공학석사학위논문

microRNA 예측을 위한 계산학적 방법

The Computational Methods for Prediction of microRNA

2004년 8월

서울대학교 대학원

협동과정 생물정보학

남 진 우

microRNA 예측을 위한 계산학적 방법

The Computational Methods for Prediction of microRNA

지도 교수 장 병 탁

이 논문을 공학석사 학위논문으로 제출함

2004년 7월
서울대학교 대학원
협동과정 생물정보학

남 진 우
남진우의 공학석사 학위논문을 인준함
2004년 7월

위원장 손 현 석 印
부위원장 장 병 탁 印
위 원 김 규 원 印

ABSTRACT

microRNA (miRNA) is a class of small non-coding RNAs and has known as the first post-transcriptional regulator which directly regulates gene expression by arresting the messenger RNA (mRNA) translation. In first, the microRNAs are transcribed as microRNA precursor of over 100 nucleotides (nt) but they are processed as mature microRNAs of ~22nt by Drosha and Dicer. Recently, miRNAs are identified by experimental methods from various species. However, the experimental identification of miRNA genes requires a number of resources.

In this paper, two-step genetic programming algorithm and probabilistic co-learning algorithm are introduced as general prediction methods which can be applied to all species' microRNA and other non-coding RNA. Our methods can not only predict microRNA precursors on genome sequence but also predict the position of mature microRNA over precursor with error rate of 2.0nt. The methods show a higher specificity and sensitivity than other methods which learn either only structure information or only sequence information. Using the methods, we screened the candidates of microRNA from human chromosome 18 and 19 with screening threshold, 0.033. In this result, we got 255 candidates of microRNA, whose expressions are verified by EST analysis, and the algorithm predicted 13 of 16 microRNAs known on chromosome 18 and 19 as positive. These results show that our approaches have a high specificity and sensitivity for microRNA prediction.

Contents

LIST OF FIGURES

LIST OF TABLES

I. Introduction

Recent advances in the small RNA research have implicated microRNAs (miRNAs) that function as anti-sense regulators of mRNA. The miRNAs constitute a large family of non-coding small RNA of about 22 nucleotides in length. The miRNA is a sort of small RNA, first known by the fact that it directly takes part in the post-transcriptional regulation (Ambros, 2001). The miRNA studies, such as the identification of miRNA, the target prediction and functional study, are actively accomplished in biology field (Ambros *et al*., 1994; Lee *et al*., 2001, Lagos-Quintana *et al*., 2001; Rhoades *et al*., 2002; Lai *et al*., 2002; Lewis *et al*., 2003; Enright *et al*., 2003; Pfeffer *et al*., 2004; Griffiths-Jones, 2004). Especially, the identification of miRNA genes is the most fundamental study of the related researches and mostly achieved by experimental approaches such as northern blot, clone library and microRNP (miRNP) separation (Lim *et al*., 2003; Dostie *et al*., 2003; Lagos-Quintana *et al*., 2003). In human, 191 miRNAs have been reported so far (Griffiths-Jones, 2004). However, a lot of miRNAs are tissue-specifically or development time dependently expressed, and some are even expressed with low level at specific time. These problems are major reasons that the experimental methods fail to identify much miRNA genes.

Another import problem of miRNA is to predict mature miRNA regions over miRNA precursors by detecting significant signals, such as Dicer binding motifs (Song *et al*., 2003, Lee *et al*., 2004), Drosha recognition motif (Lee *et al*., 2003)

and miRNP complex protein binding motifs (Dostie *et al*., 2003). The miRNA precursors are processed to generate the final products of ~22 nt single-stranded mature miRNAs by another RNase III type enzyme, Dicer in cytoplasm (Lee *et al*., 2002). Dicer and Drosha recognize the structurally conserved region around the cleavage site on stem-loop precursor and liberate miRNAs from the stem-loop precursors (Zamore *et al*., 2000; Bernstein *et al*., 2001; Lee *et al*, 2003). Computational approaches for mature miRNA prediction are useful not only to support experimental results but also to screen unknown miRNA genes which are not screened with experimental method. The computational approaches can help us to identify the miRNA genes with less time and labor-work than experimental methods. However miRNA genes are diverse and heterogeneous in the sequence patterns. The statistical information of miRNA genes is insufficient to identify miRNA genes, which makes it difficult to predict miRNAs using computational methods. Though this difficulty, some approaches are introduced in previous works of some groups.

First, it is an approach as comparative analysis. Lately, two groups developed statistical measures and comparative methods to identify homologous miRNA precursors with known miRNA precursors of related species, such as drosophila and c. elegance (Lim *et al*., 2003; Lai *et al*., 2003). Though the approaches provided convenient method to detect homologous miRNA genes, they could not find unknown miRNA genes without known homologous miRNA, since the comparative approaches search only the homologous miRNA precursors.

The second approach is to find common structure or motif of ncRNA genes. To detect the conserved primary motif or secondary structure is a straight-forward approach for identification of new targets such as gene, regulatory motif, protein, RNA and chemical (Bernardo *et al*., 2003; Klein *et al*., 2003; Griffiths-Jones *et al*., 2003). To achieve this aim, some methods have been introduced. (1): Profile HMMs such as HMMer is usually used to detect conserved primary motif likes protein motif and regulatory motif in multiple sequence alignment (Eddy, 1998). However, because profile HMMs is based on the transition probability and the emission probability of the sequences, it is not easy to show the good efficiency in the prediction problem of miRNA conserved structurally. (2): Covariance model such as INFERNAL is usually used to detect structurally conserved motif likes RNA secondary structure in structural multiple alignment (Eddy *et al*., 1994; Eddy, 2002). The success of covariance model depends on finely curated structural multiple alignments. "MARNA" is famous as method of multiple structural alignments to search conserved secondary structure of ncRNAs (Siebert *et al*., 2003).

In this study, we introduce two new approaches to predict miRNA. The first method is esRCSG, which is a method recently introduced to optimize RNA common-structural grammar using genetic programming, a kind of evolutionary algorithm (Nam *et al*., 2004a; Nam *et al*., 2004b). This method does not need multiple alignment data but use only primary sequence as input data. The second method is a probabilistic co-learning of the structure and the sequence for miRNA

genes. Probabilistic co-learning makes it possible to develop a computational method designed to effectively and generally search miRNA genes containing both highly conserved structure and weakly conserved sequence. The merit of this approach is that this probabilistic model can simultaneously provide a method to identify miRNA genes, to predict the mature region and to determine the orientation of mature miRNA over miRNA precursor.

II. Two-Step Genetic Programming

II-1. Genetic Programming

```
begin
    t = 0                                   /* generation */
    initialize P(t)                         /* population */
    evaluate P(t)
    while (not termination-condition) do
    begin
        S = S + above(P(t))
        t = t+1
        select P(t) from P(t-1)                          /* selection */
        crossover-mutate P(t) except Best                /* genetic operators */
        evaluate P(t)                                    /* fitness function */
    end
end
```

Genetic programming is an automated method for creating a working genetic program, which is called individual and generally represented by tree structure, from a high-level problem statement of a problem (Koza *et al*., 1992). A genetic tree consists of elements from a function set and a terminal set. Function symbols appear as internal nodes. Terminal symbols are used to denote actions taken by the program. Genetic programming does this by genetically breeding a population of genetic programs using the principles of Darwinian natural selection and biologically inspired operations. Genetic programming uses crossover and mutation as the transformation operators, which can endow variation to genotype, to change candidate solutions into new candidate solutions as above pseudo-code.

