Available online at www.sciencedirect.com
ScienceDirect
Plant Physiology and Biochemistry 46 (2008) 739–751
Plant Physiology and Biochemistry
www.elsevier.com/locate/plaphy

Research article

# Identification of micro-RNAs in cotton

Muhammad Younas Khan Barozai [a,b,*], Muhammad Irfan [b], Rizwan Yousaf [b], Imran Ali [b], Uzma Qaisar [b], Asma Maqbool [b], Muzna Zahoor [b], Bushra Rashid [b], Tayyab Hussnain [b], Sheikh Riazuddin [b]

[a] Botany Department, University of Baluchistan, Quetta, Pakistan
[b] Centre of Excellence in Molecular Biology, 87 West Canal Bank Road, Thoker Niaz Baig, Lahore, Pakistan



Received 28 April 2008
Available online 28 May 2008

**Abstract**

The plant genome has conserved small non-coding microRNAs (miRNAs) genes about 20–24 nucleotides long. They play a vital role in the gene regulation at various stages of plant life. Their conserved nature among the various organisms not only suggests their early evolution in eukaryotes but also makes them a good source of new miRNA discovery by homology search using bioinformatics tools. A systematic search approach was used for interspecies orthologues of miRNA precursors, from known sequences of *Gossypium* in GenBank. The study resulted in 22 miRNAs belonging to 13 families. We found 7 miRNA families (miR160, 164, 827, 829, 836, 845 and 865) for the first time in cotton. All 22 miRNA precursors form stable minimum free energy (mfe) stem loop structure as their orthologues form in *Arabidopsis* and the mature miRNAs reside in the stem portion of the stem loop structure. Fifteen miRNAs belong to the world's most commercial fiber producing upland cotton (*Gossypium hirsutum*), five are from *Gossypium raimondii* and one each is from *Gossypium herbaceum* and *Gossypium arboreum*. Their targets consist of transcription factors, cell division regulating proteins and virus response gene. The discovery of 22 miRNAs will be helpful in future for detection of precise function of each miRNA at a particular stage in life cycle of cotton.


Keywords: Cotton; Micro RNAs; Post-transcriptional gene silencing; Homology search

## 1. Introduction

Micro RNA genes that encode miRNAs reside in a specific genomic region. The non-coding RNA such as miRNA, transfer RNA (tRNA) and others constitute 3% out of the total 5% of the functional genome [1]. miRNAs are endogenous, non-coding, small RNAs about 20–24 nucleotides long [2] and are conserved in plants and animals [3,4]. They have a crucial role in post-transcriptional gene regulation [5,6]. Mature miRNAs are produced by a chain of reaction with the help of enzymes [7]. Primary transcripts of mature miRNAs (pri-miRNAs) fold into a stable stem loop structure forming miRNA precursor (pre-miRNA). The loop of pre-miRNA is cleaved, producing a short double-stranded RNA (dsRNA); a single strand of the dsRNA acts as mature miRNA [8]. The processing occurs in nucleus and is processed by a special RNaseIII-like endonuclease, Drosha and Dicer in animals [9] and Dicer-like enzyme (DCL) in plants [10], that also predominantly incorporate the mature miRNA into the RNA induced silencing complex (RISC) [8]. The RISC complex negatively regulates gene expression either by inhibiting translation

*Abbreviations:* ath, *Arabidopsis thaliana*; BLAST, basic local alignment search tool; cr_ratio, core hairpin ratio; DCL, dicer-like enzyme; dsRNA, double-stranded RNA; ESTs, expressed sequence tags; ghr, *Gossypium hirsutum*; mRNA, messenger RNA; MIR, micro-RNA; miRNAs, microRNAs; mfe, minimum free energy; pre-miRNAs, miRNAs precursor; NCBI, National Center for Biotechnology Information; osa, *Oryza sativa*; Ptc, *Populus trichocarpa*; Ptn, position of terminal nucleotide; Pfn, position of the first nucleotide; pri-miRNAs, primary transcripts of mature miRNAs; rRNA, ribosomal RNA; RISC, RNA induced silencing complex; tRNA, transfer RNA; UTRs, untranslated regions.

* Corresponding author. Centre of Excellence in Molecular Biology, 87 West Canal Bank Road, Thoker Niaz Baig, Lahore, Pakistan. Tel.: +92 (0)33 781 7319.
*E-mail address:* barozaikhan@gmail.com (M.Y. Khan Barozai).

