Front. Biol. 2010, 5(2): 171–179
DOI 10.1007/s11515-010-0009-6

RESEARCH ARTICLE

MicroRNA target prediction based on second-order Hidden Markov Model

Song GAO[1], Liangsheng ZHANG[1,2], Diangang QIN[1], Tienan FENG[1], Yifei WANG (✉)[1]

1 Department of Mathematics, School of Sciences, Shanghai University, Shanghai 200444, China
2 School of Life Sciences, Institute of Plant Biology, Fudan University, Shanghai 200433, China

Abstract MicroRNAs are one class of small single-stranded RNA of about 22 nt serving as important negative gene regulators. In animals, miRNAs mainly repress protein translation by binding itself to the 3' UTR regions of mRNAs with imperfect complementary pairing. Although bioinformatics investigations have resulted in a number of target prediction tools, all of these have a common shortcoming—a high false positive rate. Therefore, it is important to further filter the predicted targets. In this paper, based on miRNA:target duplex, we construct a second-order Hidden Markov Model, implement Baum-Welch training algorithm and apply this model to further process predicted targets. The model trains the classifier by 244 positive and 49 negative miRNA:target interaction pairs and achieves a sensitivity of 72.54%, specificity of 55.10% and accuracy of 69.62% by 10-fold cross-validation experiments. In order to further verify the applicability of the algorithm, previously collected datasets, including 195 positive and 38 negative, are chosen to test it, with consistent results. We believe that our method will provide some guidance for experimental biologists, especially in choosing miRNA targets for validation.

Keywords microRNA, target gene, experimentally supported targets, second-order Hidden Markov Model, forward algorithm

1 Introduction

MicroRNAs (miRNAs) are one class of small single-stranded RNA molecules of about 22 nt that are endogenous and non-coding RNA serving as important negative gene regulators. They mainly form RNA-induced Silencing Complex (RISC) with related protein to repress translation in animals (Barciszewski and Erdmann, 2008; Rossi et al., 2008). Previous studies have shown that miRNAs play important roles in multiple biological and metabolic processes, including developmental timing, signal transduction, cell maintenance, cell differentiation and so on. MiRNAs regulate gene expression at the post-transcriptional level by directly cleaving targeted mRNA or repressing translation (Zhang et al., 2006). Aberrant or absent expression of miRNA is often closely associated with diseases. For instance, some miRNAs may play a "cancer gene" or "tumor suppressor gene" role in tumor and cancer (Rossi et al., 2008).

MiRNAs bind to the 3' UTR of mRNA by partial or near-precise pairing to form duplex structures and regulate mRNA's translation. At present, miRBase (Griffiths-jones et al., 2006) has collected more than 9500 mature miRNAs in 103 species, including animals, plants, and viruses (miRBase Release 13.0, June 2009). The miRNA's function is heavily dependent on the formation of specific duplex structure with its target genes by two modes. In the first mode, miRNA bind to target 3' UTR of mRNA by precise or near-precise pairing, leading to direct mRNA cleavage and degradation through a mechanism involving the RNA interference (RNAi) machinery. In the other mode, the pairing of a miRNA to its target is usually less perfect in animals, but still affects the mRNA's stability, regulating the target by translational repression (Yang et al., 2008). Due to the lack of sensitive miRNA cloning methods and high-throughput experimental methods for identification of miRNA target genes, many bioinformatics methods have been developed to achieve a large number of target genes in prediction, such as miRanda, Pictar, Targetscan (Enright et al., 2003; Lewis et al., 2003, 2005; John et al., 2004; Kiriakidou et al., 2004; Rehmsmeier et al., 2004; Krek et al., 2005; Rusinov et al., 2005; Saetrom et al., 2005; Huynh et al., 2006; Kim et al., 2006; Thadanil et al., 2006; Yousef et al., 2007). Work in computational prediction of miRNA targets revealed that

Received November 19, 2009; accepted December 18, 2009

E-mail: yifei_wang@staff.shu.edu.cn

each human miRNA could potentially target hundreds of genes and at least 30% of the human genes could be targeted by miRNAs (John et al., 2004; Krek et al., 2005; Lewis et al., 2005; Yang et al., 2008).

MiRNA:target duplexes have certain disciplines which can be summarized from experimentally supported miRNA:target duplexes. Although computational prediction methods of miRNA target genes are different from each other, they are mainly based on five major characteristic properties: (1) miRNAs have good complementarity to their target mRNAs; (2) the miRNA:target duplex has a higher negative folding free energy; (3) the capacity of binding to the target genes at 5' end of miRNA is stronger than the 3' end; (4) the miRNA:target duplex structure should not contain complex secondary structure; (5) miRNA:mRNA interactions are highly conserved from species to species, particularly within the same kingdom. In addition to the five basic principles above, different prediction methods will limit and optimize their algorithms in accordance with their respective laws (Zhang et al., 2006; Xia et al., 2009). Kiriakidou et al. (2004) compared some published methods for mammalian miRNA targets prediction and found that the overlap of identical predictions from the different computational approaches varied between 10% and 50% for a common set of 79 miRNA. This indicates that false positive predictions could account for a large percentage of all the predicted miRNA target genes. Therefore, it is very necessary to filter the prediction targets by a post-processing step (Yang et al., 2008).

