Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

Assessing potential miRNA targets based on a Markov model

Hao-Yue Fu[1,2], Ding-Yu Xue[2], Xiang-de Zhang[1] and Pei-Ying Yang[2]

[1]College of Sciences, Northeastern University, Shenyang, China
[2]College of Information Science and Engineering, Northeastern University, Shenyang, China

Corresponding author: Hao-Yue Fu
E-mail: fuhaoyue@tom.com

Genet. Mol. Res. 8 (3): 848-860 (2009)
Received March 15, 2009
Accepted May 25, 2009
Published July 21, 2009

ABSTRACT. At present, studies on microRNA mainly focus on the identification of microRNA genes and their mRNA targets. Although researchers have identified many microRNA genes, relatively few microRNA targets have been identified by experimental methods. Computational programs designed for predicting potential microRNA targets provide numerous targets for experimental validation. We used a Markov model to examine base-pairing binding patterns of known microRNA targets. Using this model, potential microRNA targets in human species predicted by four well-known computational programs were assessed. Each potential target was assigned a score reflecting consistency with known target binding patterns. Targets with scores higher than the cutoff value would be identified by our model. The predicted targets identified by our model have base-pairing binding pat-

terns consistent with known targets. This model was efficient for evaluating the extent to which a potential target was accurately predicted.

Key words: Markov chain model; Machine-learning method; MicroRNA target prediction; Maximum likelihood estimation; Potential target assessment

INTRODUCTION

MicroRNA (miRNA) is a new and large class of genes that do not encode proteins. They produce roughly 22-nucleotide-long transcripts that recognize and bind to partially complementary sites in the 3' untranslated regions of target genes in animals and, by unknown mechanisms, regulate protein production of the target transcripts. The earliest identified miRNAs, lin-4 and let-7, are involved in the timing of developmental processes and bantam has been shown to affect cell proliferation and death (Brennecke et al., 2003), and mir-14 regulates the expression of the cell death pathway and fat metabolism (Xu et al., 2003). To understand the biological functions of miRNAs, it is necessary to identify their targets. Although a large number of animal miRNAs have been identified, only a few targets are known and no experimental high-throughput method for target site identification has been published yet. The prediction of miRNA targets by computational approaches provides an alternative approach to assigning biological functions (Brown and Sanseau, 2005; Zhang et al., 2006). Currently, some computational approaches (Enright et al., 2003; Stark et al., 2003; Kiriakidou et al., 2004; Rajewsky and Socci, 2004; John et al., 2004; Rehmsmeier et al., 2004; Lewis et al., 2003, 2005; Krek et al., 2005; Rusinov et al., 2005; Burgler and MacDonald, 2005; Grun et al., 2005; Robins et al., 2005; Saetrom et al., 2005) have been successfully used to select potential targets for experimental validation.

TargetScanS (Lewis et al., 2005), miRanda (John et al., 2004), PicTar (Krek et al., 2005), and DIANA-microT (Kiriakidou et al., 2004) are the four well-known computational programs designed for predicting miRNA targets in humans. However, the numbers of the predicted targets by different programs are different, with only limited overlap in the top-ranking targets. Variation in the numbers of the predicted target comes from the selection criteria and the use of numerical cutoffs. For example, the complementarity criteria in TargetScanS (Lewis et al., 2005) are consecutive 6-nucleotide-length (positions 2-7 from the 5' end of the miRNA) complementary matches between miRNAs and their targets, whereas in PicTar (Krek et al., 2005), 7-nucleotide-length matches starting at 1 or 2 of the 5' end of the miRNA are required. Incidently, all known targets have no consecutive matches at the 5' end of miRNAs. Moreover, the targets predicted by miRanda and DIANA-microT have more combinational base-pairing complementary match patterns in the 5' end of miRNAs. In order to improve the complementarity criteria, we present a Markov model, which can be learned from complementary patterns of the known miRNA targets. Using the model learned from the known targets (i.e., the training data set), 86% (110 of 128) of the training data were recovered, when the false positive was controlled within 5%. A target would be identified by our model when its scores were more than the

cutoff value. The results shows that although only 35% of the targets (5433 of 15433) predicted by miRanda were identified by the model, the identifying rate was increased to 70% (471 of 671) when the common targets predicted by both miRanda and PicTar were assessed. The improved rate proved that the targets identified by the model had higher probabilities of being true targets.

First-order non-homogeneous Markov chain model

The accuracy of computational prediction of miRNA targets is improved by the thorough refinement of training sets, rules for defining complementarity and statistical approaches (Brown and Sanseau, 2005). In this paper, a novel statistical approach, first-order non-homogeneous Markov chain, is presented to explore base-pairing complementarity patterns of miRNA targets using the known targets as the training data set.

