Acta Biochimica Polonica

Vol. 53 No. 2/2006, 303–309

on-line at: www.actabp.pl

Regular paper

An approach for the identification of microRNA with an application to *Anopheles gambiae*

Raghunath Chatterjee and Keya Chaudhuri✉

Human Genetics & Genomics Group, Indian Institute of Chemical Biology, Jadavpur, Kolkata, India;
✉e-mail: kchaudhuri@iicb.res.in

Received: 17 February, 2006; revised: 23 April, 2006; accepted: 08 May, 2006
available on-line: 13 June, 2006

MicroRNAs (miRNAs) are an abundant class of 20–27 nt long noncoding RNAs, involved in post-transcriptional regulation of genes in eukaryotes. These miRNAs are usually highly conserved between the genomes of related organisms and their pre-miRNA transcript, about 60–120 nt long, forms extended stem-loop structure. Keeping these facts in mind miRsearch is developed which relies on searching the homologues of all known miRNAs of one organism in the genome of a related organism allowing few mismatches depending on the phylogenetic distance between them, followed by assessing for the capability of formation of stem-loop structure. The precursor sequences so obtained were then screened through the RNA folding program MFOLD selecting the cut-off values on the basis of known *Drosophila melanogaster* pre-miRNAs. With this approach, about 91 probable candidate miRNAs along with pre-miRNAs were identified in *Anopheles gambiae* using known *D. melanogaster* miRNAs. Out of these, 41 probable miRNAs have 100% similarity with already known *D. melanogaster* miRNAs and others were found to be at least 85% similar to the miRNAs of various other organisms.

Keywords: microRNA, *Anopheles gambiae,* RNA folding

INTRODUCTION

MicroRNAs (miRNAs) are an abundant class of small 20 to 27 nucleotide (nt) noncoding RNAs found in diverse organisms, both plants (Bartel & Bartel, 2003) and animals (Lim *et al.*, 2003a). Many of these are known to control the expression of other genes at the post transcriptional level (Lagos-Quintana *et al.*, 2001; Lau *et al.*, 2001; Lee & Ambros, 2001; Moss & Poethig, 2002; Bartel, 2004). The founding members of this class of noncoding RNAs are the *lin-4* and *let-7* gene products of *Caenorhabditis elegans* (Lee *et al.*, 1993; Reinhart *et al.*, 2000). Both *lin-4* and *let-7* RNAs act as repressor of their respective target genes *lin-14*, *lin-28* and *lin-41* (Lee *et al.*, 1993; Moss *et al.*, 1997; Slack *et al.*, 2000). In all these cases repression was mediated by the presence of complementary miRNA sequences in the 3′ untranslated regions (UTRs) of the target mRNAs (Slack *et al.*, 2000; Lewis *et al.*, 2003).

The unique characteristics of miRNA are — first, all miRNAs are present in noncoding regions of the genome; second, when genomic sequences surrounding the identified 22 nt RNAs were examined, computer analysis predicted miRNA precursors capable of forming stem-loop structure, a single miRNA molecule ultimately accumulates from one arm of each miRNA hairpin precursor molecule; third, miRNA sequences are nearly always conserved in related organisms.

To identify novel miRNAs, several approaches have been used involving biochemical approach based on purification of RNAs after size fractionation (Lau *et al.*, 2001) or bioinformatics approach centering on the conservation of intergenic regions of DNA between two clearly related *Caenorhabditis* species (Lim *et al.*, 2003b) or between *Drosophila melanogaster* and *Drosophila pseudoobscura* species (Lai *et al.*, 2003). Both miRscan and MiRseeker extracted conserved intergenic regions between two closely related species. MiRseeker subjects conserved intronic and intergenic sequences to an RNA folding and evaluation procedure to identify evolutionarily constrained hairpin structures with features charac-

Abbreviations: *Dme, Drosophila melanogaster*; miRNA, microRNA; UTR, untranslated region.

teristic of known miRNAs (Lai *et al.*, 2003). On the other hand, miRscan evaluates conserved stem-loops as miRNA precursors by passing a 21-nt window along each conserved stem-loop, assigning a log-likelihood score to each window that measures how well its attributes resemble those of the first experimentally verified *C. elegans* miRNAs with *C. briggsae* homologs (Lim *et al.*, 2003b).

