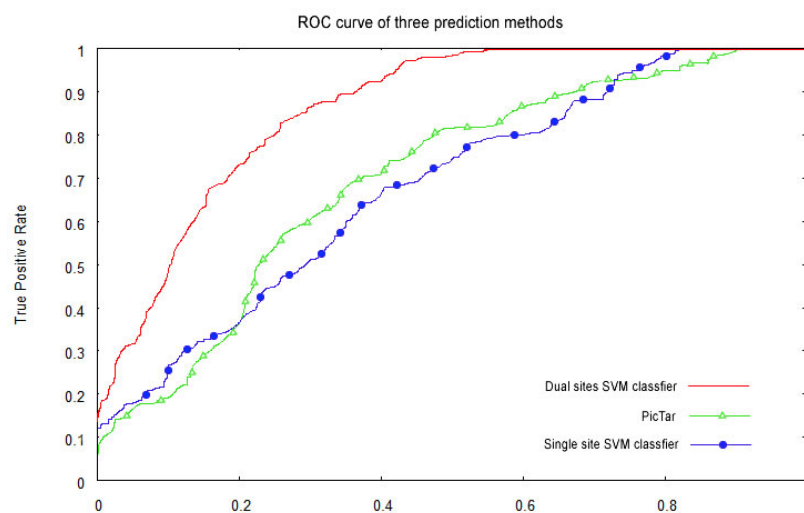
Figure 2. Optimal parameter-searching process generated by gnuplot tools.

## RESULTS

### Model evaluation by receiver operating characteristic (ROC) curve plotting

The ROC curve is recognized as one of the more reliable means for evaluating the executing efficiency of models (Bradley, 1997). To verify the feasibility of the dual interaction site design SVM model, and to compare the efficiency of the model with that of other machine-learning methods, we built an extra single-interaction site SVM model (only the primary interaction site feature was considered) and chose the commonly used PicTar machine-learning miRNA target scan algorithm (Gamazon et al., 2010). The gnuplot tools and the LibSVM library were employed to construct the ROC curves (Williams and Kelley, 1986;

Chang and Lin, 2011), and these curves will be automatically generated using commands, training datasets, and testing datasets. As shown in Figure 3, the dual site SVM model had a better true positive target recognition rate than the other two methods, and the area under the dual site SVM model curve was approximately 86.21%. This result indicates that the dual site SVM mode has improved target prediction and is a feasible and effective method.



