BIOINFORMATICS ORIGINAL PAPER Vol. 24 no. 3 2008, pages 325–332
doi:10.1093/bioinformatics/btm595

Sequence analysis

Prediction of both conserved and nonconserved microRNA targets in animals

Xiaowei Wang* and Issam M. El Naqa

Department of Radiation Oncology, Washington University School of Medicine, St. Louis, MO 63110, USA

Received on September 25, 2007; revised on November 3, 2007; accepted on November 27, 2007

Advance Access publication November 29, 2007

Associate Editor: Alex Bateman

ABSTRACT

Motivation: MicroRNAs (miRNAs) are involved in many diverse biological processes and they may potentially regulate the functions of thousands of genes. However, one major issue in miRNA studies is the lack of bioinformatics programs to accurately predict miRNA targets. Animal miRNAs have limited sequence complementarity to their gene targets, which makes it challenging to build target prediction models with high specificity.

Results: Here we present a new miRNA target prediction program based on support vector machines (SVMs) and a large microarray training dataset. By systematically analyzing public microarray data, we have identified statistically significant features that are important to target downregulation. Heterogeneous prediction features have been non-linearly integrated in an SVM machine learning framework for the training of our target prediction model, MirTarget2. About half of the predicted miRNA target sites in human are not conserved in other organisms. Our prediction algorithm has been validated with independent experimental data for its improved performance on predicting a large number of miRNA down-regulated gene targets.

Availability: All the predicted targets were imported into an online database miRDB, which is freely accessible at http://mirdb.org.

Contact: xwang@radonc.wustl.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Thousands of microRNAs (miRNAs) have been identified in recent years. These miRNAs are involved in many diverse biological processes, such as development, differentiation, apoptosis and viral infection (Ambros, 2004; Miska, 2005). miRNAs function primarily through negative regulation of the expression level of their gene targets. Both computational and experimental studies have suggested that thousands of human genes are likely to be regulated by miRNAs (Lewis *et al.*, 2005; Lim *et al.*, 2005; Miranda *et al.*, 2006). Because of their critical roles in gene expression regulation, the functional characterization of miRNAs has become one of the most active research fields in biology in recent years. However, one major issue facing miRNA research is the lack of computational tools for accurate target prediction. Although multiple computational approaches have been proposed recently (Brennecke *et al.*, 2005; Enright *et al.*, 2003; Kim *et al.*, 2006; Kiriakidou *et al.*, 2004; Krek *et al.*, 2005; Lewis *et al.*, 2005; Miranda *et al.*, 2006; Rehmsmeier *et al.*, 2004; Stark *et al.*, 2003; Wang and Wang, 2006), this remains a major challenge for bioinformatists because of very limited sequence complementarity between miRNAs and their targets, as well as the scarcity of experimentally validated gene targets to guide bioinformatics design (Rajewsky, 2006).

One strategy for target prediction is to use machine learning approaches. Machine learning methods, such as support vector machines (SVMs), attempt to extract relevant information from data automatically using computational and statistical methods. Machine learning has been applied to many diverse fields including biological research, but has not been applied to miRNA target prediction with great success to date. One major obstacle is the lack of high quality training data for building robust prediction models. There are only a limited number of validated miRNA targets from literature (Sethupathy *et al.*, 2006); in addition, most of these targets were validated because they were predicted miRNA targets by existing programs. As a result, the validation data are biased toward these algorithms and are less useful for developing new target prediction algorithms. One SVM approach has been proposed recently for miRNA target prediction based on validated miRNA targets from literature survey (Kim *et al.*, 2006); however, the algorithm comparison analysis indicates that it underperforms another non-machine learning based algorithm, TargetScan for prediction specificity (Kim *et al.*, 2006).

miRNA can regulate gene expression at both the mRNA and the protein levels. A general model has been that miRNA function primarily through suppressing the protein expression of their targets (He and Hannon, 2004). However, many recent studies have convincingly demonstrated that miRNAs can also commonly decrease the target transcript expression level (Bagga *et al.*, 2005; Jing *et al.*, 2005; Lai, 2002; Lai *et al.*, 2005; Lee and Dutta, 2007). Consistent with these observations, miRNAs and their predicted target transcripts are found to have mutually exclusive expression profiles in different tissues (Farh *et al.*, 2005; Stark *et al.*, 2005); and overexpression of miRNA leads to the downregulation of hundreds of mRNA transcripts in cell lines (Lim *et al.*, 2005). The majority of these transcripts have miRNA seed pairing sites, implying direct regulation by miRNAs. All these studies indicate that

*To whom correspondence should be addressed.

regulation of mRNA expression level is a common mechanism for miRNA function (Sontheimer and Carthew, 2005). This opens a new door to high-throughput target validation by transcriptional profiling, such as microarrays. We have previously demonstrated that microarrays are a reliable approach to quickly identify a large number of miRNA downregulated genes, and many of these downregulated genes are also miRNA targets as predicted by bioinformatics (Wang and Wang, 2006). A more recent study by Linsley *et al*. (2007) has studied the functions of 24 miRNAs by systematically analyzing their downregulated genes with microarrays. This study has characterized the specific miRNAs of interest by miRNA overexpression and at the same time has made available a large experimental dataset to guide bioinformatics target prediction in general.