For the detection of novel miRNAs in specific animals and plants, comparative genomics was used in several reports (Lim *et al.*, 2003a; 2003b; Lai *et al.*, 2003; Bonnet *et al.*, 2004; Jones-Rhoades & Bartel, 2004; Ohler *et al.*, 2004; Wang *et al.*, 2004) and for the detection of orthologs and paralogs of known miRNAs, homology searching was also used (Pasquinelli *et al.*, 2000; Lagos-Quintana *et al.*, 2001; Lau *et al.*, 2001; Lee & Ambros, 2001; Weber, 2005). Using the current sequence alignment tools like blast (Altschul *et al.*, 1990), short sequences (mature miRNAs are of about 22 nt) as a query sequence will produce large number of irrelevant hits. Pre-miRNA sequences are also used for homologue searching. But due to the non-conservation of the other parts compared with miRNA and miRNA* (the fragments on the opposite arm of the hairpin) (Lau *et al.*, 2001), it is expected that, the homology searching based on pre-miRNA sequences may produce many false positives. So, the more sensitive approach will be to consider both sequence and structure conservation. Using this strategy, ERPIN used pro-files to capture both sequence and structure information of animal miRNA precursors (Gautheret & Lambert, 2001; Legendre *et al.*, 2005). Another study have been proposed starting with the BLAST searching with known pre-miRNA sequences followed by assessment of structure information (Wang *et al.*, 2005). Since in many cases though the miRNA sequences may be conserved, the precursor sequences are much less conserved (Lau *et al.*, 2001), so, searching with pre-miRNAs might lead to under estimation of the actual miRNAs present in an organism.

The present strategy uses sequence alignment of mature miRNAs, the structure conservation and assessing the position of the mature miRNAs in the pre-miRNA. Starting with the known mature miRNAs from an organism as query, homologues were searched in a related organism allowing few mismatches depending on their phylogenetic distance. The pre-miRNA and their potentiality to form stem loop structure were assessed further and finally the stem-loop secondary structure was confirmed using MFOLD (Zuker, 2003), followed by assessing the preferable position of the miRNA in the pre-miRNA secondary structure. As an application, about 91 probable miRNA sequences have been identified in *Anopheles gambiae* genome starting from *D. melanogaster* miRNAs.

MATERIALS AND METHODS

Sources of nucleotide sequences. All available *D. melanogaster* miRNAs and the pre-miRNA sequences (79 in number) were selected from the ftp site (Griffiths-Jones, 2004) (ftp://ftp.sanger.ac.uk/pub/databases/Rfam/miRNA/).

The complete genome sequence of *A. gambiae* are arranged in 8987 (Accession No. AAAB01000001 to AAAB01008987) scaffolds for downloading at FTP site at NCBI (ftp://ftp.ncbi.nih.gov/genbank/genomes/Anopheles_Gambiae/Assembly_scaffolds/)(Holt *et al.*, 2002). Out of these 93 large scaffolds, covering 82% of the total genome, were selected. Remaining 8894 scaffolds, covering 18% of the genome were not taken into the analysis because of their small size and large number and also to minimize the miRsearch screening time.

Strategy of miRsearch. The computational screening of miRNA was executed through the program written in Perl scripts (Fig. 1), followed by the miRNA characteristics based screening algorithm (Fig. 2), the entire algorithm being named as miRsearch. Using *Drosophila* miRNA as query sequence, genome of *A. gambiae*, which belongs to the same order diptera, was searched with a user defined score (S). The score S is based on the number of mismatches and defined as:

S=2*[length of miRNA–2*(no. of mismatches)].

The number of mismatches in case of organisms belonging to same genus but different species, (for example, *C. elegans* and *C. briggsae* pair) was chosen to be zero, but we have relaxed it to allow for 3 mismatches for *D. melanogaster* and *A. gambiae* pair as they belong to the same order diptera.

Next, the searching and selection of pre-miRNAs were done using the following algorithm. The query sequence (q[i]) of size n nucleotide was placed along the column and the input sequence of same size of the query (target t[j]) was passed along the row, so as to form a n×n matrix (M). For i=j=0 to n–1, q[i] was compared with t[j] for perfect matching and assigned a score of +2 and otherwise –2, the scores were placed along M[i][j] for i=j=0 to n–1. If the trace of the matrix is greater than the user given score S, the reverse match was searched for from –80 to –1 from the first base of the target sequence, to find whether other arm of the hairpin loop precursor miRNA is in the upstream of the target sequence else from +1 to +80 from the last base of the target sequence to find whether other arm of the hairpin loop pre-miRNA is in the downstream of the target sequence. For reverse matching, the reverse complement of q[i] was searched with a different scoring system. As the pre-miRNAs are known to form a stable hairpin loop structure, so for A-T base pairing a reward of +2 and for G-C pairing a reward +3 was given otherwise the reward was taken as 0. If the reverse matching score is greater or equal to R, then the precursor sequence was reported with the lower and upper co-ordinates. The assigned score R of the reverse matching was determined by training the program to find the all known *D. melanogaster* miRNAs.

Both the forward and reverse complement sequences of the scaffolds were searched for the analysis of miRNAs. The selected sequences representing probable candidates were then examined through NCBI map viewer for their possible location (http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=7165) and those located in the exonic regions were eliminated.