**Figure 3.** ROC curves of three prediction methods were plotted with individual test sets on train sets.

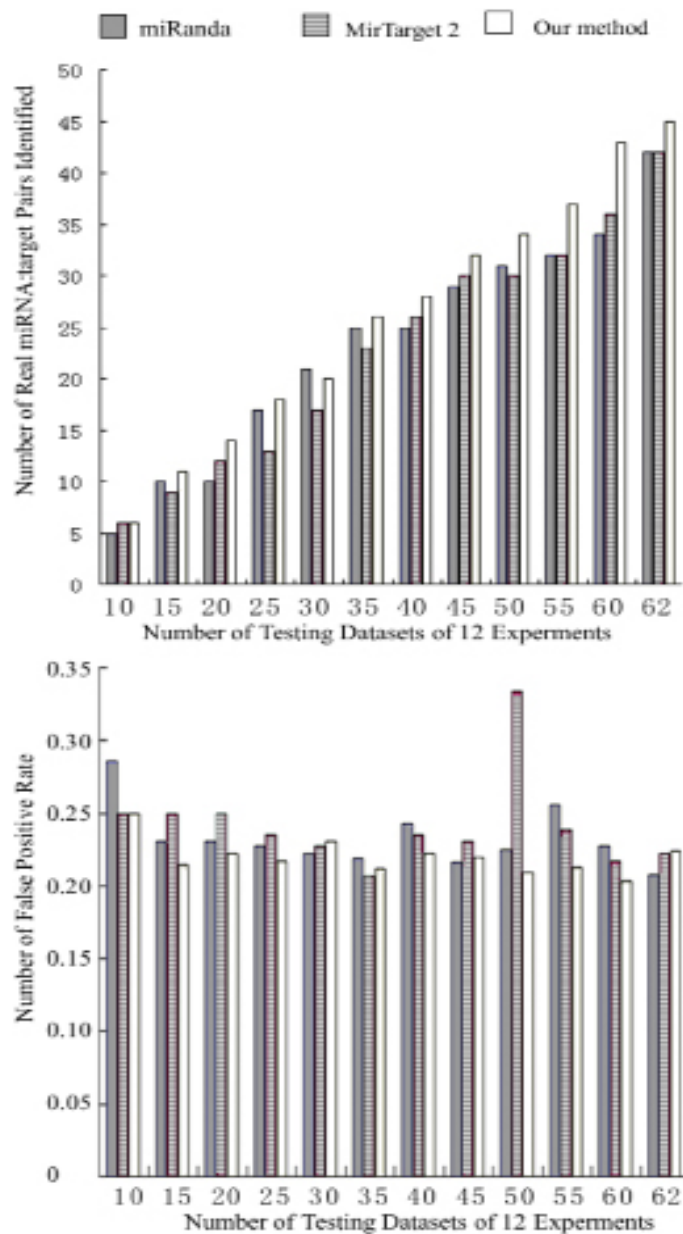
### Performance of the SVM classifier

Sensitivity is an important criterion for evaluating the quality of miRNA target-prediction methods. The most general method is calculating the identification rate of prediction methods to real miRNA:gene pairs (Sarker et al., 2011). We used 350 positive independent test samples selected from miRecords and 92 negative test samples selected from microarray data (see Material and Methods). They were randomized 100 times using a Fisher-Yates shuffle algorithm to ensure the random order of inputs. Then, the datasets were divided into 12 groups to test our method and two other previous methods, namely MirTarget2 and miRanda. Finally, the rate of the real target identification and the false positive rate of the three methods were compared. In terms of the average rate of real target identification, our method was 7.44% higher than that of miRanda and 8.75% higher than that of MirTarget2. The false positive rate was 1.26% lower than miRanda and 2.15% lower than MirTarget2. As shown in Figure 4, our method can effectively improve the identification rate based on non-reduction of classification accuracy.

Most of the training and testing data employed in this study originated from the miRecords, pSILAC, and TarBase V5.0c databases (Selbach et al., 2008; Papadopoulos et al., 2009; Xiao et al., 2009). However, an increasing number of new miRNA target genes and interaction sites have been identified with increasing numbers of miRNA studies published. Thus, we randomly selected 20 miRNA:target pairs from more recently published papers, and used the data to evaluate our method along with six other commonly employed miRNA target-prediction methods, including MirTarget2, miRanda, PITA, TargetSpy, TargetMiner, and TargetScan.



As shown in Table 2, PITA and TargetScan both identified 9 miRNA:target pairs among the samples and performed better than any of the other previous methods. Due to the dual site character, the SVM classifier we built identified 11 miRNA:target pairs, which comprised 55% of the total samples; this result was at least 10% higher than that of the other methods. Results show that the performance of our method is better than that of previous algorithms.



