Published online June 24, 2005

3570–3581 *Nucleic Acids Research, 2005, Vol. 33, No. 11*
doi:10.1093/nar/gki668

Human microRNA prediction through a probabilistic co-learning model of sequence and structure

Jin-Wu Nam[1,2], Ki-Roo Shin[3], Jinju Han[4], Yoontae Lee[4], V. Narry Kim[4] and Byoung-Tak Zhang[1,2,3,*]

[1]Graduate Program in Bioinformatics, [2]Center for Bioinformation Technology (CBIT), [3]Biointelligence Laboratory, School of Computer Science and Engineering and [4]Department of Biological Sciences, Seoul National University, Seoul 151-744, Korea

Received January 10, 2005; Revised May 27, 2005; Accepted June 6, 2005

ABSTRACT

MicroRNAs (miRNAs) are small regulatory RNAs of ~22 nt. Although hundreds of miRNAs have been identified through experimental complementary DNA cloning methods and computational efforts, previous approaches could detect only abundantly expressed miRNAs or close homologs of previously identified miRNAs. Here, we introduce a probabilistic co-learning model for miRNA gene finding, ProMiR, which simultaneously considers the structure and sequence of miRNA precursors (pre-miRNAs). On 5-fold cross-validation with 136 referenced human datasets, the efficiency of the classification shows 73% sensitivity and 96% specificity. When applied to genome screening for novel miRNAs on human chromosomes 16, 17, 18 and 19, ProMiR effectively searches distantly homologous patterns over diverse pre-miRNAs, detecting at least 23 novel miRNA gene candidates. Importantly, the miRNA gene candidates do not demonstrate clear sequence similarity to the known miRNA genes. By quantitative PCR followed by RNA interference against Drosha, we experimentally confirmed that 9 of the 23 representative candidate genes express transcripts that are processed by the miRNA biogenesis enzyme Drosha in HeLa cells, indicating that ProMiR may successfully predict miRNA genes with at least 40% accuracy. Our study suggests that the miRNA gene family may be more abundant than previously anticipated, and confer highly extensive regulatory networks on eukaryotic cells.

INTRODUCTION

MicroRNAs (miRNAs), constituting a large family of noncoding (nc) small RNAs, directly take part in post-transcriptional regulation either by arresting the translation of messenger RNAs (mRNAs) or by the cleavage of mRNAs (1). miRNAs are defined as single-stranded RNAs of ~22 nt in length (ranged 19–25 nt) generated from endogenous transcripts that can form local hairpin structures (2). miRNA genes are transcribed by RNA polymerase II (3,4). The local hairpin structures in the primary transcripts [primary microRNAs, (pri-miRNAs)] are first processed by the nuclear RNase type III enzyme, Drosha, to release the hairpin-shaped intermediates (pre-miRNAs) (5). Pre-miRNAs are typically 60–70 nt, and contain an ~22 bp double-stranded stem and an ~10 nt terminal loop. The terminal end at the opposite side of the loop contain an ~2 nt overhang at the 3′ end, which is typical of RNase III products. Pre-miRNAs are then exported to the cytoplasm by the nuclear export factor Exportin 5 and the Ran-GTP cofactor (6–8). In the cytoplasm, pre-miRNAs are cleaved by another RNase III type enzyme, Dicer, to generate an ~22 nt RNA duplex (9–13). Dicer cleaves at ~22 nt from the terminal staggered end discarding the terminal loop region. One strand of the miRNA duplex is usually selected as mature miRNA while the other strand is degraded (14,15). Therefore, two-step processing events are required for miRNA maturation: (i) the cleavage at the non-looped side of the stem by Drosha in the nucleus, followed by (ii) the cleavage at the looped end by Dicer in the cytoplasm (5,6).

Identification of novel miRNA genes is one of the most imminent problems towards the understanding of post-transcriptional gene regulation. Thus far, 227 human miRNAs have been reported (16–31). Experimental cloning efforts have successfully identified highly expressed miRNAs from various tissues. However, cloning methods are highly biased towards miRNAs that are abundantly and/or ubiquitously expressed.

On the other hand, computational prediction of miRNAs could become a robust approach for tissue-specific or lowly expressed miRNAs. Several computational methods have been developed to find close homologs among related miRNAs (29,32–35). The program MiRscan successfully predicted close homologs of *Caenorhabditis briggsae* with statistically

*To whom correspondence should be addressed. Tel: +82 2 880 1847; Fax: +82 2 875 2240; Email: btzhang@bi.snu.ac.kr

conserved patterns of *Caenorhabditis elegans* miRNAs (35). However, MiRscan failed to detect miRNAs that lack clear homologs in related species. Recently, the structural features of pre-miRNAs and the upstream sequence motif of miRNAs have been incorporated in the computations (34,36). MiRscan has been improved by defining the newly observed upstream sequence motif and the patterns of flanking sequence conservation for nematode miRNAs (34). Likewise, the program miRseeker was able to predict new *Drosophila* miRNA genes by screening closely homologous stem–loops in entire genomes (32). A profile-based method is better than the previous similarity searches and can predict close homologs in animal genomes by profiling the miRNA sequence family (33). However, these methods also frequently fail to detect miRNAs that lack detectable homologs.

In this study, we suggest a probabilistic co-learning method based on the paired hidden Markov model (HMM) to implement a general miRNA prediction method to identify close homologs as well as distant homologs. It combines both sequential and structural characteristics of miRNA genes in a probabilistic framework, and simultaneously decides if an miRNA gene and a region of mature miRNA are present by detecting the signals for the site cleaved by Drosha. We employed this method to predict novel miRNA genes and experimentally validated the candidates by examining the accumulation of pri-miRNAs in the cells depleted of Drosha.

MATERIALS AND METHODS

Datasets

We trained and validated the algorithm through 5-fold cross-validation with a positive dataset [known human miRNA precursor (pre-miRNAs)] and a negative dataset. We used previously known human pre-miRNAs consisting of 81 5′ strand and 55 3′ strand mature miRNAs as the positive dataset (available at http://www.sanger.ac.uk/Software/Rfam/mirna/search.shtml, release 4.0).

The negative dataset consisted of 1000 extended stem–loop structures randomly extracted from human chromosomes with several criteria described in the Supplementary Material (based on the NCBI build 34, version 3 of the human genome). All stem–loop structures were predicted using the Vienna RNA software package (37). These criteria were obtained through learning the common structure of human pre-miRNAs (38) and were also used for the extraction of extended stem–loop structures similar to pre-miRNAs in genome sequences.

Probabilistic co-learning model of pre-miRNAs

An pre-miRNA can be represented as a pairwise sequence. It forms an extended stem–loop structure, and this structure can be formulated as a sequence of matched base pairs (Figure 1a). The pairwise sequence starts from the non-looped side of the pre-miRNA and ends at the loop. The state of each pair can be classified on the basis of its base pairing status as (i) match, (ii) mismatch, (iii) deletion or (iv) insertion. In particular, we regard a loop structure as an ordered sequence of mismatches and insertions as in the paired HMM.

