Genomics & Informatics Vol. 3(1) 15-23, March 2005

Identification of *Caenorhabditis elegans* MicroRNA Targets Using a Kernel Method

Wha-Jin Lee[1,2], Jin-Wu Nam[1,2], Sung-Kyu Kim[1,2] and Byoung-Tak Zhang[1,2,*]

[1]Center for Bioinformation Technology (CBIT),
[2]Biointelligence Laboratory, School of Computer Science and Engineering, Seoul National University, Seoul 151-742, Korea

Abstract

Background

MicroRNAs (miRNAs) are a class of noncoding RNAs found in various organisms such as plants and mammals. However, most of the mRNAs regulated by miRNAs are unknown. Furthermore, miRNA targets in genomes cannot be identified by standard sequence comparison since their complementarity to the target sequence is imperfect in general. In this paper, we propose a kernel-based method for the efficient prediction of miRNA targets. To help in distinguishing the false positives from potentially valid targets, we elucidate the features common in experimentally confirmed targets.

Results

The performance of our prediction method was evaluated by five-fold cross-validation. Our method showed 0.64 and 0.98 in sensitivity and in specificity, respectively. Also, the proposed method reduced the number of false positives by half compared with TargetScan. We investigated the effect of feature sets on the classification of miRNA targets. Finally, we predicted miRNA targets for several miRNAs in the *Caenorhabditis elegans* (*C. elegans*) 3' untranslated region (3' UTR) database.

Conclusions

The targets predicted by the suggested method will help in validating more miRNA targets and ultimately in revealing the role of small RNAs in the regulation of genomes. Our algorithm for miRNA target site detection will be able to be improved by additional experimental-knowledge. Also, the increase of the number of confirmed targets is expected to reveal general structural features that can be used to improve their detection.

*Corresponding author: E-mail btzhang@cse.snu.ac.kr, Tel +82-2-875-2240, Fax +82-2-880-1833
Accepted 4 February 2005

Introduction

MicroRNAs (miRNAs) are endogenous ~22 nt RNAs that act as post-transcriptional regulators in animals and plants by binding to the mRNAs. They induce their cleavages or block their translation into proteins (Bartel, 2004). The first members of the miRNAs, lin-4 and let-7, were discovered in *C. elegans*, and they are components of the gene regulatory network that controls the timing of *C. elegans* larval development (Lee *et al.*, 1993; Wightman *et al.*, 1993; Ha *et al.*, 1996; Moss *et al.*, 1997; Seggerson *et al.*, 2002; Abrahante *et al.*, 2003; Lin *et al.*, 2003; Reinhart *et al.*, 2000). Recently, it was reported that miRNAs are implicated in control of cell proliferation (Brennecke *et al.*, 2003), cell death for fat metabolism in flies (Xu *et al.*, 2003), control of leaf and flower development in plants (Aukeman *et al.*, 2003; Chen, 2004; Emery *et al.*, 2003; Palatnik *et al.*, 2003), and hematopoietic lineage differentiation (Chen *et al.*, 2004), though no targets were confirmed in these studies. These suggest the a broad range of possible functions for miRNAs. The overall importance of miRNAs has been further established by the notion that many miRNAs appear to have tissue-specific or developmental stage-specific expression patterns as well as their evolutionary conservation, which is very strong within mammals and often extends to invertebrate homologs (Lagos-Quintana *et al.*, 2001; Lau *et al.*, 2001; Lee *et al.*, 2001; Lai, 2003; Lim *et al.*, 2003a; Lim *et al.*, 2003b; Pasquinelli *et al*, 2000; Aravin *et al.*, 2001; Lagos-Quintana *et al.*, 2002; Lagos-Quintana *et al*, 2003; Ambros *et al.*, 2003; Dostie *et al.*, 2003; Houbaviy *et al.*, 2003; Krichevsky *et al.*, 2003).

However, although several functions of miRNA are uncovered, the factors and the mechanisms related to function of miRNAs are still unknown. The functional annotation of miRNAs is difficult because the size of miRNAs is small and the experiments for target prediction are not efficient. Therefore, a computational method to identify the target genes that are regulated by miRNAs would greatly help the study of miRNA function in animals (Ambros, 2001).

Targets for plant miRNAs have been identified on a

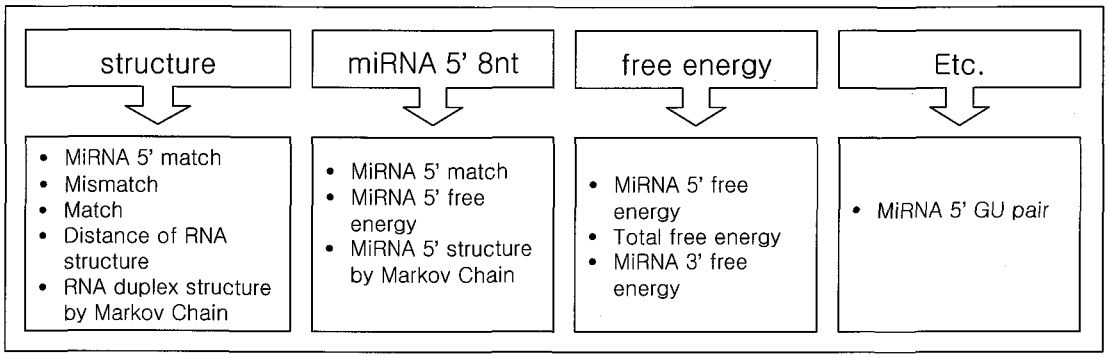

Fig. 1. Feature selection for learning.