While this manuscript was in preparation, a new version of TargetScan (release 4.0) was published (Grimson *et al*., 2007). By performing microarray experiments with miRNA overexpression, the authors have identified five general features that boost target site efficacy and a linear regression model has been developed to combine these features for miRNA target prediction. Another study was also reported recently for identifying several miRNA targeting determinants from microarray data analysis (Nielsen *et al*., 2007).

Here we present a new strategy to predict miRNA targets with machine learning. The Linsley microarray transcriptional profiling dataset (Linsley *et al*., 2007) was used for algorithm training. Statistically significant features were identified by comparing genes downregulated or unaffected by miRNA overexpression. We confirmed that several widely used prediction features, such as perfect seed pairing and target site conservation, were very significantly associated with genes downregulated by miRNA. In addition, new significant sequence features have been identified for their preferential association with miRNA downregulated genes. An SVM based machine learning algorithm has been developed by combining heterogeneous prediction features in the algorithm training process and the prediction model has also been validated with independent experimental data.

2 METHODS

2.1 Data retrieval

All mRNA sequences and gene mapping index files were downloaded from the NCBI ftp site (Benson, *et al*., 2007). The 3′-untranslated region (UTR) sequences were then parsed with BioPerl (http://www.bioperl.org) using the GenBank annotations. Orthologous gene relationships were predicted based on NCBI HomoloGene. Transcript 3′ UTR sequences from human orthologs in mouse, rat, dog and chicken were also prepared and included in our computational analysis.

Microarray data were downloaded from the NCBI GEO database (Barrett, *et al*., 2007) for two published miRNA studies (accessions GSE6838 and GSE6207). In the Linsley study, multiple miRNAs were transfected in two cell lines and the global effect of miRNA overexpression was examined by microarrays (Linsley *et al*., 2007). In the Wang study, miR-124a was transfected in HepG2 cells and changes in global gene expression profiles were evaluated by microarrays at different time points (Wang and Wang, 2006).

Predicted miRNA targets by several published algorithms were retrieved from the public websites (TargetScan, http://targetscan.org; PicTar, http://pictar.bio.nyu.edu; miRanda, http://microrna.sanger.ac.uk; MirTarget version 1, http://nar.oxfordjournals.org/cgi/content/abstract/34/5/1646). The target IDs from all these algorithms were mapped to NCBI Gene IDs. For the analysis using the Linsley dataset, the training sequences were pre-selected for their seed matching sites. Thus, predicted targets without seed matching sites or without representing array probes were not included in our analysis. For the algorithm comparison using the Wang dataset, all predicted targets with representing array probes were included in the analysis.

2.2 Identification of downregulated genes by miRNA overexpression

Two cell lines (HCT116 Dicer^{ex5} and DLD-1 Dicer^{ex5}) were included in the Linsley study. Most of the miRNAs were transfected in both cell lines to evaluate changes in global gene expression profiles [for more experimental details, please refer to (Linsley *et al*., 2007)]. Probe IDs were mapped to NCBI gene IDs with NCBI gene index files and multiple probe signals for the same gene were averaged to represent the expression level of the gene. A gene was defined as downregulated if compared to mock transfection, its expression level was reduced by at least 40% with P-value < 0.001 in either cell line; a gene was defined as unaffected (normal) if its gene expression level was at least 95%, but no more than 120% with P-value > 0.3 in both cell lines.

Six miRNAs from the Linsley dataset, let-7c, miR-103, miR-106b, miR-141, miR-15a and miR-215 were selected for model training and testing. These miRNAs were selected for their non-redundant seed (positions 2–8) sequences. In this way, the potential bias associated with any specific miRNA sequence was expected to be reduced. All except one miRNA were transfected in both HCT116 Dicer^{ex5} and DLD-1 Dicer^{ex5} cells, and candidate transcripts were selected by analyzing microarray data from both cell lines. There were 1401 downregulated genes and 16 761 normal (unaffected) genes identified in this way.

We also analyzed a dataset that we published previously (Wang and Wang, 2006). This was a temporal study evaluating the effect of miR-124a overexpression on global transcriptional profiles. The downregulated genes at each time point were identified with the same criteria as described previously (Wang and Wang, 2006).

2.3 Computational analysis

The SVM package LIBSVM was used to construct miRNA target prediction models (http://www.csie.ntu.edu.tw/∼cjlin/libsvm/). Training optimization was performed according to the recommended protocol by the program. SVM is a universal constructive learning procedure based on statistical learning theory. SVM has been applied in many diverse applications such as pattern recognition, computational biology and in our previous work on medical image analysis (El-Naqa *et al*., 2002), in which its superior performance to competing methods was demonstrated. The basic concept is to maximize the separation between two data groups in a non-linear feature space.