II-2. RNA Common Structural Grammar (RCSG)

Some research groups have developed the structural grammar as context free type or context sensitive type to express the structure and sequence of RNA (Sakakibara *et al*., 1994; Knudsen *et al*., 1999; Cai *et al*., 2003). Using the structural grammar, we can easily and formally express secondary structures as well as primary motif. Macke et al. introduced the structure definition language, which is used in RNAMotif (Thomas *et al*., 2001). The language can not only easily represent various RNA secondary structure elements such as loop, bulge, stem and mispair, but also include some sequence features such as conserved motif and mismatch number. This structure definition language allows abstraction of the structural pattern into a 'descriptor' with a pattern language so that it can give detailed information regarding base pairing, length and sequence.

On the other hand, the structural grammar and structure definition language can expressed the common structures and conserved sequences learned from multiple sequence set. We named the structural grammar describing RNA common structure for RNA Common Structural Grammar (RCSG). There are various approaches to optimize the RCSGs but we apply genetic programming, a kind of evolutionary algorithms, with focus on two steps, structural learning step and sequence learning step. Because genetic programming uses tree structure as individuals, we should transform the structural grammar to function tree to optimize RCSG.

(a)

Fun	Grammar	Variable
f1	h5 (f1 or f2) h3	minlen/maxlen, len, mispair, seq, mismatch
f2	ss	minlen/maxlen, len, seq, mismatch
root	descr	

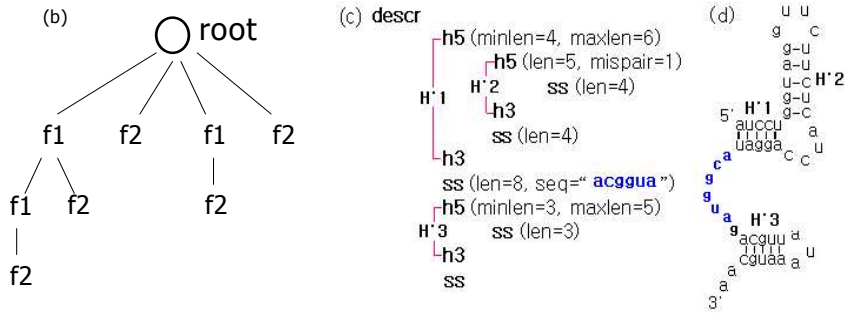

Fig. 1. RNA Common Structural Grammar and Tree Representation. (a): Function f1 generates recursively structural grammar, including one helix structure and either f1 or f2 as next deviation. Function f2 only represents ss (single strand), which means loop, bulge and single strand. Both f1 and f2 contain some variables, which measure structural information such as the length of helix (len), the number of pair (mispair) and mismatch, and sequence (seq). (b): A function tree to which can be converted into structural grammar (c): The child nodes of root in (b) conform to the first indentation of (c) and the nodes of second depth conform to the second indentation of (c). One helix that consists of the pair of h5-h3, h means helix, 5 and 3 mean 5' end and 3' end) (d): Secondary RNA structure is represented by structural grammar (c). H1, H2 and H3 in (c) and (d) are helix structures.

In order to convert structural grammars into function trees, we have defined the function f1, f2 and root as shown in Figure 1a. These functions can be formulated by some expression rules (Figure 1a). Therefore, using the expression rules, we can represent the structural grammars (Figure 1c) as function trees

(Figure 1b). However, we should consider the creation of invalid structural grammars converted from function trees. In order to avoid creation of invalid structural grammars, function trees have some constraints about the order of the function and the terminal node. First, f2 function should not appear consecutively in the same depth of the tree, contiguous f2 functions can be considered as only one. Second, f2 function can only appear as terminal node to terminate recursive generation of function tree. Finally, variables 'minlen' and 'maxlen' should always come in pair and should not coexist with variable 'len.'

II-3. Two-Step Genetic Programming (esRCSG)

To identify the putative ncRNAs in genome database, we have tried to find common-structure conserved among ncRNAs. We have implemented a program for learning of common-structure, which is devised by genetic programming. We named the program esRCSG, evolutionary search for RNA Common-Structural Grammar. The algorithm of esRCSG can be illustrated by the pseudo-code in Figure 2. The algorithm consists of two-step; a structural learning which optimizes only tree structure of grammar without sequence and a learning of sequence that specifies RCSGs of structural learning by incorporating a word, fragment of sequence, into the RCSGs.

```
begin                    /* Structural Learning */
      t = 0              /* generation */
      initialize P(t)    /* population */
      convert P(t)       /* tree to grammar*/
      evaluate P(t)
      while (not termination-condition) do
      begin
         S = S + above(P(t))                  /* Top group for Seq learning*/
         t = t+1
         select P(t) from P(t-1)              /* selection */
         crossover-mutate P(t) except Best    /* genetic operators */
         convert P(t)
         evaluate P(t)                        /* fitness function */
         if (local search)
            while (not termination-condition) do
               j = j + 1
               P_j(t) = mutate P(t)
                     if (evaluate P(t) < evaluate P_j(t) )
                           P(t) = P_j(t)
      end
end
w = wordwise(training data)
begin                                  /* Learning of Sequence */
      t = 0                            /* generation */
      initialize S(t) from S with w    /* population */
      convert S(t)
      evaluate S(t)
      while (not termination-condition) do
      begin
         t = t+1
         select S(t) from S(t-1)                    /* selection */
         mutate S(t) for only seq. except Best      /* genetic operators */
         convert S(t)
         evaluate S(t)
      end
end
```

Fig. 2. The pseudo-code for RCSG optimization algorithm. The pseudo-code for RCSG optimization algorithm. The algorithm consists of two steps: (1) learning RCSG from a set of training data, known miRNA precursors without sequence -related variables (such as “seq” and “mismatch”); (2) Optimizing RCSG which is learned in structural learning by incorporating the sequence-related variables. The "wordwise“ method randomly splits sequences of training data set into 7-mer words. Initialization of this step is accomplished by incorporating the words into RCSGs that are learned in structural learning.

Both steps (1) (2) are implementations (or instances) of genetic programming and both share many of the procedures. The common procedures of (1) and (2) can be described as follows: (a) initialize the population with randomly generated trees; (b) convert all function trees into structural grammars; (c) calculate the fitness, specificity, sensitivity, and complexity for all grammars with the positive and negative training data set; (d) evaluate all structural grammars using the sensitivity, specificity and complexity; (e) using ranking selection, select function trees that will generate offspring (next generation); (f) apply variations, such as mutation and crossover, with the selected function trees; (g) Iterate steps (b) through (f) for the user-defined number of generations. There are two differences between two steps; the structural learning step includes a local search procedure; the sequence learning step uses initializes the population with random n-mere sequence using wordwise method. In addition, structural learning step uses mutation and crossover as variation operators to change a tree structure but learning of sequence uses only mutation to change the sequence and the number of mismatch without changing tree structure.