genome-wide scale by searches which require a high degree of sequence complementarity (Rhoades *et al.*, 2003; Tang *et al.*, 2003). However, most animal miRNAs are thought to recognize their mRNA 3' UTR via partial complementarity (Lee *et al.*, 1993; Wightman *et al.*, 1993; Moss *et al.*, 1997; Reinhart *et al.*, 2000; Zeng *et al.*, 2002; Doench *et al.*, 2003). Because of this partial complementarity, simple homology-based searches have failed to uncover targets for miRNAs (Ambros *et al.*, 2003; Bartel and Bartel, 2003). Recently, carefully designed computational approaches have been used to predict mRNA targets for *Drosophila* (Stark *et al.*, 2003; Enright *et al.*, 2003) and mammalian miRNAs (Lewis *et al.*, 2003).

The methods used by Lewis *et al.* (2003) and Stark *et al.* (2003) incorporated conservation of the mRNA target site in the related organisms to separate signal from noise. However, the methods the have high false positive rate because they rely on inferences from a free energy of the miRNA/target duplex folding.

We used the kernel method to classify the miRNA targets, which is popular in modern statistical branch, particularly in probability density estimation and regression function approximation. In this paper, we propose an efficient kernel-based method that predicts *C. elegans* miRNA targets and present computational evidences that the most important factor to determine miRNA targets is the 5' region of miRNA.

Methods

Feature selection

Experimentally validated miRNA/target duplexes contain mismatches, gaps, and loops at different positions. Such structures of the duplexes make it difficult to identify targets within whole-genome or transcriptome databases, since standard alignment methods produce many false positives with such short variable sequences. Furthermore, the small number of validated examples makes it difficult the correct identification of miRNA targets by traditional classifiers.

In this paper, we analyzed the various characteristics of miRNA/target duplexes to reduce the false positive rate and classified the features into four major categories. Fig. 1 presents the four main characteristics of the miRNA/target duplexes. First, we used the structure information of miRNA/target duplexes. Even though the miRNA sequence has diverged, the secondary structure in confirmed miRNA/target pairs might be conserved. Therefore, we extracted the structural features based on the characteristic patterns of the secondary structure. Then, we calculated the distances of RNA secondary structures using the RNAdistance software (Hofacker *et al.*, 1994) and compared the similarity of secondary structure of miRNA/target duplex using a Markov Chain model. Also, we counted the number of match and mismatch sites in the miRNA/target duplexes.

Second, we used the features of the miRNA 5' region. The ability of a miRNA to translationally repress a target mRNA is largely dictated by the free energy of binding of the first eight nucleotides in the 5' region of the miRNA (Doench *et al.*, 2004). Moreover, the 5' ends of related miRNAs tend to be better conserved than the 3' ends (Lim *et al.*, 2003; Mallory *et al.*, 2004), further supporting the hypothesis that these segments are most crucial for miRNA target recognition. Therefore, we used the structure and the free energy data extracted from the miRNA 5' region/target duplexes.

Third, we used the free energy of miRNA/target duplexes formation. The pairing of the miRNA 5' region to the mRNA is sufficient to cause repression, and the free energy value of this interaction is an important determinant of activity. The 3' region of the miRNA is less critical, but can modulate activity in certain circumstances (Doench *et al.*, 2004). Therefore, we calculated the free energy of three different parts, the free energy of miRNA/mRNA duplex, miRNA 5' region/mRNA duplex and miRNA 3' region/mRNA duplex.

Table 1. The features for SVM classifier

(1) The number of matches at the 8 nt of miRNA 5' region
(2) MiRNA 5' region/mRNA duplex free energy
(3) MiRNA/mRNA duplex free energy
(4) The number of G/U wobble pairs at the 8 nt of miRNA 5' region
(5) The number of mismatch of miRNA/mRNA duplex
(6) The number of match of miRNA/mRNA duplex
(7) MiRNA 3' region/mRNA duplex free energy
(8) The distance of miRNA/mRNA duplex secondary structures using the RNAdistance program
(9) The similarity of secondary structures of miRNA/mRNA duplex using a Markov chain model
(10) The similarity of secondary structures of miRNA 5' region /mRNA duplex using a Markov chain model

Lastly, we used the G:U wobble pairing feature because G:U wobble pairing is highly detrimental to miRNA function despite its favorable contribution to RNA:RNA duplexes (Doench *et al.*, 2004). Table 1 presents all features described above.

Secondary structure prediction of RNA/RNA duplexes

Mfold (Zuker, 2003) is a program package for the RNA secondary structure prediction using nearest neighbor thermodynamic rules. We searched for the most stable binding site with RNA sequences from 3'UTR database using MFold. We calculated the free energy, the number of G:U wobble pairs and the number of mismatch and match through this program.

