Distribution of Mature MicroRNA on Its Precursor: A New Character for MicroRNA Prediction

Huiyu Xia, Fei Li, Tao He, and Yanda Li

MOE Key Laboratory of Bioinformatics / Department of Automation, Tsinghua University, Beijing, 100084, China

xiahuiyu00@mails.tsinghua.edu.cn

Abstract

Background: MicroRNA (miRNA) is a large family of 20~22 nucleotides non-coding RNA, which regulates expression of protein-coding genes. Stem-loop structure is an important character of miRNA precursor for computational identification of miRNA genes and has been proved to be close related with miRNA biogenesis.
Methods: This paper statistically analyzed the hairpin structures of 557 miRNA genes from six eukaryotic organisms. A random model was adopted to show the significance of the distribution of mature miRNAs on their precursors.
Results: The results showed that the terminals of mature miRNAs tend to locate near loop structures within 1-6nt rather than in loops or very far from loop structures (>6nt), which is quite different from expected by chance. Further more, free energies of 1-6nt flanking mature miRNAs are much higher than those of mature miRNA terminal regions.
Conclusions: These results provide a new character of stem-loop structure for miRNA prediction and indicate that this character might facilitate miRNA biogenesis.

Keyword: MicroRNA (miRNA), Stem-loop structure, free energy, distribution

I. Introduction

MicroRNAs (miRNAs) are an evolutionary conserved class of 20~22 nucleotides (nt) noncoding RNAs, which have been proved to regulate the expression of mRNAs at the posttranscriptional level by either inactivating or degrading mRNA genes. They are widespread in many organisms. Recently, with informatics methods and biological experiments, hundreds of miRNAs were identified from *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Caenorhabditis briggsae*, *Arabidopsis thaliana* and *Oryza sativa* [1-7]. In addition, five miRNA genes were also reported in Epstein-Barr virus [8].

Previous efforts have proved that stem-loop hairpin structure is an important character of miRNA precursor and informative in computational identification of miRNA genes. There are some programs, such as miRseeker and MiRscan, have been developed to computationally predict miRNA genes, which mainly rely on feature selection from stem-loop structures of miRNA precursors and comparative genome analysis [4, 6].

Moreover, biogenesis of miRNA gene is highly associated with features of stem-loop precursor. At present, though the details of miRNA biogenesis remain to be unclear, the main procedures have been worked out. It has been reported that miRNA biogenesis involves in at least three steps. First, miRNA is transcribed as long transcript, named as pri-miRNA [9]. Then nuclear RNase Ⅲ Drosha initiates processing pri-miRNA into 60~70nt miRNA precursor (pre-miRNA), which folds into

stem-loop hairpin [10]. With the involvement of Exportin-5, pre-miRNA is transported from nuclear into cytoplasm [1-3, 11], where RNase Ⅲ like enzyme, Dicer, mediates the next key step of processing pre-miRNA into mature miRNA [12-14]. Pre-miRNA is cleaved into 20-22nt duplex bearing two nucleotides single-stranded 3'extension and generally only one strand of the duplex serves as the mature miRNA [15-17]. It has been believed that the efficiency of Drosha processing and Dicer cleavage are related to the stem-loop structure [9-11, 18].

Thus, it is our great interest to systematically study the distribution of mature miRNAs on the stem-loop hairpins of their precursors. Here we report systematic analysis of stem-loop hairpins of 557 pre-miRNAs from six eukaryotic organisms as well as the distribution of their mature miRNAs on them, which manifested that mature miRNAs do not randomly distribute on the stem-loop hairpins of their precursors. Further more, free energies of the segment flanking mature miRNA are much higher than those of mature miRNA terminal regions. These results provide a new character for miRNA prediction and indicate that this character might facilitate miRNA biogenesis.

II. Materials and Methods

A. *MicroRNA Data*

The dataset of microRNA has been obtained from Rfam database (release 2.1) at http://www.sanger.ac.uk/Software/Rfam/mirna/index.shtml [19, 20]. In total, 557 pre-miRNAs from six eukaryotic organisms, *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Caenorhabditis briggsae* and *Arabidopsis thaliana*, were collected, from which 579 mature miRNAs were extracted according to the annotations in the database. This dataset is the main subject of recent active miRNA research.