RNA secondary structure stability, represented by the ΔG value, was calculated with RNAfold (Hofacker, 2003). Secondary structures predicted by RNAfold were analyzed to identify nucleotides that were base-paired in these structures. Statistical computing was performed with MATLAB (Mathworks, Inc., Natick, MA) and the R package (http://www.r-project.org/). Statistical significance (P-value) for the training features was calculated with independent t test or χ^2 test. The seed enrichment analysis was performed with the hypergeometric test using all the genes represented on microarrays as the background. We used the DREES package (http://radium.wustl.edu/drees), which is built with MATLAB, for extracting the top relevant training features. The software applies step-wise logistic regression analysis in conjunction with resampling methods to identify the most predictive features.

Table 1. 3′ UTRs with seed pairing sites from different gene groups

Seed type	Downregulated genes (%)	Normal genes (%)	Enrichment ratio
seed 8	16.5	1.4	11.9
seed 6	67.4	22.9	2.9
seed 7a; no seed 7b	9.3	4.3	2.2
seed 7b; no seed 7a	22.8	5.2	4.4
seed 7a or 7b	52.8	11.4	4.6
seed 7b	**43.5**	**7.1**	**6.1**

3 RESULTS

3.1 Selection of training sequences

The Linsley dataset was downloaded and used as the training data in our study (Linsley *et al*., 2007). In that study, multiple miRNAs were transfected in two cell lines and the global effect of miRNA overexpression was examined by microarrays. Downregulated and normal (unaffected) genes due to miRNA overexpression were identified and their 3′-untranslated region (UTR) sequences were retrieved from the NCBI databases.

A typical miRNA target site has perfect complementarity to the miRNA seed sequence. There are four major types of seed sequence: positions 1–8 (defined as seed 8 in our study), positions 1–7 (as seed 7a), positions 2–8 (as seed 7b) and positions 2–7 (as seed 6). The training data were analyzed to determine which seed type was most relevant to the identification of downregulated genes. The result is summarized in Table 1. The seed match enrichment ratio was defined as the fraction of UTRs with seed match in the downregulated genes/that in the normal genes. Seed 8 match was enriched 11.9-fold in the downregulated genes as compared to the normal genes, and this was statistically very significant ($P = 9.4E-135$). However, only a small percentage (16.5%) of the downregulated genes have seed 8 pairing sites in their 3′ UTRs. On the other hand, seed 6 matching sites were present in 67.4% of the downregulated genes. However, this seed type was not considered further because a large number of normal genes also had seed 6 matching sequences and thus the enrichment ratio was low.

Seeds 7a and 7b were examined in a similar way. Seed 7a pairing sites were enriched 2.2-fold in 3′ UTRs without any seed 7b match. In comparison, seed 7b pairing sites were enriched at a much higher level (4.4-fold) in 3′ UTRs without any seed 7a match. If seed 7b was considered alone, 43.5% of the downregulated genes were identified, and the seed enrichment ratio was higher than that from using both seeds 7a and 7b, implying better signal-to-noise ratio. Thus, we decided to consider primarily seed 7b in our target prediction. Since a seed 8 pairing site also matches to seed 7b (seed 7b plus a terminal base match is equivalent to seed 8), all seed 8 matches were evaluated in the context of seed 7b. Most seed 7a pairing sites (82.4%) were also considered in our analysis because only a small percent of type 7a seeds matched to 3′ UTRs without any seed 7b site.

A 3′ UTR sequence may contain multiple seed pairing sites, and it is challenging to determine the contribution of each binding site. In our study, 3′ UTR sequences with only single seed pairing sites were selected. In this way, 454 downregulated genes and 1017 normal genes were included in the training dataset for their single seed 7b matching sites.

Table 2. Summary of the training features for target prediction

Feature name	Fold change	*P*-value
Seed match conservation	2.47	$7.4E-29$
Terminal base match	1.98	$3.2E-13$
Seed 7a matching site	1.65	$2.7E-03$
Location >600 bases	0.52	$3.1E-10$
Location >900 bases	0.43	$1.6E-07$
Duplex hybridization ΔG	0.95	$2.0E-05$
Duplex hybridization index	1.15	$6.7E-05$
Features related to the target site		
ΔG	0.70	$1.9E-31$
GC content	0.80	$3.7E-46$
Accessibility index	0.82	$2.4E-09$
A count	1.22	$5.5E-14$
U count	1.19	$4.0E-13$
C count	0.75	$2.4E-18$
G count	0.82	$3.3E-10$
Dinucleotide counts		See Table 3
Position-specific base counts		See Table S1

Fold change is defined as the fraction of candidate sites with the selected features or the average feature values in the downregulated genes/that in the normal genes. The *P*-values were calculated with independent *t*-test or χ^2 test by comparing the feature values from the downregulated genes and normal genes.

3.2 Identification of training features

Support vector machines (SVMs) were applied to the training data to construct classifiers for miRNA target prediction. A critical step in SVM is to collect relevant training features to separate positive samples from the negative ones. The training features used in our study are summarized in Table 2.