Each position of the pairwise sequence has two properties, i.e. structural states (match/mismatch/deletion/insertion) and hidden states (information for the mature miRNA region). In the structural states, each match state, M, can emit A–U, U–A, G–C, C–G, U–G or G–U as an emission symbol. Deletion states, D, can emit •–A, •–U, •–G or •–C. Insertion states, I, can emit A–•, U–•, G–• or C–•. Mismatch states, N, can

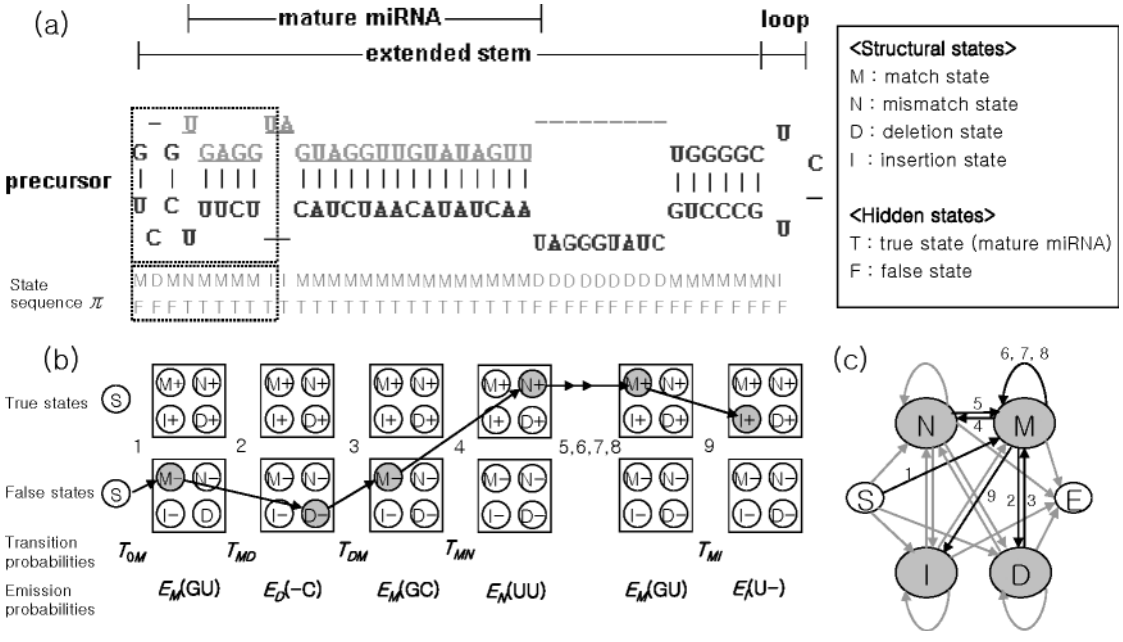

Figure 1. Pairwise representation of stem–loop structures and state sequences of pre-miRNAs, where the state of each pair includes structural information and mature miRNA region information (hidden states). (**a**) The structure of the pre-miRNA. (**b**) The transition and emission scheme of the structural states and the hidden states for pairwise sequence in the dotted rectangle shown in (a). T_{0M}, T_{DM}, T_{MN} and T_{MI} are transition probabilities. E_M(GU), E_D(−C), E_M(GC), E_N(UU), E_M(GU) and E_I(U−) are emission probabilities. (**c**) The four-state finite state automaton. Finally, the probability of the pairwise sequence is assigned by multiplication of the transition probabilities and the emission probabilities.

emit one of the remaining combinations. The possible transitions between the four structural states are shown in Figure 1c. Each emission is represented as a corresponding character in alphabetical order. In the hidden states, T means a true state, namely a region of mature miRNA, and F means a false state, the precursor region outside mature miRNA sequences (Figure 1b). The probabilities of hidden states in this sequence-structure co-learning model are estimated from the distribution of all four structural states.

Screening of pre-miRNAs

To screen miRNPs, we estimate the probability of the pairwise sequence of pre-miRNAs. To derive the probability of the pairwise sequence of the states and the symbols, we need to estimate two parameters: a transition probability and an emission probability. For the transition probability, let us call the state sequence a path, π. The probability of a state depends only on the previous state. If π_i denotes the i-th state in the path, the transition probability is defined as

$$T_{kl} = P(\pi_i = l \,|\, \pi_{i-1} = k), \quad \mathbf{1}$$

where the transition is from state $\pi_{i-1} = k$ to state $\pi_i = l$. The probability of starting in state k can be defined as T_{0k}. Let x_i denote the symbol emitted from the i-th state. Then, the emission probability of observing symbol b in state k is defined as

$$E_k(b) = P(x_i = b \,|\, \pi_i = k). \quad \mathbf{2}$$

Using the transition and emission probabilities, we can estimate the probability $P(x)$ that sequence x is generated by the probabilistic co-learning model. It is easy to define the joint probability of an observed sequence x and a state sequence π:

$$P(x, \pi) = T_{0\pi_1} \prod_{i=1}^{L} E_{\pi_i}(x_i) T_{\pi_i \pi_{i+1}}, \quad \mathbf{3}$$

where L is the window size. If we are to choose just one path for our prediction, that one, π^*, with the highest probability should be chosen as follows:

$$\pi^* = \arg\max_{\pi} P(x, \pi). \quad \mathbf{4}$$

The Viterbi algorithm (39) is a common method for finding the most probable state transition path and its probability in HMMs. However, a straightforward application of the algorithm is impossible in this case because the values of the probability returned by the algorithm are very small. In particular, when the given sequence is longer, the probability the Viterbi algorithm produces is smaller, exponentially. To use the Viterbi probability for classification, we should evaluate the Viterbi probability of the fixed-length sequence that represents the mature miRNA region instead of the entire sequence. We evaluate the Viterbi probability for the mature miRNA region as

$$P(x, \pi) = T_{0\pi_1} \prod_{i=1}^{22} E_{\pi_i}(x_i) T_{\pi_i \pi_{i+1}}. \quad \mathbf{5}$$

On a given pairwise sequence, we search for the maximum $P(x, \pi)$ value by using a sliding window, the size of which is 22 ± 2 bp—the mean length of the mature miRNAs in the pairwise representation. We evaluate two $P(x, \pi)$ values for the models of 5′ strand pre-miRNAs and 3′ strand pre-miRNAs, respectively. If the $P(x, \pi)$ values for the 5′ and 3′ strands are higher than a threshold selected in advance, then we classify the given candidate as an pre-miRNA. The threshold was determined by the receiver operator characteristic (ROC) curve analysis.

Evidence for validation of miRNP prediction

Statistical significance of minimum free energy. The thermodynamic stability and statistical significance of the secondary structures can be assessed using minimum free energy (MFE) and Monte Carlo simulations (40). Van de Peer's group (41) proposed *P*-values to assess the statistical significance of MFE values of miRNAs and ncRNAs and developed the randfold, a program for testing statistical significance. The *P*-values of miRNAs were lower than those for other ncRNAs, with statistical significance. The low *P*-value (under $P = 0.05$) provides evidence to identify putative miRNA genes from the genome sequence. However, the stability and conservation of secondary structures was stated as insufficient evidence to predict new ncRNAs as a general method in a study on the common structures of ncRNAs (42). Thus, to efficiently predict miRNA genes, the statistical stability of the MFE value should be combined with other information.

Repeat and published RNA sequences. No published human pre-miRNAs contain human repeat motifs such as *alu* sequences. Results of genome-wide miRNA prediction might thus produce false positives of repeat sequences. Therefore, candidates containing human repeat sequences should be excluded. Human repeat sequences were downloaded from the Alu sequence database of GenBank. Only 10 of the published human pre-miRNAs could be matched with published RNA sequences that include transfer, ribosomal and small nuclear RNAs, and others. Some of these were located in the intron of mRNA. Therefore, such evidence can be used as negative information to predict pre-miRNAs in the genome sequence. Here, we have aligned the sequences using BLAST to investigate whether candidates contain repeat sequences or published RNAs. The *E*-value threshold was 10^{-20}.

Prediction of a mature miRNA region and a functional strand

Mature miRNA region. To apply the miRNA maturation mechanism to our probabilistic model, we first detect the stem-end of mature miRNA and then we seek the loop-end, 22 ± 2 bp distant from the non-looped end. We introduce two hidden states indicating whether the position is a mature miRNA region. The probabilities that state whether the i-th position is true or false are computed as

$$P_t(i) = \max\{P_t(i-1) \cdot T_{\tau_{i-1}\tau_i}, P_f(i-1) \cdot T_{\upsilon_{i-1}\tau_i}\} \cdot E_{\tau_i}(x_i) \quad \mathbf{6}$$

$$P_f(i) = \max\{P_t(i-1) \cdot T_{\tau_{i-1}\upsilon_i}, P_f(i-1) \cdot T_{\upsilon_{i-1}\upsilon_i}\} \cdot E_{\upsilon_i}(x_i), \quad \mathbf{7}$$

where τ_i means that the i-th state is true and υ_i means that the i-th state is false. The initial condition is $P_t(1) = 0$, $P_f(1) = 1$.