B. *Predicting RNA Secondary Structure*

RNA secondary structures of all 557 precursor miRNAs were predicted using RNAfold software in Vienna RNA secondary structure server available at http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi [21]. For RNA fold algorithm, RNA parameters were applied choosing partition function and pair probabilities. The amounts and distributions of different kinds of structures, i.e. hairpins, bulge loops, junctions, mismatch, internal loops and continuous duplex, in the secondary structure of pre-miRNAs were analyzed.

C. *Definition of Different Regions in Stem-Loop of Pre-miRNA*

According to the distance between a terminal of mature miRNA segment and its nearest loop, we defined Dis_{loop} as:

$$Dis_{loop} = \begin{cases} D_{5'} & \text{for 5' terminal of mature miRNA} \\ D_{3'} & \text{for 3' terminal of mature miRNA} \end{cases} \quad (1)$$

where $D_{5'}$ denotes the number of nucleotides from 5' terminal of a mature miRNA to its nearest upstream loop structure and $D_{3'}$ denotes the number of nucleotides from 3' terminal of a mature miRNA to the nearest downstream loop structure. Therefore, the 5'/3' terminal of mature miRNA is defined in:

1. Without-loop region (WLR), if no loop in the stem arm
2. Cross-loop region (CLR), if Disloop<0
3. Base-loop region (BLP), if Disloop=0
4. Near-loop region (NLR), if 1nt≤Disloop≤3nt
5. Far-from-loop region (FLR), if 4nt≤Disloop≤6nt
6. Very-far-from-loop region (VFLR), if 6nt< Disloop

D. *Statistical Analysis*

A random model was adopted to show the significance of the distribution of mature miRNAs on their precursors [22]. The expected proportions of mature miRNAs' distribution were computed through direct enumeration via a sliding window. For each pre-miRNA, a window of equal size of its mature miRNA was used to scan the stem-loop structure of it. The expected proportions of 5' and 3' terminals of mature miRNAs in different regions were calculated. In general, if the mature miRNAs are randomly locate on their precursors, the observed results can fit the expected ones well and vice versa. The goodness-of-fit χ^2 test was then applied to compare the observed and expected proportions.

E. *Free Energy Analysis*

The free energies of mature miRNAs and their adjacent regions were calculated using Oligo 6.0 (National Biosciences, Inc., Plymouth, MN), which is widely accepted for PCR or hybrid primer designing. Free energy values used for calculating internal stability of RNA duplexes were adopted in previous report [23, 24].

III. Results

A. *Characteristic of Pre-miRNA Secondary Structure*

We predicted the secondary structures of 557 pre-miRNAs and found that 93.5% miRNA precursors formed into typical stem-loop hairpins. Only 36 (6.5%) pre-miRNAs had complex secondary structures with more than one hairpin. Statistical analysis showed that the average length of pre-miRNAs is 92nt, of which 71.7% (66nt) appear in duplex and 28.3% (26nt) locate in loops (i.e. hairpins, internal loops, bulges, junctions) or single strands. Averagely, there are seven loops appearing in the stem-arm of pre-miRNA and less than three loops in the mature miRNA:miRNA* duplex regions (see Table 1). In the stem-arms of pre-miRNAs, 69.5% continuous duplexes are less than 7nt (see Fig. 1).