The candidate pre-miRNAs were then filtered through the RNA folding program MFOLD (Zuker, 2003) (http://www.bioinfo.rpi.edu/applications/mfold/old/rna/form3.cgi) selecting the cut-off values on the basis of known *D. melanogaster* pre-miRNAs (Fig. 2). The characteristics observed from the MFOLD output of *D. melanogaster* pre-miRNAs are: (i) the predicted mature miRNAs are within the long helical arm of the secondary structure of pre-miRNA; (ii) $\Delta G \leq -21.0$ kcal/mole; (iii) largest helical arm contained at least 23 bp sequence. Considering these observations, the MFOLD output of each *A. gambiae* pre-miRNA was examined. A structure is accepted as probable miRNA if (a) $\Delta G \leq -21.0$ kcal/mole, (b) the longest helical arm contains at least (20–29) bp depending on miRNA sequence length and (c) the predicted miRNAs are within the long helical arm.

The search program will be available from the authors on request.

RESULTS

Computational prediction of miRNAs by miRsearch. Observations have suggested that mature miRNA sequences are phylogenetically conserved and have characteristic stem-loop secondary structure. Based on this miRsearch used a homologous sequence searching strategy to identify the primary sequence which was simultaneously examined for the capability of formation of stem-loop secondary structure and subsequently MFOLD was used for final prediction of miRNA as described in detail in methods. Using *D. melanogaster* (*Dme*) miRNA as input, it searches homologous sequences with a maximum of 3 mismatches in the scaffolds of *A. gambiae* sequences. Homologous sequences, with 3 mismatches, may be present in many places in the genome, all of which may not have the capability of forming stem-loop precursor structure characteristics of pre-miRNAs. To eliminate those sequences, which do not form typical pre-miRNA structures, reverse complement of the homologue of miRNA sequences (reverse match) were searched at a position –80 to +80 from the matched sequence. To assign a proper score value to the reverse matching sequence, the program was trained with all *Dme* miRNA sequences and we empirically set the minimum score value obtained from *Dme* sequences as the cut off score for *A. gambiae* miRNA (Fig. 1). As miRNA genes can be located on either strand, we searched each sequence in both the forward strand as well as in its reverse complement. Using a total of 79 mature miRNA sequences in the miRbase sequence database and 93 scaffolds of *A. gambiae*, we have detected 489 sequences, which had homologous miRNA sequences and their pre-miRNAs are capable of forming of stem-loop secondary structures. This total set was viewed through Mapviewer to identify their location in the *A. gambiae* genome. Only 13 of these were found to present in the exonic region of genes and were excluded from the set as miRNAs are not supposed to be present in exonic region. Further evaluation of the quality of stem-loop formation of the remaining 476 pre-miRNA sequences was assessed through the RNA folding program MFOLD and