0981-9428/$ - see front matter © 2008 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.plaphy.2008.05.009

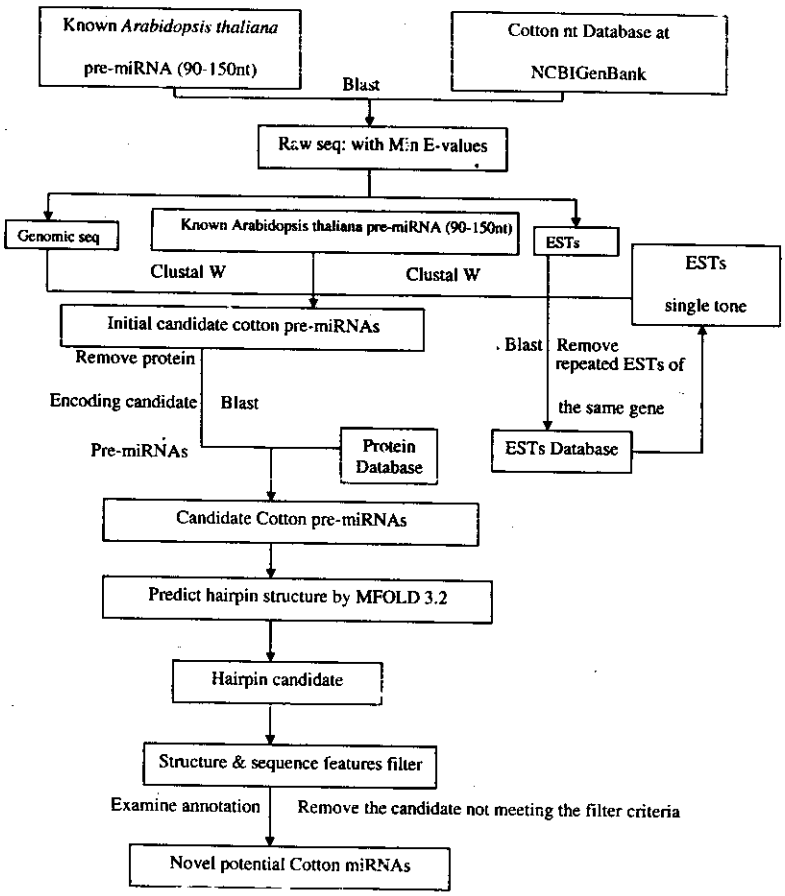


Fig. 1. Schematic representation of the cotton pre-miRNA search procedure used to identify homologues of known *Arabidopsis thaliana* pre-miRNAs.

elongation or by triggering messenger RNA (mRNA) destruction on the basis of the degree of complementarity of miRNA within its target [11,12]. Most animal-mRNA targets have many weak miRNA complementary sites, so miRNA imperfectly complement these sites and suppress gene expression [6,13]. The plant mRNA targets have single and perfect or near perfect miRNA complementary sites, so corresponding miRNAs perfectly complement these sites and trigger the mRNA degradation [14]. Usually the 3′ untranslated regions (UTRs) of the mRNA targets contain the miRNA complementary sites [6].

The first finding of miRNA in *lin-4* and *let-7* mutants of *Caenorhabditis elegans* [15,16] has helped in the discovery of miRNAs in plants [17] and animal-species [18].

The miRNAs perform versatile functions in plant and animals, e.g. in development [19], organ morphogenesis [14,19], signaling pathway [20], transgene repression [21], abiotic stresses [22,23], disease progression [24], and parasitism for the host cell invasion by viruses [25].

Most miRNAs are conserved in animals and plants and from animals to plants [4,17,26], suggesting early development of the miRNAs in the evolution. The conserved nature also indicates their conserved function in different organisms. The conserved nature of these miRNAs becomes a powerful strategy for identification of new orthologues by homology

Table 1
Raw sequences accession numbers, nature, maximum score of homology with known *Arabidopsis thaliana* pre-miRNAs and E-values

| *Arabidopsis thaliana* pre-miRNA | Raw sequences | | | |
|---|---|---|---|---|
| | Accession # | Nature of nucleotide | Max score | *E*-value |
| ath-MIR156e | BV679360 | DNA | 72 | 5e-12 |
| ath-MIR157a | CO076888 | EST | 62 | 4e-09 |
| ath-MIR157a | CO082782 | EST | 62 | 4e-09 |
| ath-miR160b | BQ401391 | EST | 40 | 0.001 |
| ath-miR160b | CO095743.1 | EST | 40 | 0.001 |
| ath-miR164c | DR461140 | EST | 40 | 0.001 |
| ath-MIR166e | DQ908436 | DNA | 38 | 0.006 |
| ath-miR171a | CO129257 | EST | 42 | 4e-04 |
| ath-MIR390a | DW238152 | EST | 64 | 1e-09 |
| ath-MIR390a | DW518163 | EST | 64 | 1e-09 |
| ath-MIR399f | AY632360 | DNA | 44 | 0.001 |
| ath-MIR399f | AY632359 | DNA | 44 | 0.001 |
| ath-miR827 | DW476866 | EST | 40 | 0.001 |
| ath-miR829 | DX525305 | DNA | 40 | 0.002 |
| ath-miR829 | DX531919 | DNA | 40 | 0.002 |
| ath-miR829 | DX53349 | DNA | 40 | 0.002 |
| ath-miR829 | DX541807 | DNA | 40 | 0.002 |
| ath-miR829 | DX545537 | DNA | 40 | 0.002 |
| ath-miR829 | DX546324 | DNA | 40 | 0.002 |
| ath-miR836 | DX561980 | DNA | 38 | 0.009 |
| ath-miR845a | AC197184 | DNA | 38 | 0.006 |
| ath-miR865-5p | DX543587 | DNA | 38 | 0.006 |