Our Markov chain model is a generative model. The 0-1 sequences generated by the model represent miRNA-target complementarity base-pairing binding patterns. In order to highlight miRNA base-pairing binding patterns, rather than to highlight its target's, a simplified 0-1 sequence was adopted to represent miRNA binding patters. Generally speaking, a base-pairing binding structure, such as a binding structure between an miRNA: 3' UUGAUAUGUUGGAUGAUGGAGU 5' and its gene target: 5'...AAACUCAGAGGUCCUCCUACCUCU...3' can be given as in Table 1a or 1b. After being simplified, a 0-1 sequence miRNA 3'11110000001110111111110 5' was adopted to represent the binding structure. Three kinds of information may be discarded in order to have the miRNA base-pairing binding patterns highlighted (see Table 2 for a further description).

Table 1. The representation of miRNA-target binding structure.

a.	target	5'	A					C	A	G	A	G	G	U				C									U	3'
				A	A	C	U								C	C	U		C	U	A	C	C	U	C			
				|	|	|	|							|	|	|		|	|	|	|	|	|	|			
				U	U	G	A								G	G	A		G	A	U	G	G	A	G			
	mRNA	3'						U	A	U	G	U	U					U								U	5'
b.	target	5'	A	A	A	C	U	C	A	G	A	G	G	U	C	C	U	C	C	U	A	C	C	U	C	U	3'
				|	|	|	|							|	|	|		|	|	|	|	|	|	|		
	mRNA	3'		U	U	G	A	U	A	U	G	U	U		G	G	A	U	G	A	U	G	G	A	G	U	5'
c.	target	5'	A	A	A	C	U	C	A	G	A	G	G	U	C	C	U	C	C	U	A	C	C	U	C	U	3'
				|	|	|	|							|	|	|		|	|	|	|	|	|	|		
	mRNA	3'		1	1	1	1	0	0	0	0	0	0		1	1	1	0	1	1	1	1	1	1	1	0	
d.	mRNA	3'		1	1	1	1	0	0	0	0	0	0		1	1	1	0	1	1	1	1	1	1	1	0	

a. and b. = general binding structures between miRNA: 3' UUGAUAUGUUGGAUGAUGGAGU 5' and its gene target: 5'...AAACUCAGAGGUCCUCCUACCUCU...3'; c. = binding structure with 1 or 0 substituted for nucleotide name at each site of miRNA corresponding to it pairing or not. d. = 0-1 sequence representation.

On the basis of the 0-1 sequence representation, a generative chain model was presented. Each 0-1 sequence can be viewed as being generated by the model with a certain

Table 2. Structure example indicating the advantage of the 0-1 sequence representation.

a.	target	5’	A					C	A	G	A	G	G	U				C								U	3’
				A	A	C	U								C	C	U		C	U	A	C	C	U	C		
				|	|	|	|								|	|	|		|	|	|	|	|	|	|		
				U	U	G	A								G	G	A		G	A	U	G	G	A	G		
	mRNA	3’						U	A	U	G	U	U					U								U	5’
b.	target	5’																								U	3’
				A	A	C	U								C	C	U		C	U	A	C	C	U	C		
				|	|	|	|								|	|	|		|	|	|	|	|	|	|		
				U	U	G	A								G	G	A		G	A	U	G	G	A	G		
	mRNA	3’																								U	5’
c.	target	5’																								U	3’
				U	U	G	A								G	G	A		G	A	U	G	G	A	G		
				|	|	|	|								|	|	|		|	|	|	|	|	|	|		
				A	A	C	U								C	C	U		C	U	A	C	C	U	C		
	mRNA	3’						U	A	U	G	U	U					U								U	5’
d.	target	5’	A	A	A	C	U	C	A	G	A	G	G	U	C	C	U	C	C	U	A	C	C	U	C	U	3’
				|	|	|	|								|	|	|		|	|	|	|	|	|	|		
	mRNA	3’		1	1	1	1	0	0	0	0	0	0		1	1	1	0	1	1	1	1	1	1	1	0	
e.	target	5’	A	A	A	C	U								C	C	U	C	C	U	A	C	C	U	C	U	3’
				|	|	|	|								|	|	|		|	|	|	|	|	|	|		
	mRNA	3’		1	1	1	1	0	0	0	0	0	0		1	1	1	0	1	1	1	1	1	1	1	0	

To highlight the miRNA base-pairing, three kinds of information can be discarded. 1) Target base alphabets. Although the target base alphabets are a necessary component in the miRNA-target binding structure, eliminating them would be helpful to highlight the miRNA binding pattern characteristics. For example, (a) and (b) have different target base alphabets, but they shared common miRNA-binding pattern: miRNA 3’ 1111000000111011111110 5’; 2) base alphabets of miRNA. We substitute 1 or 0 for nucleotide names (A, U, G, C) along with miRNA sites corresponding to their pairing or not. The base-pairing status can be highlighted disregarding the detailed nucleotide. For example, (a) and (c) have different miRNA nucleotides, but they shared a common miRNA-binding pattern: miRNA 3’ 1111000000111011111110 5’; 3) internal loop and bulge loop in miRNA-target secondary structure. Given a 0-1 sequence, such as miRNA 3’ 1111000000111011111110 5’, it is uncertain whether the subsequence 000000 comes from an internal loop of (d) or a bulge loop of (e) or other structures. That is to say, we cannot discriminate the internal loop and bulge loop corresponding to subsequences of consecutive 0’s. Actually, the discrimination of the internal loop and bulge loop just helps in the calculation of miRNA-target free energy. With respect to the base-pairing binding patterns, the internal loop and bulge loop basically show little difference. Therefore, the simplified representation we adopted (i.e., 0-1 sequences) keeps the focus on the base-pairing complementary pattern of miRNA along with its sites.