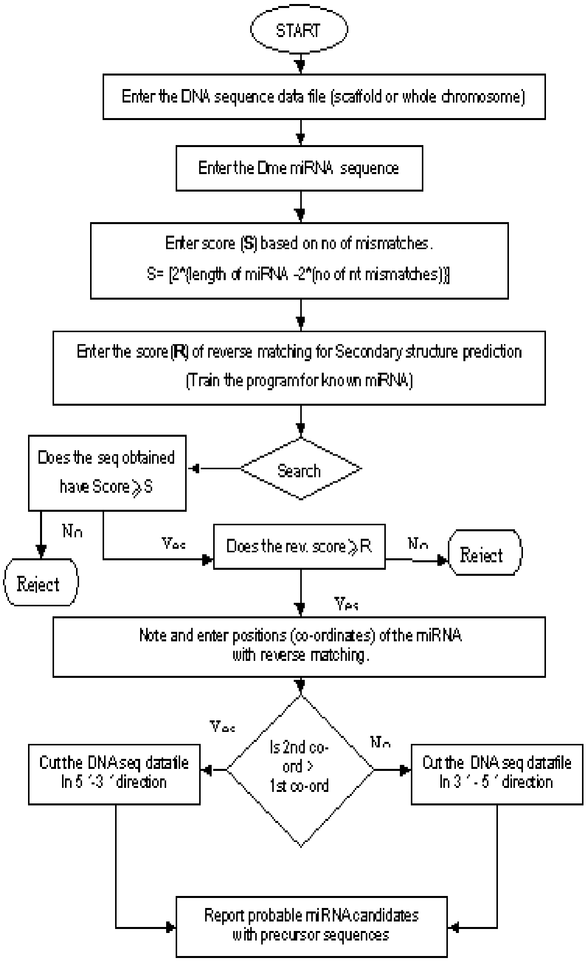

START
Enter the DNA sequence data file (scaffold or whole chromosome)
Enter the Dme miRNA sequence
Enter score (S) based on no of mismatches.
S= [2^(length of miRNA -2^(no of nt mismatches))]
Enter the score (R) of reverse matching for Secondary structure prediction
(Train the program for known miRNA)
Search
Does the seq obtained have Score ≥ S
No
Reject
Yes
Does the rev. score ≥ R
No
Reject
Yes
Note and enter positions (co-ordinates) of the miRNA with reverse matching.
Yes
Is 2nd co-ord > 1st co-ord
No
Cut the DNA seq datafile In 5 ´-3 ´ direction
Cut the DNA seq datafile In 3 ´- 5 ´ direction
Report probable miRNA candidates with precursor sequences

Figure 1. Flowchart for the identification of probable miRNA candidates with precursor sequences.

Table 1. miRsearch predicted *A. gambiae* miRNAs 100% similar to already reported miRNA

Sl No.	Position of miRNAs in *A. gambiae*			100% Homologous miRNA
	Chromosome No.	Scaffold No.	Co–ordinate	
1	2L	AAAB01008960	1144915 (–)	miR–9a
2*	2L	AAAB01008960	11605587 (+)	miR–2a miR–2bmiR–2c
3*	2L	AAAB01008960	11607193 (+)	miR–2a miR–2b miR–2c
4	3R	AAAB01008964	6508743 (+)	miR–9c miR–9a
5	3R	AAAB01008980	2232282 (+)	miR–275
6	3R	AAAB01008944	3005476 (–)	miR–133
7	3R	AAAB01008984	10212876 (+)	miR–124
8	3R	AAAB01008964	6511654 (+)	miR–79
9	3R	AAAB01008980	2241140 (+)	miR–305
10	3R	AAAB01008964	2129745 (+)	miR–125
11*	2L	AAAB01008960	3418059 (–)	miR–281–2 miR–281a miR–281b
12	2L	AAAB01008948	3223987 (–)	miR–7
13	3L	AAAB01008986	3020315 (–)	miR–8
14	3L	AAAB01008986	7425449 (–)	miR–307
15	3R	AAAB01008984	6108942 (+)	miR–14
16	3R	AAAB01008964	1830797 (+)	miR–1
17	2L	AAAB01008960	4977122 (–)	miR–219
18*	2L	AAAB01008960	4977170 (+)	miR–276a miR–276b
19	2L	AAAB01008960	5047343 (+)	miR–276a miR–276b
20*	2L	AAAB01008960	5047365 (+)	miR–276a miR–276b
21*	2L	AAAB01008960	5047385 (+)	miR–276a
22	2R	AAAB01008879	2159581 (+)	miR–315
23	2L	AAAB01008807	11606723 (+)	miR–13b miR–13a
24*	2R	AAAB01008987	3180493 (–)	miR–11
25	2R	AAAB01008850	89732 (–)	miR–10
26	2R	AAAB01008898	1736641 (–)	miR–279
27*	2R	AAAB01008859	3300849 (+)	miR–34
28	2R	AAAB01008850	675477 (+)	miR–iab–4–5p
29*	2R	AAAB01008850	675511 (+)	miR–iab–4–3p
30	2R	AAAB01008851	1388847 (–)	miR–283
31	3R	AAAB01008964	2125233 (+)	miR–100
32*	X	AAAB01008846	10052753 (+)	mir–33
33	2L	AAAB01008960	3418021 (–)	mir–281
34	2L	AAAB01008960	3551786 (+)	mir–282
35	2R	AAAB01007923	7996 (–)	mir–92b
36	2R	AAAB01007923	22576 (–)	mir–92a
37	2R	AAAB01008888	202598 (+)	mir–92a
38	2R	AAAB01008888	222094 (+)	mir–92b
39	3L	AAAB01008986	7425449 (–)	mir–307
40	3L	AAAB01008986	1063490 (+)	mir–308
41	3R	AAAB01008964	2129204 (+)	let–7

*Newly identified *A. gambiae* miRNAs.

some selection procedure set empirically by studying *Dme* miRNAs (Material and Methods, Fig. 2). A total of 91 pre-miRNA sequences have been finally identified as the probable candidate *A. gambiae* miRNA after passing through the entire miRsearch.

