

Sequence analysis

MicroRNA identification based on sequence and structure alignment

Xiaowo Wang[†], Jing Zhang[†], Fei Li, Jin Gu, Tao He, Xuegong Zhang and Yanda Li*

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

Received on March 23, 2005; revised on June 25, 2005; accepted on June 27, 2005

Advance Access publication June 30, 2005

ABSTRACT

Motivation: MicroRNAs (miRNA) are ~22 nt long non-coding RNAs that are derived from larger hairpin RNA precursors and play important regulatory roles in both animals and plants. The short length of the miRNA sequences and relatively low conservation of pre-miRNA sequences restrict the conventional sequence-alignment-based methods to finding only relatively close homologs. On the other hand, it has been reported that miRNA genes are more conserved in the secondary structure rather than in primary sequences. Therefore, secondary structural features should be more fully exploited in the homologue search for new miRNA genes.

Results: In this paper, we present a novel genome-wide computational approach to detect miRNAs in animals based on both sequence and structure alignment. Experiments show this approach has higher sensitivity and comparable specificity than other reported homologue searching methods. We applied this method on *Anopheles gambiae* and detected 59 new miRNA genes.

Availability: This program is available at <http://bioinfo.au.tsinghua.edu.cn/miralign>

Contact: daulyd@tsinghua.edu.cn

Supplementary information: Supplementary information is available at <http://bioinfo.au.tsinghua.edu.cn/miralign/supplementary.htm>

INTRODUCTION

MicroRNAs (miRNAs) are ~22 nt long endogenous non-coding RNAs that play important regulatory roles in diverse organisms (Bartel and Bartel, 2003; Bartel, 2004; He and Hannon, 2004; Lai, 2003). In animal cells, miRNA genes are first transcribed as long pri-miRNAs (Lee *et al.*, 2002) and processed to ~70 nt precursors (pre-miRNA) with stem-loop structure by the RNase III enzyme Drosha (Lee *et al.*, 2003). Then, another RNase III enzyme Dicer (Grishok *et al.*, 2001; Hutvagner *et al.*, 2001; Ketting *et al.*, 2001) cuts the pre-miRNAs to release the ~22 nt mature miRNAs (Lee *et al.*, 2002, 2003). Finally, RNA-induced silencing complexes (Hammond *et al.*, 2000) are formed to regulate the expression of target genes via complementary base pair interactions. In plants, the maturation process of miRNAs is similar to that in animals, but the length of the pre-miRNAs are more variable and their structures are more complex (Bartel, 2004).

Since the first miRNA was discovered in *Caenorhabditis elegans* (Lee *et al.*, 1993) hundreds of miRNAs have been cloned in many organisms (Lagos-Quintana *et al.*, 2001; Lau *et al.*, 2001; Lee and Ambros, 2001). However, only abundant miRNA genes can be easily detected by PCR or northern blot due to limitations of the techniques. For finding those low-expression or tissue-specific miRNA genes computational prediction provides an efficient strategy (Bartel, 2004).

Up to now, several computational approaches have been reported to identify miRNAs. Comparative genomics was adopted to find entirely novel miRNA families in specific animals and plants (Bonnet *et al.*, 2004; Jones-Rhoades and Bartel, 2004; Lai *et al.*, 2003; Lim *et al.*, 2003a,b; Ohler *et al.*, 2004; Wang *et al.*, 2004). Homologue search was used to reveal orthologs and paralogs of known miRNAs (Lagos-Quintana *et al.*, 2001; Lau *et al.*, 2001; Lee and Ambros, 2001; Maher *et al.*, 2004; Pasquinelli *et al.*, 2000; Weber, 2005). Because the mature miRNA sequences are short (~22 nt), current sequence alignment tools like BLAST (Altschul *et al.*, 1990) can only find the nearly-perfect matches due to the large number of irrelevant hits. The ~70 nt pre-miRNA sequences have also been used for homologue search. However, compared with miRNA and miRNA* (the fragments on the opposite arm of the hairpin) (Lau *et al.*, 2001) the other parts of the precursor sequence are less conserved. Therefore, sequence alignment alone may fail to detect the distant homologs that diverge in sequence but keep conservation in structure (Legendre *et al.*, 2004). So, more sensitive approaches that take the advantage of both sequence and structure conservation are needed. Recently, an ERPIN (Gautheret and Lambert, 2001) search strategy was reported (Legendre *et al.*, 2004), which used profiles to capture both sequence and structure information of animal miRNA precursors. This method increased the number of found novel miRNA candidates by 17% compared with BLAST search, but the construction of profile makes it only applicable to miRNA families with sufficient known samples.