probability. Since the 0-1 sequence stemming from biological molecular sequence (i.e., the miRNA sequence), we assume that the 0-1 sequence has a first-order Markov property. Besides, since different sites have significantly different properties with respect to base-pairing binding statistical characteristics, a non-homogeneous rather than homogeneous Markov model is adopted.

Specifically, we assign two states for the *i*th base site of 22-nucleotide-length miRNA from 5’ to 3’, state 1 and state 0. Either state can be viewed as a value of a variable s_i. State

1 stands for forming Watson-Crick pairing, while the state 0 stands for unforming pairing. From the state s_i, either the next state 1 or 0 can be transferred with probability $p^i_{s_i,0}$ or $p^i_{s_i,1}$, respectively, as shown in Figure 1. The 0-1 sequence $s_1\ s_2\ s_3\ \ldots s_{22}$ is the input of the model. $p^i_{s_i,\ s_{i+1}}$ *(i = 1...21)* is the model parameter, standing for probability of the *i*th base site with the state s_i transferring to the *i + 1*th base site with the state s_{i+1}. Therefore, the probability of the 0-1 sequence $s_1\ s_2\ s_3\ \ldots s_{22}$, such as, 01111111011100000001111 (from 5' to 3' end of miRNA), being generated by the above model is

$$P_0^{\ 1} * P_{0,1}^{\ \ 1} * P_{1,1}^{\ \ 2} * \ldots * P_{1,1}^{\ \ 7} * P_{1,0}^{\ \ 8} * P_{0,1}^{\ \ 9} * \ldots * P_{1,1}^{\ \ 21} \qquad \textbf{(Equation 1)}$$

where $P_0^{\ 1}$ is the probability of the state 0 occurring at the 1st base site and is one of the other kind of model parameter $P_{s_i}^{\ i}$ *(I = 1...22)*. It stands for the probability of the *i*th base site being with the state s_i.

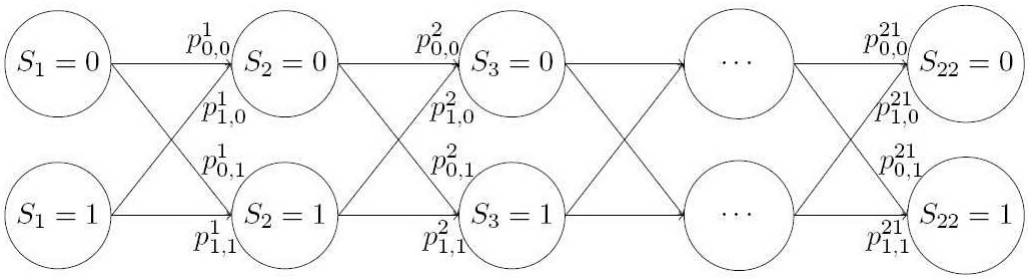

Figure 1. First-order non-homogeneous Markov chain model.

The model parameters are not only used for calculation of the probability of the input being generated by the model, but also contribute to reflecting characteristics of base-pairing binding patterns. Specifically, the higher the $p^{i-1}_{s_{i-1},s_i=1}$ (i.e., $p^{i-1}_{1,1}$ and $p^{i-1}_{0,1}$) and p^i_1 are, the more probable it is that the *i*th base site forms Watson-Crick pairing. Moreover, $p^i_{1,1}$ and $p^i_{0,0}$ reflect consecutive formed or unformed base-pairing transfers whereas $p^i_{1,0}$ and $p^i_{0,1}$ reflect non-consecutive ones.

All of the model parameters can be estimated by maximum likelihood estimation method.

$$\begin{gathered} p^i_{0,0} = \mathrm{n}^i_{0,0} / \mathrm{n}^i_0\ ,\ p^i_{0,1} = \mathrm{n}^i_{0,1} / \mathrm{n}^i_0, \\ p^i_{1,0} = \mathrm{n}^i_{1,0} / \mathrm{n}^i_1\ ,\ p^i_{1,1} = \mathrm{n}^i_{1,1} / \mathrm{n}^i_1 \\ p^i_0 = \mathrm{n}^i_0 / \mathrm{n},\ p^i_1 = \mathrm{n}^i_1 / \mathrm{n} \end{gathered} \qquad \textbf{(Equation 2)}$$

Given training data sets, $n^i_{s_i, s_{i+1}}$ is the total number of the 0-1 sequences with the *i*th state being s_i and the *i+1*th state being s_{i+1} *(i = 1...21, s_i = 0,1)*; $n^i_{s_i}$ is the total number of the 0-1 sequences with the *i*th state being s_i *(i = 1...21)* and *n* is the number of all 0-1 sequences.