Out of 91 *A. gambiae* miRNAs, 41 have 100% similarity to the *Dme* miRNA (Table 1). One miRNA, which was 85.7% similar to *Dme* miRNA (dme-mir-33), was 100% similar to the hsa-*mir-33* miRNA (*Homo sapiens*). Two of the *Dme* miRNAs viz. miR-*9a* and miR-*2a* were conserved both in sequence and in their location in chromosome 2L. Other predicted miRNAs were not conserved in their chromosomal location (not shown). The miRNA gene cluster of miR-*276a* found in chromosome 2L of *D. melanogaster* was detected in chromosome 3L in *A. gambiae*. The locations of the predicted miRNA were also recorded (Table 1, 2). Remaining 50 miRNAs are probable newly identified *A. gambiae* miRNA having potential for the formation of hairpin secondary structure with high degree of MFE (ΔG) and more than 85% sim-

Table 2. Probable newly identified *A. gambiae* miRNAs predicted by miRsearch

Sl No.	Probable miRNA (*Anopheles gambiae*)	Position (chr no., scaffold & co-ordinates)	Closest homologous miRNA
1	TCACTGGGCAAAGTTTGTCGCA	2L AAAB01008968 494152	miR-3
2	ATCACAGCCAGCTTTGAAGAGC	2L AAAB01008960 11606893	miR-2c
3	GATCACATGCAGCTTTGAGGAGA	2L AAAB01008960 7464065	miR-2b
4	TATCACAGCCAGCTTTGAAGAGC	2L AAAB01008960 11606893	miR-2
5	TCAGGCATCTGCAGTAGCGCACG	2L AAAB01008960 3748081	miR-275b
6	CAGCGAGGTATAGAGTTCCTATG	2LAAAB01008960 4977122	miR-276
7	TGTCATGGAATTGCTCTCTTTAT	2L AAAB01008960 3418021	miR-281
8	AATCTAGCCTCTTCTAGGCTTTGTCTGT	2L AAAB01008960 3551786	miR-286
9	CAGCGAGGTATAGAGTTCCTATG	2L AAAB01008960 4977122	miR-276
10	TGTGTTGAAAATCATGTGCAA	2L AAAB01008960 9175668	miR-287
11	TGTGTTGAAAATCATTTGTAA	2L AAAB01008960 3952821	miR-287
12	TGAGACAATTTTGAAAGCTGAGT	2L AAAB01008807 4023317	miR-bantam
13	TATCACAGCCAGCTTTGAAGAGC	2L AAAB01008807 11606893	miR-2
14	CCTTATTATGCTTTCGCCCCG	2R AAAB01008844 938949	miR-184
15	ATTGCACTTGTCCCGGCCTGC	2R AAAB01007923 7996	miR-92b
16	ATATTGCACTTGTCCCGGCCTAT	2R AAAB01007923 22576	miR-92a
17	TATTGCACTTGTCCCGGCCTAT	2R AAAB01008888 202598	miR-92a
18	TGGCAGTCCGGTTTGCTGGTTG	2R AAAB01008987 4283494	miR-34
19	TCGCTCCATTCGCAATCAGTGC	2R AAAB01008859 1718173	miR-285
20	AATTGCACTTGTCCCGGCCTGC	2R AAAB01007923 7996	miR-92b
21	AATTGCACTTGTCCCGGCCTGC	2R AAAB01008888 222094	miR-92b
22	AAACGGACGAAAGTCCCACCGA	3L AAAB01008986 7850785	miR-212
23	TGTGTTGAAAATCATGTGCAC	3L AAAB01008816 932527	miR-287
24	CGTGTTGAAAATCGTGTGCAA	3L AAAB01008823 2635002	miR-287
25	TGTGTTGAAAATCATTTGAAA	3L AAAB01008823 3200243	miR-287
26	TAGCACCATTCGAAATCAGTAC	3L AAAB01008986 2720706	miR-285
27	AGAGATCATTTTGCAAGATGATT	3L AAAB01008816 577447	miR-bantam
28	TTTGTTGAAAATCCTTTGCAA	3L AAAB01008986 7693438	miR-287
29	TGTGTTGAAAATCATGTGGAC	3L AAAB01008986 8813783	miR-287
30	TTATCTCAATTGGTTAGTGTGAG	3L AAAB01008966 2521934	miR-304
31	AATCACAGGAGTATACTGTGAGA	3L AAAB01008986 1063490	miR-308
32	ACAGTTTTTTCCCTCTCCTA	3R AAAB01008980 11682415	miR-14
33	TCAGTCTTTTACTCTCCACTA	3R AAAB01008980 14365242	miR-14
34	AACCCGTAGATCCGAACTTGT	3R AAAB01008964 2125233	miR-100
35	GCTTTGGTAATCTAGCTTTATGA	3R AAAB01008964 6511651	miR-9
36	TCACTGGGCAAAGTTTGTCGCA	3R AAAB01008980 6863595	miR-3
37	CATCACAGCCCAATTTGATGAGC	3R AAAB01008980 5038400	miR-2a
38	CCTTATCATTCTTTCGCCCCG	3R AAAB01008980 6487250	miR-184
39	TGGACGGAGAACTGATAAGGG	3R AAAB01008980 64872876	miR-184
40	TCGGTGGGACTTTGGTGTGTTT	3R AAAB01008984 6174056	miR-278
41	AGTTTTTATGTTATATATGATATGATA	3R AAAB01008839 416964	miR-280
42	TGACTAGACCGAACACTCGCGTC	3R AAAB01008980 6864288	miR-286
43	TGTGTTGAAAATCATGTGCAA	3R AAAB01008944 2423617	miR-287
44	TGTGTTGAAAATCATGTGCAA	3R AAAB01008980 10846168	miR-287
45	TGAGGTAGTTGGTTGTATAGT	3R AAAB01008964 2129204	miR-let7
46	TCTCTCTTTCTCTCTCTCCTA	X AAAB01008847 1241766	miR-14
47	TGTGTTGAAAATCATGTGCAA	X AAAB01008846 4269543	miR-287
48	GTGAGCAAATATTCAGGTGTG	X AAAB01008846 11047618	miR-87
49	TGGCAAGATGTTGGCATAGCTA	X AAAB01008847 3057089	miR-72
50	TGGCAAGATGTTGGCATAGCTAA	X AAAB01008847 357089	miR-72