Using only the true and the false probabilities, we cannot exactly determine mature miRNA regions, because the transition probability around the cleavage site of an miRNA is low.

Thus, we focus on the transition probability of false states and compute $S(i)$ as

$$S(i) = \frac{P_t(i-1) \cdot T_{\upsilon\tau}}{P_t(i-1) \cdot T_{\upsilon\tau} + P_f(i-1) T_{\upsilon\upsilon}}. \quad \mathbf{8}$$

Functional strand. In addition, we can determine the functional strand of a mature miRNA not only by comparing the probabilities of both the strands, but by absolute and relative internal stability of the base pairs at the 5′ ends of the pre-miRNA. The helicase initiates unwinding at the end of base pairs with lower stability and the miRNP complex is assembled with the 5′ end strand, which becomes the functional strand, at its unwinding end (15). To determine the functional strand before prediction of the mature miRNA region, the free energy values for five bases at the end region are calculated from the known 2 nt free energy value table (14).

Experimental verification

Depletion of Drosha was achieved by RNA interference experiments as previously described (see Supplementary Material) (5). Briefly, HeLa cells were incubated with small interfering RNA (siRNA) duplex (Samchully Pharm, Seoul, Korea) complementary to Drosha mRNA for 3 days. As a control, siRNAs complementary to firefly luciferase (instead of Drosha siRNA) were incubated with HeLa cells. Total RNA from both control and test cells was prepared using Trizol (Invitrogen, Carlsbad, CA) and used to synthesize a complementary DNA (cDNA) with oligo-dT primers and the Superscript II enzyme (Invitrogen). The resulting cDNA was then used for PCR amplification. PCR primers were designed to detect pri-miRNAs and are described in the Supplementary Material. The size of PCR products is ~200–280 nt. If the given miRNA gene candidate can indeed express miRNA, the PCR product is expected to accumulate when Drosha is depleted, because under normal condition pri-miRNA would be rapidly cleaved by Drosha. To quantitate the relative level of accumulation of pri-miRNAs, we also performed real-time quantitative PCR (see Supplementary Material). For real-time PCR, the relative quantity of each product is inversely proportional to the threshold cycle (C_T) value. The difference in the C_T value between the control and the test sample means the difference in relative expression level.

RESULTS

Evaluation of performance

We performed 5-fold cross-validation for various screening thresholds to plot ROC curves, which is an effective method for evaluating the performance of diagnostic tests. Figure 2 describes an ROC curve of 136 known miRNAs and a negative dataset of 1000 sequences according to change of threshold. Thresholds are provided as a parameter to predict putative miRNAs in genome-wide search. When selecting thresholds, the trade off between sensitivity and specificity should be considered. We chose the threshold ($P = 0.033$) for the classification of pre-miRNA candidates at the point that shows 73% sensitivity and 96% specificity, on average.

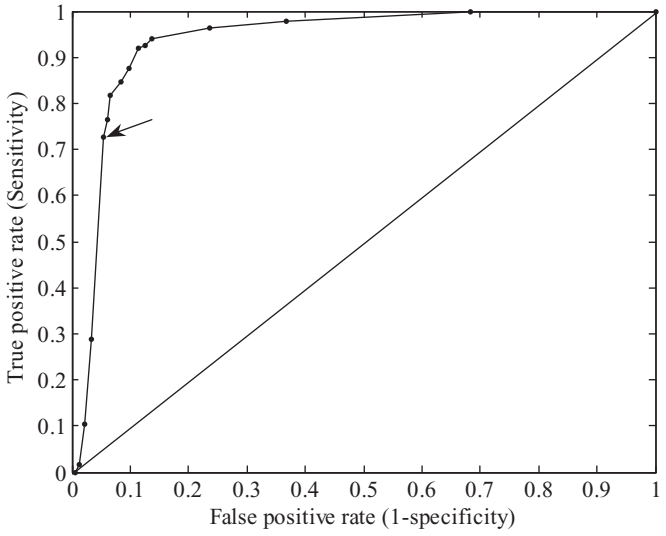

Figure 2. The ROC curve, which is defined as a plot of test sensitivity as the *y*-coordinate, versus the false positive rate (FPR; 1 − specificity) as the *x*-coordinate. The area under the ROC curve is 0.936 by non-parametric estimation. The arrow indicates the point of threshold, where $P = 0.033$, specificity is 96% and sensitivity is 73%.

We have performed additional evaluation with miRNAs, recently reported by the Poy group, to reconfirm the efficiency of our method. In the validation, we could predict the most recently reported miRNAs, hsa-mir-376a, hsa-mir-377, hsa-mir-378, hsa-mir-381, hsa-mir-382, hsa-mir-384, hsa-mir-423 and hsa-mir-424, which are unrelated to our original training data. This result indicates that, unlike previously reported methods, our approach may be sensitive enough to identify unrelated miRNA genes.

Screening for miRNAs on human chromosomes 16, 17, 18 and 19

To perform genome-wide screening for miRNA genes, we extracted 65539, 68458, 34853 and 62229 sequences of stem–loop structures on chromosomes 16, 17, 18 and 19, respectively, using the extracting method of stem–loops mentioned above (Figure 3a and Table 1, see Supplementary Material). Next, to verify the expression of the extracted stem–loops, we performed a human expressed sequence tag (EST) database search using BLAST (NCBI Entrez EST database; February 6, 2003). From the EST database search, 8153, 9367, 3135 and 7765 stem–loops on chromosomes 16, 17, 18 and 19, respectively, were matched with human ESTs and the *E*-values were below 1.0×10^{-30} (Figure 3b, Table 1).

From these 28420 expressed stem–loops, our model 'ProMiR' classified 817 pre-miRNA candidates. The candidates included two of the three known miRNA sequences (mir-138-2, mir-140) on chromosome 16, eleven (mir-21, mir-22, mir-195, mir-196a-1, mir-10a, mir-132, mir-152, mir-195, mir-301, mir-338, mir-108) of the 16 known miRNAs on chromosome 17, four (mir-1-2, mir-122a, mir-133a-1 and mir-187) of the four known miRNA sequences on chromosome 18 and seven (has-let-7e, mir-7-3, mir-27a, mir-150, mir-199a-1, mir-330 and mir-371) of the 14 known miRNAs on

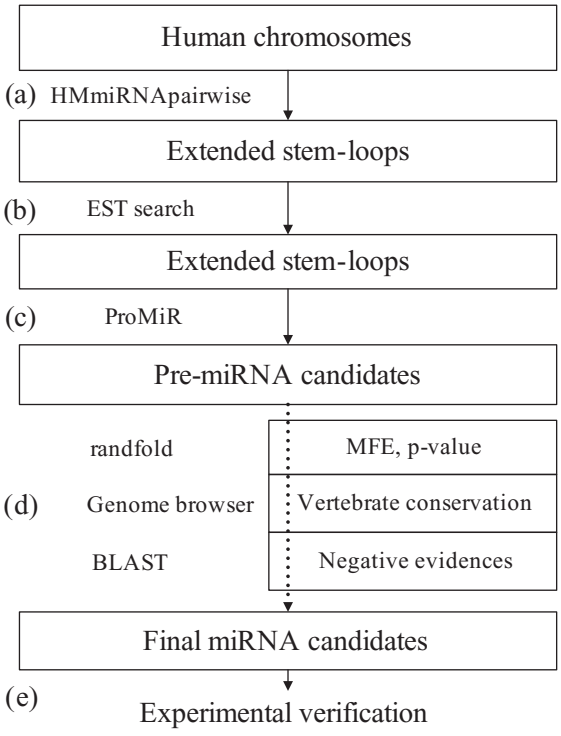

Figure 3. Flow chart for human miRNA gene finding. (a) The program, HMmiRNApairwise using an RNAfold algorithm extracts extended stem–loops with several criteria described in the Supplementary Material; (b) human EST database search; (c) ProMiR predicts pre-miRNA candidates, the region of mature miRNA and the location of a functional strand; (d) screening by additional evidence—MFE values, vertebrate conservations and negative evidences; (e) experimental verification.

chromosome 19 (Figure 3c, Table 1). Because the candidate sets still included many false positives, we selected sequences using the additional evidence mentioned above to identify the more reliable candidates. The miRNA candidates were further screened as follows. We selected candidates with thermodynamically stable stem–loop structures by testing the statistical significance for each candidate's MFE value using the randfold (Figure 3d). Then we screened the candidates with conserved patterns among vertebrates. Recently, computational phylogenetic shadowing showed that the stems of pre-miRNAs are strongly conserved in whole genome alignments, whereas most terminal loop sequences are only loosely conserved (43). The conservation of the flanking region of the conserved pre-miRNA is rapidly decayed (43). Thus, using the UCSC genome browser we investigated whether the putative pre-miRNAs show similar conservation patterns among vertebrates (Figure 3d, http://genome.ucsc.edu, based on NCBI Build 35). Finally, we screened the 23 representative candidates using negative evidence to determine if the candidates matched with human repeats or published RNA sequences (Figure 3d, Tables 1 and 2).