In this paper, we present a novel computational approach named miRAlign that detects new miRNAs based on both sequence and structure alignment. Two main characters make miRAlign distinct from existing homologue search methods: firstly, to be able to find distant homologs, miRAlign requires neither the sequence conservation of the whole pre-miRNA sequence nor the nearly-perfect match of the ~22 nt mature part, but just assumes relatively loose conservation of the mature miRNA sequence. Secondly, more properties of miRNA structure conservation are considered. And unlike profile search methods, which need relatively large family members

*To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

to construct the profile, miRAlign introduces a structure alignment strategy and can use each single miRNA as a query to do homology search. In our experiments, miRAlign outperforms conventional BLAST search and ERPIN search by higher sensitivity and comparable specificity. Using this method, more distant miRNA homologs or orthologs can be revealed.

MATERIALS AND METHODS

miRNA reference sets and genomic sequences

The miRNAs and their precursor sequences were downloaded from the MicroRNA Registry (Ambros *et al.*, 2003; Griffiths-Jones, 2004) release 5.0 (<http://www.sanger.ac.uk/Software/Rfam/mirna/index.shtml>). This set contains 1298 miRNAs (generated by 1345 pre-miRNAs) from 12 species (*C.elegans*, *Caenorhabditis briggsae*, *Drosophila melanogaster*, *Drosophila pseudoobscura*, *Danio rerio*, *Gallus gallus*, *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, Epstein Barr virus, *Arabidopsis thaliana* and *Oryza sativa*), of which 1054 are from animals. Most of these miRNAs were identified or verified by experiments, and others were computationally predicted as their close homologs. These 1054 animal miRNAs and their precursors (1104 pre-miRNAs) composed our raw training set Train_All. In order to illustrate miRAlign performance on *C.briggsae* genomic sequences, two sub-training sets were constructed. Train_Sub_1 consists of all the known animal miRNAs except for those from *C.briggsae*. It contains 976 miRNAs and 1025 pre-miRNA sequences. Train_Sub_2 was constructed by removing both the *C.briggsae* and *C.elegans* miRNAs from Train_All, leaving 859 miRNAs and 909 precursors.

Genomic sequences of six species were used in this work, including *Anopheles gambiae* (MOZ2a, <http://www.ensembl.org/>), *C.briggsae* (cb25.agp8, http://www.sanger.ac.uk/Projects/C_briggsae/), *C.elegans* (WormBase release WS 140, www.wormbase.org), *D.melanogaster* (release 4, <http://www.fruitfly.org/annot/release4.html>), *D.pseudoobscura* (freeze1, <http://www.hgsc.bcm.tmc.edu/projects/drosophila/>) and *G.gallus* (WASHUC1, <http://www.genome.wustl.edu/projects/chicken/>).

Extraction of miRNA candidate sequences (preprocessing)

Several preprocessing steps are taken to draw miRNA candidate sequences from the genome. Firstly, all the known pre-miRNAs in the training set are used as queries to BLAST search against the genome with a sensitive BLAST parameter setting (word-length 7 and *E*-value cutoff 10). Next, sequence segments of the potential regions are cut from the genome with 70 nt flanking sequences to each end and scanned by a 100 nt-sliding window with a step of 10 nt. The sequences overlapping with repeat sequences are discarded and the remains are treated as miRNA candidates to be scored by miRAlign.

miRAlign

miRAlign is designed to find new miRNAs using both sequence information and structural characteristics of known miRNAs. For a given ~70 nt candidate miRNA precursor sequence, miRAlign assigns it a similarity score by following five steps.