3.2.1 Seed conservation The seed pairing site in a target 3′ UTR is often conserved across multiple species. This is a primary selection filter in most existing target prediction algorithms. A candidate site is usually rejected if it does not meet a threshold level of evolutionary conservation. In our training process, seed conservation was also considered, although not as a requirement. Transcript 3′ UTR sequences from human gene orthologs in mouse, rat, dog and chicken were analyzed to identify miRNA seed matches, and the level of seed conservation was recorded. The candidate sites from the downregulated genes were significantly more conserved than those from the normal genes ($P = 7.4E-29$). This result is in agreement with many previous studies indicating the importance of seed conservation in target prediction.

3.2.2 Other seed types A seed 7b site is also a site for seed 8 if the terminal base is a match. As described earlier, 3′ UTRs in the downregulated genes were highly enriched in seed 8 binding sites. Therefore terminal base match was recorded as a training feature in our analysis. The presence of seed 7a may also be important to miRNA target identification, and thus the total

Table 3. Statistical significance of dinucleotide composition

Dinucleotide	Fold change	*P*-value
UU	1.55	1.8*E*−14
GC	0.63	4.1*E*−14
AA	1.53	7.4*E*−13
CC	0.63	3.8*E*−11
UA	1.42	1.2*E*−10
AU	1.34	1.3*E*−09
GG	0.64	8.5*E*−09
CA	0.82	2.5*E*−05
UC	0.81	2.9*E*−05
CG	0.59	5.0*E*−05
GA	1.20	1.1*E*−03
CU	0.88	2.4*E*−03

Fold change is defined as the average dinucleotide count in the downregulated genes/that in the normal genes. The *P*-values were calculated with χ^2 test by comparing the feature values from the downregulated genes and normal genes.

count of seed 7a sites in a 3′ UTR (except at the seed 7b site) was recorded as another training feature.

3.2.3 Base composition Local 3′ UTR regions next to the seed pairing sites were analyzed for their base composition. Compared to candidate sites in the normal genes, the sites in the downregulated genes had a significantly lower GC content in general ($P = 3.7E-46$) and this was in agreement with previous studies (Grimson *et al*., 2007; Lewis *et al*., 2005; Nielsen *et al*., 2007). All four base counts were significantly different between candidate sites in the downregulated and normal genes, with C as the most under-represented base in the downregulated gene sites ($P = 2.4E-18$, Fig. S1a). In addition to mononucleotide counts, the frequencies of all 16 dinucleotides were determined. Many of the dinucleotide counts were statistically significant, with UU, GC, AA and CC ranked at the top of the list (Table 3).

Base composition at each individual position in the local target site was also analyzed (see Table S1 with the relative seed positions defined as 0–6). Many of these position-specific base counts were statistically significant, reflecting the overall site requirement for low GC content. However, there were also a few interesting observations that may be related to specific functional requirements. For example, C was significantly absent at positions −4 to −1 (immediately upstream of the seed-binding site) in the downregulated gene sites. This cannot be explained by the low GC content requirement since G count difference was much smaller at the same positions. A typical miRNA/target-binding duplex has a loop or bulge structure in this region, and the absence of Cs might be related to this requirement. Another example is the A counts at positions −5 and 7 (immediately downstream of the seed binding site). Although there was no difference in T counts at these positions, A counts were significantly higher in candidate sites from the downregulated genes. These position-specific base counts (20 bases surrounding the seed binding site) were recorded as SVM training features.

3.2.4 Secondary structure A target site is not likely to be functional if it is inaccessible to miRNA-binding (Kertesz *et al*., 2007; Long *et al*., 2007; Robins *et al*., 2005; Zhao *et al*., 2005). RNA secondary structure prediction is still a challenging task to date and in general the prediction accuracy decreases dramatically as the sequence length increases. In our study, a short length of 25 nts on each side of the seed matching site was included in the structural calculation. Other various short sequence lengths were also tested and the results were similar to the one presented here. Local secondary structure of a candidate site was calculated with RNAfold (Hofacker, 2003). The propensity for secondary structure formation (measured by ΔG) was significantly higher in the candidate sites from the normal genes than those from the downregulated genes (Fig. S1b, P-value $= 1.9E-31$). Besides the overall accessibility of a candidate site, the base-pairing of individual nucleotides was evaluated and significant positions were combined (the accessibility index) to capture potential position specific effect.

Potential miRNA/target hybridization structures were calculated to determine the level of nucleotide base-pairing at each position (Fig. S2). Nucleotides 13–17 in a miRNA were more likely to be base-paired to candidate sites in the downregulated genes than to those in the normal genes. This observation was in agreement with previous studies showing that limited base-pairing in the 3′ region of a target site contributes to target recognition by miRNA (Grimson *et al*., 2007). On the other hand, position 10 was more unstructured in the downregulated genes. A typical miRNA-binding duplex has a bulge/loop following the seed-binding region, which likely results in an exposed base at position 10. These significant base positions were combined (the hybridization index) to evaluate the overall position-specific base-pairing of the binding duplex. The overall sequence alignments between miRNAs and their potential target sites were also determined using the Smith-Waterman algorithm, but no significant enrichment in alignment was associated with the downregulated genes.