Distance of RNA secondary structure

We used the RNAdistance software of the Vienna RNA package (Hofacker *et al.*,1994) to calculate distances between the analysed RNA secondary structures. RNAdistance accepts structures in bracket format, where matching brackets symbolize base pairs and unpaired bases are represented by a dot '.', or coarse grained representations where hairpins, interior loops, bulges, multi loops, stacks and external bases. We calculated the distance by

$$score = \frac{\sum^{i=n_p} RNAdistance\,(str_i, str_{query})}{n_p} \quad (1)$$

where n_p is the number of positive training data, str_{query} is a query structure, and str_i is a structure of positive training data. (see the eighth feature in Table 1)

Comparison of RNA secondary using a Markov Chain probability

Markov Chain is a random process which has the property that the next state is conditionally independent of the past given the current state. It is useful for biological structure analysis because of their ability to incorporate biological information in their structure. We calculated the Markov Chain probability according to the frequency of states that are given in Fig. 2. The structural probability was calculated by

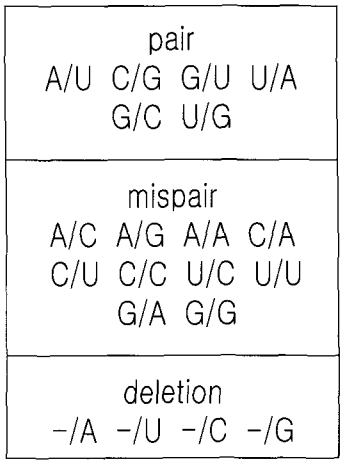

Fig. 2. Definition of States for a Markov Chain

$$Score(x_{ij}) = \log\left(\frac{f(x_{ij}) + s(con)}{p(x_{ij})}\right) \quad (2)$$

where $f(x_{ij})$ is the structural probability for position *i* and *j*, $s(con)$ is a low-valued constant to prevent log going to zero and $p(x_{ij})$ is the background probability. We constructed the Markov Chain as equation (3) and (4) when *s* and *t* is a given state.

$$a_{st} = P[(x_i = t)|(x_{i-1} = s)] \quad (3)$$

$$P(x) = a_{0x_1}\prod_{i=1}^{L} a_{x_i x_{i+1}} \quad (4)$$

We used the Markov Chain probability to compare a RNA secondary structure of miRNA/mRNA duplex. (see the ninth and tenth features in Table 1)

Kernel method for target identification

We used a support vector machine (SVM) to classify miRNA targets from mRNA 3' UTR database. This method has attracted a lot of attention by its successful application in pattern recognition (Scholkopf *et al.*, 1999). The kernel trick used in SVM is applicable not only for classification but also for other linear techniques (Vapnik, 1998). SVM is a method of obtaining the optimal boundary of two sets in a vector space independently on probabilistic distributions of training vectors in the sets.

Table 2. The miRNA target prediction for *lin-4* in the *C.elegans* 3'UTR database

Genes	products	accession number	miRNA 5' ΔG	total ΔG
lin-41A	LIN-41A	AC: CC085391	-7.6	-19.6
lin-28	LIN-28	AC: CC013576	-11.5	-18.6
lin-14	LIN-14A	AC: CC013457	-11.5	-17.5
nhr-11	nuclear receptor NHR-11	AC: CC073419	-11.5	-16.5
hbl-1	hunchback-related protein	AC: CC073457	-7.6	-12.5

These predicted miRNA targets contain all known *lin-4* targets: *lin-41, lin-14, hbl and lin-28.* This table was sorted by total free energy.

Its fundamental idea is locating the boundary that is most distant from the vectors nearest to the boundary in both of the sets. Let x be a vector in a vector space. A boundary hyperplane is expressed in the form of $f(x) = w^t x + b$ where w is a weight coefficient vector and b is a bias term. The distance between a training vector x_i and the boundary, called *margin*, is expressed and reduced as maximization of $1/\ w\ ^2$. Consequently, the optimization is formalized as

$$minimize \frac{1}{2}\ w\ ^2 \qquad (5)$$
$$\text{subject to } y_i\,(w \cdot z_i - b) \geq 1, i = 1, ..., n.$$

This conditional optimization can achieved by Largrange's method of indeterminate coefficient.

$$\max_a \sum_{i=1}^{n} a_i - \frac{1}{2}\sum_{i,j=1}^{n} a_i a_j y_i y_j k(x_i, x_j) \qquad (6)$$
$$\text{subject to } a_i \geq 0, i = 1, ..., n, \textstyle\sum_{i=1}^{n} a_i y_i = 0$$

where $k(x_i, x_j)$ is a kernel function. If the sets are not linearly separable, the slack variable ζ_i is allowed to exist in a limited region in the erroneous side along the boundary. Also, in this paper, we used the polynomial kernel $K(x, x') = (x^T x' + 1)^p$and sequential minimal optimization (SMO) algorithm (Keerthi *et al.*, 2001; Platt *et al.*, 1999) to learn our SVM. The SMO that can be viewed as the most extreme case of decomposition methods is the most popular optimization algorithms for SVM. It allows for fast convergence with small memory requirements even on large problem.

The learning methods used in this study were obtained from the WEKA machine learning package. We used the SMO algorithm with the complexity parameter $C = 4$ and the polynomial kernel exponent = 5. All other parameters were default values.