RESULTS

Estimation of model parameters based on known miRNA targets

Using 128 0-1 sequences, corresponding to 128 known human miRNA-target base-pairing bindings structures downloaded from TarBase (Sethupathy et al., 2006; Supplementary Table 1), the model parameters $p^i_{s_i, s_{i+1}}$ and $p^i_{s_i}$ were estimated. Their values are shown in Figures 2 and 3.

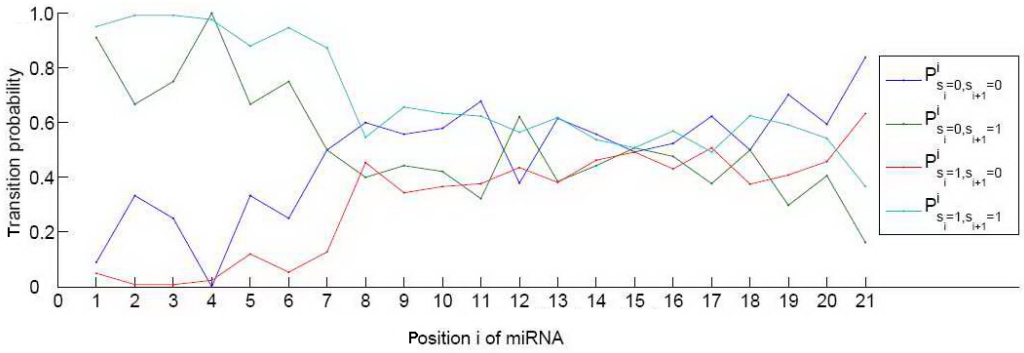

Figure 2. The learned model parameters of transition probability $p^i_{s_i s_{i+1}}$ from *i* = 1 to *i* = 21.

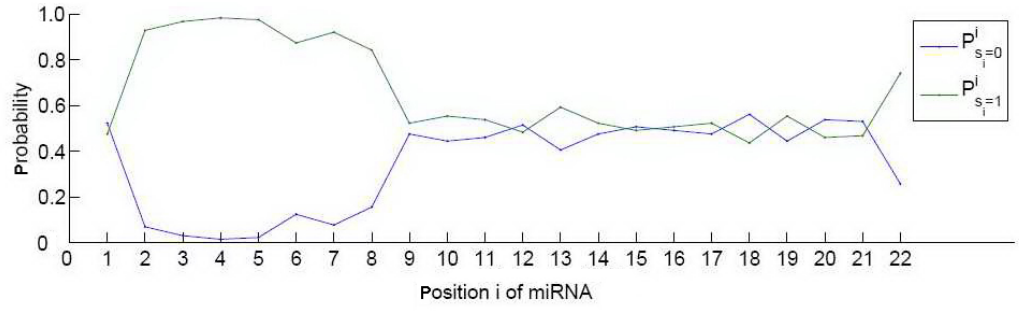

Figure 3. The learned model parameters of probability $p^i_{s_i}$ from *i* = 1 to *i* = 22.

The characteristics of base-pairing binding patterns reflected by these parameters are also illustrated. See Figure 4 for a further description.