Some of the new miRNA candidates are found in clusters as is often observed in known miRNA gene loci (Table 3). One cluster contains two miRNAs (NC19-5 and NC19-6) spaced ~900 bp apart and another cluster includes two miRNAs (NC16-1 and NC16-2) spaced 5320 bp apart (Figure 4c). Paralogs are also found: NC17-5 appears to be a paralog of NC16-1 with variation in loop sequences (Table 3).

Table 1. Results of genome-wide screening of human miRNA

Chr	Extracted stem–loop structures	Expressed stem–loops	pre-miRNA candidates	Detected known miRNA
16	65539	8153	253	2
17	68458	9367	274	11
18	34853	3135	83	4
19	62229	7765	207	7

The second column shows the number of stem–loop structures extracted from each chromosome by the program RNAfold; the third column shows the number of stem–loops matched by EST search; the fourth column shows the number of pre-miRNA candidates screened from the stem–loops by ProMiR and the fifth column shows the number of published miRNA genes detected by ProMiR and the number of all known miRNA genes.

To provide experimental confirmation, we sought to detect the putative pri-miRNAs by RT–PCR (Figure 3e) (44). Because pri-miRNAs, which are the primary precursors for mature miRNA, are rapidly cleaved by the processing enzyme Drosha, the authentic pri-miRNAs would accumulate when Drosha is depleted in cells. Because this assay is based on PCR amplification, it can detect miRNA genes that are expressed at a relatively low level (Figure 4a) (44). Nine putative miRNPs were confirmed using this method (Figure 4b and Table 3). Seven of the 14 remaining candidates were not detected in the PCR experiment, which may have been because of their low abundance in HeLa cells. For further confirmation, cells from different tissues and developmental stages should be examined. The rest (seven) of the candidates did not accumulate, suggesting that these candidates are unlikely to be authentic miRNA genes.

The miRNA candidates confirmed by RT–PCR experiment have relatively distant homologous patterns and show diverse sequence patterns, compared with previously published miRNAs. However, these new miRNAs are not more diverse than other ncRNAs. These results are provided by the phylogenetic analysis for pre-miRNA sequences (Figure 5).

The results of real-time quantitative PCR indicate that some of the new miRNAs are expressed at low levels (C_T value 27–35, Table 2), and some miRNAs such as NC16-2 and NC19-5 may be expressed at higher levels (C_T value under 25). Low-abundant miRNAs may be difficult to detect with less sensitive methods such as northern blotting and may have escaped the conventional cloning (Table 2, Figure 6). Interestingly, the accumulation folds of pri-miRNAs vary significantly between the different miRNA genes. For instance, pri-miRNAs for Let7a-1, NC16-1, NC16-2 and NC18-3 accumulate more dramatically compared with those for miR-345 and NC17-5. This result suggests that pri-miRNA processing by Drosha may be differentially regulated in different miRNA genes.

Mature miRNA region prediction

We evaluated the accuracy of mature miRNA region prediction through 5-fold cross-validation with 136 known miRNAs (Table 4). The measures for assessment are the means of absolute distances and the square root of the mean of the squares. We found that for the 5′ strands ProMiR predicts the cleavage site at the non-loop side by Drosha more precisely than the cleavage site at the loop side by Dicer (mean absolute

Table 2. The real-time PCR results of final candidates

Candidates	Chr	C_T Mean (Luc)	C_T Mean (Dro)	Folds	C_T SD (Luc)	C_T SD (Dro)	MFE	*P*-value
NC16-1	16	34.1	29.3	27.9	0.66	0.32	−35.4	0.001
NC16-2	16	31.3	24.47	113.8	0.12	0.17	−45.2	0.002
NC16-3	16	39.37	35.59	13.7	0.76	1.4	−31.3	0.001
NC16-4	16	40.37	39.48	1.9	0.14	1.2	−31.6	0.008
NC16-5	16	—	—	—	—	—	−33.5	0.001
NC17-1	17	—	—	—	—	—	−30.0	0.001
NC17-2	17	29.1	28.14	1.9	0.12	0.12	−38.1	0.007
NC17-3	17	34.15	34.2	1.0	0.19	0.07	−40.9	0.014
NC17-4	17	—	—	—	—	—	−41.8	0.002
NC17-5	17	29.59	27.91	3.2	0.04	0.12	−26.6	0.004
NC17-6	17	33.28	32.27	2.0	0.58	0.21	−32.1	0.003
NC17-7	17	34.33	33.2	2.2	0.32	0.12	−33.5	0.002
NC17-8	17	—	—	—	—	—	−31.4	0.003
NC17-9	17	28.71	26.48	4.7	0.14	0.2	−48.8	0.001
NC18-1	18	34.41	34.47	1.0	1	—	−50.1	0.001
NC18-2	18	38	35.8	4.6	1.1	0.64	−47.2	0.001
NC18-3	18	39.6	33.1	90.5	1	0.89	−26.9	0.001
NC19-1	19	—	—	—	—	—	−41.1	0.034
NC19-2	19	26.78	25.68	2.1	1.03	0.01	−37.5	0.003
NC19-3	19	—	—	—	—	—	−40.5	0.005
NC19-4	19	—	—	—	—	—	−33.9	0.002
NC19-5	19	27.52	25.19	5.0	0.11	0.17	−26.7	0.013
NC19-6	19	30.17	28.23	3.8	0.23	0.09	−37.7	0.001
Let7a-1		29.89	24.58	39.7	0.44	0.13	−36.2	0.001
mir-345		31.42	30.1	2.5	0.24	0.33	−51.3	0.002
GAPDH		13.92	13.95	1.0	0.11	0.2	—	—

The gray rows indicate pre-miRNA candidates with meaningful differences >2.5-fold, mir-345's fold. The '—' indicate no PCR products. The first column gives indexes of the candidates; the third and fourth columns show the mean threshold cycle values (C_T) for each candidate of control cells and Drosha knocked down cells, respectively; the fifth column shows the ratio of relative quantity between control and Drosha knocked down cells; the sixth and seventh columns show the standard deviations of the C_T values for the control cells and the Drosha knocked down cells, respectively; the eighth column indicates the minimum folding energy (MFE, kcal/mol); the last column gives their *P*-values calculated by the randfold.

distance 1.96 versus 2.47 nt). For the 3′ strand, however, the end at the loop side is predicted more precisely than the end at the non-loop side (mean absolute distance 1.60 versus 2.13 nt). This indicates that the 3′ protruding ends generated by RNase III may be more variable than the 5′ recessive ends. Alternatively, the 3′ protruding ends may be subjected to additional modification in cells (decay or addition of extra nucleotides). The statistical signal of the cleavage site at the non-loop side of mature miRNA is relatively more dominant over the one at the loop side. This suggests that Drosha may be more important in determining the sequences of mature miRNA than Dicer.

We correctly predicted the orientation of the mature miRNA region for the 57/81 5′ sense strand and 41/55 3′ antisense strand pre-miRNAs. The mean accuracy was 72%.