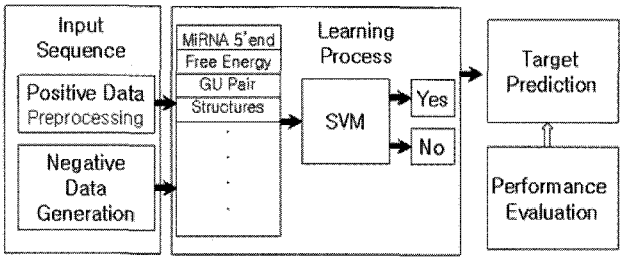

Fig. 3. The program process for target identification

Results

The framework and implementation of our method

The overall process of our method is described in Fig.3. We extracted structural information, free energy of miRNA/target site interaction and 8 nucleotides (nt) information of miRNA 5' end as input data from negative and positive datasets. Then, input data was classified by the kernel method implemented by SVM. The computational experiments in this study were performed by the WEKA machine learning package (Witten and Frank, 1999) (http://www.cs.waikato.ac.nz/~ml/weka/).

Datasets

The training data for the SVM classifier is a set of RNA/RNA pairs divided into positive and negative samples. The positive training set consists of 39 experimentally defined miRNA/target pairs (Slack *et al.*, 2000; Lin *et al.*, 2003; Banerjee *et al.*, 2002; Poy *et al.*, 2004; Stark *et al*, 2003). The positive samples include seven pairs of *lin-14/ cel-let-7*, three pairs of *lin-14/ cel-lin-4*, one pair of *lin-28/cel-lin-4*, one pair of *lin-28/cel-let-7*, one pair of *lin-41/cel-lin-4*, one pair of *lin-41/cel-let-7*, five pairs of *daf-12/cel-let-7*, one pair of *hbl/cel-lin-4*, ten pairs of *hbl/cel-let-7*, four pairs of *hid/dme-bantam*, one pair of *HLHm3/dme-miR-7*, one of *hiary/dme-miR-7*, one pair of *rpr/dme-miR-2*, one pair of *grim/dme-miR-2* and one pair of *Mtpn/mir-375.*

The negative training set consists of 1022 random sequence/target pairs. We searched for the high-affinity

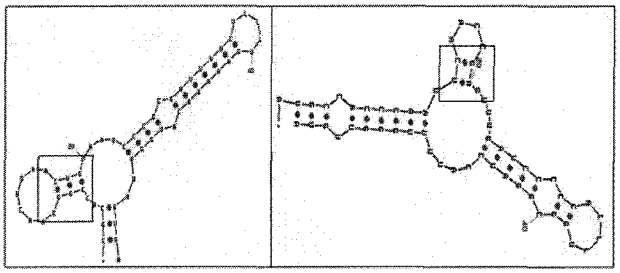

Fig. 4. Structures not found from miRNA/target pairs.
Shown are the RNA/RNA duplex structures as predicted by MFOLD. All sequence pairs containing additional branch structures were removed from negative data.

Table 3. The miRNA target prediction for *let-7* in the *C.elegans* 3'UTR database

Genes	products	accession number	miRNA 5' ΔG	total ΔG
lin-41A	LIN-41A	AC: CC085391	-9.8	-24.2
daf-12	DAF-12 A2	AC: CC073740	-9.8	-23.4
lin-14	LIN-14A	AC: CC013457	-9.6	-23
hbl-1	hunchback-related protein	AC: CC073457	-8.6	-22.2
lin-28	LIN-28	AC: CC013576	-8.6	-20.2
unc-129	UNC-129	AC: CC054258	-9.8	-20
CePqM96	paraquat-inducible protein	AC: CC013365	-6.9	-19.8
-	nuclear receptor NHR-43	AC: CC125308	-7.8	-19.5
skr-21	SKR-21	AC: CC181290	-9.6	-18
Mio	Mlx interactor	AC: CC125253	-9.8	-17.2
ces-2	CES-2	AC: CC013549	-9.6	-16.2
-	-	AC: CC013431	-9.8	-14.8
unc-16	UNC-16	AC: CC181137	-6.7	-14.6
wrk-1	immunoglobulin domain-containing protein WRK-1C	AC: CC181358	-9.8	-14.5
pip-1	PIP-1	AC: CC228823	-8.9	-14
daf-4	BMP receptor	AC: CC013410	-8	-13.9
daf-16	DAF-16	AC: CC046572	-8.6	-13.7
-	histone H1.Q	AC: CC085543	-9.6	-13.7
unc-2	High voltage activated calcium channel alpha-1	AC: CC230280	-7	-12.2
-	histone H1.1	AC: CC012659	-9.8	-11.6
-	sodium-calcium exchanger	AC: CC049470	-5.2	-10.1
unc-115	putative actin-binding protein UNC-115	AC: CC060593	-6.9	-9.7

The predicted miRNA targets contain all known *let-7* targets: *lin-41, daf-21, lin-14, hbl and lin-28.* Tables are sorted in ascending order by total free energy.

binding sites with random sequences from *C. elegans* 3'UTR database (ftp://bighost.ba.itb.cnr.it/pub/Embnet/Database/UTR/data/) to make the negative training set. The random sequences similar in length to miRNAs (approximately 18-22 nt) were produced by site-independent sampling. We extracted the random sequence/target pairs that have more than six perfect Watson-Crick pairs of 8 nt from the miRNA 5' end, and that have a thermodynamicaly stable free energy less then -8.5 kcal/mole. Also all sequence pairs with structures like Fig.4, consisting of additional branch structures, were removed from the negative data since those structures were not found in the positive datasets.