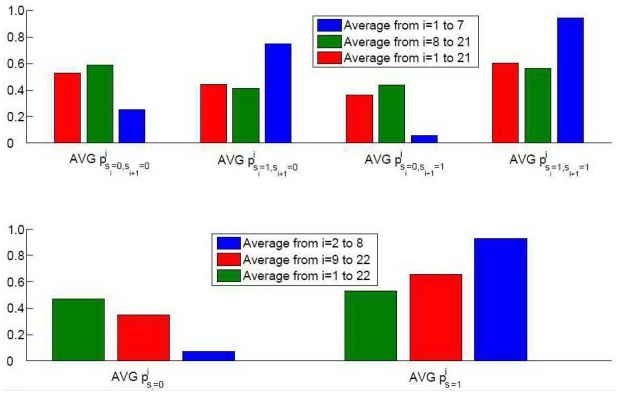

Figure 4. The average values of the model parameters. The upper graph shows the average (AVG) values of transition probabilities AVG ($p^i_{s_i s_{i+1}}$) and the lower graph shows the average values of probabilities AVG ($p^i_{s_i}$). The figure shows that the probabilities of forming Watson-Crick pairing in the 5' end of miRNA sites are higher than those in the rest of miRNA sites. These kinds of higher probabilties are obvious since the probabilities $p^i_{s_i=1}$, $p^i_{s_i=1, s_{i+1}=1}$ and $p^i_{s_i=0, s_{i+1}=1}$ are higher when i = 1 to 7 than those when i = 8 to 21. In other words, the average values of these probabilities in the 5' end of miRNA sites, AVG ($p^i_{s_i=1}$) = 0.93, AVG ($p^i_{s_i=1, s_{i+1}=1}$) = 0.94 and AVG ($p^i_{s_i=0, s_{i+1}=1}$) = 0.75 (sites from i = 1 to i = 7) are significantly higher than AVG ($p^i_{s_i=1}$) = 0.53, AVG ($p^i_{s_i=1, s_{i+1}=1}$) = 0.56 and AVG ($p^i_{s_i=0, s_{i+1}=1}$) = 0.41 (sites from i = 8 to 21), respectively. On the other hand, the tendency toward consecutiveness can be shown by comparing $p^i_{1,1}$ and $p^i_{0,0}$ with $p^i_{1,0}$ and $p^i_{0,1}$. AVG ($p^i_{s_i=0, s_{i+1}=0}$) = 0.53 and AVG ($p^i_{s_i=1, s_{i+1}=1}$) = 0.60 are both significantly higher than the average ones of which sites have the different states with the next sites, that is to say, AVG ($p^i_{s_i=0, s_{i+1}=1}$) = 0.44 and AVG ($p^i_{s_i=1, s_{i+1}=0}$) = 0.36 in the whole range. It shows that the consecutive Watson-Crick pairing in the whole range is more probable than the non-consecutive one.

Owing to the asymmetry of the 5' end and the 3' end, the probability parameters of the different ends have different significance. Figure 5 indicates that the probability parameters of base sites at the 5' ends of miRNAs are more efficient for recognizing the known miRNA targets. See Figure 5 for a further description.

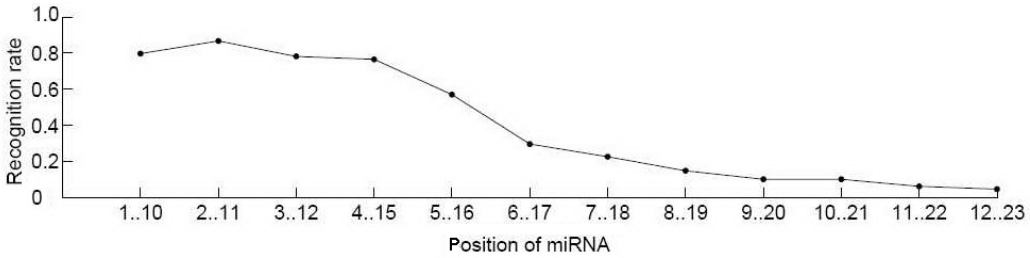

Figure 5. The recognition rates with the same 5% false-positive rate at different miRNA site region. The 10-length 0-1 subsequences have 2^{10} = 1024 different permutations. Each one of them has 1/1024 odds to be selected under entirely random condition. On the other hand, each 0-1 subsequence has a probability of being generated by the Markov model. Therefore, when we select the top 5% of all the permutations ordered by the probability values, the false-positive rate is 5%. Each 0-1 subsequence in the top 5% of all the permutations was a pattern identified by the Markov model and all the identified patterns made up an identified pattern set. A 0-1 subsequence of the known miRNA target was recognized by the Markov model when it was in the identified pattern set. The recognition curve shows that when the 0-1 subsequences of base sites from 2 to 11 were adopted, the recognition rate achieved a maximum value of 87%.

Recognizing 86% of the known targets with only 5% false positive

As mentioned above, considering the asymmetry of 5' end and 3' end of miRNA, we do not have to use all transition probabilities from site 1 to 21. As shown in Figure 5, using the probability parameters from site 2 to 11 (simply denoted as 2..11) achieved the maximal recognition rates of 87%.

There are troubling 10-nucleotide-length 0-1 subsequences, such as A and B in Table 3, which are in the identified pattern set, but are not consistent with any of the known miRNA target's 0-1 subsequence. In other words, if a candidate potential miRNA target has the 0-1 subsequence of A or B, it would be identified by the Markov model although it is not consistence with any of the known miRNA targets. This kind of false identifying was eliminated when the 0-1 subsequences were restricted within base sites from 2 to 8, with almost the same recognition rate as within sites from 2 to 11. Table 4 shows that when the 0-1 subsequences was restricted from 2 to 8, the identified pattern set contained altogether six patterns and all six patterns were included in the known miRNA targets' pattern set.

Table 3. Some special 0-1 subsequence patterns.

	5'	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	3'
A		*	1	1	1	1	1	0	0	1	0	1	*	*	*	*	*	*	*	*	*	*	*	
B		*	1	1	1	1	1	0	1	0	0	1	*	*	*	*	*	*	*	*	*	*	*	
C		*	1	1	1	1	1	1	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	
D		1	1	1	1	1	1	1	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	

Table 4. Numbers of distinct high-scored patterns not included in the known target pairing pattern of the 5' end of miRNA.