3.2.5 Location in 3′ UTR Previous studies have suggested that the function of a target-binding site is related to its location in 3′ UTR, and a site in the middle of a long UTR is less likely to be functional (Gaidatzis *et al*., 2007; Grimson *et al*., 2007). In our analysis, candidate sites from the downregulated genes and normal genes were examined to determine whether they were at least 600 or 900 nucleotides away from both ends of the UTRs. Significant difference was observed between sites in the downregulated genes and normal genes (Table 2).

3.3 Target prediction model

An SVM-based miRNA target prediction algorithm was developed by integrating 131 heterogeneous prediction features described above. Overall, 454 downregulated genes (positive samples) and 1017 normal genes (negative samples) were analyzed. The LIBSVM package was used to build miRNA target classifiers. The training model parameters were optimized by multiple rounds of cross-validation to minimize the overtraining risk.

The DREES package, which utilizes multivariate logistic regression analysis, was used to evaluate prediction models with

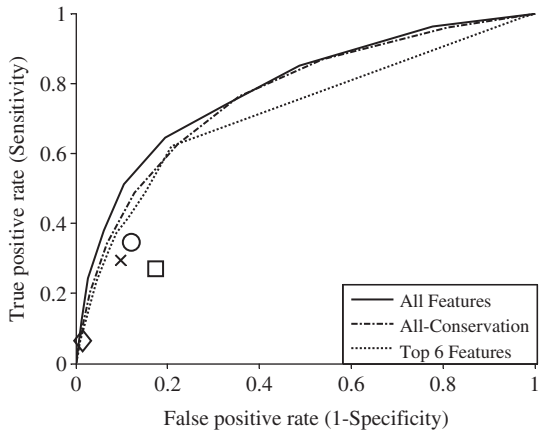

Fig. 1. Receiver operating characteristic (ROC) curves to evaluate target prediction models. Three models were built with: (1) all the features; (2) all except the cross-species seed match conservation feature; (3) the top six features. Four published programs were represented by data points: open diamond for MirTarget v1, open circle for TargetScan, cross mark for PicTar and open square for miRanda.

different feature sets. The model building process is composed of two steps: (1) selection of the model order; (2) estimation of the model parameters. In our case, 10-fold cross-validation was chosen for the model order selection in stepwise regression analysis. The correlation with true class labels was evaluated using Spearman's rank correlation (Rs).

Receiver operating characteristic (ROC) curves were used to evaluate prediction sensitivity and specificity. Figure 1 shows a 10-fold cross-validation prediction result, with an ROC area of 0.79. Stepwise regression was performed to identify features that contributed most to the prediction model (Table S2). The *P*-values were estimated with Wald statistics. The cross-species seed match conservation feature was ranked at the top of the list with the largest correlation coefficient ($Rs = 0.28$). The contribution of this conservation feature was further evaluated by removing it from the feature set. As a result, the ROC area decreased to 0.77. The algorithmic improvement from the conservation feature was more evident when the false positive rate was relatively low. An alternative SVM model was also built based on the top six features, leading to a reduced ROC area of 0.72 (Fig. 1).

We also compared the performance of our prediction model to a few existing prediction algorithms, TargetScan (Grimson *et al.*, 2007), PicTar (Krek *et al.*, 2005) and miRanda (Griffiths-Jones *et al.*, 2006). These algorithms were chosen because they are widely used by miRNA researchers and all the predicted targets are publicly available. In addition, we also included an algorithm we developed previously, MirTarget version 1 in the analysis (Wang and Wang, 2006). All public algorithms included in our analysis have cutoff values for prediction score assignment, and over 90% of all human genes were not assigned any score. As a result, ROC curves cannot be completely constructed for these algorithms. Instead, the overall prediction performance of these algorithms could only be represented by selected cutoff points on the ROC plot.

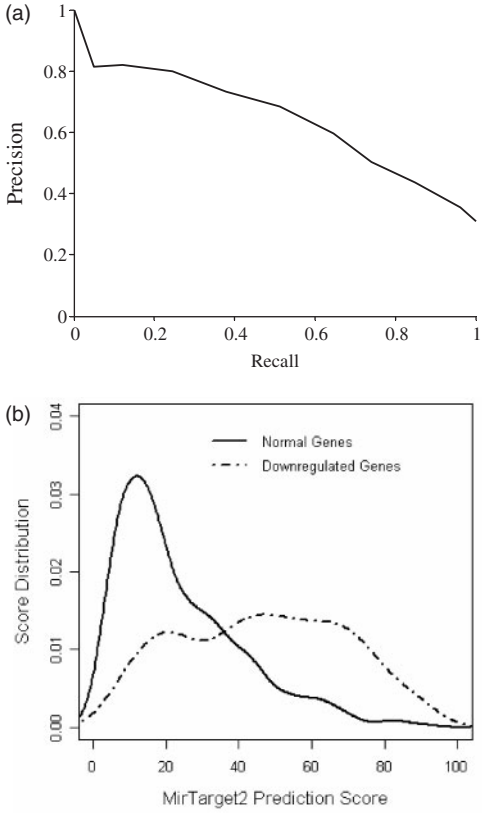

Fig. 2. Evaluation of MirTarget2 prediction algorithm. (**a**) Precision-recall curve to evaluate prediction sensitivity and specificity. (**b**) Target prediction score distributions for the candidate sites in downregulated and normal genes.