**Figure 4.** Comparison of the identification rate and false positive rate of three methods.

**Table 2.** Predicted results of seven methods by employing samples published recently.

miRNA	Target gene	Predicted results						
		Our method	MirTarget2	miRanda	PITA	TargetSpy	TargetMiner	TargetScan
hsa-miR-us4	ERAP1	N/A	N/A	N/A	N/A	Matched	N/A	N/A
hsa-miR-223	CMYC	Matched	N/A	N/A	N/A	N/A	N/A	N/A
hsa-miR-101	EZH2	N/A	Matched	N/A	Matched	N/A	N/A	Matched
hsa-miR-335	SOX4	Matched	N/A	N/A	N/A	N/A	Matched	N/A
hsa-miR-221/222	TRPS1	Matched	N/A	Matched	Matched	Matched	Matched	N/A
hsa-miR-33a/b	PIM1	Matched	Matched	N/A	Matched	Matched	N/A	Matched
hsa-miR-200bc	ETS1	Matched	Matched	Matched	N/A	N/A	N/A	Matched
hsa-miR-199a-3p	PAK4	N/A	Matched	N/A	N/A	N/A	Matched	N/A
hsa-miR-101	MYCN	Matched	Matched	N/A	Matched	Matched	Matched	Matched
hsa-miR-96	GPC3	N/A	Matched	N/A	Matched	Matched	N/A	Matched
hsa-miR-150	NOTCH3	N/A	N/A	Matched	N/A	N/A	N/A	Matched
hsa-miR-15a	CCNT2	N/A	Matched	Matched	Matched	Matched	N/A	N/A
hsa-miR-373	MEF2C	Matched	N/A	Matched	Matched	Matched	N/A	N/A
hsa-miR-214	PLXNB1	Matched	N/A	N/A	N/A	N/A	N/A	Matched
hsa-miR-107	AIP	N/A	N/A	N/A	N/A	N/A	N/A	N/A
hsa-miR-375	KLF5	N/A	N/A	Matched	N/A	N/A	N/A	N/A
hsa-miR-9	CAMTA1	N/A	N/A	Matched	N/A	N/A	N/A	N/A
hsa-miR-17	CAMTA1	Matched	N/A	Matched	Matched	Matched	N/A	Matched
hsa-miR-34a	CD44	Matched	N/A	N/A	N/A	N/A	N/A	N/A
hsa-miR-145	FSCN1	Matched	Matched	N/A	Matched	N/A	N/A	Matched

## DISCUSSION

In this study, a novel miRNA target-prediction approach is proposed based on SVM by considering the overall regulation relationships between miRNA and the target gene. To improve the sensitivity of traditional prediction methods, we designed two potential binding site-searching processes and abstracted all of the eigenvectors of these two sites. To obtain the actual dimensional structures without increasing computing complexity, we extended the sequence length of the sites located at the original gene sequence. Meanwhile, the use of high-quality training datasets that are experimentally verified or supported by microarray data also ensures the executing efficiency of this model. Finally, we used independent testing datasets to evaluate our method and to compare its performance with other commonly used methods, and plotted ROC curves to verify the accuracy of our model. The evaluation results showed that our method could effectively improve identification rate on the basis of non-reduction of classification accuracy. In conclusion, compared with traditional methods, the design of the double potential binding site can identify more potential miRNA target genes. This model can therefore help to provide new evidence for the theory that many genes have several target sites for either one miRNA or a few different miRNAs. However, the dual sites design also results in massive computing complexity. To overcome this disadvantage, we will continue to collect published miRNA:target pairs with experimental verifications in order to extend and refine our training database.

For biology researchers' convenience, we made our method an online tool. Users may now visit the website via "<http://hpabws.s87.cnaaa7.com/>" to use our prediction application.

## ACKNOWLEDGMENTS

Research supported by the Sciences and Technologies Projects of the Science and

Information Technology Bureau of GuangZhou, China (Grant #2005Z12E4023).