Table 1. Amounts of different kinds of loops in stem-loop structures of pre-miRNAs and mature miRNA segments

	Stem-loop structure		Mature miRNA segments	
Loop	Average Number	Percentage	Average Number	Percentage
Hairpins	1	14.3%	<< 0.1	-
Internal Loop	4	57.1%	2	>80.0%
Bulges	2	28.6%	<0.5	<20.0%
Junctions	<<1	-	-	-
Total	7	-	<2.5	-

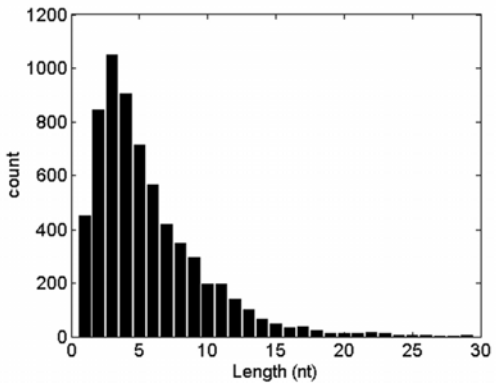

Fig. 1. The histogram of the length distribution of continuous duplexes without interruption of loop(s)/mismatch(es) in arms of stem-loop hairpins

B. Position Distribution of Mature MiRNA

Position distributions of 5' and 3' terminal of mature miRNA segment were analyzed separately. A random model was adopted to show the significance of the distributions. For 5' terminals of mature miRNA segments, statistical analysis showed that 46.11% 5' terminals are in NLR, which is much higher than expected by chance (33.62%). On the other hand, only 6.39% are in VFLR and 2.59% are in CLR, much lower than expected 17.24% and 10.52% in the random model (see Fig. 2 (a), p-value=4.34×10^{-23} by χ^2 test).

The distribution of 3' terminals of mature miRNAs exhibits similar behavior as that of 5' terminals. There are 49.91% of 3' terminals that are adjacent to loop structures (1-3nt) but only 33.85% are expected in the random model. In VFLR and CLR the proportions are 7.25% and 2.59% respectively, whereas 17.27% and 10.36% are expected in the random model respectively (see Fig. 2(b), p-value=1.08×10^{-23} by χ^2 test).