Population initialization. An initial population is randomly created with some constraints about function tree as described above. The initial population contains various function trees because there is no limitation in the number of nodes and the width of tree. That makes it possible to cover a wide range for searching start point. The broad coverage at start point is one of the major reasons the esRCSG is

efficient for searching optimal solution.

Fitness function. The fitness function (Equation 1) is defined by using specificity, sensitivity and the complexity that are defined at Equation (4).

$$Fitness = spC * Specificity + stC * Sensitivity - Complexity \quad (\mathbf{1})$$

$$spC + stC = 1 \quad (\mathbf{2})$$

Two parameters, namely *spC* and *stC* were added as a way to regulate the effects of specificity and sensitivity. To normalize the fitness, the sum of *spC* and *stC* is always 1 (Equation 2). The parameters decide the trade-off between the specificity and the sensitivity on the fitness function.

The complexity, which is a negative effecter in fitness function, controls the growing of tree. Without the complexity term, the tree would not convert to the minimum size where the tree has best efficiency, but it would grow indefinitely during the evolution. To overcome that the over-sizing problem, we make $Comp_i^j$ include the node number and the depth of *j*th tree on *i*th generation (Equation 3). Equation (4) describes the definition of *Complexity* of *j*th tree on *i*th generation.

$$Comp_i^j = TreeDepth_i^j \times 10 + NodeNum_i^j \quad (\mathbf{3})$$

$$Complexity_i^j = \frac{1}{(NS+PS)^2} \times \frac{Comp_i^j}{Comp_{i-1}^{best}} \quad (\mathbf{4})$$

where *Complexity* is normalized by square of the number of training data set (*NS* + *PS*). *NS* is the number of negative training data set (see Table 1) and *PS* is the

number of positive training data set (see Table 1). $Complexity_i^j$ depends on $Comp_{i-1}^{best}$ which is *Comp* of the best individual (tree) on (*i*-1)th generation. That dependency makes the trees have the minimum *Comp* as the progress of generation. Finally, the size of best tree on last generation is converged into minimum length as the principle of Occam's razor (Zhang *et al*., 1997).

Variation. The variation operators are applied so that each descendent will have a different tree structure relative to the parents. The first operator to perturb the tree is the mutation. The mutation changes the value of the function variable by a random variable drawn from Poisson distribution. The crossover exchanges each sub-tree in two parent trees via single-point recombination to generate two new offspring trees. In the crossover, two parent function trees are selected at random from the population and then each single recombination point is selected at random from the each parent.

II-4. Results

Dataset

For each step, training data set and test data set are described in Table 1. In order to extract extended stem-loop structures to be used as negative data set of scoring model in human genome, we predicted RNA secondary structures using RNAfold

program (available at http://rna.tbi.univie.ac.at/ cgi-bin/RNAfold.cgi) for human chromosome 18 and 19. The stem-loop structures are selected under some criterions obtained through learning common structure of human miRNA precursors, which are sequence length (64 ~ 90 nucleotides), stem length (above 22 nucleotides), bulge size (under 15 nucleotides), loop size (3 ~ 20 nucleotides) and free energy (under -25 kcal).

Table 1. Training and test data set for two-step genetic programming

Steps	Class	Training number (Test number)	
Learning	Positive	50 (102)	http://www.sanger.ac.uk/software/
RCSG	Negative	200	Primary sequences, hairpins, RNA peuoknots, IRE, bulges and internal loops

Learning RCSG of miRNA precursor

The genetic programming succeeded in optimizing the RCSG of miRNA precursors (Figure 3a). The conditions and results of the experiment are described in Figure 3c. In this experiment, we tried a local search with the words of miRNA precursors. The optimal RCSG with the word 'gcaggga' allows one mismatch and has lower sensitivity than itself without the word. Figure 4 shows the plots of the fitness and the specificity of the best individual on each generation. Because using the elitism, the fitness readily increased or equaled according to growing generation.

(a)

```
descr (fitnss=0.93, specfty=0.96, sensty=0.71)
h5(mispair=3)
        ss(minlen=5, maxlen=24)
        h5(len=5, mispair=1)
                h5(minlen=8, maxlen=14, seq='gcaggga', mismatch=1)
                        ss(minlen=5, maxlen=27)
                        h5(mispair=5)
                                ss(minlen=5, maxlen=15)
                    h3
                h3
        h3
h3
```

(b)

```
descr (fitnss=0.74, specfty=0.84, sensty=0.84)
h5(mispair=5)
        h5(mispair=1)
          ss(len=27, seq='gaguaaa', mismatch=0)
  h3
  ss(len=22)
h3
```

(c)

Population size	30
Generation number	30
Iteration number for local search	30
spC	0.95
stC	0.05

Fig. 3. (a): The optimal RCSG of miRNA precursors. (c): The setting of parameters and the measures. (b): The RCSG is more general and more sensitive than the RCSG of (a) but less specific.

In order to evaluate the RCSGs for miRNA precursors, we performed the test with the test set consisting of 102 miRNA precursors (excluding training set) and 100 the negative data. In the results, we could measure sensitivity (= 0.71) with detecting 72 of 102 miRNA precursors and specificity (= 0.92) with 72 true positive of 78 positive candidates. The miRseeker of Lay group showed the validation results of the 75% (18/24) sensitivity and about 50% of specificity (Lai *et al*., 2003). Our approach made a more specific identifier than miRseeker and

could reduce the false positives. However, it seems that our strategy is less sensitivity than miRseeker (Lai *et al*., 2003) If using alternative optimal RCSGs, such as Figure 3b, together, we think that it is enough to cover the low sensitivity.

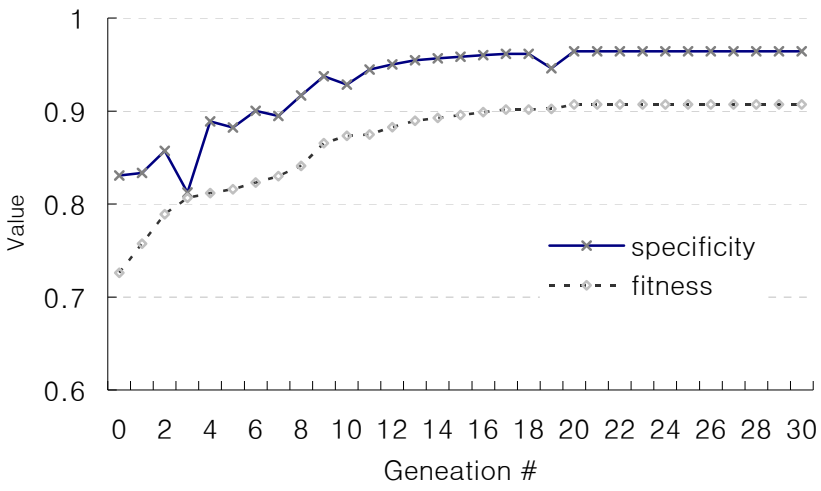

Fig. 4. The plot of the best RCSG for learning miRNA precursor.