(1) *Secondary structure prediction.* Both the candidate sequence and its reverse complementary are analyzed by miRAlign. The secondary structures of both strands of the candidate are predicted by RNAfold (Hofacker *et al.*, 1994) using minimum free energy (MFE) rule (Zuker and Stiegler, 1981). Only the strands with MFE lower than -20 kcal/mol are kept for further analysis.

(2) *Pairwise sequence alignment.* The strands of the candidate sequences that pass Step 1 are pairwise aligned to all the ~22 nt known miRNA sequences in the training set. Sequence similarity score (*mature_seq_sim*) between the candidate and each known mature miRNAs is calculated by CLUSTALW (Thompson *et al.*, 1994). The candidate-to-known miRNA

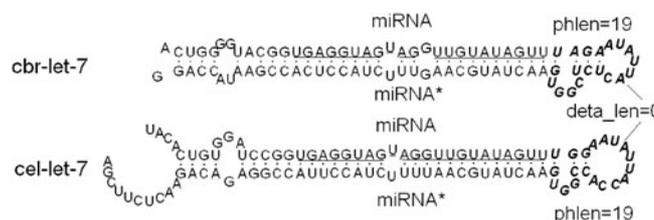


Fig. 1. Definition of *hplen* and δ_len . The nucleotides highlighted with underline are the mature miRNAs, and their complementary on the other arm of the stem-loop structure are miRNAs*. *hplen* is defined as the sequence length of the nucleotide in boldface italics and δ_len is defined as the absolute difference *hplen* value of two miRNA precursors.

pairs are kept as potential homologue pairs for further analysis only if their *mature_seq_sim* exceed a user-defined threshold *min_seq_sim*. Here, *min_seq_sim* = 70 is selected as the default parameter for miRAlign in our experiments. More than 98.1% of the known animal miRNA homologs have sequence similarity higher than this value (Supplementary Figure S1). For each of those potential homologue pairs, the ~22 nt sub-sequence on the candidate that aligns to the known miRNA is regarded as potential miRNA.

(3) *miRNA's position on the stem-loop structure.* Three properties for the ~22 nt miRNA's position on the stem-loop structure derived from the miRNA reference set are considered by miRAlign for each of the potential homologue pairs: (a) the ~22 nt potential miRNA sequence should not locate on the terminal loop of the hairpin structure; (b) potential miRNA should locate on the same arm of the stem-loop structure as its known homologs and (c) the position of the potential miRNA sequence on the stem-loop structure should not differ too much from its known homologs.

The position difference of the mature miRNAs (or potential mature miRNAs) on the stem-loop structure between precursor A and B is calculated by

$$\delta_len(A, B) = |hplen(A) - hplen(B)|,$$

where *hplen* denotes the number of the nucleotides between the ~22 nt miRNA and miRNA* (Fig. 1). A δ_len value close to zero indicates that there are few deletions or insertions occurred between the miRNA homologs. In our experiments, a non-stringent δ_len cutoff 15 was used as the default parameter. Over 97.5% of the known animal miRNA homologue pairs have δ_len less than this value (Supplementary Figure S2).

(4) *RNA secondary structure alignment.* To measure the secondary structure conservation of the potential homologue pairs, RNAforester (Hochsmann *et al.*, 2003) is used to compute pairwise structure alignment. This tool implements a tree alignment model. It can calculate local structure alignment and give a similarity measure of two structures. Here, the parameter '-RIBOSUM' is used, and RNAforester calculates structure alignment considering both sequence and structure information by using the base and base pair substitution matrix RIBOSUM85-60 as described in Klein and Eddy (2003). Using this log-odd position independent substitution matrix, this tool can use a single sequence as a query to align against the target sequence (while profile-based scoring scheme like ERPIN needs large family members to construct the profile and compute position dependent log-odds scores).