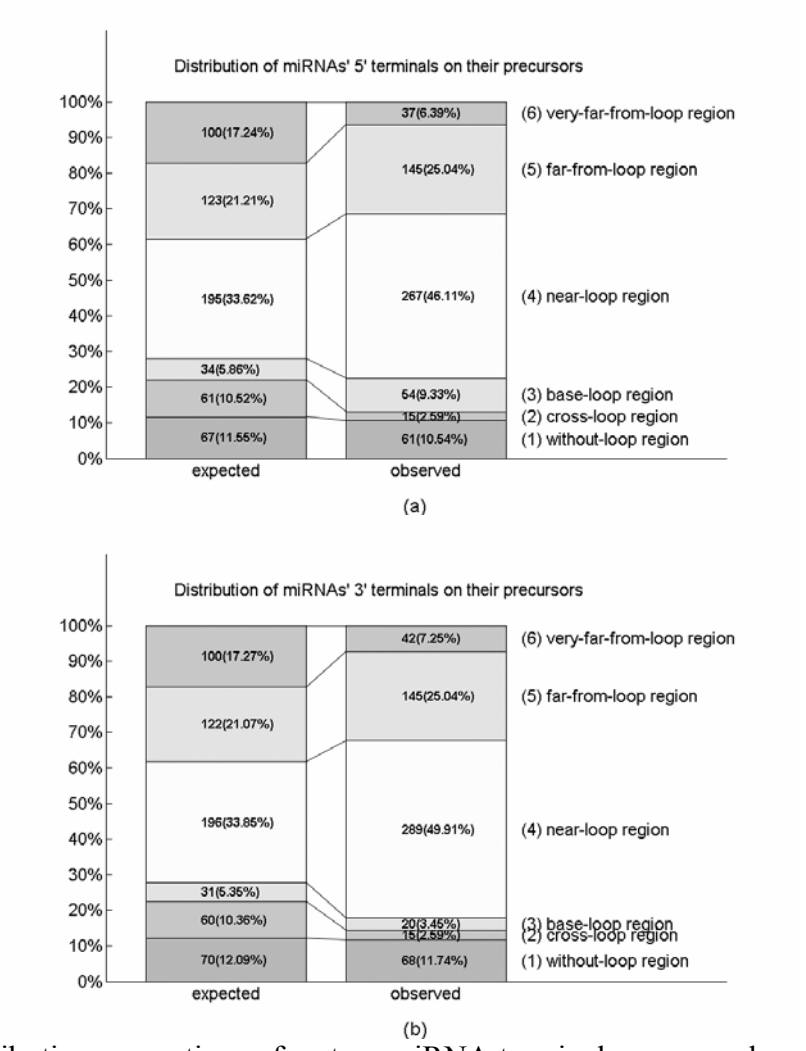

Fig. 2. The distribution proportions of mature miRNA terminals on secondary structure of pre-miRNAs. The numbers of each situation are shown and the percentages of the total are in parentheses. (a) The distribution proportions of 5' terminals of mature miRNAs on their precursors; (b) the distribution proportions of 3' terminals of mature miRNAs on their precursors

Further analysis showed that there are loop structures near both 5' and 3' terminals of mature miRNA segments within 1-6nt in 292 pre-miRNAs (50.43%), which is also much

higher than expected in the random model (p-value=1.94×10^{-31}, by $\chi2$ test, see Table 2 and Table 3).

Table 2. Counts of 5' and 3' terminals of mature miRNA in different regions

Regions		3' terminal		
		CLR+BLR	NLR+FLR	EFLR+WLR
5' terminal	CLR+BLR	11 (1.90%)	66 (11.40%)	6 (1.04%)
	NLR+FLR	47 (8.12%)	292 (50.43%)	73 (12.61%)
	EFLR+WLR	3 (0.52%)	76 (13.13%)	5 (0.86%)

Table 3. Expected counts of 5' and 3' terminals of mature miRNA in different regions in random model

Regions		3' terminal		
		CLR+BLR	NLR+FLR	EFLR+WLR
5' terminal	CLR+BLR	17 (2.94%)	68 (11.74%)	30 (5.18%)
	NLR+FLR	76 (13.13%)	175 (30.22%)	66 (11.40%)
	EFLR+WLR	41 (7.08%)	75 (12.95%)	31 (5.35%)

Apparently, the 5' and 3' terminals of mature miRNAs prefer locating near loop structures. The statistical significant difference implies that positions of mature miRNAs on their precursors have been evolutional selected rather than randomly distributed.

C. *Free Energy of MiRNA*

It has been suggested that thermodynamic properties are related with function of miRNA genes [25]. Since an unexpected position distribution bias of mature miRNAs on their precursors has been observed, we then further analyzed the free energy distribution of pre-miRNAs. Statistically, free energies of 1-6nt segment flanking the mature miRNA:miRNA* duplex are much higher than those of mature miRNA terminal regions (see Fig. 3).

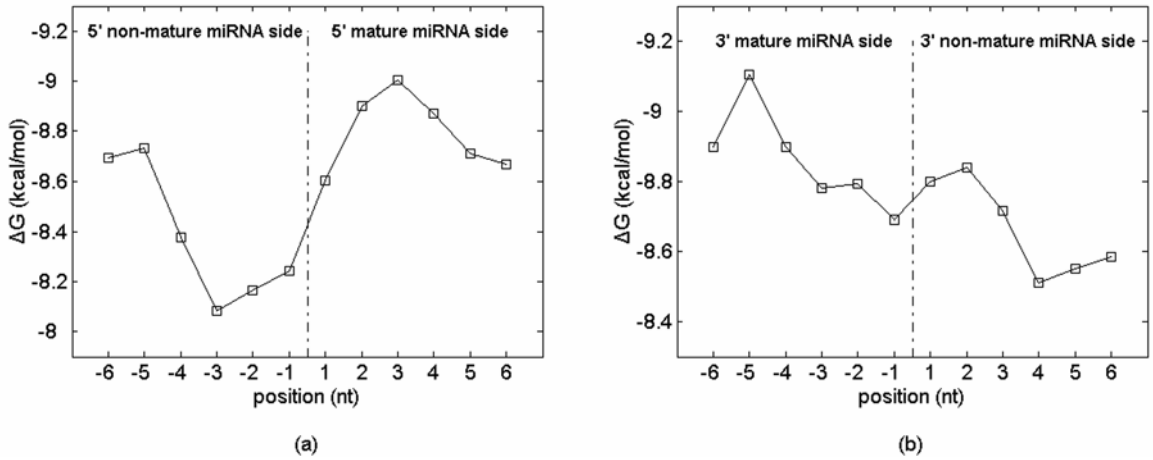

Fig. 3. Predicted average internal stability profiles for mature miRNA terminal regions. (a) Average internal stability profiles of 5' terminal regions of mature miRNA; (b) average internal stability profiles of 3' terminal regions of mature miRNA