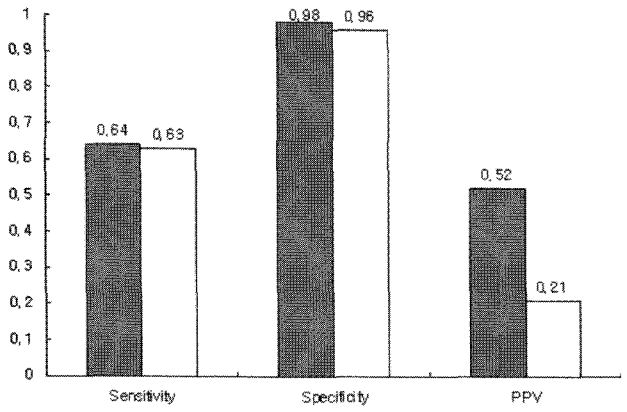

Fig. 5. The performance comparison with TargetScan
The grey box indicates the efficiency by SVM and the white box indicates the efficiency by TargetScan. The sensitivity and specificity obtained by SVM are similar or better than TargetScan program.

Performance of the kernel method

The objective of this study is to construct classifiers that can correctly classify the miRNA target genes from the 3'UTR database. The performance of our classification method was evaluated by five-fold cross-validation. That is, the whole data set was partitioned into five subsets. The four of the subsets were used as a training test, and the rest were used as a test set, and this process was repeated five times. Table 2 shows the sensitivity and the specificity of the five-fold cross-validation in classifying the miRNA target genes using the kernel machine. The sensitivity was 0.64 and the specificity was 0.98.

Also, we compared the performance to TargetScan (Lewis *et al.*, 2003) with 32 of 39 training data and 1751 random negative data. TargetScan combines thermodynamics-based modelling of RNA/RNA duplex interactions with comparative sequence analysis to predict miRNA targets conserved across multiple genomes. Fig.5 shows that

Table 4. The performance of 5 fold cross-validation for training data set

TP	FN	FP	TN	Sensitivity	Specificity	PPV
25	14	23	999	0.64	0.98	0.52

The performance is presented in terms of three statistical measures: sensitivity = TP/(TP+FN); specificity = TN/(TN+FP) and PPV = TP/(TP+ FP) where TP is the number of true positives, TN is the number of true negative, FP is the number of false positives and FN is the number of false negatives.

Table 5. The miRNA target predictions for *miR-228, miR-229, miR-230 and miR-231* in the *C.elegans* 3'UTR database

Genes	products	accession number	miRNA 5' ΔG	total ΔG
cel-miR-228				
unc-75	putative RNA-binding protein	AC: CC266578	-8.9	-25.6
vab-2	VAB-2	AC: CC085460	-8.8	-16.5
vab-10	VAB-10A protein	AC: CC231603	-7.5	-16.2
lin-9	LIN-9L	AC: CC103462	-7.5	-16.1
pcr55	transmembrane protein	AC: CC013339	-8.9	-15.7
pme-1	poly ADP-ribose metabolism enzyme-1	AC: CC181762	-9.1	-15.1
tim9b	small zinc finger-like protein	AC: CC084885	-8.9	-13.4
ehs-1	EHS-1	AC: CC126116	-7.5	-12.8
-	methuselah-like protein MTH-1	AC: CC279142	-6.9	-11.6
cel-miR-229				
-	Na/Ca,K-exchanger	AC: CC054777	-6.8	-13.2
	cel-miR-230			
mab-21	mab-21 protein	AC: CC103635	-5.5	-13.6
cel-miR-231				
klp-12	kinesin like protein KLP-12	AC: CC121004	-7.7	-10.8
mom-1	MOM-1	AC: CC012616	-7	-21.6
let-413	LET-413 protein	AC: CC103637	-8.5	-12.7
let-23	tyrosine kinase	AC: CC013452	-8.5	-10.9

the sensitivity and the specificity of our method using SVM are similar to or better than TargetScan program. However, the more important thing is that the number of false positives is much lower than TargetScan, that is, 35 false positives by our method compared with 77 in TargetScan. This shows that our method is more efficient and correct than TargetScan.

MicroRNA target prediction in C. elegans

We applied our method to *let-7* and *lin-4* miRNA genes, whose targets are known (Wightman *et al.*, 1993; Ha *et al.*, 1996; Moss *et al.*, 1997; Olsen *et al.*, 1999; Seggerson *et al.*, 2002; Slack *et al.*, 2000; Abrahante *et al.*, 2003; Lin *et al.*, 2003; Banerjee *et al.*, 2002) (Tables 3 and 4). These predicted miRNA targets contain all the known *let-7* targets: *lin-41, daf-21, lin-14, hbl* and *lin-28*. This shows that most miRNA/target pairs can possibly be detected by an SVM classifier with low specificity in genome-wide searches. Table 5 presents the miRNA targets of miR-228, miR-229, miR-230 and miR-231.