The raw structure alignment score σ_{local} computed by RNAforester is the summation of all base (base pair) match (insertion, detention, etc.) scores in the alignment. It cannot be used directly to measure the similarity for the structures with different sizes. We define the normalized similarity score of structure *C* and *m* as

$$str_sim(C, m) = \frac{\sigma_{local}(C, m)}{\sigma(m, m)} \times 100,$$

where *C* is the candidate sequence, *m* is the known pre-miRNA, $\sigma_{local}(C, m)$ denotes the raw local alignment score of two structures *C* and *m*, and $\sigma(m, m)$ is the self-alignment score of *m*. After the normalization by $\sigma(m, m)$, *str_sim*

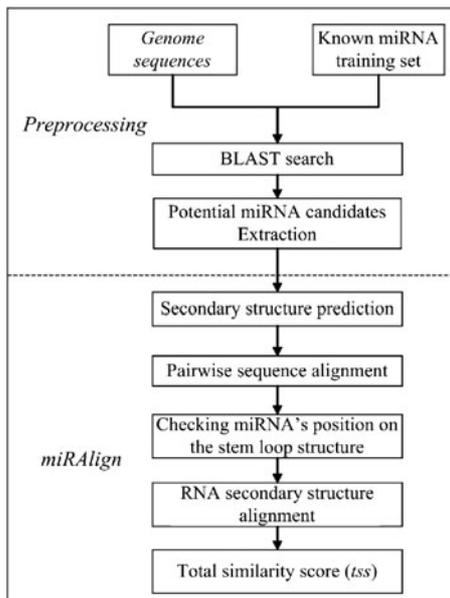


Fig. 2. Overview of miRAlign, a systematic computational approach to identify miRNAs in animals. See text for details.

ranges from 0 to 100. The higher this value, the more similar are the two structures.

(5) *Total similarity score.* After aligning all the potential homologue pairs, a total similarity score (*tss*) is assigned to the candidate sequence. This *tss* is defined as

$$tss(C, R) = \begin{cases} \max_{P \in R} (str_sim(C, P)) & \text{if } R \neq \emptyset \\ 0 & \text{if } R = \emptyset \end{cases}$$

where *C* denotes the candidate sequence, *R* is the set composed of all the *C*'s miRNA homologs identified in Step 3. *tss* is defined to be zero if *R* is empty. As the two strands of the candidate sequence are analyzed by miRAlign independently, the strand with higher score is chosen as the final result for the candidate sequence.

An overview description of the miRAlign procedures is shown in Figure 2.

RESULTS AND DISCUSSION

Application on *C.briggsae*

MicroRNA Registry 5.0 contains 79 *C.briggsae* miRNAs, all of which were predicted by the homologous search. So we first applied our method on *C.briggsae* data with the training set *Train_Sub_1* to check the sensitivity and specificity of miRAlign. Then *Train_Sub_2* was used as the training set to investigate the ability of miRAlign to find new miRNAs between distantly related species.

Detection of miRNA homologs. Using training set *Train_Sub_1*, which contains all the animal miRNAs except for those from *C.briggsae*, ~96 500 candidate sequences passed the preprocessing step and all of them were scored by miRAlign. Candidates belonging to same regions on the genome were merged and represented by the one with the highest *tss* score. In total, ~1050 non-overlapping candidates got none-zero *tss*, including 74 miRNAs among the 79 reported *C.briggsae* miRNA genes (Fig. 3A). With a *tss* cutoff of 35, 90 putative miRNA sequences were detected, and a sensitivity of