IV. Discussion and Conclusion

Our studies showed that positions of mature miRNAs on their precursors are not randomly distributed, which infers that these distributions are evolutional selected. Each terminal of mature

miRNA tends to locate near a loop structure rather than very far from a loop or in a loop. This is a new character of stem-loop structure of miRNA beyond those structure characters used in present miRNA predicting programs. Taking MiRscan as an example, we defined a new rule for scoring as:

$$S = S_{MiRscan} + \omega P(D_{min} = d) \qquad (2)$$

where $S_{MiRscan}$ is the MiRscan score of a candidate miRNA, D_{min} is the minimum between $D_{5'}$ and $D_{3'}$ of this candidate miRNA, $P(D_{min}=d)$ is the frequency of D_{min} equaling to d in our dataset above and ω is the weight coefficient. We scored 60 true *C. elegans* miRNAs in our dataset and the ~36,000 *C. elegans* sequences used in the analysis of MiRscan as a candidate set of miRNAs [6]. MiRNAs whose score are higher than some certain cut-off score are considered as top candidate miRNAs and might be selected for experimental validation. Then we compared the results with MiRscan in the sensitivity of true miRNAs using different cut-off score and the related percentage of top candidate miRNAs (see Fig. 4). The results show that our new scoring rule can get smaller number of top candidate miRNAs than MiRscan based on the same sensitivity of true miRNAs, which can provide more accurate clues for experimental validation of miRNA.

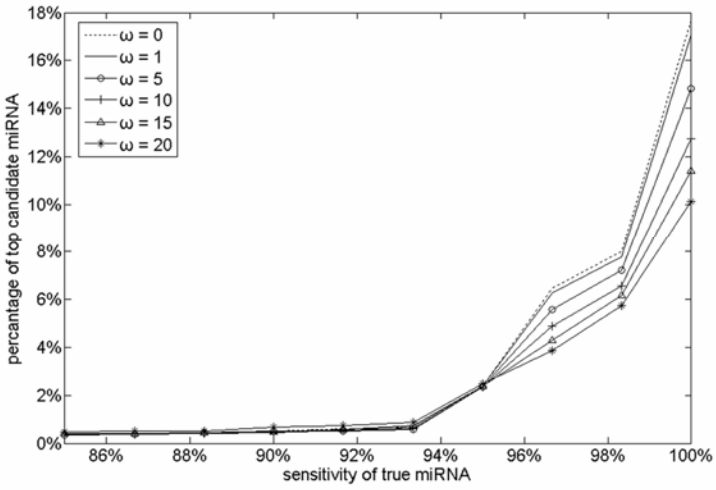

Fig. 4. Sensitivity of true miRNAs and related percentage of top candidate miRNAs according to different weight coefficient (ω)

Since a loop structure results in an unstable region with higher free energy in the secondary structure, our observation can also incurred us to deduce that loop structures near the mature miRNA terminals might be closely related to miRNA biogenesis. And so, existing of a loop structure near each terminal of mature miRNA might facilitate the cleavage of mature miRNA:miRNA* duplex. Site-directed-mutagenesis experiments were performed to remove the internal loops and bulges at or near the cleavage sites of mature miRNA [10]. These mutants were processed efficiently, indicating that the internal loops and bulges in miRNA:miRNA* duplex are not essential for processing at least for miR-30. Mature miRNA can be processed as long as there are loops near its terminals within 6nt. This might infer that the distance between each terminal of mature miRNA and its nearest loop has robust to a certain extent. Apparently, a hairpin stem-loop with too stable secondary structure should not be an optimal substrate for Drosha or Dicer RNase. It remains possible that the loop structure near either terminal of mature miRNA:miRNA* is associated with the miRNA biogenesis, which facilitates the cleavage of mature miRNA duplex.