ilar with the already predicted miRNAs (Table 2). In some cases, same predicted miRNA was obtained using different *Dme* miRNAs as the query sequence (Table 1 and 2).

Interestingly, although the miRNA sequences of *Drosophila* and *Anopheles* were 100% similar (Table 1), the sequences and structures of corresponding pre-miRNAs from *Drosophila* and *Anopheles* were not conserved to that extent (Fig. 3). This holds good for

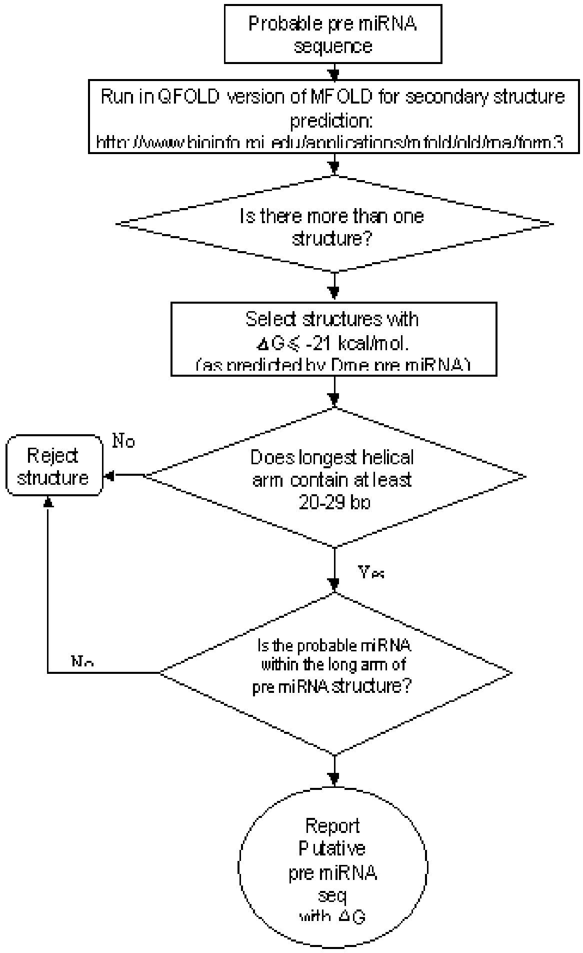

Probable pre miRNA sequence
Run in QFOLD version of MFOLD for secondary structure prediction:
http://www.bioinfo.rpi.edu/applications/mfold/old/rna/form3
Is there more than one structure?
Select structures with ΔG ≤ -21 kcal/mol.
(as predicted by Dme pre miRNA)
Does longest helical arm contain at least 20-29 bp
No
Reject structure
Yes
Is the probable miRNA within the long arm of pre miRNA structure?
No
Report Putative pre miRNA seq with ΔG