## REFERENCES

- Amari S and Wu S (1999). Improving support vector machine classifiers by modifying kernel functions. *Neural Networks* 12: 783-789.
- Bandyopadhyay S and Mitra R (2009). TargetMiner: microRNA target prediction with systematic identification of tissue-specific negative examples. *Bioinformatics* 25: 2625-2631.
- Barrett T, Suzek TO, Troup DB, Wilhite SE, et al. (2005). NCBI GEO: mining millions of expression profiles - database and tools. *Nucleic Acids Res.* 33: D562-D566.
- Bartel DP (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 116: 281-297.
- Bradley AP (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recogn.* 30: 1145-1159.
- Carrillo H and Lipman D (1988). The multiple sequence alignment problem in biology. *SIAM J. Appl. Math.* 48: 1073-1082.
- Chang C-C and Lin C-J (2011). LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* 2: 1-26.
- Cortes C and Vapnik V (1995). Support-vector networks. *Mach. Learn.* 20: 273-297.
- Elemento O and Tavazoie S (2005). Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach. *Genome Biol.* 6: R18.
- Elemento O and Tavazoie S (2007). Fastcompare: a nonalignment approach for genome-scale discovery of DNA and mRNA regulatory elements using network-level conservation. *Methods Mol. Biol.* 395: 349-366.
- Fung G and Mangasarian OL (2004). A feature selection newton method for support vector machine classification. *Comp. Optim. Appl.* 28: 185-202.
- Furey TS, Cristianini N, Duffy N, Bednarski DW, et al. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* 16: 906-914.
- Gamazon ER, Im HK, Duan S, Lussier YA, et al. (2010). Exprtarg: an integrative approach to predicting human microRNA targets. *PLoS One* 5: e13534.
- Gregory RI (2006). *MicroRNA Protocols*, Humana Press.
- Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, et al. (2006). miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.* 34: D140-D144.
- Grün D, Wang YL, Langenberger D, Gunsalus KC, et al. (2005). microRNA target predictions across seven *Drosophila* species and comparison to mammalian targets. *PLoS Comput. Biol.* 1: e13.
- Hammell M (2010). Computational methods to identify miRNA targets. *Semin. Cell Dev. Biol.* 21: 738-744.
- Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, et al. (1994). Fast folding and comparison of RNA secondary structures. *Monatshfte für Chemie/Chemical Monthly* 125: 167-188.
- John B, Enright AJ, Aravin A, Tuschl T, et al. (2004). Human MicroRNA targets. *PLoS Biol.* 2: e363.
- Kertesz M, Iovino N, Unnerstall U, Gaul U, et al. (2007). The role of site accessibility in microRNA target recognition. *Nat. Genet.* 39: 1278-1284.
- Larry Wall TC and Orwant J (2000). Programming Perl, O'Reilly Media, Sebastopol.
- Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP, et al. (2003). Prediction of mammalian microRNA targets. *Cell* 115: 787-798.
- Lewis BP, Burge CB and Bartel DP (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* 120: 15-20.
- Lim LP, Lau NC, Garrett-Engele P, Grimson A, et al. (2005). Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* 433: 769-773.
- Lytle JR, Yario TA and Steitz JA (2007). 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.
- Nguyen GH, Phung SL and Bouzerdoum A (2010). Efficient SVM Training with Reduced Weighted Samples. IJCNN. IEEE, Barcelona.
- Papadopoulos GL, Reczko M, Simossis VA, Sethupathy P, et al. (2009). The database of experimentally supported targets: a functional update of TarBase. *Nucleic Acids Res.* 37: D155-D158.
- Rehmsmeier M, Steffen P, Hochsmann M and Giegerich R (2004). Fast and effective prediction of microRNA/target duplexes. *RNA* 10: 1507-1517.
- Rhoades MW, Reinhart BJ, Lim LP, Burge CB, et al. (2002). Prediction of plant microRNA targets. *Cell* 110: 513-520.
- Sarker R, Bandyopadhyay S and Maulik U (2011). An overview of computational approaches for prediction of miRNA genes and their targets. *Curr. Bioinformatics* 6: 129-143.

- Selbach M, Schwanhaussner B, Thierfelder N, Fang Z, et al. (2008). Widespread changes in protein synthesis induced by microRNAs. *Nature* 455: 58-63.
- Sturm M, Hackenberg M, Langenberger D and Frishman D (2010). TargetSpy: a supervised machine learning approach for microRNA target prediction. *BMC Bioinformatics* 11: 292.
- Wang X and El Naqa IM (2008). Prediction of both conserved and nonconserved microRNA targets in animals. *Bioinformatics* 24: 325-332.
- Williams T and Kelley C (1986). GNUPLOT: An Interactive Plotting Program. Official GnuPlot Documentation. Available at [<http://sourceforge.net/projects/gnuplot>]. Accessed March 11, 2012.
- Xiao F, Zuo Z, Cai G, Kang S, et al. (2009). miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Res.* 37: D105-D110.
- Zhang B, Pan X, Wang Q, Cobb GP, et al. (2006). Computational identification of microRNAs and their targets. *Comput. Biol. Chem.* 30: 395-407.