Permutation test for the learning model

From the results so far, we can conclude that ProMiR effectively detects the cleavage signal recognized by Drosha. However, it is difficult to judge where the major cleavage signal originated. We designed a random permutation test to investigate whether the high specificity obtained by our model is caused by the base composition or the structure. Thus, we compared the change of efficiency for the trained model during 10 random permutations of the base pairs and bases in the stem, respectively. To measure the effect of base mutation, we randomly changed the base without changing the base pair; the effect of structure was measured by changing the base pair. Figure 7 presents the result of this study. The ln(*P*) value produced by structural permutation rapidly decayed to far below the threshold [ln(0.033)] even when the number of permutations was only one. In contrast, ln(*P*) values measured by sequence permutation reduced a little at the first permutation and then fluctuated near the threshold value. Thus, the specificity of our algorithm is influenced more by the conserved structural signals, such as match, mismatch, deletion and insertion, than by conserved sequence information.

Comparison of efficiency with other approaches

Detection of the conserved primary motif or secondary structure is a straightforward approach for identifying new ncRNAs, especially, miRNAs (45,46). Several methods to search the common motifs in RNA sequences or protein sequences have been introduced. Profile HMM tools such as HMMer are based on the frequency and the transition probability of the sequences and are usually used to detect conserved primary motifs such as those for proteins and regulatory regions in multiple sequence alignment (47). This method shows the effectiveness of searching distantly homologous sequences. Covariance models such as INFERNAL are usually used to detect structurally conserved motifs (48,49). The success of covariance models depends on finely curated structural multiple alignments. In this experiment, we used MARNA, a method of multiple structural alignments to search for conserved secondary structures of ncRNAs (50). The esRCSG is a method recently introduced to optimize RNA common structural grammar using genetic programming, a form of evolutionary algorithm (38). This method does not need multiple alignment data but uses only primary sequence as input data.

Table 3. Secondary structures by mfold of the new pre-miRNAs verified experimentally; the underline fonts indicate the mature miRNA regions predicted by ProMiR

ID	New pre-miRNA structures	Notes
NC16-1	TCGT-| C AA AC GA TTTCCAT TAC GCAGGG AATGAGGG TTTTGGGGGCA TGTG ₩ ATG CGTTCT TTATTCCT AAAATCCCCGT ATAC T TCCTT^ A CG A- A- TATCACC	Clustered with NC16-2 Paralog to NC17-5
NC16-2	-| GA- CT GT T GT G A AG TTAT TG GG GTG CTCAGAA CGGG TTTGAGGGC AG TG T ₩ AC TC TAC GGGTTTT GCCC AAACTCCCG TC AC A G G^ AGG T- TG C TG G A CT TTTT	Clustered with NC16-1
NC16-3	TAA----| T - CC AG CTG TCCCTGCTT TATTTGT AGCTT CAA CTTTTG ₩ AGGGACGAA ATAAACA TTGAA GTT GAAAAT G AAGAAAC^ T G AT AG AAT	
NC17-5	GACA--------| A AC C ----- T T GCAAGAA AATGAGGG TTT AGGGGCA GCTG GTT T CGTTCTT TTATTCCT AAA TCCCCGT TGAC CAG C GGGCTACGTGGA^ G AA - AATAC T T	Paralog to NC16-1
NC17-9	TAA---| T AG- G A CTAT AGGAAGT AGGCTGAGGGGC AGA CGAG CTTTT ₩ TCCTTCG TCTGACTCCCCG TCT GCTC GAAAA T GCCCAA^ T GAG G - CCTT	
NC18-2	TTTTA | A - CCAA A AAAAAATTACTGGTGTCCA GCTCCACCC TGGA TT A TTTTTTAATGACTACAGGT CGGGGTGGG ATCT AG A AAAA- ^ C G CTCG T	
NC18-3	AGTGAGA| T- A C T AAAGAAATTTTTTGT GTT CAAAA ACCTTTTT A TTTCTTTGAAAAACA CGA GTTTT TGGAAAAA A AA-----^ TC - A C	
NC19-5	AAAAATAAAAATAA-| AAT - ---- TAA TG CA ATA TGTG GTAATTGTGTAACCAC AC T GT ₩ TGT ACGT CGTTAACACATTGGTG TG G CG A CGCGCCCAATAAAAC^ GT- A ATAG TTA GT AT	Clustered with NC19-6
NC19-6	----| - A A - ATTG CGTG GTA TTGTGTAACCACTATTATAATC CA CTA ₩ GCGT CGT AACACATTGGTGATAGTGTTAG GT GAT G GTGT^ A C - C AACT	Clustered with NC19-5

We compared the efficiency of our ProMiR model with four different approaches including a previous miRNA prediction method, which relies on the characteristic feature that the known miRNAs derive from conserved stem–loop structures (Table 5, see Supplementary Material). To perform a fair comparison, we trained each model with various amounts of data. This made it possible to search for the optimal result of each method. In this comparison, the HMMer method by multiple sequence alignment showed very low accuracy. This might have been caused by inappropriate alignment results, because miRNA genes have conserved structural motifs rather than conserved sequence motifs. The results for INFERNAL showed higher sensitivity when there were more training data, but the specificity decreased to 0.18. This low specificity might

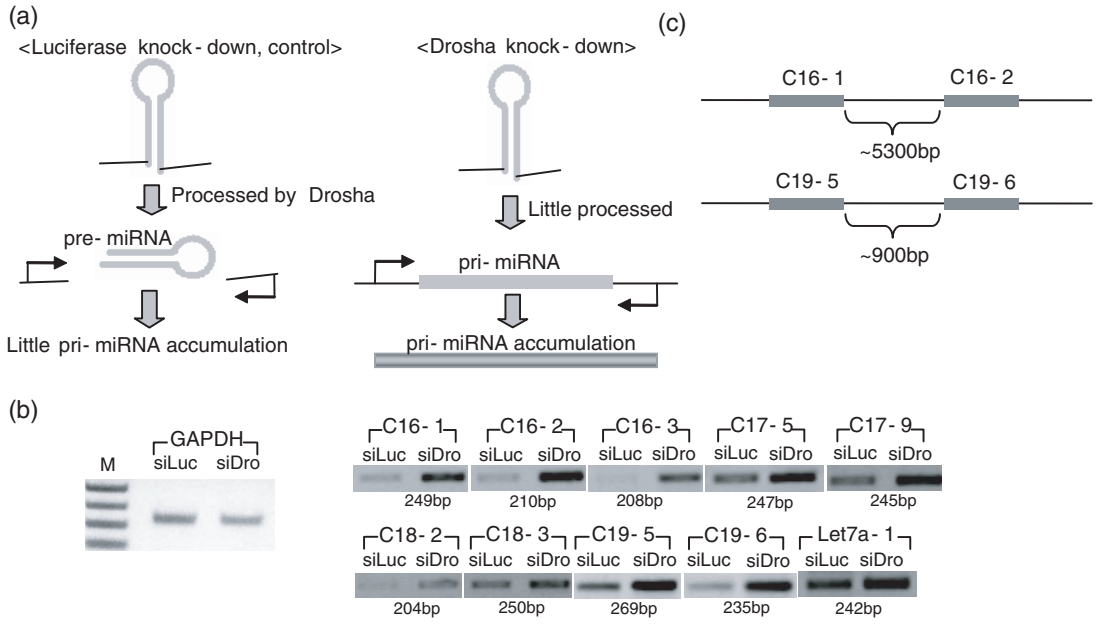

Figure 4. Experimental verification of the candidates predicted by ProMiR. (**a**) Comparison of the expected PCR results for the candidates in control and Drosha knockdown HeLa cells (**b**) The left gel image shows the PCR intensity of glyceraldehyde-3-phosphate dehydrogenase (GAPDH) for each cDNA. The first lane is the PCR result following silencing (si) RNA treatment for luciferase in HeLa cells for 3 days as the control and the second lane is the PCR result following siRNA for luciferase in HeLa cells for 3 days. (**c**) NC16-1 and 2, and NC19-5 and 6 are on a transcript that contains two pre-miRNAs, respectively.

be because, when given more training data, the method learns to recognize structures that are more common. Next, contrary to our expectation, a method based on conservation did not show improved performance. The lower sensitivity clearly demonstrates that it is not suitable to predict unrelated (not conserved) miRNAs. Finally, the results using esRCSG showed more effective prediction than the other methods using sequence or structural alignment. However, esRCSG gave lower sensitivity (0.67) than ProMiR (0.73) using probabilistic co-learning of sequence and structure.