In this paper, we present a machine-learning algorithm based on second-order Hidden Markov Model (HMM2) that can be used as a post-processing method for filtering the predicted targets by other prediction programs, such as miRanda, Pictar and Targetscan. This method not only reduces the number of prediction targets, but also decreases the false positive rate. The prediction algorithm is trained with the experimentally supported animal miRNA targets found in TarBase (Sethupathy et al., 2006), containing 244 positive and 49 negative miRNA: target interaction pairs. Each miRNA:target duplex is mapped into states transition and symbols emission of Hidden Markov Model (HMM). Three states are defined, including match, mismatch, and insertion. Under each state, there are five symbols emission such as A, U, C, G, and – (gap). Here we adopt the 10-fold cross-validation method and use second-order Hidden Markov Model to train the optimal state transition matrix and the symbol emission matrix, achieving a sensitivity of 72.54% and specificity of 55.10%. In order to further validate the applicability of our algorithm, we also use 195 positive and 38 negative data collected by Yang (Yang et al., 2008) to test the model by the 10-fold cross-validation method, obtaining a sensitivity of 83.08% and specificity of 60.53%, which is consistent with Yang's (Yang et al., 2008).

2 Materials and methods

2.1 Data

The experimentally supported miRNA:target interactions are downloaded from the TarBase version 5.0 (Sethupathy et al., 2006), including translationally repressed targets and cleaved ones. Only miRNA:target interactions with reported binding duplexes are extracted from worm, fruit fly, zebrafish, rat, mouse, and human. By removing duplicated entries as well as entries with incomplete binding diagrams, a result containing 244 positive and 22 negative miRNA:target duplex sequences is obtained. As the number of negative data is small and highly different from positive data, we go on searching for the target sites of those known negative interactions without duplexes being reported. Therefore, we can increase 17 negative data from Yang's paper predicted by PicTar and RNAhybrid software (Yang et al., 2008). In addition, we also add 10 negative data that have been validated negative interactions but without miRNA:target duplexes from six papers (Landais et al., 2007; Skalsky et al., 2007; Duursma et al., 2008; Hébert et al., 2008; Luo et al., 2008; Sengupta et al., 2008). Five of the experimentally validated negative interactions are from Hébert's (Hébert et al., 2008) and Sengupta's (Sengupta et al., 2008) paper and duplexes were predicted by Pictar. The other five negative data are from Luo's (Luo et al., 2008), Duursma's (Duursma et al., 2008), Landais's (Landais et al., 2007), and Skalsky's (Skalsky et al., 2007) papers, using RNA22 (Huynh et al., 2006) to predict the binding sites. At last, there are a total of 49 negative samples. All the targets of 244 positive and 49 negative data come from the prediction of the miRNA listed in Table 1.

2.2 HMM construction

The miRNA:target interaction is a pair of matching sequences. We define the pair sequences beginning at 3' end and terminating at 5' end of the miRNA (Fig. 1a). Taking miRNA sequences as the reference, the miRNA: target duplex is divided into the seed and non-seed area. The seed area are the 8 bases in the front of miRNA 5' end, needing at least six base pairings and including G = U base pairing. In non-seed area, HMM is constructed and hidden states are defined as (1) match state (AU/UA/CG/ GC/GU/ UG), (2) mismatch state (AC/CA/AG/GA/UC/CU/A-/U-/ C-/G-), and (3) insertion state (-A/-T/-C/-G). The possible transitions between the three states are shown in Fig.1c. Under the definition of hidden states, corresponding target sequences are regarded as symbol emission sequences and contain five possible symbols, such as A, U, C, G, -(gap), in each state (Fig.1a and b). Thus, an HMM is constructed containing the necessary hidden state and symbol emission.

Table 1 List of miRNAs appeared in the training sets

	miRNA	species
positive set	let-7, lin-4, lsy-6, miR-273, miR-61, miR-84	worm
	bantam, let-7, miR-1, miR-11, miR-2, miR-278, miR-2a-1, miR-4, miR-7, miR-79	fruit fly
	let-7	zebrafish
	miR-9	rat
	miR-125b, miR-134, miR-181a, miR-206, miR-7b	mouse
	let-7, let-7b, let-7e, miR-1, miR-101, miR-103-1, miR-10a, miR-122, miR-125a, miR-125b, miR-127, miR-130, miR-132, miR-133, miR-136, miR-140, miR-141, miR-143, miR-145, miR-15, miR-155, miR-16, miR-17-5p, miR-181, miR-189, miR-192, miR-196, miR-199b, miR-19a, miR-1b, miR-20, miR-206, miR-21,miR-221, miR-222, miR-223, miR-23, miR-23a, miR-24, miR-26, miR-27b, miR-29, miR-32, miR-34, miR-375, miR-376a-5p, HSVl-miR-LAT, Edited-miR-376a-5p, miR-431, miR-433-3p, miR-433-5p, miR-434-3p, miR-434-5p, miR-98	human
negative set	let-7	worm
	mir-276b, mir-278, mir-286, mir-287, mir-288, mir-303, mir-316, mir-317, mir-318	fruit fly
	mir-124, miR-34, mir-375	mouse
	let-7b, let-7e, miR-15, miR-16, miR-24, miR-103, miR-141, miR-145, miR-1, miR-19a, miR-34, miR-133, miR124a, miR-148, miR15a, miR-19b, miR-29c, miR-92-2, miR-155	human