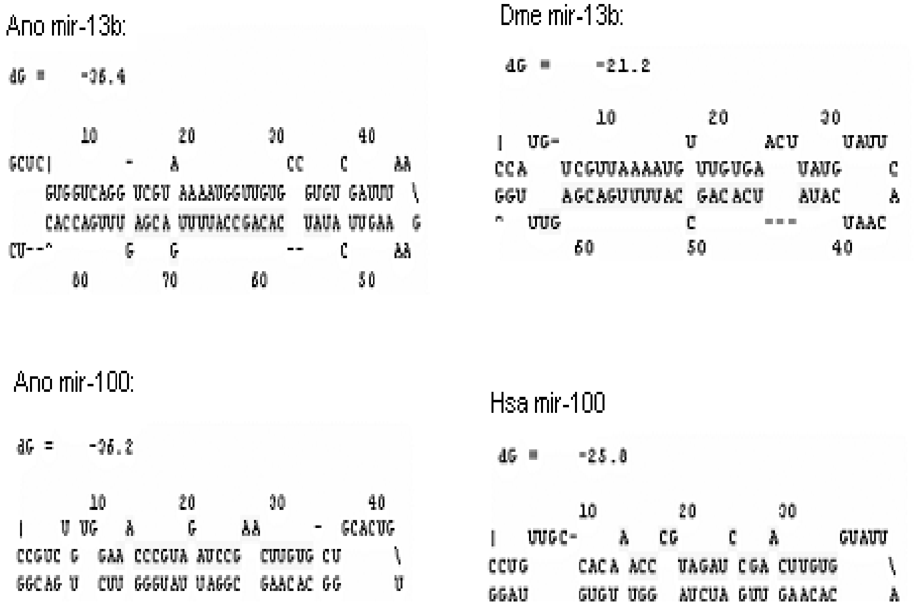
Figure 2. Flowchart for prediction of putative miRNAs based on the miRNA characteristics.

each of the 30 miRNA sequences. This analysis supports our strategy for the miRNA detection, which is based on homology search with the mature miRNA sequences rather than precursor miRNA sequences.

Out of the total 83 *A. gambiae* miRNA reported so far (Griffiths-Jones *et al.*, 2004), we have detected 31 miRNAs. A set of 11 *Anopheles* miRNAs were found to be 100% similar to *Dme* miRNAs, was missed in our study, which is due to exclusion of scaffolds covering 18% of the genome in our study. Moreover, we have identified 10 new miRNAs having 100% similarity with the other species miRNAs. This data suggest that, use of only *Dme* miRNAs as query data set would have detected at least 42 of the predicted miRNAs, new 10 miRNAs and 50 putative novel miRNAs. However, use of all the known miRNA sequences, proper choice of mismatches for each species and considering the whole genome sequences will enhance the efficiency of miRsearch for the identification of new miRNAs.

DISCUSSION

Informatics approach used so far to identify miRNA involves alignment of genomes of two closely related species to find conserved regions followed by identification of stem-loop precursor transcripts capable of processing and forming about 22 nt mature RNA (Lai *et al.*, 2003; Lim *et al.*, 2003a). In our approach we have eliminated the whole genome alignment step and instead have used the following steps: (i) searching for homologues of known mature miRNA in one organism (*Dme*) to the genome of another related organism (*Aga*), after allowing some mismatches, depending on phylogenetic distance between them, (ii) assessing the capability to form stem-loop precursors or structures and finally (iii) observing the preferable position of the mature miRNA in the secondary structure of pre-miRNA. Such an approach is most useful when the complete set of miRNA in one organism is available along


Ano mir-13b:

```
dG =    -36.4

       10        20        30        40
GCUC|        -    A         CC    C     AA
    GUGGUCAGG UCGU AAAAUGGUUGUG  GUGU GAUUU  \
    CACCAGUUU AGCA UUUUACCGACAC  UAUA UUGAA  G
CU--^        G    G         --    C     AA
      80        70        60        50
```

Dme mir-13b:

```
dG =    -21.2

        10         20         30
|  UG-           U       ACU    UAUU
CCA   UCGUUAAAAUG UUGUGA    UAUG    C
GGU   AGCAGUUUUAC GACACU    AUAC    A
^  UUG           C       ---    UAAC
      60         50          40
```

Ano mir-100:

```
dG =    -36.2

        10       20        30        40
|    U UG   A      G     AA      -  GCACUG
CCGUC G  GAA CCCGUA AUCCG  CUUGUG CU      \
GGCAG U  CUU GGGUAU UAGGC  GAACAC GG      U
```

Hsa mir-100

```
dG =    -25.0

         10         20         30
|   UUGC-    A   CG      C   A       GUAUU
CCUG     CACA ACC  UAGAU CGA CUUGUG       \
GGAU     GUGU UGG  AUCUA GUU GAACAC       A
```

Figure 3. MFOLD generated secondary structure of pre-miRNA corresponding to *D. melanogaster* miRNA (*Dme mir-13b*) and *H. sapiens* miRNA (*Hsa mir-100*) and their corresponding 100% similar *A. gambiae* miRNA predicted by miRsearch.

with the genome sequence of a related organism. The chances of getting false positive are little. While comparing with the other methods for the detection of *A. gambiae* miRNAs, miRAlign (Wang *et al.*, 2005) used the complete set of miRNAs available in the database and the whole genome sequence of *A. gambiae*, which is available at NCBI through BLAST search and MapViewer, whereas we have used only the *Dme* miRNAs and 82% of the genome of *A. gambiae*. MiRAlign detected 59 putative miRNAs, out of these 37 (44.6%) are already predicted (Griffiths-Jones, 2004). Our study have detected 91 total miRNAs, out of these 41 are already predicted. For comparison if we study the whole genome of *A. gambiae*, we would have detected at least 52 (63%) already predicted miRNAs. As we have suggested earlier, the search for homologues by BLAST starting with pre-miRNAs as query may miss many of the miRNAs as the pre-miRNA sequences of 100% similar mature miRNAs differ considerably (Fig. 3).