search in other species. Weber found 35 new human and 45 new mouse miRNAs by homology search [4]. miRNAs are non-coding mRNAs, as introns or complementary strand of mRNAs, so they may be present in expressed sequence tags (ESTs). Zhang et al. identified 481 miRNAs belonging to 37 miRNA families in 71 different plant species from EST sequences in plants by homology [27]. Qiu et al. [28] and Zhang et al. [29] reported 37 and 30 miRNAs belonging to 21 and 22 miRNAs families in cotton with similar approach.

Here we report 22 putative miRNAs, belonging to 13 miRNA families, by a systematic search of the known cotton ESTs and genomic sequences, using the strategy described by Zhang et al. [27] with slight modifications as illustrated in Fig. 1. The seven miRNAs families are reported for the first time in cotton. In the 22 new miRNAs, 15 miRNAs belong to the world's most commercial fiber producing upland cotton (*Gossypium hirsutum*), five are from *Gossypium raimondii*, one is from *Gossypium herbaceum* and one from *Gossypium arboreum*.

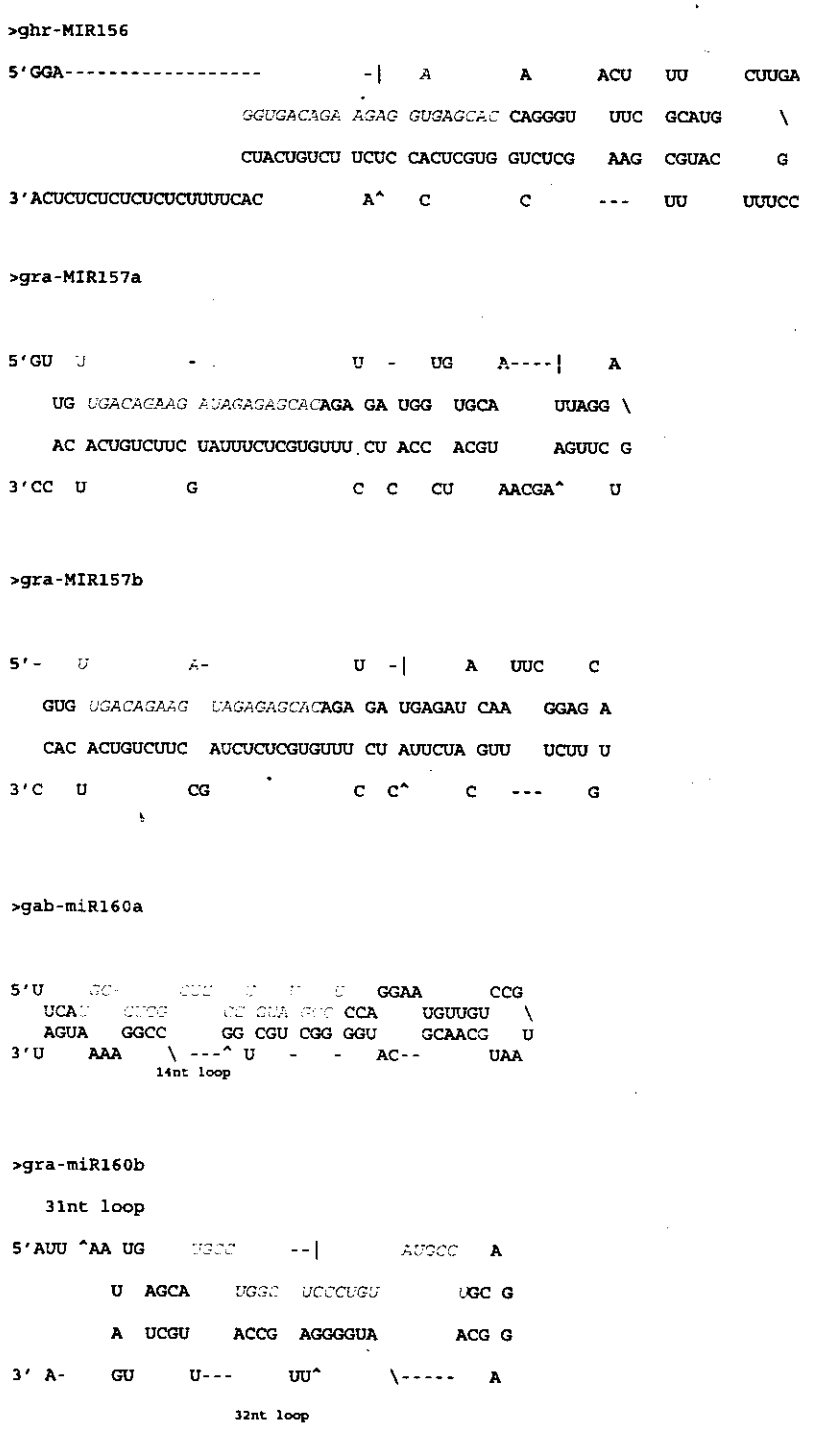


Fig. 2. Cotton pre-miRNAs secondary structures showing the mature miRNAs in stem portion, highlighted in red and italic.

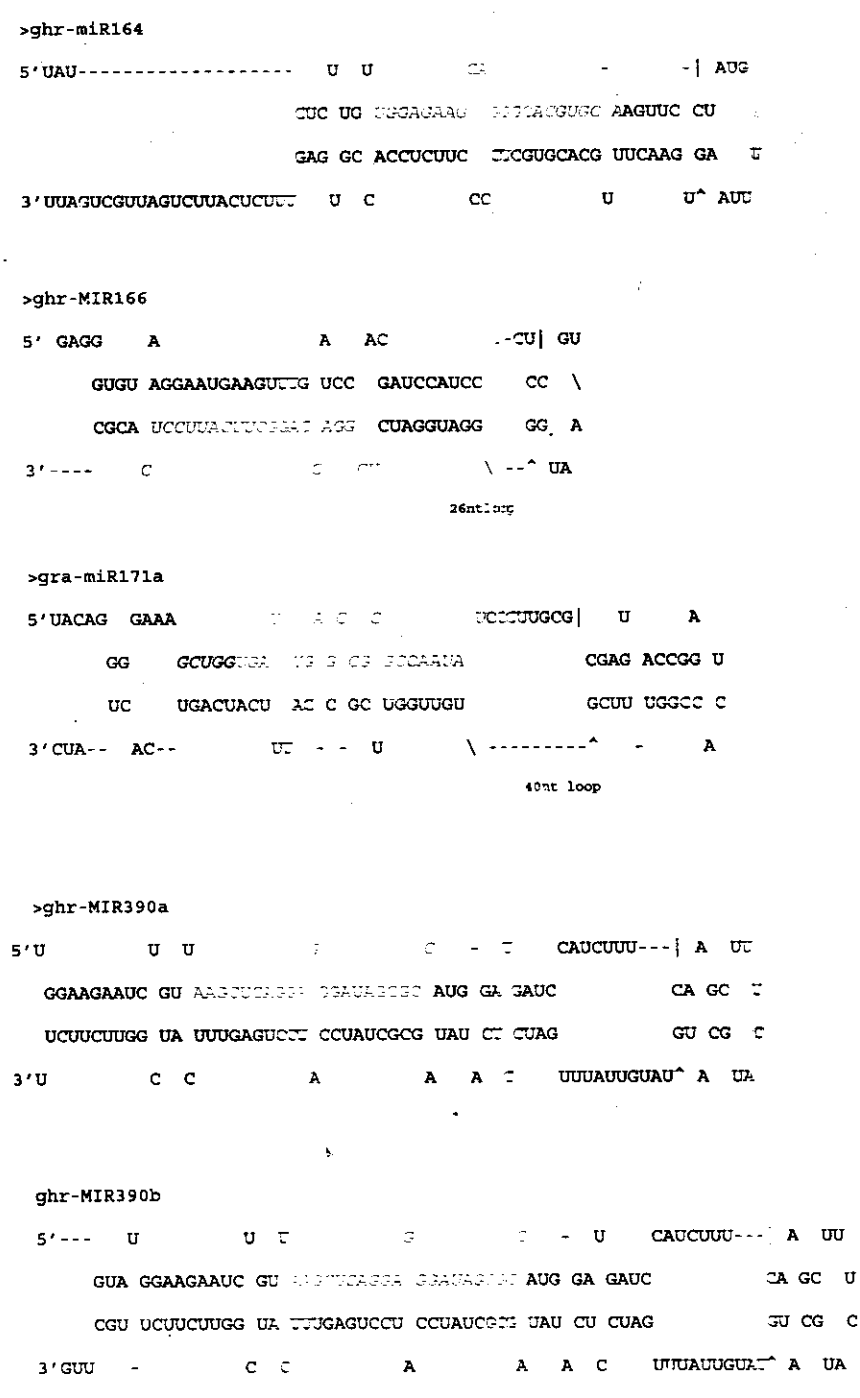


Fig. 2. (*continued*).