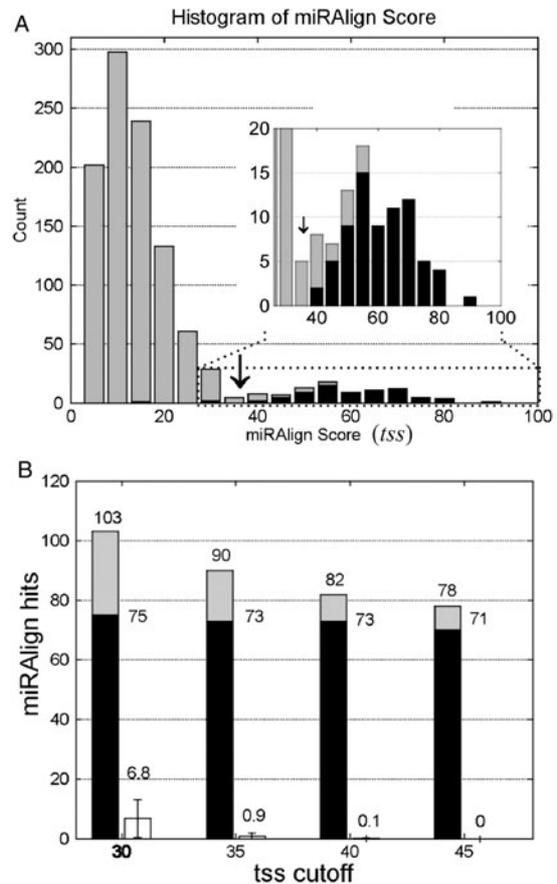


Fig. 3. (A) Histogram of the *tss* of *C.briggsae* miRNA candidate sequences using the training set *Train_Sub_1*. This test set contains ~1050 independent sequences that got non-zero *tss*. The known miRNA genes captured by miRAlign are represented using black bar. With *tss* > 35, 90 putative miRNA genes were identified, including 73 copies of 71 known *C.briggsae* miRNAs. (B) miRAlign hits with different cutoff. Gray bars denote all the detected putative miRNAs in *C.briggsae* and the black bars represent the known miRNAs hits. The white bars are the average numbers of false positive hits estimated based on 10 shuffled genomes.

89.9% was achieved by finding 73 copies of 71 known *C.briggsae* miRNA genes. Supplementary Table S1 lists the novel putative miRNAs with *tss* > 35. Eight genuine *C.briggsae* miRNA genes with *tss* cutoff 35 were not identified by miRAlign. Four of them were missed during the BLAST search in the preprocessing step; one's mature part overlaps with the hairpin-loop; and the other three got *tss* < 35. For the four 'BLAST-missing' miRNAs, we used their reported precursor sequences directly as the input of miRAlign and three of them got *tss* > 35.

Identification of miRNAs in distantly related species. All *C.briggsae* miRNAs in MicroRNA Registry 5.0 were predicted based on close homologs to verified miRNAs from *C.elegans*. Obviously, the *C.elegans* miRNAs in the training set greatly contribute to the homologue search results. In order to investigate the ability of miRAlign to identify miRNAs in distantly related species, we performed the experiment described above by using the training set *Train_Sub_2*, in which both the *C.briggsae* and *C.elegans* miRNAs are excluded. There were 18 putative miRNAs that got *tss* > 35, but

Table 1. Summary of the performance of miRAlign and BLAST search on *C.briggsae* genomic sequences

Training set	Method	Cut off	Total hits ^a	Known miRNA hits ^b	Sensitivity	Average FP hits ^c
<i>Train_Sub_1</i>	miRAlign	<i>tss</i> 35	90	71	89.9% (71/79)	0.9
	BLAST	<i>E</i> -value 0.01	88	66	83.5% (66/79)	7.1
<i>Train_Sub_2</i>	miRAlign	<i>tss</i> 35	18	8	10.1% (8/79)	0.8
	BLAST	<i>E</i> -value 0.01	17	5	6.3% (5/79)	5.9

^aTotal number of putative miRNAs detected by the methods.

^bThe number of known miRNAs detected by the approach.