However, in this approach we may miss some of the miRNAs, which are exceptionally divergent and may not be homologous at all to the available miRNAs. The above program may be accommodated to identify miRNAs in not so related organism also (as many of the miRNAs are evolutionarily conserved) by increasing the number of mismatches in miRsearch, although the chances of getting a large number of false positives will be high. To reduce this, further filtering techniques need to be devised, which is currently under investigation.

Acknowledgement

R.C. is indebted to Council of Scientific and Industrial Research (C.S.I.R) for Research fellowship.

REFERENCES

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* **215**: 403–410.

Bartel B, Bartel DP (2003) MicroRNAs: at the root of plant development? *Plant Physiol* **132**: 709–717.

Bartel DP (2004) MicroRNAs: genomics biogenesis mechanism and function. *Cell* **116**: 281–297.

Bonnet E, Wuyts J, Rouze P, Van de Peer Y (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.

Gautheret D, 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**: D109–D111.

Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, Nusskern DR *et al.* (2002) The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* **298**: 129–149.

Jones-Rhoades MW, Bartel DP (2004) Computational identification of plant microRNAs and their targets including a stress-induced miRNA. *Mol Cell* **14**: 787–799.

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

Lai EC, Tomancak P, Williams RW, Rubin GM (2003) Computational identification of *Drosophila* microRNA genes. *Genome Biol* **4**: R42.

Lau NC, Lim LP, Weinstein EG, Bartel DP (2001) An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans. Science* **294**: 858–862.

Lee RC, Ambros V (2001) An extensive class of small RNAs in *Caenorhabditis elegans. Science* **294**: 862–864.

Lee RC, Feinbaum RL, Ambros V (1993) The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell* **75** 843–854.

Legendre M, Lambert A, Gautheret D (2005) Profile-based detection of microRNA precursors in animal genomes. *Bioinformatics* **21**: 841–845.

Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP, Burge CB (2003) Prediction of mammalian microRNA targets. *Cell* **115**: 787–798.

Lim LP, Glasner ME, Yekta S, Burge CB, Bartel DP (2003a) Vertebrate microRNA genes. *Science* **299**: 1540.

Lim LP, Lau NC, Weinstein EG, Abdelhakim A, Yekta S, Rhoades MW *et al.* (2003b) The microRNAs of *Caenorhabditis elegans. Genes Dev* **17**: 991–1008.

Moss EG, Lee RC, Ambros V (1997) The cold shock domain protein LIN-28 controls developmental timing in *C. elegans* and is regulated by the lin-4 RNA. *Cell* **88**: 637–646.

Moss EG, Poethig RS (2002) MicroRNAs: something new under the sun. *Curr Biol* **12**: R688–R690.

Ohler U, Yekta S, Lim LP, Bartel DP, Burge CB (2004) Patterns of flanking sequence conservation and a characteristic upstream motif for microRNA gene identification. *RNA* **10**: 1309–1322.

Pasquinelli AE, Reinhart BJ, Slack F, Martindale MQ, Kuroda MI, Maller B *et al.* (2000) Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature* **408**: 86–89.

Reinhart BJ, Slack FJ, Basson M, Pasquinelli AE, Bettinger JC, Rougvie AE *et al.* (2000) The 21-nucleotide *let-7* RNA regulates developmental timing in *Caenorhabditis elegans*. *Nature* **403**: 901–906.

Slack FJ, Basson M, Liu Z, Ambros V, Horvitz HR, Ruvkun G (2000) The lin-41 RBCC gene acts in the *C. elegans* heterochronic pathway between the let-7 regulatory RNA and the LIN-29 transcription factor. *Mol Cell* **5**: 659–669.

Wang X, Zhang J, Li F, Gu J, He T, Zhang X *et al* (2005) MicroRNA identification based on sequence and structure alignment. *Bioinformatics* **21**: 3610–3614.

Wang XJ, Reyes JL, Chua NH, Gaasterland T (2004) Prediction and identification of *Arabidopsis thaliana* microRNAs and their mRNA targets. *Genome Biol* **5**: R65.

Weber MJ (2005) New human and mouse microRNA genes found by homology search. *FEBS J* **272**: 59–73.

Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **31**: 3406–3415.