As shown in Figure 1, our SVM models had more robust performance than the published algorithms at the selected cutoff points. MirTarget v1 had a very close false positive rate (0.01) at a sensitivity of 0.07 when compared to the SVM models, partly because of its very conservative strategy of predicting only the most confident candidate sites (296 predicted targets versus over 1500 by other published programs).

Precision-recall curves were constructed to evaluate the prediction accuracy. Precision-recall curves are commonly used in machine learning to evaluate prediction precision (the fraction of true positives among all predicted positives) in relation to the recall rate (the fraction of true positives among all positive samples). As shown in Figure 2a, the precision rate was ~80% when the recall rate was below 20%.

A scoring system was developed to assign scores to all 3′ UTRs with seed matching sites. A small fraction of 3′ UTRs had multiple candidate sites, and all sites in one UTR were combined to compute the UTR score as the following:

$$S = 100 \times (1 - \prod_{i=1}^{n} P_i)$$

where n represents the number of candidate target sites in one UTR and P_i represents the statistical significance *P*-value for

Table 4. Seed match conservation of predicted human target sites

Only in human	In 2 orthologs	In 3 orthologs	In 4 orthologs	In 5 orthologs
78 027	29 383	24 495	14 010	5172

There were 151 087 predicted human miRNA targets. The numbers in the table represent the number of targets with different levels of seed match conservation across five organisms.

each of these candidate sites as estimated by SVM. The scores for single-site UTRs were calculated using the same equation with $n = 1$. These scores were used to assign ranks to evaluate the relative significance of the predicted targets.

The score distributions for the downregulated and normal genes were compared to evaluate the predictive power of the scoring system (Fig. 2b). The score distribution for the normal genes was significantly separated from that for the downregulated genes. On average, the scores for candidate sites from the normal genes were much lower, with a major peak below 20. The downregulated gene sites had a more spread-out score distribution, reflecting the fact that ~50% of the sites had scores higher than 50.

The prediction algorithm was implemented as a Perl program, which we named MirTarget2, for genome-wide miRNA target prediction. There were 151 087 predicted miRNA targets from 15 059 unique genes in human. Among these predicted human target sites, 52% were not conserved in seed pairing in orthologous 3′ UTRs from other organisms (Table 4).

3.4 Algorithm evaluation with independent dataset

The performance of MirTarget2 was evaluated with an independent dataset that we published previously (Wang and Wang, 2006). In that experiment, miR-124a was overexpressed in HepG2 cells, and changes in global expression profiles were evaluated by microarrays at multiple time points. miR-124a had not been included in the SVM training process. In addition, its seed sequence is distinct from those in the training miRNAs. Downregulated genes were identified at 8, 16, 24 and 32 h after miR-124a overexpression.

MirTarget2 was compared to four other algorithms, TargetScan, PicTar, MirTarget version 1 and miRanda for miR-124a target prediction. For all the genes represented on the microarrays with detectable expression levels (Present calls), 283, 482, 372, 54 and 408 genes were predicted miR-124a targets by MirTarget2, TargetScan, PicTar, MirTarget v1 and miRanda, respectively. No complete ROC curves could be drawn for any of these algorithms (including MirTarget2) because they assign prediction scores to ≤10% of all human genes. In our analysis, the algorithm performance was compared by evaluating the overall prediction specificity and sensitivity for miRNA downregulated genes at multiple time points. First we counted the number of downregulated genes that were also predicted targets for miR-124a. Overall, MirTarget2 and TargetScan predicted similarly higher