III. Probabilistic Co-Learning of Structure and Sequence

III-1. Probabilistic Model for Co-Learning

Fig. 5. Pairwise representation of stem-loop structures and state sequences of miRNA precursor, where the state of each pair includes match information and mature miRNA region information.

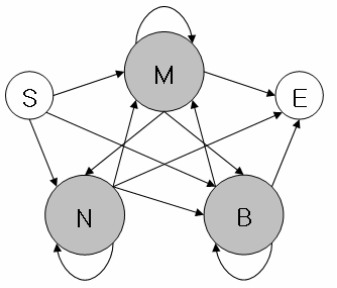

Fig. 6. State transition diagram, where S is start, E is end, M is match state, N is mismatch state and B is bulge state.

A miRNA precursor can be represented as a pairwise sequence. It forms an extended stem-loop structure and the stem-loop structure can be formulated as a sequence of matched base pairs (Figure 5). The pairwise sequence starts from the outside of the primary sequence, and ends at the inside loop. The state of each pair can be match, mismatch, or bulge. Especially, we represent a loop structure as a sequence of mismatches and bulges in order.

Each position of the pairwise representation has two properties, i.e. structural information (match/mismatch/bulge) and mature miRNA region information (true/false). Match states, M can emit A-U, U-A, G-C, C-G, U-G or G-U. Bulge states, B can emit A-•, U-•, G-•, C-•, •-A, •-U, •-G, or •-C. Mismatch states, N can emit one of the rest. The possible transitions among 3 structural states are shown in Figure 6. We represent each emission as a corresponding character in alphabetical order.

III-2. Prediction of microRNA Precursor

The Viterbi algorithm (Forney, 1973) is the common method for finding the most probable state transition path and its probability in Hidden Markov models (HMMs). We introduce a variant of the Viterbi algorithm. The practical problem of the Viterbi algorithm is that the value of the probability returned by the algorithm is very small. Particularly, the longer the given sequence is, the exponentially smaller is the probability the Viterbi algorithm produces. In order 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 determine the threshold probability using the training data that consists of 136 known miRNA precursors. We can control the trade-off of sensitivity and specificity with the threshold.

We evaluate the Viterbi probability for the mature miRNA region by

$$P = E_{s(q1)}(q_1) \cdot \prod_{i=2}^{22} \{T_{s(q_{i-1})s(q_i)} \cdot E_{s(q_i)}(q_i)\} \qquad (1)$$

where $E_s(q)$ is the probability which state s emits character q, T_{ab} is the transition probability from state a to state b, $s(q)$ is the match state of character q and q_i is the i-th character of the given sequence. E and T are the results of training HMMs.

On the given pairwise sequence we search the maximum P value using a sliding window, the size of which is 22±2 base pairs; the mean length of the mature miRNAs in the pairwise representation is 22 base pairs. We evaluate two P values for the model of 5' sense strand miRNA precursors and for the model of 3' sense strand miRNA precursors respectively. If either of them is bigger than the selected threshold, then we classify the given candidate as a miRNA precursor. In addition, we can determine the orientation of its mature miRNA by comparing the both P values.

III-3. Prediction of mature miRNA region

As a result of the screening, we get the miRNA precursor candidates. However, regional information of the screening algorithm is not accurate since the rest of the entire sequence is not considered in the screening. We introduce two hidden states indicating whether the position is mature miRNA region or not. The probabilities which state that the *i*-th position is true or false are computed as

$$P_t(i) = \max\{P_t(i-1) \cdot T_{\tau(q_{i-1})\tau(q_i)}, P_f(i-1) \cdot T_{\upsilon(q_{i-1})\tau(q_i)}\} \cdot E_{\tau(q_i)}(q_i) \quad (2)$$

$$P_f(i) = \max\{P_t(i-1) \cdot T_{\tau(q_{i-1})\upsilon(q_i)}, P_f(i-1) \cdot T_{\upsilon(q_{i-1})\upsilon(q_i)}\} \cdot E_{\upsilon(q_i)}(q_i) \quad (3)$$

where $\tau(q)$ is the true state of character q, $\upsilon(q)$ is the false state of character q and the initial condition is $P_t(1) = 0$, $P_f(1) = 1$.

Using the original Viterbi algorithm, we can not determine mature miRNA region because most probabilities of the transitions to the mature miRNA region states are not sufficient. We focus on the signal of the transition to false states and compute $S(i)$ as follows:

$$S(i) = \frac{P_t(i-1) \cdot T_{\tau\upsilon}}{P_t(i-1) \cdot T_{\tau\upsilon} + P_f(i-1) T_{\upsilon\upsilon}} \quad (4)$$

The equations given above describe the forward processing. We can find the loop-side cleavage site (3' end) by searching the position of the maximum forward signal. We perform the algorithm not only forward but also backward to

determine the stem-side cleavage site (5' start). In the backward processing, the initial condition is $P_t(length) = 0$, $P_f(length) = 1$ and we evaluate P_t, P_f and $S(i)$ from the last position to the first position backward in the same manner as the forward processing. We first determine the 3' end of mature miRNA and then we seek the 5' start with the backward signal of the partition that is 22±5 bp distant from the 3' end, because the forward signal is dominant.

III-4. Results

Datasets

We trained probabilistic co-learning model with previously known human miRNA precursor data that consist of 81 5'-sense strand and 55 3'-antisense strand miRNA precursors (available at http://www.sanger.ac.uk/Software/Rfam/mirna /search.shtml). The models of the both datasets are used for screening miRNA and for determining which orientation is more probable; whether mature miRNA is located in 5'-sense strand or 3'-antisense strand. The model of the entire dataset is used for the prediction of the mature miRNA region.

We tested the algorithm on human chromosome 18 and 19 (available at NCBI GenBank human genomes). The numbers of stem-loop structures extracted from chromosome 18 and 19 are 34853 and 62229 respectively. The stem-loops were considered as negative data set because majority of them must be false data.

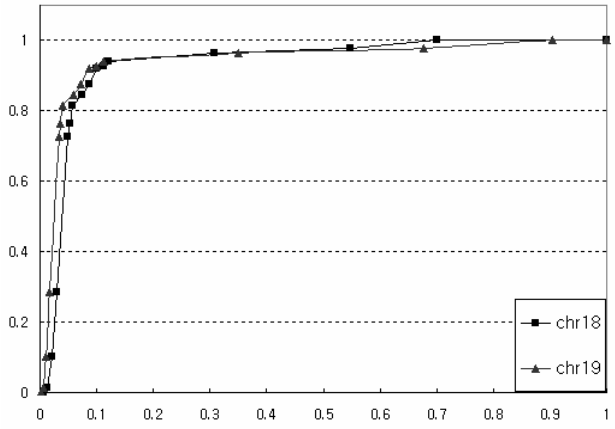

Fig. 7. Curves that represent the performance of the screening. The vertical shows the accuracy of the screening for 136 known miRNA precursor with 5-fold cross-validation, i.e. the sensitivity. The horizon shows, threshold, the percentage of the screening result for stem-loops extracted from chromosome 18 and 19

Training of probabilistic model and determining of screening threshold

First, we randomly split the 136 training data into 5 subsets. We performed 5-fold cross-validation for various thresholds, which range is from 0 to 1 and which is the ratio of the screened stem-loop structures of overall stem-loop structures extracted from genome. The threshold should be selected at the value maximizing true positives but minimizing false positives. It is provided as a parameter to predict putative miRNAs in genome-wide search. We evaluated the threshold with stem-loop structures extracted from human chromosome 18 and 19. In genome-wide screening of miRNA genes, it is not an easy problem to define the negative

dataset. However, instead of true negative dataset, we can use all stem-loop structures extracted in the order 10^4 as negative data, since the expected number of actual miRNA precursors in a chromosome is extremely small in the order of 10. Figure 7 describes screening ratio of 136 known miRNAs according to change of the thresholds. Considering the trade-off of sensitivity and specificity we chose the threshold (P=0.033) for classification of miRNA precursor candidates at the point that shows 72.8% sensitivity and 95.9% specificity on the average. Here, the specificity, which is the ability to reject "false positive" matches, was calculated by TN/(TN+FP) and the sensitivity, which is the ability to detect "true positive" matches, was calculated by TP/(TP+FN).