^cAverage number of false positive hits per genome, estimated by 10 times shuffling of the *C.briggsae* genome. Both strands were searched.

only 8 known *C.briggsae* miRNAs genes were included (Table 1). We investigated these known miRNAs that failed to be detected in this experiment and found that most of them were lost in the preprocessing step. If we directly use miRAlign to score the known *C.briggsae* pre-miRNAs with a sensitive parameter *mature_seq_sim* 60, 20 known miRNAs could get *tss* > 35. This suggests that miRAlign can be used to find miRNAs in relatively distant related species with the help of more sensitive sequence alignment approaches in the preprocessing step. This character will facilitate the extension of known miRNA set from the current limited number of species to a wide range. However, we also noticed that there still remain some miRNA homologs failed to be detected by miRAlign because of their greater divergence in either primary sequences or secondary structures.

tss cutoff selection. As shown in Figure 3A, we may detect more putative miRNAs by using a lower *tss* cutoff. However, the increase in sensitivity will decrease the specificity. So it is essential to find out a *tss* cutoff that has reasonable selectivity and satisfactory sensitivity. Because the number of false positives is inaccessible directly without experimental verification, we estimated it based on the miRAlign hits on 10 randomly shuffled *C.briggsae* genomes using the training set *Train_Sub_1*. When choosing a *tss* of 45 as a cutoff, 86.1% sensitivity is achieved and no false positive hits are found. When the cutoff decreases to 35, the sensitivity increases to 89.9% while the average number of false positive hits only increases to 0.9 per random genome. As shown in Figure 3B, typically, a *tss* cutoff between 35 and 45 can reasonably balance the sensitivity and selectivity. Though the number of false positives is influenced by many factors like GC content of the genome, in our experiments, it is mainly related to the size of the training set and the number of candidate sequences (which is principally determined by the BLAST *E*-value in the preprocessing step) in the test set. Here, the BLAST *E*-value 10 was selected in the preprocessing step and the training set contains ~1000 miRNAs. Based on this estimation, one can choose a proper *tss* cutoff for his own application.

Comparison with other methods

We compared miRAlign with BLAST search and ERPIN search (Legendre *et al.*, 2004) on several genomic datasets. All the known pre-miRNA sequences in training sets *Train_Sub_1* and *Train_Sub_2* were used as BLAST search queries against the *C.briggsae* genome, with a non-stringent BLAST parameter setting (word length = 7, *E*-value cutoff = 0.01). Only the BLAST hits that can be fold with MFE lower than -20 kcal/mol were treated as putative miRNAs. Table 1 shows the summary of the results of BLAST search and

Table 2. Summary of miRAlign and ERPIN search performance on *C.briggsae* genomic sequences using the 18 animal miRNA training sets provided by ERPIN website

Species	Method	Total hits	Known miRNAs hits	Ave FP hits ^a
<i>C.briggsae</i>	ERPIN	16	16	0.3
	miRAlign	23	21	0
<i>C.elegans</i>	ERPIN	24	23	0.2
	miRAlign	25	25	0
<i>D.melanogaster</i>	ERPIN	31	31	0.2
	miRAlign	31	31	0.2
<i>D.pseudoobscura</i>	ERPIN	22	22	0.1
	miRAlign	28	28	0.2
<i>G.gallus</i>	ERPIN	54	51	0
	miRAlign	59	54	0

^aAverage number of false positive hits per genome, estimated by 10 times shuffling of the genome (one time for *G.gallus*). Both strands were searched.

miRAlign search. With *Train_Sub_1*, 66 and 71 known miRNAs genes were detected by BLAST search and miRAlign, respectively, 63 of which are identical. The average numbers of false positives are 7.1 and 0.9 using the two methods, respectively. With *Train_Sub_2*, known miRNAs detected by BLAST search and by miRAlign sharply decreased to 5 and 8, respectively. And the average numbers of false positives per *C.briggsae* genome slightly reduced to 5.9 and 0.8. In this case, all the five BLAST-hits were also detected by miRAlign. This result illustrates that miRAlign outperforms BLAST search in both sensitivity and selectivity, and furthermore, nearly all the known miRNAs hit by BLAST can be detected by miRAlign also.