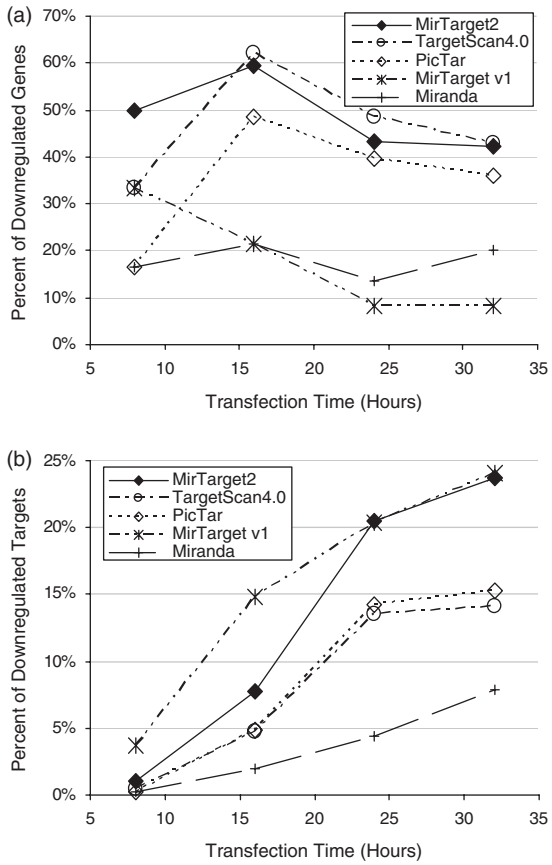

Fig. 3. Comparison of target prediction algorithms. Five algorithms were compared, MirTarget2, TargetScan, PicTar, miRanda and MirTarget v1. These algorithms were evaluated with the miR-124a microarray data (Wang and Wang, 2006) for predicting downregulated genes. Gene downregulation was analyzed at 8, 16, 24 and 32 h after miR-124a overexpression. (**a**) Percentage of downregulated genes that were also predicted miR-124a targets. (**b**) Percentage of predicted targets that were downregulated by miR-124a.

percentages of downregulated genes than PicTar, MirTarget v1 or miRanda (Fig. 3a).

The percentages of predicted targets that were downregulated by miR-124a were also determined. As shown in Figure 3b, the percentages for all the algorithms were relatively low at early time points. This suggests that target downregulation by miRNA in general may not be a rapid process. Among these algorithms, MirTarget v1 was most selective at predicting downregulated gene targets. MirTarget2 performed similarly to MirTarget v1, especially at 24 and 32 h when large numbers of genes were downregulated. Both algorithms were more selective at identifying downregulated genes than other algorithms included in the analysis.

4 DISCUSSION

4.1 Training data selection

At present, target prediction is limited by the scarcity of experimentally validated targets to guide bioinformatics design.

In our study, a large microarray dataset was used to guide the selection of target prediction features. miRNA downregulated genes were identified from high-throughput microarray experiments, and thousands of transcript sequences were analyzed statistically to identify relevant prediction features. A similar feature selection strategy has also been used and proved to be highly effective in the latest version of TargetScan (Grimson *et al*., 2007). Our analysis has demonstrated that many widely used prediction features, such as perfect seed pairing, target site conservation and structural accessibility were highly over-represented in the sequences of miRNA downregulated genes. This general agreement in feature selection with previous studies provides additional support to the validity of using transcriptional profiling data for model training. Besides these well-known features, new sequence features were identified by rigorous statistical analysis. All these significant features were quantitatively combined and analyzed together in a common computational framework. In this way, the lack of certain features could be compensated by the presence of other significant features for target identification. In contrast, many existing algorithms focus on a few selected features and the predicted targets are required to possess all these features.

One potential limitation of using transcriptional profiling data alone for algorithm training is that some targets could be excluded from our analysis if they are principally down-regulated at the protein level. This would potentially reduce target detection sensitivity because we did not include protein expression data in the training process. However, it might not be a major issue because many of the identified top selection features, such as perfect seed pairing, target site conservation and structural accessibility, are also most useful features to date for the prediction of targets that are downregulated at the protein level. In addition, a recent study suggested that there may not be specific determinants for target downregulation only at the protein level (Grimson *et al*., 2007). This issue may be further clarified when proteomic data are available for high-throughput target validation. The accuracy of target prediction may be further improved by integratively analyzing transcriptional profiling and proteomic data together when they are available in the future.

4.2 Target prediction model

Among all the selected features for target prediction, some are more important than others. Therefore, they should not be considered with equal weight. We have previously developed a prediction algorithm (MirTarget v1) to heuristically assign weights to individual features based on past research experience. With the availability of large datasets from microarray experiments, it is now possible to rely on machine learning algorithms to computationally assign weights and combine all the features together in a non-linear way.

The robust performance of MirTarget2 was mainly the result of using a large high-quality training dataset and the integration of many statistically significant known and novel features for model building. Although the contribution of each individual feature may be small, the improvement on the algorithm performance was more evident when these features were considered collectively (Fig. 1). MirTarget2 was based on an SVM framework in which various overlapping and non-overlapping features were combined. In this way, non-linear interactions among the features can be captured and used for model improvement. In addition, MirTarget2 was able to integrate features that were heterogeneous in nature and both numerical and categorical features were combined and analyzed within a common computational framework.

Although our target prediction model compared favorably with other algorithms, it still needs to be significantly improved in future. The algorithm correctly identified over 90% of the negative training samples; however, it only identified half of the positive training samples. It is likely that many of the undetected positive training sequences represent true miRNA targets. Further algorithmic improvement may be possible by incorporating new training data and identifying new relevant features.

ACKNOWLEDGEMENTS

We thank Lance Ford for reading the manuscript. This research was supported by a startup fund from Washington University School of Medicine in St. Louis.

Conflict of Interest: none declared.