Table 2. Results of mature miRNA region prediction for 5-fold cross-validation. The last row of the table shows the result except 20 prediction failures. Prediction failures mean that we can not make a decision because the defined signal $S(i)$ is too weak.

	Mean of absolute distance				Square root of the mean of the squares			
	5' sense		3' anti-sense		5' sense		3' anti-sense	
	start	end	Start	End	start	End	start	end
Total	2.83	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

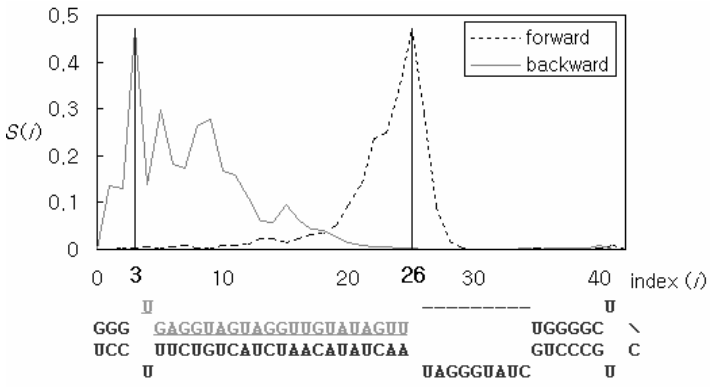

Fig. 8. The forward and backward signals for prediction of mature miRNA region. The sample is human miRNA precursor hsa-let-7a-3. The actual mature miRNA region is [4,25].

Mature miRNA region prediction

We evaluated the accuracy of mature miRNA region prediction through 5-fold cross-validation with 136 known miRNA (Table 2). The criterions of assessment are the mean of absolute distances and the square root of the mean of the squares. The result shows that HMmiRNA finds 3' end more precisely than 5' start for sense strands; HMmiRNA finds 5' start more precisely than 3' end for anti-sense strands. In general, the statistical signal of the 3' cleavage site of mature miRNA is relatively more dominant.

We exactly predicted the orientation of mature miRNA region for 57 of 81 5' sense strand miRNA precursors and for 41 of 55 3' anti-sense strand miRNA precursors. The accuracy is 72.1% on the average, which is not as high as we expected. However, the orientation determination may be not foremost but supplementary.

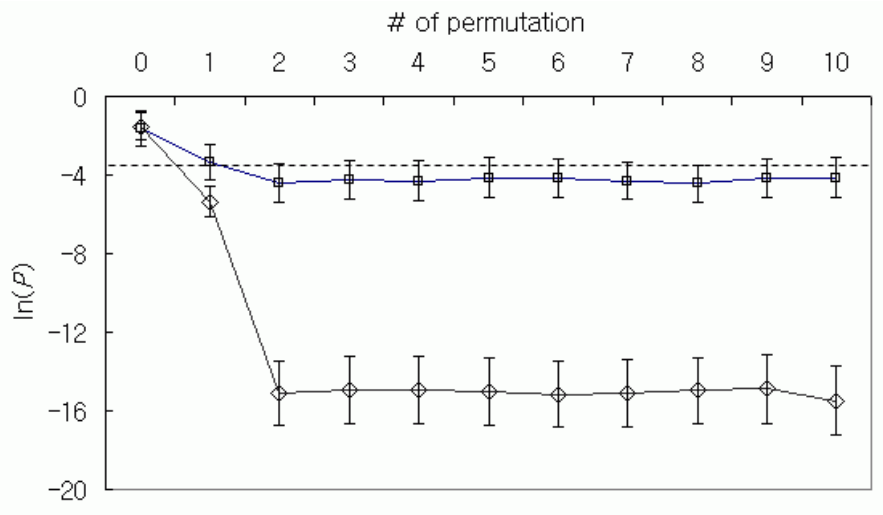

Fig. 9. Permutation test for structure and sequence of mature miRNA

Permutation test for learning model

From the results so far, we can conclude that HMmiRNA effectively detects the cleavage signal recognized by Dicer as shown in Figure 9. However, it is difficult to judge where the major cleavage signal was stemmed from, through these results. We designed the random permutation test to investigate if the high specificity obtained by our model is due to 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 or bases in the stem respectively. To measure the effect of base mutation, we randomly change the base without changing the base-pair. Also, the effect of structure is measured by changing the base-pair. Figure 6 presents the result of this study. ln*(P)* by structural permutation

was rapidly decayed to far under the threshold (= ln(0.033)) when the number of permutation was even one. On the contrary, ln(*P*) by sequence permutation a little went down at the first permutation and fluctuated at near to threshold value. We performed T-test to estimate the significant probability for this result and p-value was under 0.00005. The results clearly demonstrate that the specificity of our algorithm is influenced more by the conserved structural signals, such as match, mismatch and bulge than by the conserved sequence information.

IV. Genome-Wide Screening of Human microRNA

IV-1. Pipeline for Screening

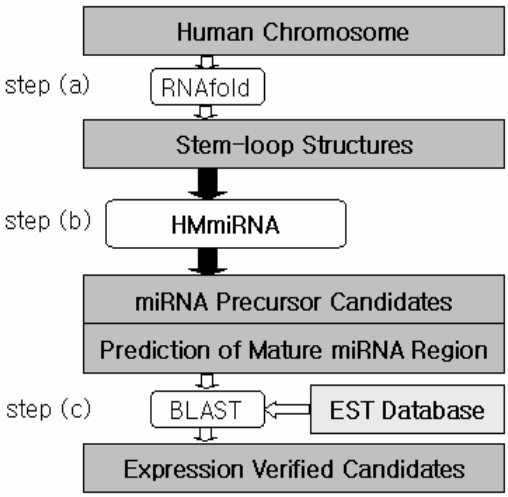

Fig. 10. Systematic overview of the entire processing flow for identification of human miRNA genes

To identify the miRNA genes from human genome, we developed an algorithm named HMmiRNA. The algorithm also predicts the mature miRNA regions over miRNA precursors. The input data of the algorithm are stem-loop structures that are extracted from human genome using RNAfold (Hofacker, 2003) (Figure 10 step (a)). HMmiRNA evaluates the likelihood of each stem-loop by HMMs trained with known human miRNA precursors and classifies human miRNA precursors (Figure 10 step (b)). After the classification of HMmiRNA, we verify

expression of the predicted miRNA precursors through human expressed sequence tag (EST) analysis to validate the reliability of predicted results (Figure 10 step (c)).