From the discussion above, we can see that the distribution of mature miRNA on its precursor is not randomly distributed. This character provides a new criterion for miRNA prediction. Further more, it also indicates that the distribution of mature miRNA on its precursor might facilitate miRNA biogenesis.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (Grand No. 60405001, 60234020) and Chinese Postdoctoral Science foundation (No. 2003034023). The authors wish it to be known that the first three authors contributed equally to this work and should be regarded as joint first authors.

References

[1] M. Lagos-Quintana, R. Rauhut, W. Lendeckel, T. Tuschl, "Identification of novel genes coding for small expressed RNAs", Science, vol.294, pp.853-858, 2001.

[2] N. C. Lau, L. P. Lim, E. G. Weinstein, D. P. Bartel, "An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans", Science, vol.294, pp.858-862, 2001.

[3] R. C. Lee, V. Ambros, "An extensive class of small RNAs in Caenorhabditis elegans", Science, vol.294, pp.862-864, 2001.

[4] E. C. Lai, P. Tomancak, R. W. Williams, G. M. Rubin, "Computational identification of Drosophila microRNA genes", Genome Biol. vol.4, pp.R42, 2003.

[5] L. P. Lim, M. E. Glasner, S. Yekta, C. B. Burge, D. P. Bartel, "Vertebrate microRNA genes", Science, vol.299, pp.1540, 2003.

[6] L. P. Lim, N. C. Lau, E. G. Weinstein, A. Abdelhakim, S. Yekta, M. W. Rhoades, C. B. Burge, D. P. Bartel, "The microRNAs of Caenorhabditis elegans", Genes Dev. vol.17, pp.991-1008, 2003.

[7] J. F. Wang, H. Zhou, Y. Q. Chen, Q. J. Luo, L. H. Qu, "Identification of 20 microRNAs from Oryza sativa", Nucleic Acids Res. vol.32, pp.1688-1695, 2004.

[8] S. Pfeffer, M. Zavolan, F.A. Grasser, M. Chien, J.J. Russo, J. Ju, B. John, A.J. Enright, D. Marks, C. Sander, T. Tuschl, "Identification of virus-encoded microRNAs", Science, vol.304, pp.734-736, 2004.

[9] Y. Lee, K. Jeon, J. T. Lee, S. Kim, V. N. Kim, "MicroRNA maturation: stepwise processing and subcellular localization", Embo. J. vol.21, pp.4663-4670, 2002.

[10] Y. Lee, C. Ahn, J. Han, H. Choi, J. Kim, J. Yim, J. Lee, P. Provost, O. Radmark, S. Kim, V. N. Kim, "The nuclear RNase III Drosha initiates microRNA processing", Nature, vol.425, pp.415-419, 2003.

[11] R. Yi, Y. Qin, I. G. Macara, B. R. Cullen, "Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs", Genes Dev. vol.17, pp.3011-3016, 2003.

[12] E. Bernstein, A. A. Caudy, S. M. Hammond, G. J. Hannon, "Role for a bidentate ribonuclease in the initiation step of RNA interference", Nature, vol.409, pp.363-366, 2001.

[13] R. H. Nicholson, A. W. Nicholson, "Molecular characterization of a mouse cDNA encoding Dicer, a ribonuclease III ortholog involved in RNA interference", Mamm. Genome, vol.13, pp.67-73, 2002.

[14] Y. S. Lee, K. Nakahara, J. W. Pham, K. Kim, Z. He, E. J. Sontheimer, R. W. Carthew, "Distinct Roles for Drosophila Dicer-1 and Dicer-2 in the siRNA/miRNA Silencing Pathways", Cell, vol.117, pp.69-81, 2004.