	Number of distinct known target patterns with score higher than cutoff	Number of all permutation patterns with score higher than cutoff	Number of distinct high scored patterns not included in the known target pairing patterns	Recognition rate with 5% false positive
2..11	20	51	51-20 = 31	87%
2..10	13	26	26-13 = 13	86%
2..9	10	12	12-10 = 2	84%
2..8	6	6	6-6 = 0	86%

At last, we adopt probability parameters from site 2 to 8 of miRNAs. The probability of the 0-1 subsequence 5' $s_2s_3s_4s_5s_6s_7s_8$ 3' produced by our generative Markov chain model is

$$P(s_2s_3s_4s_5s_6s_7s_8) = p^2_{s_2} * p^2_{s_2,s_3} * p^3_{s_3,s_4} * p^4_{s_4,s_5} * p^5_{s_5,s_6} * p^6_{s_6,s_7} * p^7_{s_7,s_8} \quad \textbf{(Equation 3)}$$

Next, we defined the score of the 0-1 subsequence as the log value of the probability. Thus, the cutoff score was -1.68 when we restricted the false-positive rate at 5%. With a false rate of 5, 86% of the known miRNA targets were recognized. The six 0-1 subsequences altogether in

the identified pattern set are listed in Table 5. Among the 128 known miRNA targets, 110 targets of them (86% targets of 128 targets) were identified by our model, such as let-7b and its target Lin28 (Kiriakidou et al., 2004) and miR-375 and its target Mtpn (Poy et al., 2004). These two targets have consecutive base-pairing from site 2 to 8 of the miRNAs. Whereas binding between miR-23 and its target HES1 (Kawasaki and Taira, 2005) has nonconsecutive base-paring from site 2 to 8 of the miRNA, but their base-pairing binding pattern is identical to one of the six patterns listed in Table 5 and was identified by our model. Binding between miR-181 and its target Tcl1(Pekarsky et al., 2006) has unpaired base at site 2 and 6 of the miRNA and is one of the 18 targets, which were not identified by our model since the score of its base-pairing binding pattern in our model was -3.23, which was lower than the cutoff score -1.68. See Supplementary Table 1 for the detailed information of all known miRNA targets identified by our model.

Table 5. The six kinds of patterns learned from the training data sets.

	5'	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	3'	Score
A		*	1	1	1	1	1	1	1	*	*	*	*	*	*	*	*	*	*	*	*	*	*		-0.19
B		*	1	1	1	1	1	1	0	*	*	*	*	*	*	*	*	*	*	*	*	*	*		-1.02
C		*	1	1	1	1	0	0	1	*	*	*	*	*	*	*	*	*	*	*	*	*	*		-1.40
D		*	1	1	1	1	0	0	0	*	*	*	*	*	*	*	*	*	*	*	*	*	*		-1.40
E		*	1	1	1	1	0	1	1	*	*	*	*	*	*	*	*	*	*	*	*	*	*		-1.63
F		*	1	1	1	1	1	0	1	*	*	*	*	*	*	*	*	*	*	*	*	*	*		-1.68

The oldest miRNA target prediction program TargetScan uses sites from 2 to 8 in the 5' end as seed. That is to say, all predicted targets have perfect base pairing from 2 to 8 sites and its corresponding 0-1 sequence is shown in Table 5, A. With our model, the pattern represented by this 0-1 subsequence is scored the highest. Furthermore, the relaxed seed in the upgraded version TargetScanS requires perfect base pairing from 2 to 7 sites in the 5' end and its corresponding 0-1 subsequences is shown in Table 3, C. Note that its corresponding 0-1 subsequences from 2 to 8 sites are as shown in Table 5, A and B. With our model, these two patterns score the top two in all six patterns.

Both seeds of TargetScan and the upgraded version TargetScanS are identified by our model. This fact verified the ability of our model. Moreover, our model does not require perfect consecutive base-pairing sites any more. As we have shown, there are altogether six different combinational patterns that the binding between the miRNAs and their targets may follow. The diversity reflected by our model is more consistent with base-pairing binding characteristics of the known miRNA targets.

Assessing potential targets predicted by computational programs

TargetScanS, miRanda, PicTar, and DIANA-microT are the four well-known computional programs designed for predicting miRNA targets. As mentioned above, TargetScanS predicted targets with perfect base-pairing binding from 2 to 7 sites, corresponding to two 0-1 subsequence as shown in Table 5, A and B. That means, 100% of the potential targets predicted by TargetScanS can be identified by our model with the highest score. Furthermore, almost all potential targets predicted by PicTar have "perfect nuclei" 7-nucleotide-length perfect base-

pairing sites as shown in Table 3, D. Corresponding to sites from 2 to 8, the patterns are also those listed in Table 5, A and B. "Perfect nuclei" also refer to 7-nucleotide-length perfect base-pairing sites as shown in Table 5, A. Apparently, all these patterns were identified by our model. In fact, not only "perfect nuclei" but also "imperfect nuclei" can be identified by our model with scoring higher than the cutoff value.