Extraction of stem-loop structures from human genome

In order to extract stem-loop structures that are similar to miRNA precursors in human genome, we predicted RNA secondary structures using RNAfold program (available at http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi) for human chromosome 18 and 19. We first subdivided chromosome into 10 contigs each, with no overlap at either end and with masking exon regions. We then scanned all contigs with 90 nucleotides window size and with 80 nucleotides of overlap at both ends. RNA secondary structures of the window fragments are predicted with their reverse complementary sequences. The stem-loop structures are selected under some criterions obtained through learning common structure of human miRNA precursors (Nam *et al*., 2004), which are sequence length (64 ~ 90 nucleotides), stem length (above 22 nucleotides), bulge size (under 15 nucleotides), loop size (3 ~ 20 nucleotides) and free energy (under -25 kcal).

IV-2. Results

Table 3. Results of genome-wide screening of human miRNA. The second column includes size of each chromosome; the third column includes the number of stem-loop structures extracted by RNAfold; the fourth column includes the number of miRNA precursor candidates classified from the stem-loops by HMmiRNA; the fifth column means specificity of HMmiRNA; the sixth column includes the number of candidates verified by EST analysis; the seventh and eighth column includes the number of known miRNA genes and the number of miRNA genes detected by HMmiRNA; the ninth column includes the number of candidates that have either paralogous or orthologous;

Chr	Size of chr. (Mbp)	Stem-loop	Precursor Candidates	Percentage (%)	Expression Verified	Known miRNA	Detected miRNA	Homologous
18	56.7	34853	2253	6.46	84	4	4	22
19	75.7	62229	2065	3.32	171	11	5	42

miRNAs on human chromosome 18 and 19

To perform genome-wide screening for miRNA genes, we extracted 34853 and 62229 sequences of stem-loop structure on chromosome 18 and 19 each, under the conditions as mentioned above. Under the selected threshold that shows 72.8% sensitivity and 95.9% specificity on the average, HMmiRNA classified total 4318 miRNA precursor candidates from extracted stem-loops (Table 3). Similarity search between the candidates and the known human miRNAs showed that the candidates include four of four known miRNA, mir-1-2, mir-122a, mir-133a-1 and mir-187, on chromosome 18 and five, has-let-7e, mir-27a, mir-150, mir -199a-1 and mir-330, of 11 known miRNA on chromosome 19 (Table 3).

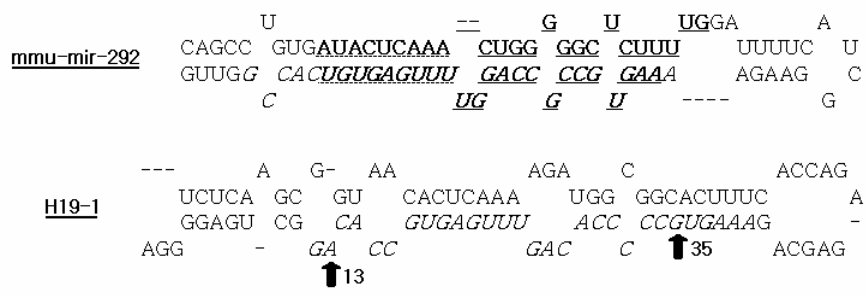

Fig. 11. A putative human miRNA, H19-1 (see supplementary document for detail), is detected by BLAST search between the candidates and known mouse miRNAs; matched with mmu-mir-292. In the miRNA precursors, the italic letters are the regions that are aligned by BLAST with E value 8e-04 and underlined bases are mature miRNA. HMmiRNA predicts that the orientation of the mature miRNA is 3'-antisense and that the region is [13,35].

Unfortunately, HMmiRNA does not directly guarantee that the candidates are actual miRNA precursors. miRNA candidates to be annotated as a novel miRNA should be proved by detecting either very close homologs or expression of 22nt mature miRNA (Ambros V, 2003). First, to detect very close homologs, we performed the BLAST search between the candidates and the known miRNAs of other species such as *Mus musculus*, *Caenorhabditis elegans* and *Drosophila melanogaster*. In the result, we identified a putative miRNA, which is homologous with mouse (Figure 11). The mature miRNA region of the putative miRNA is close to the result of our region prediction. In addition, we confirmed that 22 and 42 of candidates had either paralogs or orthologos. Second, to verify the expression of candidates, we fulfilled human EST analysis for the candidates

instead of experimental validation. This result is shown in next section.

EST analysis of miRNA precursor candidates

In order to verify expression of the predicted miRNA precursors, we performed human EST analysis for them (NCBI Entrez EST database; Feb. 6. 2003). Smalheiser reported that 36 characterized human miRNA precursors are homologous to human ESTs (Smalheiser, 2003). Actually, 40 of 152 known human miRNAs are matched with human ESTs, where E values are under 1.0e-15; 32 of them are perfectly matched. It means that many human miRNA transcripts can be still discovered in human EST database.

As a result of EST analysis, 84 of 2253 miRNA precursor candidates on chromosome 18 and 171 of 2065 miRNA precursor candidates on chromosome 19 were matched with human ESTs; E values were under 1.0e-30 (Table 1). In order to investigate whether the numbers of candidates matched by EST are statistically significant, we estimated p-value under null hypothesis, where the average of the numbers of candidates matched by EST is identical with the average of the numbers of random sequence matched by EST. The p-values were 0.0344 at chromosome 18 and 0.0102 at chromosome 19. Thus, we could reject the null hypothesis and accept the alternative hypothesis of it.

The other hand, we can estimate the number of miRNA genes in EST database using Bayes' theorem likes below equation (T. Bayes, 1763).

$$P(miRNA|EST) = \{P(EST|miRNA)P(miRNA)/P(EST)\}R, \quad R: \textit{redundancy of EST}$$

where $P(EST)$ is the ratio of genome region covered by EST contigs and the value is 0.0849. $P(miRNA)$ is the ratio of golden-standard miRNA genes on genome and the value is 1/100000. $P(EST|miRNA)$ is the ratio of known miRNA matched by EST and the value. P(*Candidates*) is the ratio of genome region covered by candidates and the value is 0.0036. R is the redundancy of EST and EST database has about 150 fold redundancy. Therefore, $P(miRNA|EST)$ is calculated by given values and the value is about 2.06E-07. Regarding the size of EST database is 2.9×10^9, the number of miRNA genes in EST database is about 600, more than we expected. This result may support that the matched candidates are expressed in human cells are more probable human miRNA genes than the others.