The latest version of ERPIN 4.2.5 and the 18 animal miRNA training sets (Legendre *et al.*, 2004) were downloaded from <http://tagc.univ-mrs.fr/erpin/>, and the recommended parameters for each training set were also got from the ERPIN website. These datasets contain 180 pre-miRNAs from *C.briggsae* (15), *C.elegans* (23), *D.melanogaster* (31), *H.sapiens* (54) and *M.musculus* (57). For the sake of comparison, we performed ERPIN and miRAlign search using these 180 miRNAs as training data on five genomes (*C.elegans*, *C.briggsae*, *D.melanogaster*, *D.pseudoobscura* and *G.gallus*) and used ERPIN *E*-value 0.01 and miRAlign *tss* 35 as our threshold for the putative miRNAs. Table 2 shows the statistics of miRAlign and ERPIN results. Despite that the training set is favorable for ERPIN search (all these miRNAs in the training set can be grouped in

large families to construct the profile), miRAlign achieved comparable specificity and higher sensitivity than ERPIN search on all five genomes used in this study, and identified all the known miRNAs searched by ERPIN. Though on these large miRNA families the improvement on sensitivity is not very dramatic, in practical applications, greater improvement can be expected because miRAlign can be applied on all miRNA families but ERPIN only works for larger ones.

Prediction of miRNAs in *A.gambiae*

A.gambiae, which has diverged from *D.melanogaster* ~250 million years ago, is a highly adapted, successful dipterans species (Gaunt and Miles, 2002; Yeates and Wiegmann, 1999). This mosquito is a major vehicle for the transmission of malaria. Up to now, no experimentally identified or computationally predicted *A.gambiae* miRNAs are reported. We applied our approach on the *A.gambiae* data using Train_all as the training set. There are totally ~1100 non-overlapped candidates getting non-zero *tss*, of which 59 putative miRNA sequences are detected with *tss* > 35. During our revision of this paper, 38 *A.gambiae* miRNAs were reported in the MicroRNA Registry release 6.0, and 37 of them were covered by our predictions. This result validates the sensitivity and accuracy of our method. Supplementary Table S2 lists the putative novel miRNAs predicted by miRAlign.

CONCLUSION

We developed a new computational approach called miRAlign to predict new miRNA genes that are homologs or orthologs to known miRNAs. Combining sequence and structure alignments, miRAlign has better performance than previously reported homologue search methods. With the help of this method and other high-throughput computational approaches, systematic genome-wide annotation will gradually reveal the panorama of the miRNAs and help us to further understand the functions and evolutionary footprint of miRNAs.

It should be noted that our experiments in this work are mainly on animal data. We also investigated the feasibility of the application of miRAlign for plant miRNA homologue search. A sensitivity of 70% was achieved by detecting 28 of the 40 known *Zea mays* miRNAs genes (Maher et al., 2004) with *A.thaliana* and *O.sativa* miRNAs as the training set (data not shown). This result indicates that miRAlign can be also applied in plants and further investigation is undergoing.

ACKNOWLEDGEMENTS

We appreciate Matthias Hochsmann for providing the source code for RNAforester. Thanks to Ying Huang, Jun Cai and Chenghai Xue for helpful discussions. We also thank Prof. Yunshen Sun for checking the language. This project is supported in part by NSFC (grants 60171038, 60234020, 60405001) and the National Basic Research Program (2004CB518605) of China.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. et al. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
 Ambros,V. et al. (2003) A uniform system for microRNA annotation. *RNA*, **9**, 277–279.
 Bartel,B. and Bartel,D.P. (2003) MicroRNAs: at the root of plant development? *Plant Physiol.*, **132**, 709–717.
 Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.