REFERENCES

Ambros,V. (2004) The functions of animal microRNAs. *Nature*, **431**, 350–355.
Bagga,S. *et al*. (2005) Regulation by let-7 and lin-4 miRNAs results in target mRNA degradation. *Cell*, **122**, 553–563.
Barrett,T. *et al*. (2007) NCBI GEO: mining tens of millions of expression profiles—database and tools update. *Nucleic Acids Res*., **35**, D760–D765.
Benson,D.A. *et al*. (2007) GenBank. *Nucleic Acids Res*., **35**, D21–D25.
Brennecke,J. *et al*. (2005) Principles of microRNA-target recognition. *PLoS Biol*., **3**, e85.
El-Naqa,I. *et al*. (2002) A support vector machine approach for detection of microcalcifications. *IEEE Trans. Med. Imaging*, **21**, 1552–1563.
Enright,A.J. *et al*. (2003) MicroRNA targets in Drosophila. *Genome Biol*., **5**, R1.
Farh,K.K. *et al*. (2005) The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. *Science*, **310**, 1817–1821.
Gaidatzis,D. *et al*. (2007) Inference of miRNA targets using evolutionary conservation and pathway analysis. *BMC Bioinformatics*, **8**, 69.
Griffiths-Jones,S. *et al*. (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res*., **34**, D140–D144.
Grimson,A. *et al*. (2007) MicroRNA Targeting specificity in mammals: determinants beyond seed pairing. *Mol. Cell*, **27**, 91–105.
He,L. and Hannon,G.J. (2004) MicroRNAs: small RNAs with a big role in gene regulation. *Nat. Rev. Genet*., **5**, 522–531.
Hofacker,I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res*., **31**, 3429–3431.
Jing,Q. *et al*. (2005) Involvement of microRNA in AU-rich element-mediated mRNA instability. *Cell*, **120**, 623–634.
Kertesz,M. *et al*. (2007) The role of site accessibility in microRNA target recognition. *Nat. Genet*., **39**, 1278–1284.
Kim,S.K. *et al*. (2006) miTarget: microRNA target gene prediction using a support vector machine. *BMC Bioinformatics*, **7**, 411.
Kiriakidou,M. *et al*. (2004) A combined computational-experimental approach predicts human microRNA targets. *Genes Dev*., **18**, 1165–1178.
Krek,A. *et al*. (2005) Combinatorial microRNA target predictions. *Nat. Genet*., **37**, 495–500.
Lai,E.C. (2002) Micro RNAs are complementary to 3′ UTR sequence motifs that mediate negative post-transcriptional regulation. *Nat. Genet*., **30**, 363–364.
Lai,E.C. *et al*. (2005) Pervasive regulation of Drosophila Notch target genes by GY-box-, Brd-box-, and K-box-class microRNAs. *Genes Dev*., **19**, 1067–1080.
Lee,Y.S. and Dutta,A. (2007) The tumor suppressor microRNA let-7 represses the HMGA2 oncogene. *Genes Dev*, **21**, 1025–1030.

Lewis,B.P. *et al*. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, **120**, 15–20.

Lim,L.P. *et al*. (2005) Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature*, **433**, 769–773.

Linsley,P.S. *et al*. (2007) Transcripts targeted by the microRNA-16 family cooperatively regulate cell cycle progression. *Mol. Cell Biol.*, **27**, 2240–2252.

Long,D. *et al*. (2007) Potent effect of target structure on microRNA function. *Nat. Struct. Mol. Biol.*, **14**, 287–294.

Miranda,K.C. *et al*. (2006) A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. *Cell*, **126**, 1203–1217.

Miska,E.A. (2005) How microRNAs control cell division, differentiation and death. *Curr. Opin. Genet. Dev.*, **15**, 563–568.

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

Rajewsky,N. (2006) microRNA target predictions in animals. *Nat. Genet.*, **38** (**Suppl**), S8–S13.

Rehmsmeier,M. *et al*. (2004) Fast and effective prediction of microRNA/target duplexes. *RNA*, **10**, 1507–1517.

Robins,H. *et al*. (2005) Incorporating structure to predict microRNA targets. *Proc. Natl Acad. Sci. USA*, **102**, 4006–4009.

Sethupathy,P. *et al*. (2006) TarBase: A comprehensive database of experimentally supported animal microRNA targets. *RNA*, **12**, 192–197.

Sontheimer,E.J. and Carthew,R.W. (2005) Silence from within: endogenous siRNAs and miRNAs. *Cell*, **122**, 9–12.

Stark,A. *et al*. (2005) Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3′UTR evolution. *Cell*, **123**, 1133–1146.

Stark,A. *et al*. (2003) Identification of Drosophila MicroRNA targets. *PLoS Biol.*, **1**, E60.

Wang,X. and Wang,X. (2006) Systematic identification of microRNA functions by combining target prediction and expression profiling. *Nucleic Acids Res.*, **34**, 1646–1652.

Zhao,Y. *et al*. (2005) Serum response factor regulates a muscle-specific microRNA that targets Hand2 during cardiogenesis. *Nature*, **436**, 214–220.