In contrast to the previous methods, the result of the comparison clearly demonstrates that our probabilistic method is more effective than previous methods and can be applied to identify miRNA with close homology as well as unrelated miRNAs (miRNAs with distant homology). Main reasons for the advanced results can be explained by several criteria. (i) The pairwise model considering structure and sequence of extended stem–loops. (ii) A sensible negative dataset; an approach to conserved structural motifs and MFE assessment, realistically using the MFE as evidence for miRNP. (iii) Additional evidence such as sequence conservation among vertebrates.

DISCUSSION

We have shown that probabilistic co-learning of structure and sequence is an effective method for the identification of miRNA genes with close or distant homology. It could also spontaneously predict mature miRNA regions. Another merit of the probabilistic model is that it provides a common method for identifying human miRNA genes as well as those from other species.

In the genome-wide screening of human miRNA, we determined a *P*-value threshold (0.033) from the screening performance ROC curve shown in Figure 2, where the sensitivity was 72.8% and the specificity was 95.9%, to minimize the number of false positive findings and maximize the number of true positive findings. The major reason for the requirement of high specificity is that genomes are very complex with noisy sequences such as repeat sequences, palindromic sequences, pseudogenes and transposons. Thus, the screening method should have a stringent classifier. Of course, we can adjust this threshold to fine-tune miRNA prediction. We performed human EST analysis and MFE tests to select the more probable candidates. Additional resources, such as vertebrate conservations and human repeat sequences, made it possible to screen for more reliable candidates. Finally, we verified the candidates by detecting the accumulation of pri-miRNA. The main purpose of the experiment was to determine whether the pre-miRNA candidates are indeed authentic Drosha substrates. Therefore, the experiments clearly demonstrate that our candidates are true positives (Figure 4). However, we still do not know whether the candidates not verified by the experiment are truly negative, because the candidates might be expressed in specific cell types but not in the tested cells.

The mean error of the mature miRNA region prediction results was 2.7 nt and the mean variation except for the 20 prediction failures was 2.0 nt. The main reasons for the error may have originated from inaccuracy of the cleavage of the pre-miRNA by Dicer, which bears an error of 1 nt and from overhanging ends of 2 nt at the 3′ end (51). In addition, there are several instances of incompatible data for the locations of mature miRNAs in the miRNA database (http://www.sanger.

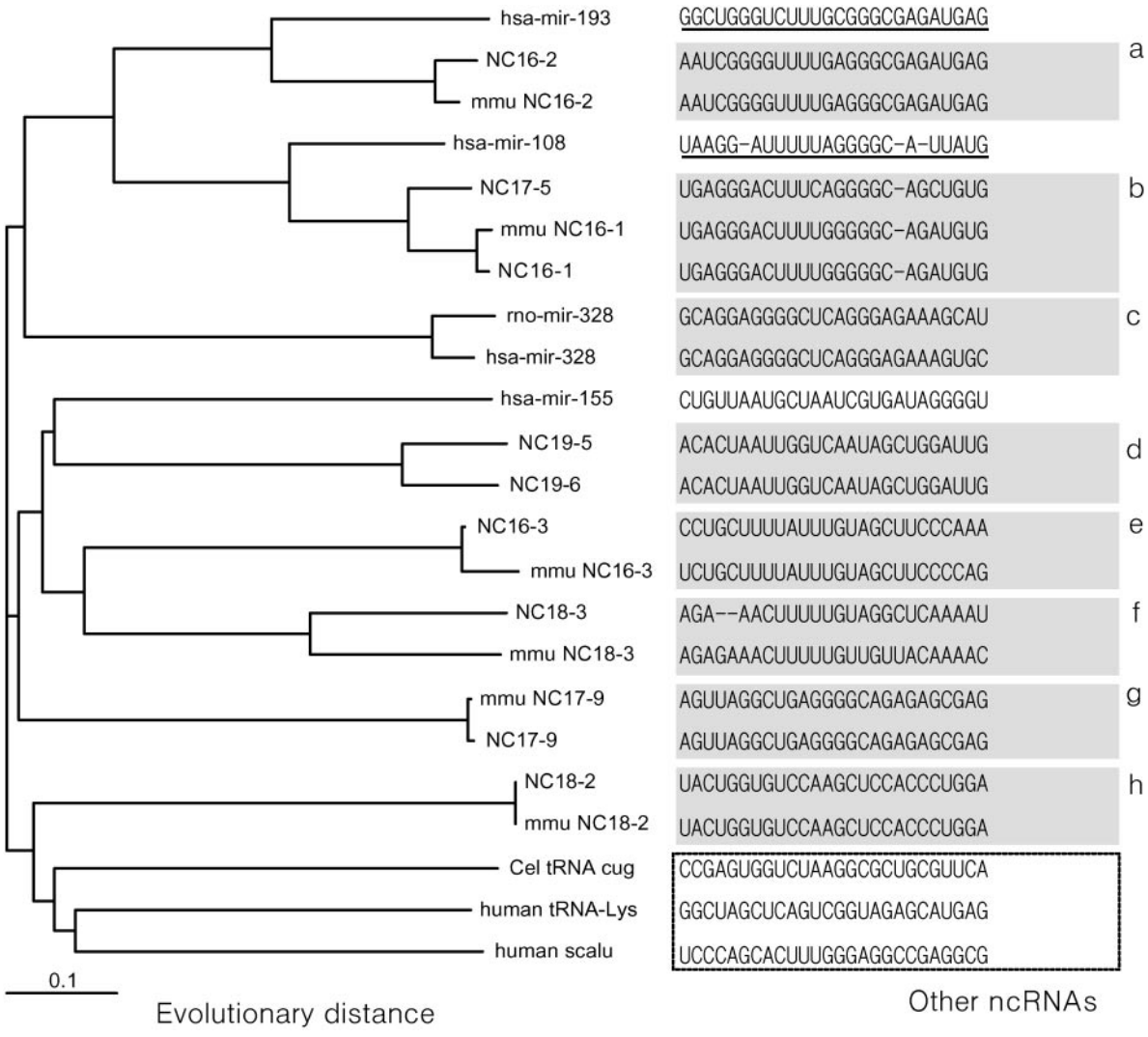

Figure 5. Phylogenetic tree for the nine new pre-miRNAs, several published miRNAs and other ncRNAs. The tree was drawn using the neighbor-joining method. The gray boxes in sequences indicate closely homologous (orthologs + paralogs) members. The dotted line box contains other ncRNA sequences, i.e. tRNAs and scAlu RNA. The new pre-miRNAs have distant homologous patterns to published pre-miRNAs—'a' homologous group shows distantly homologous pattern with has-mir-193 and the 'b' homologous group is a distant homolog to has-mir-108; however, they are relatively closer together than other ncRNAs.

ac.uk/Software/Rfam/mirna/) that may lead to some errors in mature miRNA region prediction. When these limitations are considered, the result indicates that our algorithm gives meaningful results for the prediction of mature miRNA regions over pre-miRNAs.

The prediction of a functional strand on precursors is a problem related with region prediction. The internal stability of 5′ terminal base pairs for mature miRNA improved the effectiveness of prediction of the functional strand (14,15). However, ~25% of the prediction results were false despite the clear criterion of internal stability. This is caused by the exceptional characteristics of pre-miRNA. Most mature miRNAs on the precursors are located in either the 5′ strand or the 3′ strand. However, some of the known miRNAs exist in both the strands simultaneously. In addition, most pre-miRNAs have extended stem–loop structures, but a few pre-miRNAs have branched stem–loop secondary structures. These exceptional features of miRNAs make it difficult to predict the orientation as well as the region of mature miRNAs. Moreover, because even 1 nt error in predicting the mature region can result in the change of functional strands, it is more difficult to verify the mature miRNAs of the correct length. We firmly feel that the improvement of our predictive model to overcome the limitations in terms of mature miRNA prediction may make it possible to address the problems.