Although the targets predicted by DIANA-microT and miRanda passed through our model with 92 and 35% recognition rate, respectively, rather than 100%, these data just show that all targets predicted by these two programs are not consistent with the currently known target binding patterns. The recognition rates of the four computational programs identified by our model are listed in Table 6.

Table 6. Percentages of predicted miRNA targets identified by our model in the four computational programs.

	Number of predicted targets	Number of predicted targets identified by our model	Percentage of predicted targets identified by our model
TargetScanS	15,000	15,000	100%
PicTar	66,484	66,484	100%
DIANA-microT	94	87	92%
miRanda	15,433	5,433	35%

As we all know, the common potential targets predicted by two different computational programs will have higher probabilities of being true targets than by only one program. Our model highlighted this kind of higher probabilities by the fact that 70% (473 of 671) of common potential targets predicted by both MiRanda and PicTar were identified by the model, while 35% (5,433 of 15,433) of targets were predicted by only one program, miRanda (as shown in Table 7). This shows that our model was an efficient method in evaluating the extent to which a potential target was accurately predicted. The common 473 identified targets predicted by both MiRanda and PicTar are listed in Supplementary Table 2.

Table 7. Identification rate of common targets predicted by two computational programs.

	Total number	Targets identified by our model	Identification rate
Predicted by miRanda	15,433	5,433	5,433/15,433 = 35%
Predicted by both miRanda and PicTar	671	473[1]	473/671 = 70%

[1]Note: each of the common 671 targets predicted by two computational programs might have different predicted binding patterns due to different programs. Therefore, 473 of 671 binding sites predicted by miRanda are identified by our model while 100% of 671 binding patterns predicted by PicTar are identified by our model.

Known miRNA target library

TarBase (Sethupathy et al., 2006) is a database of experimental supported miRNA targets in 8 species. We queried the database on all miRNA targets that satisfied such conditions: 1) belonging to human species; 2) being true miRNA targets as determined by experimental testing, and 3) inhibition by translational repression mechanism. All targets are listed in the Supplementary Table 1.

Predicted potential targets by computational programs

Predicted miRNA targets using DIANA-microT include targets for 10 human miRNAs and we downloaded the "Supplementary Table 1" from the paper by Kiriakidou et al. (2004), which contains these predicted targets from http://www.genesdev.org/cgi/content/full/1184704/DC1.

Predicted human miRNA targets using miRanda include genome-wide targets conserved cross mammals and we downloaded "Table S3" from the paper by John et al. (2004), which contains genes targeted by miRNA from http://biology.plosjournals.org/perlserv/?request=get-document&doi=10.1371/journal.pbio.0020363#JOURNAL-PBIO-0020363-ST002.

Predicted miRNA targets using PicTar include genome-wide targets up to eight vertebrate species and we downloaded the "Supplementary Table 3" from the paper by Krek et al. (2005), which contains targets conserved in human, chimpanzee, mouse, rat, and dog for a set of 168 microRNAs from http://www.nature.com/ng/journal/v37/n5/abs/ng1536.html.

Predicted miRNA targets using TargetScanS include genome-wide targets conserved cross five vertebrate species and we downloaded the "Supplementary Table S2" from the paper by Lewis et al. (2005), which contains targets conserved in human, mouse, rat, dog, and chicken from http://www.cell.com/cgi/content/full/120/1/15/DC1/.

DISCUSSION

When there are already so many miRNA target prediction computational programs and a large number of potential targets to be verified by experiments, a model that can limit the set of predicted targets to a small one, which is more consistent with known targets, is a valuable tool.

The known miRNA targets provided valuable information on miRNA-target interaction patterns. In this paper, a first-order non-homogeneous Markov chain was devised for representing the characteristics of the base-pairing complementarity binding patterns between miRNAs and their targets and the model parameters were estimated based on the known miRNA targets.

In order to identify microRNA target sites, Rajewsky and Socci (2004) proposed a computational method with which GC and AU relative weights were learned from known miRNA targets. MicroRNA GC content was further studied by Davis et al. (2008). Compared to GC content, miRNAs and their targets base-pairing binding pattern characteristics represent more critical rules to be followed in the process of identifying miRNA targets.

We developed the idea of learning from known miRNA targets to construct base-pairing binding patterns and learned the patterns based on the first-order non-homogeneous Markov chain model. The more known miRNA targets are validated, the more accurately the patterns learned from our model represent the true characteristics of miRNA-target binding.

The fact that our model identified a greater percentage of common targets predicted by two computational programs than by only one program shows that our model is efficient in identifying targets consistent with known miRNA targets.

Extracting a large amount of statistical data, such as structural features, thermodynamic features and position-based features, miTarget (Kim et al., 2006) showed a support vector machine-learning classifier method to predict miRNA targets. Compared to this method, we strongly emphasized on the miRNA-targets base-pairing binding pattern characteristics,

which are the most critical feature among all kinds of features in the process of miRNA targets prediction. Using the only one feature, we successfully identified potential targets that are consistent with the training data sets with respect to miRNA-targets base-pairing binding pattern. In the future research, we will develop our model to add more features to be considered as well as definitely discriminating contributions of different features in the model.