V. Discussions

V-1. Comparison of Methods

Table 4. Comparison of the efficiency for miRNA prediction. [a]: results by sequence and structure multiple alignment, [b]: results by sub-optimal RCSG, [c]: results by 5-fold cross validation

	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)
esRCSG	50	0.36 (0.67)[b]	0.96 (0.89)
HMmiRNA	68	0.69	0.94
	5-fold[c]	0.73	0.96

We compared the efficiency of miRNA prediction with four different approaches (Table 4). To perform more fair comparison, we trained each model with various numbers of data. This made it possible to search the optimal results of each method. The HMMer method by multiple sequence alignment shows very low efficiency. This result might be caused by inappropriate alignment results because miRNA genes are structurally conserved rather than sequencially conserved. The INFERNAL by multiple structural alignments showed the higher sensitivity according to increasing the number of training data, whereas the specificity decreased until 0.18. This low specificity might be caused by only structural learning. In the previous work, we introduced esRCSG to detect common

structural-grammar. The results by this method showed more effective prediction than methods using sequence or structural alignment. However esRCSG (= 0.67) gives lower sensitivity than HMmiRNA (= 0.73) using probabilistic co-learning of sequence and structure. These results clearly show that the HMmiRNA gives more reliable prediction results than other methods.

V-2. Conclusion

We have shown that two-step genetic programming and probabilistic co-learning of structure and sequence are effective methods for the identification of miRNA genes and for the prediction of mature miRNA region without comparative analysis with the sensitivity and the specificity comparable to or better than other approaches. Two-step genetic programming is an approach with respectively optimizing structure and sequence to search for RCSGs from miRNA sequences. When applying the human miRNA precursors, we could find the distinctive RCSGs from known human microRNAs. The identified RCSGs effectively reflected the common structures of miRNA precursors. Therefore, we have proven the possibility to learn common-structural grammar from structurally unknown sequences through genetic programming. We believe that the approach, two-step genetic programming, can be applied for various applications such as RNA similarity search and putative RNA identification. The high specificity of the approach is caused by applying the information of evolutionarily conserved specific sequence in the second learning step.

HMmiRNA is the program designed with probabilistic co-learning model and implemented on the website. Though the sensitivity and specificity of HMmiRNA depend on the screening threshold, we should have selected a threshold to be able to minimize the number of false positive and maximize the number of true positive. We determined threshold (=0.033) showing that the sensitivity was 72.8% and the specificity was 95.9% in the screening performance curve of Figure 9. The major reason requiring high specificity is due to human genome complexity. Human genome includes much noisy sequences such as repeat sequence, palindromic sequence, pseudo-gene and transposon, and the screening method should have a stringent classifier with high specificity. Of course, we can more or less stringently adjust a threshold according to the aim of miRNA prediction. In genome-wide screening of human miRNA, the candidates included 9 of 15 known miRNAs on chromosome 18 and 19. The sensitivity is less than the sensitivity estimated from the screening performance curve. It seems to be caused by the neighboring noisy sequence. Two among the candidates were paralogs of the known human miRNAs (supplementary information). In order to select the more probable candidates, we performed human EST analysis and verified their expression. Interestingly, three (C19-53, C19-56, C19-65) of the candidates are located in the introns of genes. C19-53 and C19-56 are placed in the introns that are caused by alternative splicing (supplementary information).

The error of the mature miRNA region prediction results was 2.7 nucleotides on the average and the variation except 20 prediction failures was 2.0 nucleotides

on the average. It has been reported that the biological cleavage of the miRNA precursor by Dicer bears the error of one nucleotide and makes two nucleotides of 3' overhang ends (Zamore, 2002). It means that our algorithm gives meaningful results for the prediction of mature miRNA regions over miRNA precursors. For the signal $S(i)$, 3' cleavage site of a mature miRNA shows higher than 5' cleavage site. We can confirm that 3' cleavage site is statistically more strictly conserved, since HMMs characterize statistical information of the conserved sequence and structure on cleavage sites over miRNA precursors. Also it can be regarded that 3' cleavage site may be the important region to interact with trans-acting elements likes RNase III type enzyme or miRNP complex that takes part in the processing of mature miRNAs or in the target regulation.

The prediction of mature miRNA orientation on precursors is the other problem with region prediction. Most mature miRNAs over precursors are located in either of 5'-sense strand or 3'-antisense strand. However, some of the known miRNAs exist in both strands simultaneously. Though most miRNA precursors have extended stem-loop structures, a few miRNA precursors have the branched stem-loop secondary structures. These exceptional phenomena in miRNAs might make it difficult to predict the region and orientation of mature miRNAs. We did not include the exceptional data to make the algorithm simple and efficient.

VI. References

Ambros,V., and Moss,E. (1994) The heterochronic genes and developmental timing in C. elegans. *Trends in Genetics,* **10**:123-127.

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

Ambros,V., Bartel,B., Bartel,D.P., Burge,C.B., Carrington,J.C., Chen,X., Dreyfuss,G., Eddy,S.R., Griffiths-Jones,S., Marshall,M., Matzke,M., Ruvkun,G. and Tuschl,T. (2003) A uniform system for microRNA annotation. *RNA*, **9**:227-279.

Bernstein,E., Caudy,A.A., Hammond,S.M. and Hannon,G.J. (2001) Role for a bidenate ribonuclease in the initiation step of RNA interference. *Nature*, **409**: 363-366.

Cai L., Malmberg R.L., Wu Y. (2003) Stochastic modeling or RNA pseudoknotted structures: a grammatical approach. *Bioinformatics,* **19**:i66-i73.

Dostie,J., Mourelatos,Z., Yang,M., Sharma,A. and Dreyfuss,G. (2003) Numerous microRNPs in neuronal cells containing novel microRNAs. *RNA*, **9**:180-186.

Eddy,S.R. Durbin,R. (1994) RNA sequence analysis using covariance models. *Nucleic Acids Res.,* **22**:2079-2088.

Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**:755-763.

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.

Enright,A., John,B., Gaul,U., Tuschl,T., Sander,C. and Marks,D. (2003) MicroRNA Targets in Drosophila. *Genome Biology*, **4**:P8.

Forney,G.D.,Jr. (1973) The Viterbi Algorithm. *Proc IEEE,* **61**(3):268-278.

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.

Griffiths-Jones,S. (2004) The microRNA Registry. *Nucleic Acids Res.,* **32**:Database Issue, D109-111.

Hofacker,I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res.*, **31**(13):3429-3431.

Klein,R.J. and Eddy,S.R. (2003) RSEARCH: Finding homologs of single structured RNA sequences. *BMC Bioinformatics*, **4**:44.

Knudsen B. and Hein J. RNA (1999) secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics*, **15**:446-454.

Koza J.R. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press, (1992).

Lai,E.C. (2002) Micro RNAs are complementary to 3' UTR sequence motifs that mediate negative post-transcriptional regulation. *Nat Genet*., **30**:363-4.

Lai,E.C., Tomancak, P., Williams, R.W. and Rubin, G.M. (2003) Computational identification of Droshophila microRNA genes. *Genome Biology*, **4**:R42.

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

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

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

Lee,Y., Jeon,K., Lee,J., Kim,S. and Kim,V. (2002) MicroRNA maturation: stepwise processing and subcellular localization. *EMBO J*., **21**(17):4663-70.

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.

Lee,Y.S., Nakahara,K., Pham,J.W., Kim,K., He,Z., Sontheimer,E.J., Carthew,R.W. (2004) Distinct roles for Drosophila Dicer-1 and Dicer-2 in the siRNA/miRNA silencing pathways. *Cell*, **117**(1):1-3.