## 2. Methods

### 2.1. Identification of raw sequences

The same methodology described by the Zhang et al. [27] with modifications such as additional sequence and structural features filter, as illustrated in Fig. 1, was used. We used *Arabidopsis thaliana* pre-miRNAs instead of mature sequences. The pre-miRNA length (90–140 nt) is 5–7 times more than the mature miRNA length (20–24 nt). Consequently, use of known pre-miRNAs in alignment would bring more precision in orthologue sequence identification. We identified raw sequences, ESTs and genomic sequences containing candidate pre-miRNAs, of cotton by taking previously known *Arabidopsis thaliana* pre-miRNAs from the microRNA Registry Database (version Rfam 9.1, released February 2007 and release 11.0, April 2008) [28] and BLAST against publicly available cotton ESTs and genomic sequences at http://cottondb.org/blast/blast.html using blastn. Adjusted blast parameter settings were as follows: expected values were set at 1000; low complexity was chosen as the sequence filter; and all other parameters were used as default. The FASTA formats of all the

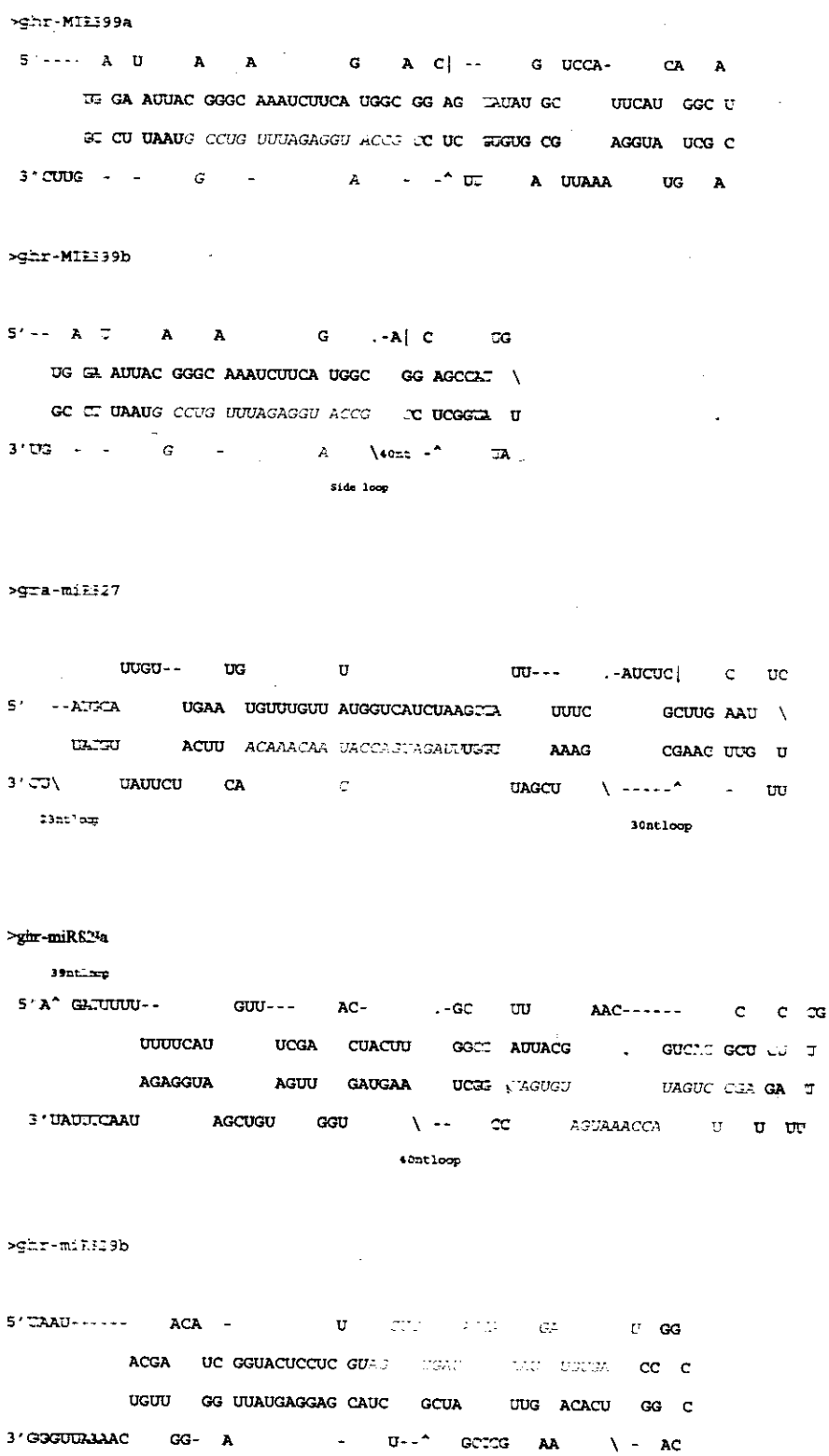

Fig. 2. (*continued*).

sequences having maximum *E*-values were saved. The ESTs created from the same mRNA were found by BLAST against the cotton ESTs Database using blastn with default parameters. The repeated ESTs from the same gene were removed and created a single tone EST.