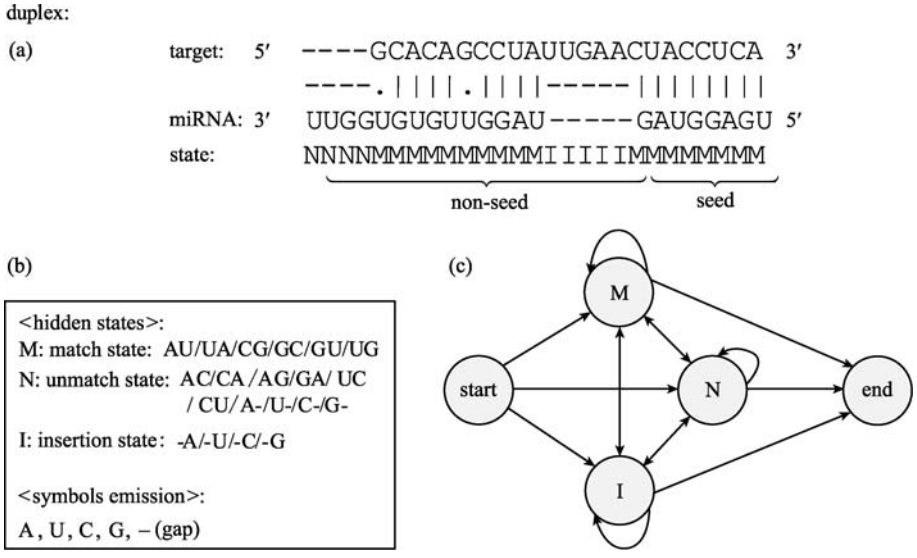

Fig. 1 MiRNA:target duplex and the definition of Hidden Markov Model (HMM). (a) MiRNA:target duplex and hidden states. (b) The definition of hidden states and symbols emission in HMM. (c) The three-hidden-state finite state automaton.

2.3 Implementation of HMM2 algorithm

Du and Shi (2001, 2004, 2007) have formulated and proved HMM2's Baum-Welch reevaluation algorithm of multiple observations in theory, and the computational formulas are as follows:

$$\overline{\pi}_i = \frac{\sum_{w=1}^{W} c_w P(O^{(w)}|\lambda) \sum_{j=1}^{N} \sum_{k=1}^{N} \xi_2^{(w)}(i,j,k)}{\sum_{w=1}^{W} c_w P(O^{(w)}|\lambda)} \quad 1 \leqslant i \leqslant N \tag{2.3.1}$$

$$\overline{a}_{ij} = \frac{\sum_{w=1}^{W} c_w P(O^{(w)}|\lambda) \sum_{k=1}^{N} \xi_2^{(w)}(i,j,k)}{\sum_{w=1}^{W} c_w P(O^{(w)}|\lambda) \sum_{j=1}^{N} \sum_{k=1}^{N} \xi_2^{(w)}(i,j,k)} \quad 1 \leqslant i,j \leqslant N \tag{2.3.2}$$

$$\overline{a}_{ijk} = \frac{\sum_{w=1}^{W} c_w P(O^{(w)}|\lambda) \sum_{t=2}^{T_w-1} \xi_t^{(w)}(i,j,k)}{\sum_{w=1}^{W} c_w P(O^{(w)}|\lambda) \sum_{t=2}^{T_w-1} \sum_{k=1}^{N} \xi_t^{(w)}(i,j,k)} \quad 1 \leqslant i,j,k \leqslant N \tag{2.3.3}$$

$$\overline{b}_i(l)=\frac{\sum_{w=1}^{W}c_w P(O^{(w)}|\lambda)\sum_{t=2}^{T_w-1}\sum_{j=1}^{N}\sum_{k=1}^{N}\xi_t^{(w)}(i,j,k)\delta_{o_t^{(w)},v_l}}{\sum_{w=1}^{W}c_w P(O^{(w)}|\lambda)\sum_{t=2}^{T_w-1}\sum_{j=1}^{N}\sum_{k=1}^{N}\xi_t^{(w)}(i,j,k)}$$

$$1\leqslant i\leqslant N,1\leqslant l\leqslant M \qquad (2.3.4)$$

In the formulas, $\overline{\pi}_i$ is the initial state vector; $\overline{a}_{ij}$ is the first-order state transition matrix; $\overline{a}_{ijk}$ stands for the second-order state transition matrix; $\overline{b}_i(l)$ stands for the symbol emission matrix; λ represents HMM2 containing the parameters above; c_w represents the value of weight; $\xi_t^{(w)}(i,j,k)$ denotes the conditional probability when the time is $t-1$、t、$t+1$ and the corresponding state is i、j、k, under known HMM2 λ and symbols emission sequence $O^{(w)}$ conditions. The definition is

$$\xi_t^{(w)}(i,j,k)=P(s_{t-1}=i,s_t=j,s_{t+1}=k|O^{(w)},\lambda)$$

$$=\alpha_t(i,j)\overline{a}_{ijk}\frac{\overline{b}_k(o_{t+1})\beta_{t+1}(j,k)}{P(O^{(w)}|\lambda)}$$

$$2\leqslant t\leqslant T-1,\quad 1\leqslant i,j,k\leqslant N \qquad (2.3.5)$$

Among them, $\alpha_t(i,j)=P(o_1,\cdots,o_t,s_{t-1}=i,s_t=j|\lambda)$ stands for forward variable in Forward Algorithm of HMM2 (Shi et al., 2001). $\beta_t(i,j)=P(o_{t+1},\cdots,o_T|s_{t-1}=i,s_t=j,\lambda)$ represents backward variable in Backward Algorithm of HMM2 (Shi et al., 2001).