Bonnet,E. et al. (2004) Detection of 91 potential conserved plant microRNAs in *Arabidopsis thaliana* and *Oryza sativa* identifies important target genes. *Proc. Natl Acad. Sci. USA*, **101**, 11511–11516.
 Gaunt,M.W. and Miles,M.A. (2002) An insect molecular clock dates the origin of the insects and accords with palaeontological and biogeographic landmarks. *Mol. Biol. Evol.*, **19**, 748–761.
 Gautheret,D. and Lambert,A. (2001) Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles. *J. Mol. Biol.*, **313**, 1003–1011.
 Griffiths-Jones,S. (2004) The MicroRNA Registry. *Nucleic Acids Res.*, **32** (Database issue), D109–D111.
 Grishok,A. et al. (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.
 Hammond,S.M. et al. (2000) An RNA-directed nuclease mediates post-transcriptional gene silencing in *Drosophila* cells. *Nature*, **404**, 293–296.
 He,L. and Hannon,G.J. (2004) MicroRNAs: small RNAs with a big role in gene regulation. *Nat. Rev. Genet.*, **5**, 522–531.
 Hochsmann,M., Toller,T., Giegerich,R. and Kurtz,S. (2003) Local similarity in RNA secondary structures. In *Proceedings of the IEEE Computational Systems Bioinformatics Conference*, Stanford, CA, pp. 159–168.
 Hofacker,I. et al. (1994) Fast folding and comparison of RNA secondary structures. *Monatshfte f. Chemie*, **125**, 167–188.
 Hutvagner,G. et al. (2001) A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. *Science*, **293**, 834–838.
 Jones-Rhoades,M.W. and Bartel,D.P. (2004) Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. *Mol. Cell*, **14**, 787–799.
 Ketting,R.F. et al. (2001) Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in *C. elegans*. *Genes Dev.*, **15**, 2654–2659.
 Klein,R.J. and Eddy,S.R. (2003) RSEARCH: finding homologs of single structured RNA sequences. *BMC Bioinformatics*, **4**, 44.
 Lagos-Quintana,M. et al. (2001) Identification of novel genes coding for small expressed RNAs. *Science*, **294**, 853–858.
 Lai,E.C. (2003) MicroRNAs: runts of the genome assert themselves. *Curr Biol.*, **13**, R925–R936.
 Lai,E.C. et al. (2003) Computational identification of *Drosophila* microRNA genes. *Genome Biol.*, **4**, R42.
 Lau,N.C. et al. (2001) An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science*, **294**, 858–862.
 Lee,R.C. and Ambros,V. (2001) An extensive class of small RNAs in *Caenorhabditis elegans*. *Science*, **294**, 862–864.
 Lee,R.C. et al. (1993) The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, **75**, 843–854.
 Lee,Y. et al. (2002) MicroRNA maturation: stepwise processing and subcellular localization. *EMBO J.*, **21**, 4663–4670.
 Lee,Y. et al. (2003) The nuclear RNase III Drosha initiates microRNA processing. *Nature*, **425**, 415–419.
 Legendre,M. et al. (2004) Profile-based detection of microRNA precursors in animal genomes. *Bioinformatics*, **21**, 841–845.
 Lim,L.P. et al. (2003a) Vertebrate microRNA genes. *Science*, **299**, 1540.
 Lim,L.P. et al. (2003b) The microRNAs of *Caenorhabditis elegans*. *Genes Dev.*, **17**, 991–1008.
 Maher,C., Timmermans,M., Stein,L. and Ware,D. (2004) Identifying microRNAs in plant genomes. In *Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference (CSB 2004)*, Stanford, CA, pp. 718–723.
 Ohler,U. et al. (2004) Patterns of flanking sequence conservation and a characteristic upstream motif for microRNA gene identification. *RNA*, **10**, 1309–1322.
 Pasquinelli,A.E. et al. (2000) Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature*, **408**, 86–89.
 Thompson,J.D. et al. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
 Wang,X.J. et al. (2004) Prediction and identification of *Arabidopsis thaliana* microRNAs and their mRNA targets. *Genome Biol.*, **5**, R65.
 Weber,M.J. (2005) New human and mouse microRNA genes found by homology search. *FEBS J.*, **272**, 59–73.
 Yeates,D.K. and Wiegmann,B.M. (1999) Congruence and controversy: toward a higher-level phylogeny of *Diptera*. *Annu Rev Entomol.*, **44**, 397–428.
 Zuker,M. and Stiegler,P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, **9**, 133–148.