Among the several rules for identifying miRNA targets, the critical one is the rule of limiting base-pairing binding pattern. Most of the previous computational prediction programs first searched in the whole genome for the targets satisfying certain base-pairing binding pattern rules between miRNA and its targets and then applied other rules in this relative small searched target set than the whole genome. This shows that well-designed base-pairing binding rules are strongly necessary for miRNA target computational programs. However, the base-pairing binding pattern rules in all of the present computational programs require fixed patterns, rather than flexible ones determined by known miRNA target patterns. Our method constructed these flexible base-pairing rules using the Markov chain model.

Our method can be used to perform a second search after having a candidate miRNA target set predicted by any available computational program. The second search results contained targets, which were more consistent with the known miRNA target base-pairing patterns. For those researchers who are accustomed to certain miRNA target prediction computational programs, our method is a convenient and efficient tool to obtain a more refined candidate target set.

REFERENCES

Brennecke J, Hipfner DR, Stark A, Russell RB, et al. (2003). Bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene hid in *Drosophila*. *Cell* 113: 25-36.

Brown JR and Sanseau P (2005). A computational view of microRNAs and their targets. *Drug Discov. Today* 10: 595-601.

Burgler C and MacDonald PM (2005). Prediction and verification of microRNA targets by MovingTargets, a highly adaptable prediction method. *BMC Genomics* 6: 88.

Davis N, Biddlecom N, Hecht D and Fogel GB (2008). On the relationship between GC content and the number of predicted microRNA binding sites by MicroInspector. *Comput. Biol. Chem.* 32: 222-226.

Enright AJ, John B, Gaul U, Tuschl T, et al. (2003). MicroRNA targets in *Drosophila*. *Genome Biol.* 5: R1.

Grun 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.

John B, Enright AJ, Aravin A, Tuschl T, et al. (2004). Human microRNA targets. *PLoS Biol.* 2: e363.

Kawasaki H and Taira K (2005). Identification of target genes of microRNAs in retinoic acid-induced neuronal differentiation. *Pure Appl. Chem.* 77: 313-318.

Kim SK, Nam JW, Rhee JK, Lee WJ, et al. (2006). miTarget: microRNA target gene prediction using a support vector machine. *BMC Bioinformatics* 7: 411.

Kiriakidou M, Nelson PT, Kouranov A, Fitziev P, et al. (2004). A combined computational-experimental approach predicts human microRNA targets. *Genes Dev.* 18: 1165-1178.

Krek A, Grun D, Poy MN, Wolf R, et al. (2005). Combinatorial microRNA target predictions. *Nat. Genet.* 37: 495-500.

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.

Pekarsky Y, Santanam U, Cimmino A, Palamarchuk A, et al. (2006). Tcl1 expression in chronic lymphocytic leukemia is regulated by miR-29 and miR-181. *Cancer Res.* 66: 11590-11593.

Poy MN, Eliasson L, Krutzfeldt J, Kuwajima S, et al. (2004). A pancreatic islet-specific microRNA regulates insulin secretion. *Nature* 432: 226-230.

Rajewsky N and Socci ND (2004). Computational identification of microRNA targets. *Dev. Biol.* 267: 529-535.

Rehmsmeier M, Steffen P, Hochsmann M and Giegerich R (2004). Fast and effective prediction of microRNA/target duplexes. *RNA* 10: 1507-1517.

Robins H, Li Y and Padgett RW (2005). Incorporating structure to predict microRNA targets. *Proc. Natl. Acad. Sci. U. S. A.* 102: 4006-4009.

Rusinov V, Baev V, Minkov IN and Tabler M (2005). Micro-Inspector: a web tool for detection of miRNA binding sites in an RNA sequence. *Nucleic Acids Res.* 33: W696-W700.

Saetrom O, Snove O Jr and Saetrom P (2005). Weighted sequence motifs as an improved seeding step in microRNA target prediction algorithms. *RNA* 11: 995-1003.

Sethupathy P, Corda B and Hatzigeorgiou AG (2006). TarBase: a comprehensive database of experimentally supported animal microRNA targets. *RNA* 12: 192-197.

Stark A, Brennecke J, Russell RB and Cohen SM (2003). Identification of *Drosophila* microRNA targets. *PLoS Biol.* 1: E60.

Xu P, Vernooy SY, Guo M and Hay BA (2003). The *Drosophila* microRNA Mir-14 suppresses cell death and is required for normal fat metabolism. *Curr. Biol.* 13: 790-795.

Zhang B, Pan X, Wang Q, Cobb GP, et al. (2006). Computational identification of microRNAs and their targets. *Comput. Biol. Chem.* 30: 395-407.