If directly use formulas (2.3.1)–(2.3.5) to implement the algorithm, the result is numerically unstable. Especially with the length T of sequence increasing, the forward variable $\alpha_t(i,j)$ and backward variable $\beta_t(i,j)$ will become smaller and smaller in implementing the Forward Algorithm or Backward Algorithm to calculate the probability score $P(O^{(w)}|\lambda)$ of symbol emission sequences. They are possibly swallowed zero by a computer, resulting in not training model. In order to solve the problem of unstable training, we normalize the forward variable $\alpha_t(i,j)$ and backward variable $\beta_t(i,j)$ by each step. In the program, the Forward Algorithm is modified as:

Step 1 Initialization:

$$\alpha_2(i,j)=\pi_i b_i(o_1)a_{ij}b_j(o_2) \qquad 1\leqslant i,j\leqslant N$$
$$S(2)=sum(\alpha_2(:,:))$$
$$\alpha_2'(i,j)=\alpha_2(i,j)/S(2) \qquad 1\leqslant i,j\leqslant N$$

Step 2 Iterative computation:

$$\alpha_{t+1}(j,k)=\sum_{i=1}^{N}\alpha_t'(i,j)\overline{a}_{ijk}\overline{b}_k(o_{t+1}) \qquad 1\leqslant j,k\leqslant N$$
$$S(t+1)=sum(\alpha_{t+1}(:,:)), \qquad 2\leqslant t\leqslant T-1$$
$$\alpha_{t+1}'(j,k)=\alpha_{t+1}(j,k)/S(t+1) \qquad 1\leqslant j,k\leqslant N$$

Step 3 Calculating the probability score of sequence:

$$P(O^{(w)}|\lambda)=S(2)S(3)\cdots S(T)$$

or

$$\log\,[P(O^{(w)}|\lambda)]=\log\,[S(2)]+\log\,[S(3)]+\cdots\log\,[S(T)]$$

Do the corresponding change with the Backward Algorithm. First, $\beta_t(i,j)(t=T-1,\cdots,2)$ is calculated by original iterative algorithm of HMM2, and then further process is done by every step, which is $\beta_t'(i,j)=\beta_t(i,j)/S(t+1)$. So the probability score of sequence is also $P(O^{(w)}|\lambda)=S(2)S(3)\cdots S(T)$.

Now, here

$$\xi_t^{(w)}(i,j,k)=\frac{\alpha_t'(i,j)\overline{a}_{ijk}\overline{b}_k(o_{t+1})\beta_{t+1}'(j,k)}{S(t+1)} \qquad (2.3.6)$$

Formula (2.3.6) can overcome numerically unstable problem of $\alpha_t(i,j)$ and $\beta_t(i,j)$, $2\leqslant t\leqslant T$ in algorithm implementation. Thus the algorithm will be fit for more and longer sequences to train the optimal model.

2.4 HMM2 training method

In order to predict potential miRNA target sites from a set of candidate sequences that are obtained from other prediction softwares, such as miRanda, Pictar, and Targetscan, we implement the algorithm and build two second-order HMM modules respectively based on positive and negative data. One is called the True Target Binding Site Module and the other is called the False Target Binding Site Module. Let $\lambda^{(t)}$ stand for the True Target Binding Site Module trained by the positive training set and $\lambda^{(f)}$ stand for the False Target Binding Site Module trained by the negative training set. Once $\lambda^{(t)}$ and $\lambda^{(f)}$ are trained, we need to calculate the probability score of the candidate sequence respectively. Let O denote a miRNA's candidate target sequence. Assume that O is the true miRNA target, setting symbol variable $Y=1$; otherwise, $Y=0$. Let $P(O|Y=1,\lambda^{(t)})$ represent the prior probability of candidate target sequence that will be a true miRNA target site and $P(O|Y=0,\lambda^{(f)})$ stand for the prior probability of candidate target sequence that will be a false miRNA target site (Nam et al., 2005; Xu et al., 2005). Both the prior probabilities are calculated by the modified Forward Algorithm that we present in this paper. Thus the posterior probability of the candidate target sequence being or not being a miRNA target site can be calculated by Bayes's theorem.

$$P(Y=1|O,\lambda^{(t)})=\frac{P(O|Y=1,\lambda^{(t)})P(Y=1)}{P(O)} \qquad (2.4.1)$$

$$P(Y=0|O,\lambda^{(f)})=\frac{P(O|Y=0,\lambda^{(f)})P(Y=0)}{P(O)} \qquad (2.4.2)$$

where,

$$P(O) = P(O|Y = 1,\lambda^{(t)})P(Y = 1) + P(O|Y = 0,\lambda^{(f)})P(Y = 0).$$

Here, we choose $P(Y=1)=P(Y=0)=0.5$.