### 2.2. Prediction of initial candidate's pre-miRNAs in cotton

Each *Arabidopsis thaliana* pre-miRNA was aligned against the corresponding raw sequences, genomic sequences and

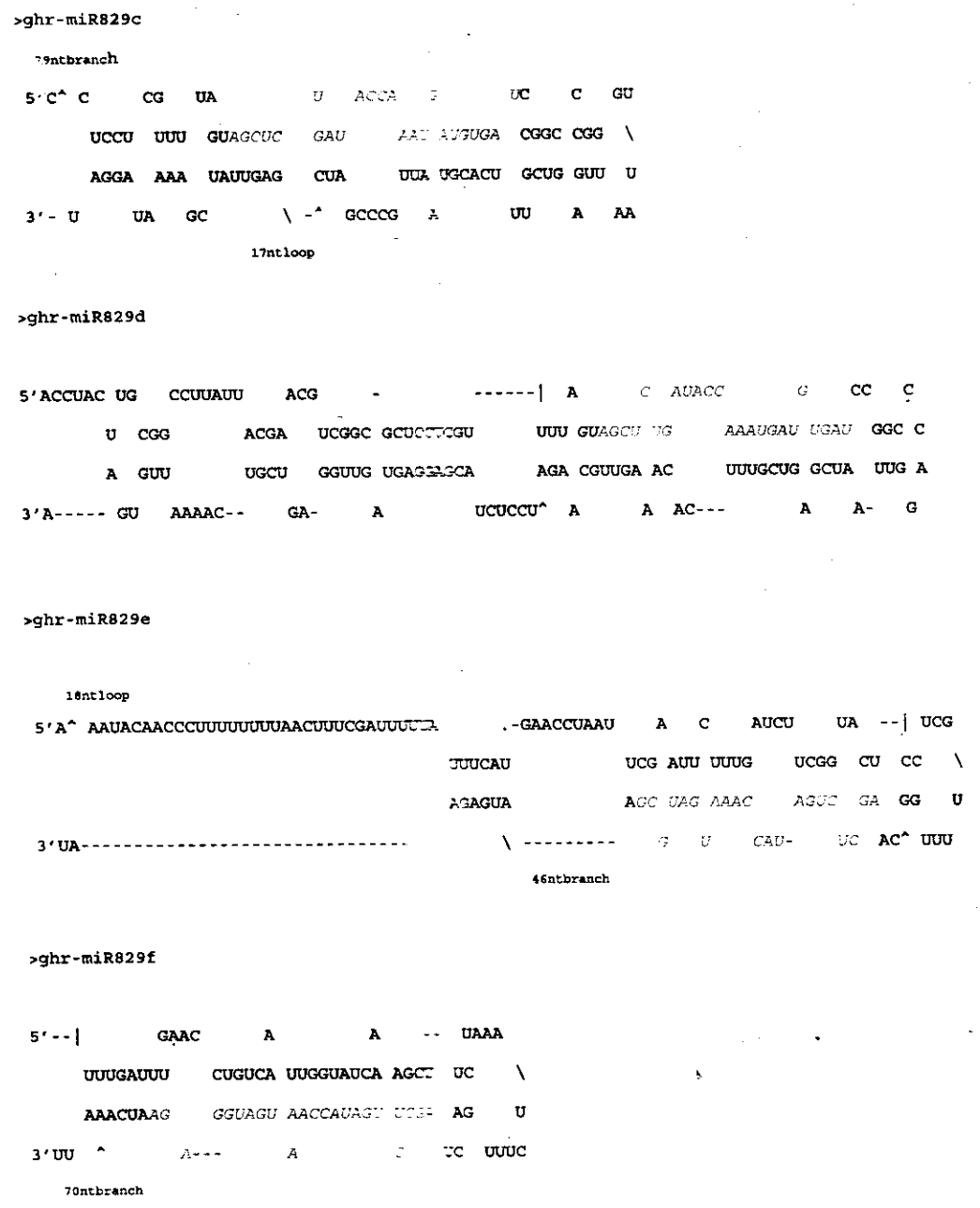


Fig. 2. (*continued*).

single tone ESTs using Clustal W (1.83), a multiple sequence alignment tool with default parameters, publicly available at http://www.ebi.ac.uk/clustalw/. The aligned portion of the raw sequence was considered as an initial candidate miRNA. The initial candidate cotton miRNAs were copied from the raw sequences and saved in FASTA format, and checked for mature sequences with their orthologues of *Arabidopsis thaliana* in the range of 0–4 mismatches.