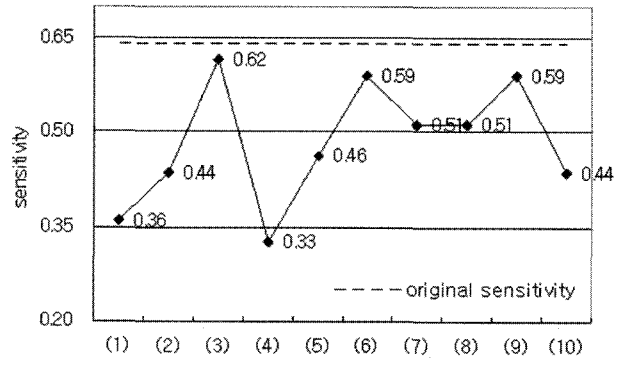

Fig. 6. Feature's influence on performance.
The x-axis means the feature number of Table 1. Each SVM classifier is trained without the feature of the corresponding number. The feature with lower sensitivity is assumed to be a more important factor in the identification of miRNA target.

Analysis of Feature Set

In this section, we investigate the effect of feature set on the performance of miRNA target classifier. We excluded a feature from the entire feature set one by one and examined how much each feature contributes to the performance of the classifier. Fig. 6 presents each feature's influence on the performance. The feature having lower sensitivity will be a more important factor to decide miRNA targets. The top-three features having the lowest sensitivity were the number of G/U wobble pairs at the 8 nt of miRNA 5' region (feature (4)), the number of matches at the 8 nt of miRNA 5' region (feature (1)), and the similarity of secondary structure of miRNA 5' region /mRNA duplex using a Markov chain model (feature (10)). All of them are related to the miRNA 5' region. Every experiment has a similar specificity (97% ~ 98%). Moreover, Fig. 7 shows that features the miRNA 5' region are the most important information to decide whether miRNA target or not, further supporting the hypothesis

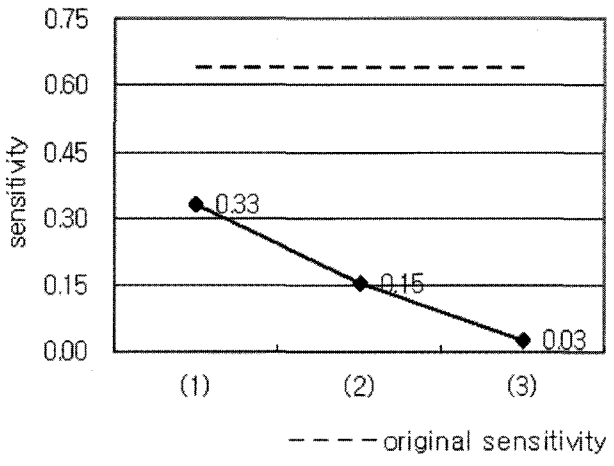

Fig. 7. Feature's influence on performance. (1) SVM classifier trained without information about free energy. (2) SVM classifier trained without RNA structure information. (3) SVM classifier trained without information of miRNA 5' region. This shows that features about miRNA 5' region are the most important information to decide whether miRNA target or not.

that the miRNA 5' region is most critical for miRNA target recognition (Doench *et al.*, 2004).

Discussion

The application of a computational approach to the prediction of miRNA targets is often hindered by the small size of miRNA sequences. The prediction of miRNA targets is more complicated due to the tendency of imperfectness in miRNA/mRNA pairings in animals. We presented a kernel-based classification method to overcome these problems and tried to identify potential miRNA targets with an RNA-folding program that evaluates the structural and thermodynamic plausibility of the predicted pairs and distinguishes the real from the random matches. Kernel-based statistical learning methods have a number of advantages for the analysis of not only vectorial and matrix data which are common in classic statistical analysis but also more exotic data types such as string, trees and graphs. The ability to handle such data is clearly essential in the biological domain. The kernel-based method provides significant opportunities for the incorporation of more specific biological knowledge and unlabelled data.

We demonstrated that the SVM classifier for the computational identification of miRNA target sites can detect miRNA target sites with high specificity. The result of this method will be able to provide better understanding of how miRNAs bind their targets. To help distinguish the false positives from potentially valid targets, we identify the features shared by valid targets. Also, the method can be applied to other species as well, because many of these miRNAs are phylogenetically conserved, suggesting strong evolutionary pressure. The functional target sites are conserved in homologous genes from related species (Moss and Tang, 2003), so we can improve the performance through the analysis of the orthologous 3'UTRs. Efforts to find more animal miRNA targets will be greatly helpful because of the deeper understanding of structural and biochemical nature of miRNA/mRNA pairing. The targets predicted by the proposed method will help in validating more miRNA targets and ultimately in revealing the role of these small RNAs in the regulation of the genome.

Acknowledgements

This work was supported by the Korea Ministry of Science and Technology under the NRL program and Systems Biology Program and by the Korea Ministry of Education and Human Resources under the BK21-IT program. The ICT at Seoul National University provided research facilities for this study.

References

Abrahante, J.E., Daul, A.L., Li, M., Volk, M.L., Tennessen, J.M., Miller, E.A., and Rougvie, A.E. (2003). The *Caenorhabditis elegans* hunchback-like gene *lin-57/hbl-1* controls developmental time and is regulated by microRNAs. *Dev. Cell* 4, 625-637.

Ambros, V. (2001). MicroRNAs: Tiny regulators with great potential. *Cell* 107, 823-826.