If $P(Y=1|O,\ \lambda^{(t)}) \geqslant P(Y=0|O,\ \lambda^{(f)})$, then we can assert that the sequence O will be a potential miRNA target site; otherwise, O will be not. Fig. 2 gives out the flowchart of the whole prediction process.

2.5 Model evaluation

The effect of the model is very crucial to assessing whether the mode can be applied in the real world. For a sample containing positive and negative sets, the predicted results include the following four types: correct prediction of the number of positive samples TP and negative samples TN, the number of false positive samples FP and false negative samples FN of prediction. Based on these values, the sensitivity (Se), specificity (Sp) and classified accuracy (ACC) of the model can be calculated,

$$\begin{cases} \mathrm{Se} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN}) \\ \mathrm{Sp} = \mathrm{TN}/(\mathrm{TN} + \mathrm{FP}) \\ \mathrm{ACC} = (\mathrm{TP} + \mathrm{TN})/(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}) \end{cases} \quad (2.5.1)$$

3 Results and discussion

The HMM2 algorithms are implemented based on the Matlab platform. First, we use all the 244 positive and 49 negative data to train and test the model with the 10-fold cross-validation method, achieving a sensitivity of 72.54% (177 out of 244), specificity of 55.10% (27 out of 49), and accuracy of 69.62%. The detailed prediction results are displayed in Table 2.

The less negative and unbalanced data may influence the specificity. Therefore, we also randomly select 30, 50, 75, 100, 150, 200 and 230 pairs from all the positive miRNA: target data respectively and train them combining with 49 negative data through the 10-fold cross-validation method. When positive data is taken as 30, 50, 75, which approach the negative data, a higher specificity of 77.55% is

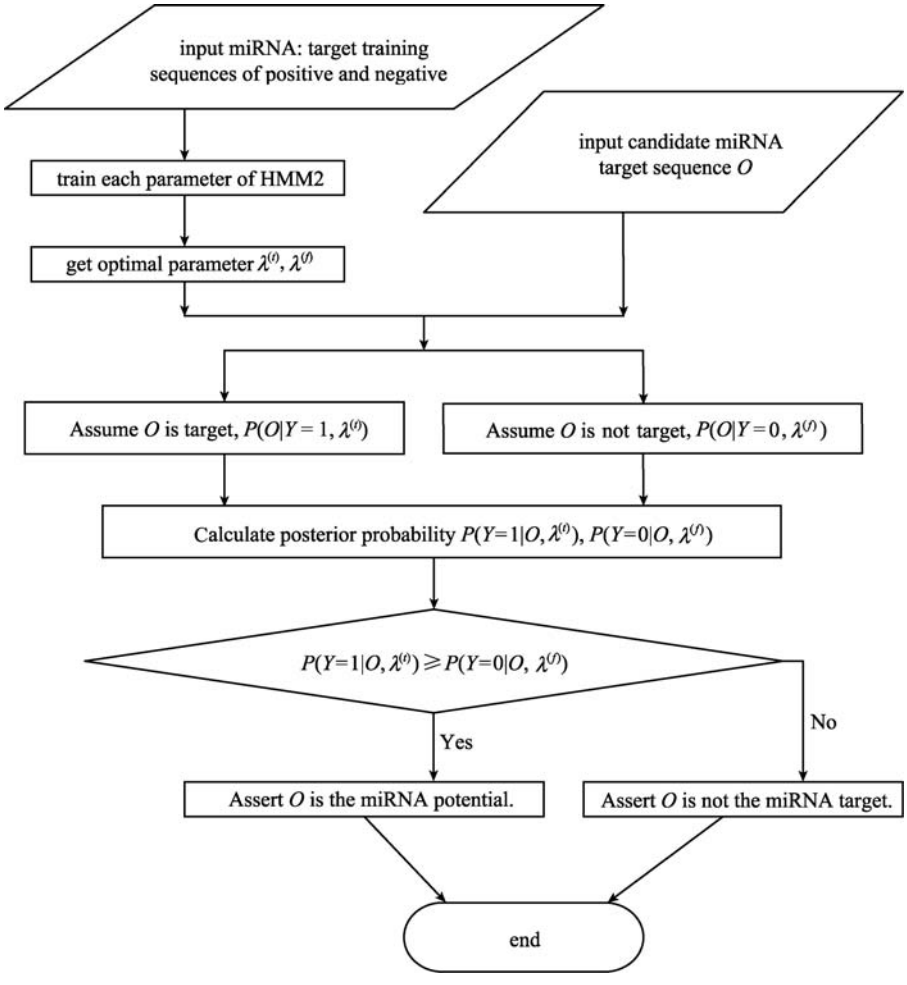

Fig. 2 Flowchart of HMM2 training and recognizing miRNA target

achieved. It is better than Yang's (Yang et al., 2008) result of 73.68% using the SVM method. The predicted results also show that there is a gradual growth in sensitivity (60%–72.54%) and decline in specificity (77.55%–55.1%) with the increase of the positive pairs, illustrating that lower prediction percentage of specialty may attribute to the lack of negative miRNA:target. Further detailed results are given in Table 3.