Our study identified at least nine novel miRNA gene candidates from four chromosomes. Importantly, these novel miRNA candidates are not related to previously reported miRNAs. These new miRNAs may define a novel subfamily of miRNA genes as they have homologs and/or paralogs in many vertebrate genomes. Further refinement of our model and a more extensive screening is likely to yield a significant number of novel miRNAs from various organisms. Effective miRNA gene mining will greatly enhance our knowledge of small RNA-mediated regulatory networks.

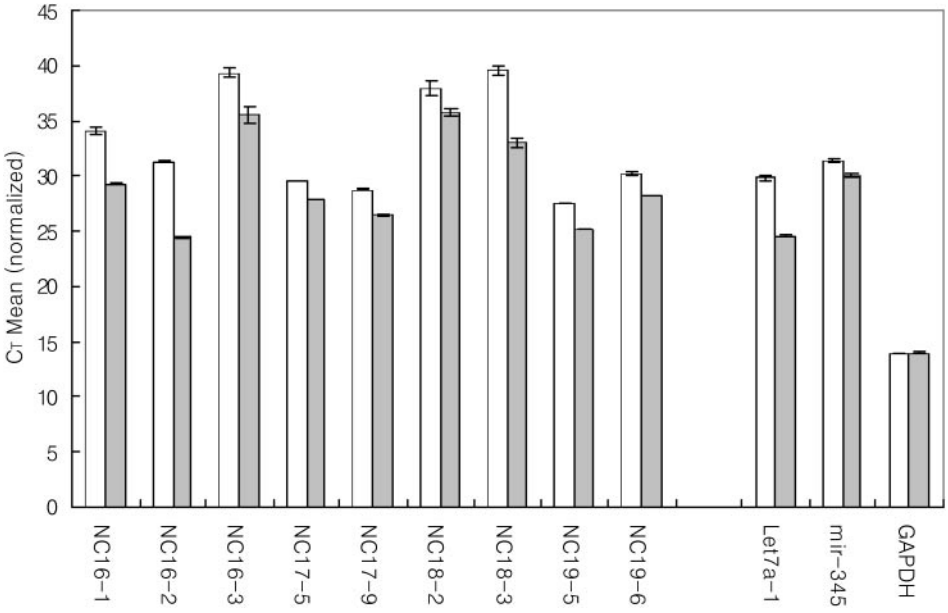

Figure 6. The differences of threshold cycle (C_T) between control and Drosha knocked down cells.

Table 4. Results of mature miRNA region prediction for 5-fold cross-validation

	Mean of absolute distance				Square root of the mean of the squares			
	5′ Strand		3′ Strand		5′ Strand		3′ Strand	
	Unlooped	Loop	Unlooped	Loop	Unlooped	Loop	Unlooped	Loop
Total	2.83 (nt)	3.31	2.42	2.15	4.16	5.11	3.32	3.65
Total except failures (68 + 48)	1.96	2.47	2.13	1.60	2.56	3.26	2.70	2.14

Total number of the cross-validation set is 136 published miRNAs. The last row of the table shows the result except for 20 prediction failures. Prediction failures imply that a decision cannot be made because the defined signal $S(i)$ is too weak.

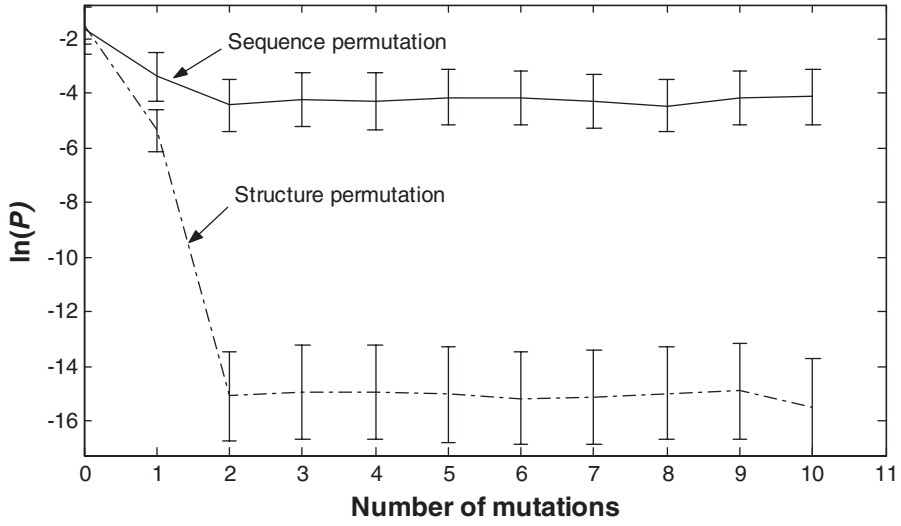

Figure 7. Permutation test for the structure and sequence of mature miRNA. The solid line indicates the change of probability P according to base permutation. The dotted line indicates the change of probability according to base pair permutation.

Table 5. Comparison of the efficiency of miRNA prediction

	Training data	Sensitivity	Specificity
HMMer	10	0.03	1.00
	30	0.00	0.00
	50	0.00	0.00
	68	0.00	0.00
INFERNAL	30	0.68 (0.00)[a]	0.50 (0.00)
	50	0.91 (0.00)	0.30 (0.00)
	68	0.94 (0.00)	0.18 (0.00)
Conservation[b]	68	0.34	0.87
esRCSG	50	0.36 (0.67)[c]	0.96 (0.89)
ProMiR	68	0.69	0.94
	5-fold cross validation	0.73	0.96

[a]Results by sequential and structural multiple alignment.
[b]Results by observing conservation of pre-miRNA.
[c]Results by adding suboptimal RCSGs.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

This work was supported by the National Research Laboratory programs (M10412000095-04J0000-03610 and M1050000010905J000010910) and the Systems Biology project (M10309000002-03B5000-00110) of the Korea Ministry of Science and Technology and the BK21-IT program of the Korean Ministry of Education. The ICT at Seoul National University provided research facilities for this study. Funding to pay the Open Access publication charges for this article was provided by the Korea Ministry of Science and Technology under the NRL program.

Conflict of interest statement. None declared.