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

Lewis,B.P., Shih,I., Jones-Rhoades,M.W., Bartel,D.P. and Burge,C.B. (2003) Prediction of Mammalian MicroRNA Targets. *Cell*, **115**:787-798.

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

Nam,J.W., Joung,J.G., Ahn,Y.S. and Zhang,B.T. (2004a) Two-step genetic programming for optimization of RNA common-structure. *Lecture Notes in Computer Science*, **3005**:73-83.

Nam,J.W., Lee,W.J. and Zhang,B.T. (2004b) Computational Methods for Identification of Human microRNA precursors *Lecture Notes in Artificial Intelligence*, In press.

Pfeffer,S., Zavolan,M., Grasser,F.A., Chien,M., Russo,J.J., Ju,J., John,B., Enright,A.J., Marks,D., Sander,C. and Tuschl,T. (2004) Identification of virus-encoded microRNAs. *Science*, **304**(5671):734-6.

Rhoades,M.W., Reinhart,B.J., Lim,L.P., Burge,C.B., Bartel,B. and Bartel,D.P. (2002) Prediction of plant microRNA targets. *Cell*, **110**:513-20.

Sakakibara Y., Brwon M., Hughey, R., Mian I.S., Sjolander K., Underwood R.C. and Haussler D. (1994) Stochastic context-free grammars for tRNA modeling. *Nucleic Acids Research*, **22**:5112-5120.

Siebert,S. and Backofen,R. (2003) MARNA: A Server for Multiple Alignment of RNAs. *In Proceedings of the German Conference on Bioinformatics*, 135-140.

Smalheiser,N.R. (2003) EST analyses predict the existence of a population of chimeric microRNA precursor-mRNA transcripts expressed in normal human and mouse tissues. *Genome Biology*, **3**:403

Song,J., Liu,J., Tolia,N., Schneiderman,J., Smith,S., Martienssen,R., Hannon,G. and Joshua-Tor,.L. (2003) The crystal structure of the Argonaute2 PAZ domain reveals an RNA binding motif in RNAi effector complexes. *Nat Struct. Biol.*, **10**(12):1026-32.

Thomas J. Macke, David J. Ecker, Robin R. Gutell, Daniel Gautheret, David A. Case and Rangarajan Sampath. (2001) RNAMotif, an RNA secondary structure definition and search algorithms. *Nucleic Acids Research*, **29**:4724-4735.

Zamore,P.D., Tuschl,T., Sharp,P. and Bartel,D. (2000) RNAi: Double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. *Cell*, **101**:25-33.

Zamore,P.D. (2002) Ancient Pathways programmed by small RNAs. *Science*, **296**:1265-1269.

Zhang B.-T., Ohm P., and Mühlenbein H. (1997) Evolutionary neural trees for modeling and predicting complex systems. *Engineering Applications of Artificial Intelligence*, **10**:473-483.

초 록

microRNA는 세포질에 약 21 nucleotides로 존재하는 small RNA의 한 종류로, 최근 세포내의 유전자 발현을 위한 새로운 조절 물질로 규명된 microRNA의 존재가 알려지면서, 여러 종의 유전체에 존재하는 small RNA의 동정에 많은 연구가 이루어 지고 있다. microRNA 동정 연구는 대부분 실험상으로 이루어 지고 있으며, 몇몇 연구그룹에서 비교 유전체학 방법을 통해, 기존의 알려진 microRNA의 유사 서열을 가진 새로운 microRNA를 동정하였다. 그러나 많은 microRNA가 조직특이적 또는 발생특이적으로 발현되거나 소량만 발현되는 이유로 실험을 통한 동정이 한계에 이르고 있다. 이러한 문제점을 해결할 수 있는 방법이 *In-Silico* 방법을 통한 동정이다.

본 논문에서 microRNA 예측을 위해 소개된 새로운 계산학적 방법들이 non-coding RNA 예측에 응용될 수 있는 일반 알고리즘임을 보이며, 특히 인간의 18번 19번 염색체에서 예측한 2253개와 2065개의 microRNA 후보와 이들 중 발현이 되는 255개의 후보들에 대한 결과를 제시한다. 두 방법은 microRNA 예측을 위한 다른 방법들과의 비교에서도 더 좋거나 동등한 특이도와 민감도를 나타냈다. 특히 완숙한 microRNA 위치 예측에서는 평균 2.0nt의 오차로 의미 있는 결과를 보였다. Permutation 실험의 결과는 두 알고리즘의 높은 특이도가 구조에 기반한 학습과 병행된 특이성 있는 염기서열의 학습에서 기인한다는 사실을 높은 통계적 유의성으로 ($p=0.00005$) 말해주고 있다. 이런 결과로 본 연구에서 소개된 새로운 방법들이 microRNA를 예측하는 좋은 접근 방법임에 틀림없다.

주제어: HMmiRNA, esRCSG, small RNA, microRNA 예측, RCSG, 유전자 프로그래밍, 확률모델, 구조학습

학번: 2002-23333

감사의 글

이 지면을 통해 그 동안 격려해 주시고 도움을 주신 분들께 감사의 인사를 드리고자 합니다.

우선, 많은 연구실 인원을 부양하고, 학문적 호기심을 계속 던져 주신 장병탁 지도교수님께 감사 드립니다. 또한 논문심사를 기꺼이 맡아주신 김규원 교수님, 손현석 교수님, 양진산 박사님께 감사 드립니다. 더불어, 이 논문에 조언을 아끼지 않으신 김빛내리 교수님께도 감사 드립니다.

이 논문이 나오기까지 연구실의 많은 분들께 도움을 받았지만, 무엇보다도 제균형과, 기루, 화진이가 없었다면, 이 연구를 완성하지 못했을 겁니다. 처음 microRNA 연구를 시작하면서 몇 일 동안 아이디어 논의를 하며 많은 것을 가르쳐 주었던 제균형, 그리고 두 학기 동안 microRNA 연구를 같이 하며 힘이 되어준 화진, 그리고 번뜩이는 두뇌로 알고리즘을 구현해준 기루. 진심으로 감사의 말 전하고 싶습니다.

또한, 연구실에 활력을 넣어주며 족구의 실력이 눈부시게 발전한 민호, 힘든 일 도맡아 하며, 힘이 되준 병희와, 언제나 학구열에 불타는 상근, 그리고 생물정보학에 뛰어든 우리 신입생들 성규, 제근. 연구실의 활기와 끈끈한 단합을 지금처럼 계속 이어주길 바랍니다. 더불어 생물정보학을 함께 공부했던 생물정보학 협동과정 2기 동기생들과 선배 후배들에게도 감사의 말 전하고 싶습니다.

든든한 마음의 지지자들, 박장군, 동글이, 땡칠이, 고양이 그 외 느릅나무 동기들에게 변함없는 우정을 바라며, 저를 믿고 지지해준 우리가족 형, 형수님, 선우, 귀여운 조카 기철, 윤지 그리고 나의 사랑하는 순두부 채영에게 진심으로 감사의 마음을 전합니다. 끝으로 변함없이 아껴주고, 무한한 사랑으로 키워주신 사랑하는 부모님께 이 논문을 바칩니다.