Ambros, V., Lee, R.C., Lavanway, A., Williams, P.T., and Jewell, D. (2003). MicroRNAs and other tiny endogenous RNAs in *C. elegans*. *Curr. Biol.* 13, 807-818.

Aravin, A.A., Naumova, N.M., Tulin, A.A., Rozovsky, Y.M., and Gvozdev, V.A. (2001). Double-stranded RNA-mediated silencing of genomic tandem repeats and transposable elements in *Drosophila* melanogaster germline. *Curr. Biol.* 11, 1017-1027.

Aukeman, M.J. and Sakai, H. (2003). Regulation of flowering time and floral organ identity by a MicroRNA and its APETALA2-like target genes. *Plant Cell* 15, 2730-2741.

Banerjee, D. and Slack, F. (2002). Control of developmental timing by small temporal RNAs: a paradigm for RNA-mediated regulation of gene expression. *BioEssays* 24, 119-129.

Bartel, B. and Bartel, D.P. (2003). MicroRNAs: At the root of plant development? *Plant Physiol.* 132, 709-717.

Bartel, D.P. (2004). MicroRNAs:genomics, biogenesis, mechanism, and function. *Cell* 116, 281-97

Brennecke, J., Hipfner, D.R., Stark, A., Russell, R.B., and Cohen, S.M. (2003). Bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the pro-apoptotic gene hid in *Drosophila*. *Cell* 113, 25-36.

Chen, C.Z., Li, L., Lodish, H.F., and Bartel, D.P. (2004). MicroRNAs modulate hematopoietic lineage differentiation. *Science* 303, 83-86.

Chen, X. (2004). A microRNA as a translational repressor of APETALA2 in *arabidopsis* flower development. *Science* 303, 2022-2025.

Doench, J.G., Petersen, C.P., and Sharp, P.A. (2003). siRNAs can function as miRNAs. *Genes Dev.* 17, 438-442.

Doench, J.G. and Sharp, P.A. (2004). Specificity of microRNA target selection in translational repression. *Genes. Dev.* 18, 504-11.

Dostie, J., Mourelatos, Z., Yang, M., Sharma, A., and Dreyfuss, G. (2003). Numerous microRNPs in neuronal cells containing novel microRNAs. *RNA* 9, 631-632.

Emery, J.F., Floyd, S.K., Alvarez, J., Eshed,Y., Hawker, N.P., Izhaski, A., Baum, S.F., and Bowman, J.L. (2003). Radial patterning of *Arabidopsis* shoots by class III HD-ZIP and KANADI genes. *Curr. Biol.* 13, 1768-1774.

Enright, A.J., John, B., Gaul, U., Tuschl, T., Sander, C., and Marks, D.S. (2003). MicroRNA targets in *Drosophila*. *Genome Biol.* 5, R1.

Ha, I., Wightman, B., and Ruvkun, G. (1996). A bulged *lin-4/lin-14* RNA duplex is sufficient for Caenorhabditis elegans *lin-14* temporal gradient formation. *Genes. Dev.* 10, 3041-3050.

Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, S., Tacker, M., and Schuster, P. (1994). Fast folding and comparison of RNA secondary structures. *Monatsh Chem.* 125, 167-188.

Houbaviy, H.B., Murray, M.F., and Sharp, P.A. (2003). Embryonic stem cell-specific microRNAs. *Dev. Cell* 5, 351-358.

Keerthi, S.S., Shevade, S.K., Bhattacharyya, C., and Murthy, K.R.K. (2001). Improvements to Platt's SMO algorithm for SVM classifier design. *Neural Computation* 13, 637-649.

Krichevsky, A.M., King, K.S., Donahue, C.P., Khrapko, K., and Kosik, K.S. (2003). A microRNA array reveals extensive regulation of microRNAs during brain development. *RNA* 9, 1274-1281.

Lagos-Quintana, M., Rauhut, R., Lendeckel, W., and Tuschl, T. (2001). Identification of novel genes coding for small expressed RNAs. *Science* 294, 853-858.

Lagos-Quintana, M., Rauhut, R., Yalcin, A., Meyer, J., Lendeckel, W., and Tuschl, T. (2002). Identification of tissue-specific microRNAs from mouse. *Curr. Biol.* 12, 735-739.

Lagos-Quintana, M., Rauhut, R., Meyer, J., Borkhardt, A., and Tuschl, T. (2003). New microRNAs from mouse and human. *RNA* 9, 175-179.

Lai, E.C. (2003). MicroRNAs: runts of the genome assert themselves. *Curr. Biol.* 13, R925-R936.

Lau, N.C., Lim, L.P., Weinstein, E.G., and Bartel, D.P. (2001). An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science* 294, 858-862.

Lee, R.C., Feinbaum, R.L., and Ambros, V. (1993). The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell* 75, 843-854.

Lee, R.C. and Ambros, V. (2001). An extensive class of small RNAs in *Caenorhabditis elegans*. *Science* 294, 862-864.

Lewis, B.P., Shih, I.H., Jones-Rhoades, M.W., Bartel, D.P., and Burge, C.B. (2003). Prediction of mammalian microRNA targets. *Cell* 115, 787-798.