REFERENCES

1. Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.
2. Kim,V.N. (2005) Small RNAs: classification, biogenesis, and function. *Mol. Cells*, **19**, 1–15.
3. Lee,Y., Kim,M., Han,J., Yeom,K.H., Lee,S., Baek,S.H. and Kim,V.N. (2004) MicroRNA genes are transcribed by RNA polymerase II. *EMBO J.*, **23**, 4051–4060.
4. Cai,X., Hagedorn,C.H. and Cullen,B.R. (2004) Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *RNA*, **10**, 1957–1966.
5. Lee,Y., Ahn,C., Han,J., Choi,H., Kim,J., Yim,J., Lee,J., Provost,P., Radmark,O., Kim,S. and Kim,V.N. (2003) The nuclear RNase III Drosha initiates microRNA processing. *Nature*, **425**, 415–419.
6. Lund,E., Guttinger,S., Calado,A., Dahlberg,J.E. and Kutay,U. (2004) Nuclear export of microRNA precursors. *Science*, **303**, 95–98.
7. Yi,R., Qin,Y., Macara,I.G. and Cullen,B.R. (2003) Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. *Genes Dev.*, **17**, 3011–3016.
8. Bohnsack,M.T., Czaplinski,K. and Gorlich,D. (2004) Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs. *RNA*, **10**, 185–191.
9. Bernstein,E., Caudy,A.A., Hammond,S.M. and Hannon,G.J. (2001) Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature*, **409**, 363–366.
10. Grishok,A., Pasquinelli,A.E., Conte,D., Li,N., Parrish,S., Ha,I., Baillie,D.L., Fire,A., Ruvkun,G. and Mello,C.C. (2001) Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control *C. elegans* developmental timing. *Cell*, **106**, 23–34.
11. Hutvagner,G., McLachlan,J., Pasquinelli,A.E., Balint,E., Tuschl,T. and Zamore,P.D. (2001) A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. *Science*, **293**, 834–838.
12. Ketting,R.F., Fischer,S.E., Bernstein,E., Sijen,T., Hannon,G.J. and Plasterk,R.H. (2001) Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in *C.elegans*. *Genes Dev.*, **15**, 2654–2659.
13. Knight,S.W. and Bass,B.L. (2001) A role for the RNase III enzyme DCR-1 in RNA interference and germ line development in *Caenorhabditis elegans*. *Science*, **293**, 2269–2271.
14. Khvorova,A., Reynolds,A. and Jayasena,S.D. (2003) Functional siRNAs and miRNAs exhibit strand bias. *Cell*, **115**, 209–216.
15. Schwarz,D.S., Hutvagner,G., Du,T., Xu,Z., Aronin,N. and Zamore,P.D. (2003) Asymmetry in the assembly of the RNAi enzyme complex. *Cell*, **115**, 199–208.
16. Lagos-Quintana,M., Rauhut,R., Lendeckel,W. and Tuschl,T. (2001) Identification of novel genes coding for small expressed RNAs. *Science*, **294**, 853–858.
17. 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.
18. Mourelatos,Z., Dostie,J., Paushkin,S., Sharma,A., Charroux,B., Abel,L., Rappsilber,J., Mann,M. and Dreyfuss,G. (2002) miRNPs: a novel class of ribonucleoproteins containing numerous microRNAs. *Genes Dev.*, **16**, 720–728.
19. Calin,G.A., Dumitru,C.D., Shimizu,M., Bichi,R., Zupo,S., Noch,E., Aldler,H., Rattan,S., Keating,M., Rai,K. *et al.* (2002) Frequent deletions and down-regulation of micro-RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. *Proc. Natl Acad. Sci. USA*, **99**, 15524–15529.
20. Houbaviy,H.B., Murray,M.F. and Sharp,P.A. (2003) Embryonic stem cell-specific MicroRNAs. *Dev. Cell*, **5**, 351–358.
21. Dostie,J., Mourelatos,Z., Yang,M., Sharma,A. and Dreyfuss,G. (2003) Numerous microRNPs in neuronal cells containing novel microRNAs. *RNA*, **9**, 180–186.
22. Michael,M.Z., O'Connor,S.M., van Holst Pellekaan,N.G., Young,G.P. and James,R.J. (2003) Reduced accumulation of specific microRNAs in colorectal neoplasia. *Mol. Cancer Res.*, **1**, 882–891.
23. Lim,L.P., Glasner,M.E., Yekta,S., Burge,C.B. and Bartel,D.P. (2003) Vertebrate microRNA genes. *Science*, **299**, 1540.
24. Lagos-Quintana,M., Rauhut,R., Meyer,J., Borkhardt,A. and Tuschl,T. (2003) New microRNAs from mouse and human. *RNA*, **9**, 175–179.
25. Kim,J., Krichevsky,A., Grad,Y., Hayes,G.D., Kosik,K.S., Church,G.M. and Ruvkun,G. (2004) Identification of many microRNAs that copurify with polyribosomes in mammalian neurons. *Proc. Natl Acad. Sci. USA*, **101**, 360–365.
26. Kasashima,K., Nakamura,Y. and Kozu,T. (2004) Altered expression profiles of microRNAs during TPA-induced differentiation of HL-60 cells. *Biochem. Biophys. Res. Commun.*, **322**, 403–410.
27. Suh,M.R., Lee,Y., Kim,J.Y., Kim,S.K., Moon,S.H., Lee,J.Y., Cha,K.Y., Chung,H.M., Yoon,H.S., Moon,S.Y. *et al.* (2004) Human embryonic stem cells express a unique set of microRNAs. *Dev. Biol.*, **270**, 488–498.
28. Poy,M.N., Eliasson,L., Krutzfeldt,J., Kuwajima,S., Ma,X., Macdonald,P.E., Pfeffer,S., Tuschl,T., Rajewsky,N., Rorsman,P. and Stoffel,M. (2004) A pancreatic islet-specific microRNA regulates insulin secretion. *Nature*, **432**, 226–230.
29. Weber,M.J. (2005) New human and mouse microRNA genes found by homology search. *FEBS J.*, **272**, 59–73.
30. Xie,X., Lu,J., Kulbokas,E.J., Golub,T.R., Mootha,V., Lindblad-Toh,K., Lander,E.S. and Kellis,M. (2005) Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature*, **434**, 338–345.
31. Griffiths-Jones,S. (2004) The microRNA registry. *Nucleic Acids Res.*, **32**, D109–D111.
32. Lai,E.C., Tomancak,P., Williams,R.W. and Rubin,G.M. (2003) Computational identification of Drosophila microRNA genes. *Genome Biol.*, **4**, R42.

33. Legendre,M., Lambert,A. and Gautheret,D. (2005) Profile-based detection of microRNA precursors in animal genomes. *Bioinformatics*, **21**, 841–845.
34. Ohler,U., Yekta,S., Lim,L.P., Bartel,D.P. and Burge,C.B. (2004) Patterns of flanking sequence conservation and a characteristic upstream motif for microRNA gene identification. *RNA*, **10**, 1309–1322.
35. Lim,L.P., Lau,N.C., Weinstein,E.G., Abdelhakim,A., Yekta,S., Rhoades,M.W., Burge,C.B. and Bartel,D.P. (2003) The microRNAs of *Caenorhabditis elegans*. *Genes Dev.*, **2**, 2.
36. Krol,J., Sobczak,K., Wilczynska,U., Drath,M., Jasinska,A., Kaczynska,D. and Krzyzosiak,W.J. (2004) Structural features of microRNA (miRNA) precursors and their relevance to miRNA biogenesis and small interfering RNA/short hairpin RNA design. *J. Biol. Chem.*, **279**, 42230–42239.
37. Hofacker,I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res.*, **31**, 3429–3431.
38. Nam,J.W., Joung,J.G., Ahn,Y.S. and Zhang,B.T. (2004) Two-step genetic programming for optimization of RNA common-structure. *LNCS*, **3005**, 73–83.
39. Forney,G.D.,Jr (1973) The Viterbi Algorithm. *Proc. IEEE*, **61**, 268–278.
40. Chen,J.H., Le,S.Y., Shapiro,B., Currey,K.M. and Maizel,J.V. (1990) A computational procedure for assessing the significance of RNA secondary structure. *Comput. Appl. Biosci.*, **6**, 7–18.
41. Bonnet,E., Wuyts,J., Rouze,P. and Van de Peer,Y. (2004) Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics*, **20**, 2911–2917.
42. Rivas,E. and Eddy,S.R. (2000) Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics*, **16**, 583–605.
43. Berezikov,E., Guryev,V., van de Belt,J., Wienholds,E., Plasterk,R.H. and Cuppen,E. (2005) Phylogenetic shadowing and computational identification of human microRNA genes. *Cell*, **120**, 21–24.
44. Lee,Y., Jeon,K., Lee,J.T., Kim,S. and Kim,V.N. (2002) MicroRNA maturation: stepwise processing and subcellular localization. *EMBO J.*, **21**, 4663–4670.
45. Griffiths-Jones,S., Bateman,A., Marshall,M., Khanna,A. and Eddy,S.R. (2003) Rfam: an RNA family database. *Nucleic Acids Res.*, **31**, 439–441.
46. Klein,R.J. and Eddy,S.R. (2003) RSEARCH: finding homologs of single structured RNA sequences. *BMC Bioinformatics*, **4**, 44.
47. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
48. Eddy,S.R. and Durbin,R. (1994) RNA sequence analysis using covariance models. *Nucleic Acids Res.*, **22**, 2079–2088.
49. Eddy,S.R. (2002) A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. *BMC Bioinformatics*, **3**, 18.
50. Siebert,S. and Backofen,R. (2003) MARNA: A server for multiple alignment of RNAs. In Mewes,H.-W., Heun,V., Frishman,D. and Kramer,S., (eds). *Proceedings of the German Conference on Bioinformatics. GCB 2003*. Vol. 1 belleville Verlag Michael Farin, München, pp. 135–140.
51. Zamore,P.D. (2002) Ancient pathways programmed by small RNAs. *Science*, **296**, 1265–1269.