[15] G. Hutvagner, J. McLachlan, A. E. Pasquinelli, E. Balint, T. Tuschl, P. D. Zamore, "A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA", Science, vol.293, pp.834-838, 2001.

[16] R. F. Ketting, S. E. Fischer, E. Bernstein, T. Sijen, G. J. Hannon, R. H. Plasterk, "Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in C. elegans", Genes Dev. vol.15, pp.2654-2659, 2001.

[17] D. P. Bartel, "MicroRNAs: genomics, biogenesis, mechanism, and function", Cell, vol.116, pp.281-297, 2004.

[18] L. He, G.J. Hannon, "MicroRNAs: small RNAs with a big role in gene regulation", Nat. Rev. Genet. vol.5, pp.522-531, 2004.
[19] V. Ambros, B. Bartel, D. P. Bartel, C. B. Burge, J. C. Carrington, X. Chen, G. Dreyfuss, S. R. Eddy, S. Griffiths-Jones, M. Marshall, M. Matzke, G. Ruvkun, T. Tuschl, "A uniform system for microRNA annotation", RNA, vol.9, pp.277-279, 2003.
[20] S. Griffiths-Jones, "The microRNA Registry", Nucleic Acids Res. 32, pp.D109-111, 2004.
[21] I. L. Hofacker, "Vienna RNA secondary structure server", Nucleic Acids Res. vol.31, pp.3429-3431, 2003.
[22] E. V. Kriventseva, I. Koch, R. Apweiler, M. Vingron, P. Bork, M. S. Gelfand, S. Sunyaev, "Increase of functional diversity by alternative splicing", Trends Genet. vol.19, pp.124-128, 2003.
[23] S. M. Freier, R. Kierzek, M. H. Caruthers, T. Neilson, D. H. Turner, "Free energy contributions of G.U and other terminal mismatches to helix stability", Biochemistry, vol.25, pp.3209-3213, 1986.
[24] S. M. Freier, R. Kierzek, J. A. Jaeger, N. Sugimoto, M. H. Caruthers, T. Neilson, D. H. Turner, "Improved free-energy parameters for predictions of RNA duplex stability", Proc. Natl. Acad. Sci. U.S.A. vol.83, pp.9373-9377, 1986.
[25] A. Khvorova, A. Reynolds, S. D. Jayasena, "Functional siRNAs and miRNAs exhibit strand bias", Cell, vol.115, pp.209-216, 2003.

Huiyu Xia received the B.S. degree in automation from Tsinghua University, Beijing, China, in 2000. She is currently working toward the Ph.D. degree in pattern recognition and intelligent systems at Tsinghua University. Her research interests are pattern recognition and bioinformatics, especially in alternative splicing and MicroRNA.
E-mail: xiahuiyu00@mails.tsinghua.edu.cn

Fei Li received the B.S. degree in Entomology and the Ph.D. degree in Insect biochemistry and molecular biology from Nanjing Agricultural University, Nanjing, China, in 1996 and 2003, respectively. From 2003 to 2005, he worked as a Postdoc fellow in Tsinghua University, Beijing, China. His research interests are molecular biology and bioinformatics, especially in MicroRNA. He is currently a professor of College of Plant Protection, Nanjing Agricultural University, Nanjing, China.
E-mail: flee@tsinghua.edu.cn

Tao He received the B.S. degree in automation from Xidian University, Xi'an, China, in 2002. He is currently working toward the Ph.D. degree in pattern recognition and intelligent systems at Tsinghua University, Beijing, China. His research interests are pattern recognition and bioinformatics, especially in MicroRNA.
E-mail: ht02@mails.tsinghua.edu.cn

Yanda Li received the B. S. degree in automatic control from Tsinghua University, Beijing, China, in 1959. From 1979 to 1981, he was a Visiting Scholar in the Department of Electrical Engineer and Computer Science, Massachusetts Institute of Technology (MIT), Cambridge, MA. His research interests are bioinformatics, application of complex systems, and signal processing. He is a professor of the Department of Automation, Tsinghua University, Beijing, China and a member of Chinese Academy of Sciences.
E-mail: daulyd@tsinghua.edu.cn