Lim, L.P., Glasner, M.E., Yekta, S., Burge, C.B., and Bartel, D.P. (2003a). Vertebrate microRNA genes. *Science* 299, 1540.

Lim, L.P., Lau, N.C., Weinstein, E.G., Abdelhakim, A., Yekta, S., Rhoades, M.W., Burge, C.B., and Bartel, D.P. (2003b). The microRNAs of *Caenorhabditis elegans*. *Genes Dev.* 17, 991-1008.

Lin, S.Y., Johnson, S.M., Abraham, M., Vella, M.C., Pasquinelli, A. *et al.* (2003). The *C. elegans* hunchback homolog, *hbl-1*, controls temporal patterning and is a probable microRNA target. *Dev. Cell* 4, 639-650.

Mallory, A.C., Reinhart, B.J., Jones-Rhoades, M.W., Jang, G., Zamore, P.D., Barton, M.K., and Bartel, D.P. (2004). MicroRNA control of PHABULOSA in leaf development: importance of pairing to the microRNA 5' region. *EMBO J.* 23, 3356-3364.

Moss, E.G., Lee, R.C., and Ambros, V. (1997). The cold shock domain protein LIN-28 controls developmental timing in *C. elegans* and is regulated by the *lin-4* RNA. *Cell* 88, 637-646.

Moss, E.G. and Tang, L. (2003). Conservation of the heterochronic regulator *lin-28*, its developmental expression and microRNA complementary sites. *Dev. Biol.* 258, 432-442.

Olsen, P.H. and Ambros, V. (1999). The lin-4 regulatory RNA controls developmental timing in *Caenorhabditis elegans* by blocking LIN-14 protein synthesis after the initiation of translation. *Dev. Biol.* 216, 671-680.

Palatnik, J.F., Allen, E., Wu, X., Schommer, C., Schwab, R., Carrington, J.C., and Weigel, D. (2003). Control of leaf morphogenesis by microRNAs. *Nature* 20, 257-263.

Pasquinelli, A.E., Reinhart, B.J., Slack, F., Martindale, M.Q., Kuroda, M., Maller, B., Srinivasan, A., Fishman, M., Hayward, D., Ball, E. *et al.* (2000). Conservation across animal phylogeny of the sequence and temporal regulation of the 21 nucleotide *let-7* heterochronic regulatory RNA. *Nature* 408, 86-89.

Platt, J.C. (1999). Fast training of support vector machines using sequential minimal optimization. *In* Advances in kernel methods, support vector learning. (MIT Press).

Poy, M.N., Ellasson, L., Krutzfeldt, J., Kuwajima, S., Ma, X., MacDonald, P.E., Pfeffer, S., Tuschi, T., Rajewsky, N., Rorsman, P., and Stoffel, M. (2004). A Pancreatic islet-specific microRNA regulates insulin secretion. *Nature* 432, 226-30.

Reinhart, B.J., Slack, F.J., Basson, M., Bettinger, J.C., Pasquinelli, A.E., Rougvie, A.E., Horvitz, H.R., and Ruvkun, G. (2000). The 21 nucleotide *let-7* RNA regulates developmental timing in *Caenorhabditis elegans*. *Nature* 403, 901-906.

Rhoades, M.W., Reinhart, B.J., Lim, L.P., Burge, C.B., Bartel, B., and Bartel, D.P. (2002). Prediction of plant microRNA targets. *Cell* 110, 513-520.

Scholkopf, B., Smola, A., and Muller, K. (1999). Kernel principal component analysis. *B. Scholkopf et al. In* Advances in kernel methods, support vector learning,

eb. (MIT Press). pp.327-352.

Seggerson, K., Tang, L., and Moss, E.G. (2002). Two genetic circuits repress the *Caenorhabditis elegans* heterochronic gene *lin-28* after translation initiation. *Dev. Biol.* 243, 215-225.

Slack, F.J., Basson, M., Liu, Z., Ambros, V., Horvitz, H.R. et al. (2000). The lin-41 RBCC gene acts in the C. elegans heterochronic pathway between the let-7 regulatory RNA and the LIN-29 transcription factor. *Mol. Cell* 5, 659-669.

Stark, A., Brennecke, J., Russell, R.B., and Cohen, S.M. (2003). Identification of *Drosophila* microRNA targets. *PLOS Biol.* 1, E60.

Tang, G., Reinhart, B.J., Bartel, D.P., and Zamore, P.D. (2003). A biochemical framework for RNA silencing in plants. *Genes. Dev.* 17, 49-63.

Vapnik, V.N. (1998). Statistical Learning Theory, Wiley, New York.

Wightman, B., Ha, I., and Ruvkun, G. (1993). Posttranscriptional regulation of the heterochronic gene *lin-14* by *lin-4* mediates temporal pattern formation in *C. elegans*. *Cell* 75, 855-862.

Witten, I.H. and Frank, E. (1999). Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann.

Xu, P., Vernooy, S.Y., Guo, M., and Hay, B.A. (2003). The *Drosophila* microRNA miR-14 suppresses cell death and is required for normal fat metabolism. *Curr. Biol.* 13, 790-795.

Zeng, Y., Wagner, E.J., and Cullen, B.R. (2002). Both natural and designed microRNAs can inhibit the expression of cognate mRNAs when expressed in human cells. *Mol. Cell* 9, 1327-1333.

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