### 2.3. Validation of cotton miRNAs as a non-protein encoding sequences

The miRNA validation as non-protein encoding sequences is an important criterion for bioinformatics based identification of miRNAs. For this purpose a homology base search of cotton pre-miRNAs with known protein sequences is required. We use the sequences of our identified cotton pre-miRNAs for protein homology search. The predicted pre-miRNA sequences in FASTA format were BLAST against protein database at NCBI using blastx with default parameter [30].

### 2.4. Generation of hairpin structures

The hairpin structure of the initial candidate's sequences were generated using the Zuker folding algorithm with MFOLD (version 3.2) [31], publicly available at http://www.bioinfo.rpi.edu/applications/mfold/rna/form1.cgi. The parameters were

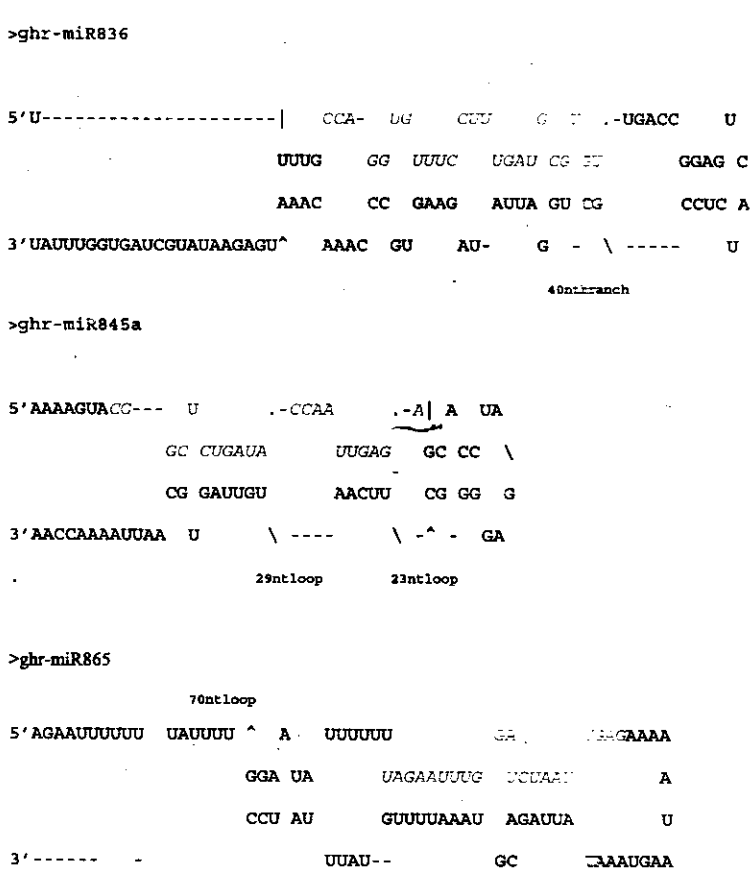


Fig. 2. (*continued*).

adjusted as RNA sequence (linear), folding temperature (37 °C), ionic condition (1 M NaCl with no divalent ions), percent suboptimility number (5); maximum interior/bulges loop size (30), and all others with default values. The lowest free energy structures were selected for manual inspection, as described by Reinhart et al. [17]. The threshold values used to select a miRNA were same as described by Zhang et al. [27]. The stem portion of the hairpin were checked for the mature sequences with at least 16–17 base pairs involved in Watson–Crick or G/U base pairing between the mature miRNA and the opposite strand (miRNA*).

### 2.5. Sequence and structural features filtration

For the sequence and structural features filter, the GC content, core mfe, hairpin mfe and ch_ratio were calculated as described by Li et al. [32] with a little modification for core mfe calculation, as the mature miRNA sequences in plant pre-miRNAs were located away from the Pfn-24 and Ptn+26, so, they were included in the core portion of the hairpin structure. The mfe for core and hairpin structures were calculated by MFOLD (version 3.2) [31], publicly available at http://www.bioinfo.rpi.edu/applications/mfold/rna/form1.cgi. The parameters were adjusted the same as described earlier. For ch_ratio calculation, we divided the core mfe by the hairpin mfe, and the quotient is referred to as the ch_ratio.

### 2.6. Conservation analysis of cotton miRNAs

The conserved nature of miRNAs makes them a bioinformatics resource for orthologue discovery. Consequently we also analyzed the cotton miRNA conservation with their orthologues. The conservation of miRNAs was analyzed by fetching the FASTA format of pre-miRNAs sequences for *Arabidopsis thaliana*, rice (*Oryza sativa*) and *Populus trichocarpa*, from the microRNA Registry Database (versionRfam 9.1, released February 2007) [33] and aligned with cotton miRNAs using Clustal W, as described earlier. The results of Clustal W were saved. For more stringent analysis of cotton miRNAs conservation the raw sequences were BLAST at NCBI using blastn [30] with default parameters as described earlier. The results were saved.