In order to further validate the applicability of the algorithm, we also use 195 positive and 38 negative data collected by Yang (Yang et al., 2008) to test through the 10-fold cross-validation method, obtaining a sensitivity of 83.08% (162 out of 195), specificity of 60.53% (23 out of 38) and accuracy of 79.40% (Table 4). The results are consistent with Yang's (Yang et al., 2008) (i.e., a sensitivity of 83.59% (163 out of 195), specificity of 73.68% (28 out of 38) and accuracy of 81.79% by the SVM method), which shows our method is effective. However, when the positive data approach the negative, HMM2 method can get a better specificity of 77.55%.

At present, the methods of predicting miRNA targets are based on similar characteristics of base pairing derived from experimentally supported miRNA:target duplexes and improved with the validated data increasing. Although these methods are fundamentally similar, each of them has merits and faults and the predictions are not in common, due to the small experimentally supported data and incomplete statistic characteristics on miRNA:target. In addition, these prediction methods usually predict hundreds of target sites for a given miRNA, which indicates that the false positive predictions could account for a large percentage among all predicted targets. Using the known miRNA:target interaction pairs, HMM2 classifiers can obtain the new characteristics of true targets comparing with the negative by machine-learning. Such as, in true targets, miRNA 3' end may be easier to emerge the insert state as a beginning base; the consecutive states transition of insert-match-mismatch (insert-match-mismatch indicates that the states transition is from insert state to match and then transfers to the mismatch state like the second-order Markov chain), match-insert-insert and match-match-insert could be more probable to occur in non-seed area of the true miRNA:target pairs; the base of G may be most likely to appear in true target genes and so on.

The method of HMM2 should not be considered as a general way to predict miRNA target by searching 3' UTR sequences directly, but a post-processing filter for the miRNA:target duplex produced by other computational methods, such as miRanda, Pictar, and Targetscan. Because HMM2 classifiers are trained on validated miRNA:target interactions, both positive and negative duplexes include the strong seed complementarities at 5'end of miRNA and high binding energy of secondary structure. So the candidate target sequence, without

Table 2 Prediction results of 244 positive and 49 negative

positive set			negative set			
TP	FN	Se/%	TN	FP	Sp/%	ACC/%
177	67	72.54	27	22	55.10	69.62

TP stands for correctly predicted positive miRNA:target pairs; FN stands for wrongly predicted positive miRNA:target pairs; TN stands for correctly predicted negative miRNA:target pairs; FP stands for wrongly predicted negative miRNA:target pairs. Se: the sensitivity; Sp: specificity; ACC: classified accuracy.

Table 3 Prediction results of among different positive data and 49 negative

number of positive	positive set			negative set			
	TP	FN	Se/%	TN	FP	Sp/%	ACC/%
30	18	12	60.00	38	11	77.55	70.89
50	32	18	64.00	38	11	77.55	70.71
75	49	26	65.33	35	14	71.43	67.74
100	68	32	68.00	32	17	65.31	67.11
150	106	42	70.70	28	21	57.14	67.73
200	144	56	72.00	28	21	57.14	69.08
230	166	64	72.17	28	21	57.14	69.53

TP stands for correctly predicted positive miRNA:target pairs; FN stands for wrongly predicted positive miRNA:target pairs; TN stands for correctly predicted negative miRNA:target pairs; FP stands for wrongly predicted negative miRNA:target pairs. Se: the sensitivity; Sp: specificity; ACC: classified accuracy.

Table 4 Prediction results of 195 positive and 38 negative

positive set			negative set			
TP	FN	Se/%	TN	FP	Sp/%	ACC/%
162	33	83.08	23	15	60.53	79.40

TP stands for correctly predicted positive miRNA:target pairs; FN stands for wrongly predicted positive miRNA:target pairs; TN stands for correctly predicted negative miRNA:target pairs; FP stands for wrongly predicted negative miRNA:target pairs. Se: the sensitivity; Sp: specificity; ACC: classified accuracy.

prescreening by those seed-sensitive programs and not having high free energy, is not fit for this method.

To show how HMM2 method is applied to judge miRNA target genes in reality, Table 5 lists a comparative result between HMM2 method and other three types of most often used softwares, Pictar, miRanda, and Targetscan. Like MiRTif (Yang et al., 2008), we also choose hsa-mir-224 as the example. This miRNA plays the most importantly up-regulatory role in hepatocellular carcinoma patients (Wang et al., 2008). There are only eight predicted candidate target genes in Table 5. The table also includes information on the function annotation, the ranking of the corresponding prediction software, MiRTif score (Yang et al., 2008) and the possibility under True Target Binding Site Module, respectively. Among the predictions, apoptosis inhibitor-5 (API5) was experimentally validated as miRNA-224 specific target (Wang et al., 2008). Our method also confirms the API5 to be the target genes with a highest probability of 0.9230, which is more efficient than MiRTif. We believe that the discriminant scores of HMM2 and MiRTif will provide experimental biologists a new insight or proposal, especially at the time of choosing miRNA targets for validation.

The HMM, which has the fundamentals of statistical theory and efficient learning algorithm, has been developed in recent decades as a double stochastic process. By introducing the local learning probability, HMM permits to compensate for the insertion and deletion state in a uniform way, and can be attained from the original data by learning directly. Current HMM has been widely applied in bioinformatics research, such as the modeling of DNA coding area, multiple sequence alignment, and protein super-family (Churchill et al., 1989; Rabiner et al., 1989; Gough et al., 2002). Borodovsky et al. (1986) discovered that coding and non-coding sequences follow Markov chain properties. So second-order Hidden Markov Model is considered in this paper to process miRNA and its targets, which is the first time to be applied in miRNA target prediction, obtaining good results.

Although HMM has achieved success in many aspects, there still remains two problems. Firstly, it often contains many structureless parameters. When only small data are available, lacking sufficient information may turn out to be a serious problem. Secondly, first-order or second-order HMM is limited by the properties of Markov, which cannot reveal the dependent relation of farther distance. For these two limitations, the lack of experimentally validated miRNA:target currently may lead to getting less optimal target sequence characteristics, affecting the accuracy of the prediction. We believe that, with the increase of validated data and by combining with higher order HMM, our method will achieve a better predicted result.

In addition, all the programs of our method have been implemented based on the Maltab platform and formed a program package. If needed, we can provide all of them, and at the same time, they can also be downloaded from http://project.kernelstudio.org/share/miRTarHMM2/.

Acknowledgements We would like to thank Dr. Xinjun PENG of Shanghai Normal University for critical reading of the manuscript. This work is supported by The National Natural Science Foundation of China (Grant No. 30871341), the grants from the National Key S&T Special Project of China (Nos. 2008ZX10002-017, 2008ZX10002-020, and 2009ZX09103-686), Shanghai Key Discipline of China (No. S30104), and Education Commission Key Discipline Construction Project (No. J50101).

References

Barciszewski J, Erdmann V A (2008). Noncoding RNAs: Molecular Biology and Molecular Medicine (in Chinese, Trans. Zheng X F).

Table 5 A comparison among HMM2, three usual prediction softwares, Pictar, miRanda, and Targetscan, and one filter software MiRTif

gene	gene description	PicTar ranking	miRanda ranking	targetscan ranking	MiRTif score	probability true HMM2
H3F3B	H3 histone, family 3B (H3.3B)	1	Na	21	+ 1.69	0.3156
API5	apoptosis inhibitor 5	2	Na	26	+ 1.32	0.9230
ARMCX2	Armadillo repeat containing, X-linked 2	4	198	Na	− 0.98	0.3587
CDK9	cyclin-dependent kinase 9	9	67	40	− 0.15	0.5935
NCOA6	nuclear receptor coactivator 6	10	192	8	− 0.94	0.7071
ATF2	activating transcription factor 2	64	5	85	+ 1.27	0.7951
NUP153	nucleoporin 153 kDa	168	789	118	− 1.46	0.2633
FOSB	FOJ murine osteosarcoma viral oncogene homolog B	222	951	212	+ 0.11	0.1387

The genes in Table 5 correspond to the predicted target of hsa-mir-224. Ranking indicates the rank predicted by the score from high to lower using respective prediction software. "Na" means that the particular program did not predict an interaction between the corresponding target and hsa-mir-224. The greater probability in True Target Binding Site Module, the more confident judgment in true target.

Beijing: Chemical Industry Press, 104–119

Borodovsky M, Sprizhitskii Y, Golovanov E, Aleksandrov A (1986a). Statistical patterns in primary structures of functional regions in the *E. coli* genome. I. Oligonucleotide frequencies analysis. Mol Biol, 20: 826–833

Borodovsky M, Sprizhitskii Y, Golovanov E, Aleksandrov A (1986b). Statistical patterns in primary structures of functional regions in the *E. coli* genome. II. Non-homogeneous Markov models. Mol Biol, 20: 833–840

Borodovsky M, Sprizhitskii Y, Golovanov E, Aleksandrov A (1986c). Statistical patterns in primary structures of functional regions in the *E. coli* genome. III. Computer recognition of coding regions. Mol Biol, 20: 1145–1150

Churchill G A (1989). Stochastic models for heterogeneous DNA sequences. Bull Mathem Biol, 51: 79–94

Du S P (2007). The Baum-Welch Algorithm of HMM2 with Multiple Observations. J Biomathem, 22(4): 685–690 (in Chinese)

Du S P, Li H (2004). Second-order Hidden Markov Models and Its Application to Computational Linguistics. J Sichuan Uni (Nat Sci Edi), 41(2): 284–289 (in Chinese)

Duursma A M, Martijn K, Mariette S, Carlos L S, Reuven A (2008). miR-148 targets human DNMT3b protein coding region. RNA, 14 (5): 872–877

Enright A J, John B, Gaul U, Tuschl T, Sander C, Marks D S (2003). MicroRNA targets in Drosophila. Genome Biol, 5(1): Article Rl

Gough J, Chothia C (2002). SUPERFAMILY: HMMs representing all proteins of known structure, SCOP sequence searches, alignments, and genome assignments. Nucl Acids Res, 30(1): 268–272

Griffiths-Jones S, Grocock R J, van Dongen S, Bateman A, Enright A J (2006). miRBase: microRNA sequences, targets and gene nomenclature. Nucl Acids Res, 34: D140–D144

Hébert S S, Horré K, Nicolaï L, Papadopoulou A S, Mandemakers W, Silahtaroglu A N, Kauppinen S, Delacourte A, De Strooper B (2008). Loss of microRNA cluster miR-29a/b-1 in sporadic Alzheimer's disease correlates with increased BACE1/beta-secretase expression. Proc Natl Acad Sci USA, 105(17): 6415–6420

Huynh T, Miranda K, Tay Y, Ang Y S, Tam W L, Thomson A M, Lim B, Rigoutsos I (2006). A pattern- based method for the identification of microRNA-target sites and their corresponding RNA/RNA complexes. Cell, 126: 1203–1217

John B, Enright A J, Aravin A, Uschl T, Sander C, Marks D S (2004). Human MicroRNA Targets. PLoS Biology, 2(11): 1862–1879

Kim S K, Nam J W, Rhee J K, Lee W J, Zhang B T (2006). MiTarget: microRNA target gene prediction using a support vector machine. BMC Bioinformatics, 7: 411

Kiriakidou M, Nelson P T, Kouranov A, Fitziev P, Bouyioukos C, Mourelatos Z, Hatzigeorgiou A (2004). A combined computational-experimental approach predicts human microRNA targets. Genes Dev, 18: 1165–1178

Krek A, Grün D, Poy M N, Wolf R, Rosenberg L, Epstein E J, MacMenamin P, da Piedade I, Gunsalus K C, Stoffel M, Rajewsky N (2005). Combinatorial microRNA target predictions. Nat Genet, 37: 495–500

Landais S, Landry S, Legault P, Rassart E (2007). Oncogenic potential of the miR-106-363 cluster and its implication in human T-cell leukemia. Cancer Res, 67(12): 5699–5707

Lewis B P, Burge C B, Bartel D P (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell, 120: 15–20

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

Luo X B, Lin H X, Pan Z W, Xiao J N, Zhang Y, Lu Y J, Yang B F, Wang Z G (2008). Down-regulation of miR-1/miR-133 Contributes to Re-expression of Pacemaker Channel Genes HCN2 and HCN4 in Hypertrophic Heart. J Biol Chem, 283(29): 20045–20052

Nam J W, Shin K R, Han J, Lee Y, Kim V N, Zhang B T (2005). Human microRNA prediction through a probabilistic co-learning model of sequence and structure. Nucl Acids Res, 33(11): 3570–3581

Rabiner L R, Juang B H (1986). An introduction to hidden Markov models. In: IEEE Acoustics, Speech & Signal Processing Magazine, 3: 4–16

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

Rossi J J, Hannon G J (2008). MicroRNA Methods. Beijing: Science Press, 1–83

Rusinov V, Baev V, Minkov I N, Tabler M (2005). MicroInspector: a web tool for detection of miRNA binding sites in an RNA sequence. Nucl Acids Res, 33: W696–W700

Saetrom O, Ola S J, Saetrom P (2005). Weighted sequence motifs as an improved seeding step in microRNA target prediction algorithms. RNA, 11: 995–1003

Sengupta S, den Boon J A, Chen I H, Newton M A, Stanhope S A, Cheng Y J, Chen C J, Hildesheim A, Sugden B, Ahlquist P (2008). MicroRNA 29c is down-regulated in nasopharyngeal carcinomas, up-regulating mRNAs encoding extracellular matrix proteins. Proc Natl Acad Sci USA, 5(15): 5874–5878

Sethupathy P, Corda B, Hatzigeorgiou A G (2006). TarBase: A comprehensive database of experimentally supported animal microRNA targets. RNA, 12(2): 192–197

Shi X X, Wang T J, He Z Y (2001). The Learning Algorithm of the Second Order HMM and Its Relationship with the First Order HMM. J Appl Sci, 19(1): 29–32 (in Chinese)

Skalsky R L, Samols M A, Plaisance K B, Boss I W, Riva A, Lopez M C, Baker H V, Renne R (2007). Kaposi's sarcoma-associated herpesvirus encodes an ortholog of miR-155. J Virol, 81(23): 12836–12845

Thadanil R, Tammi M T (2006). MicroTar: predicting microRNA targets from RNA duplexes. BMC Bioinformatics, 7(Suppl 5): S20

Wang Y, Lee A T, Ma J Z, Wang J, Ren J, Yang Y, Tantoso E, Li K B, Tan P, Lee C G L (2008). Profiling microRNA expression in hepatocellular carcinoma reveals microRNA-224 up-regulation and apoptosis inhibitor-5 as a microRNA-224-specific target. J Biol Chem, 283(19): 13205–13215

Xia W, Cao G J, Shao N S (2009). Research approach of microRNA target gene in search and indentification. Sci China, C: Life Sci, 39 (1): 121–128 (in Chinese)

Xu D, Liu H J, Wang Y F (2005). BSS-HMM^{3}s: An improved HMM method for identifying transcription factor binding sites. DNA Sequence, 16(6): 403–411

Yang Y C, Wang Y P, Li K B (2008). MiRTif: a support vector machine-based microRNA target interaction filter. BMC Bioinformatics, 9 (Suppl 12): S4

Yousef M, Jung S, Kossenkov A V, Showe L C, Owe M K Sh (2007). Naïve Bayes for MicroRNA Target Predictions Machine Learning for MicroRNA Targets. Bioinformatics, 23(22): 2987–2992

Zhang B H, Pan X P, Wang Q L, Cobb G P, Anderson T A (2006). Computational identification of microRNAs and their targets. Comput Biol Chem, 30: 395–407