### 2.7. Prediction of cotton miRNA targets

We predict the cotton miRNA targets using the miRU software publicly available at http://bioinfo3.nobie.org/miRNA/miRU.htm [34]. The parameters were adjusted as: Score for each 20 nucleotides (3), G:U Wobble pairs (6), Indel (0) and Other Mismatches (3). *Arabidopsis thaliana* is used as a reference organism for the gene function conservation. The results were saved.

Table 2
Cotton pre-miRNAs length, mfe, mature seq, mature seq: arm, homology, mature seq: length and nucleotides difference with *Arabidopsis* orthologues

| Cotton miRNAs | Length | mfe | Mature seq | Mature seq arm | Homology (%) | Mature seq: length | Nt diff: |
|---|---|---|---|---|---|---|---|
| ghr-miR156a | 117 | −49.9 | UGACAGA AGAGAGUGAGCAC | 5′ | 100 | 20 | Nil |
| gra-miR157a | 101 | −44.8 | UUGACAGAAGAUAGAGAGCAC | 5′ | 100 | 21 | Nil |
| gra-miR157b | 91 | −44.5 | UUGACAGAAGAUAGAGAGCAC | 5′ | 100 | 21 | Nil |
| gab-miR160 | 91 | −23.6 | UGCCUGGCUCCCUGUAUGCCU | 5′ | 95 | 21 | 1 |
| gar-miR160 | 124 | −30.0 | UGCCUGGCUCCCUGUAUGCCU | 5′ | 95 | 21 | 1 |
| ghr-miR164 | 104 | −43.1 | UGGAGAAGCAGGGCACGUGCA | 5′ | 95 | 21 | 1 |
| ghr-miR166 | 115 | −48.2 | UCGGACCAGGCUUCAUUCCUC | 3′ | 95 | 21 | 1 |
| gra-miR171 | 145 | −42.0 | UGAUUGAGCCGCGCCAAUAUC | 5′ | 100 | 21 | Nil |
| ghr-miR390a | 124 | −54.4 | AAGCUCAGGAGGGAUAGCGCC | 5′ | 100 | 21 | Nil |
| ghr-miR390b | 132 | −54.1 | AAGCUCAGGAGGGAUAGCGCC | 5′ | 100 | 21 | Nil |
| ghr-miR399a | 124 | −39.3 | CGCCAAUGGAGAUUUGUCCGG | 3′ | 85 | 21 | 3 |
| ghr-miR399b | 128 | −47.3 | CGCCAAUGGAGAUUUGUCCGG | 3′ | 85 | 21 | 3 |
| gra-miR827 | 195 | −61.2 | UUAGAUGACCAUCAACAAACA | 3′ | 95 | 21 | 1 |
| ghr-miR829a | 218 | −47.4 | AGCUCUGAUACCAAAUGAUGUGAU | 3′ | 91 | 24 | 2 |
| ghr-miR829b | 127 | −32.3 | AGCUCUGAUACCAAAUGAUGUGAU | 5′ | 91 | 24 | 2 |
| ghr-miR829c | 146 | −35.4 | AGCUCUGAUACCAAAUGAUGUGAU | 5′ | 91 | 24 | 2 |
| ghr-miR829d | 149 | −38.3 | AGCUCUGAUACCAAAUGAUGUGAU | 5′ | 91 | 24 | 2 |
| ghr-miR829e | 204 | −29.7 | AGCUCUGAUACCAAAUGAUGCGAU | 3′ | 91 | 24 | 2 |
| ghr-miR829f | 150 | −36.2 | AGCUCUGAUACCAAAUGAUGGAGA | 3′ | 91 | 24 | 2 |
| ghr-miR836 | 135 | −22.5 | CCAGGUGUUUCCUUUGAUGCGUGU | 5′ | 80 | 24 | 4 |
| ghr-miR845a | 120 | −24.6 | CGGCUCUGAUACCAAUUGAGA | 5′ | 91 | 21 | 2 |
| ghr-miR865 | 151 | −24.7 | UAGAAUUUGG AUCUAAUUGAG | 5′ | 91 | 21 | 2 |

### 2.8. Minimizing false positives

Minimizing false positives is a compulsory practice for validation of novel miRNAs; identified on bioinformatics basis. We also have taken various steps to remove false positives, as described by Zhang et al. [27]. In the first step we used known pre-miRNAs for orthologue discovery. The long length of pre-miRNAs made them more suitable for the new miRNAs orthologue identification on conservation basis with the mature sequences in a range of 0–4 mismatches. In the second step we removed the repeated ESTs from the same gene to obtain a single representative from a gene. In the third step we removed the protein coding cotton pre-miRNAs. In the fourth step we used the sequence and structure filter to validate them as candidates of pre-miRNAs in cotton. In the fifth step we used the hairpin structure parameters as described earlier [3,4,27]. All these steps were taken to remove the false positives and confirm the real nature of our identified miRNAs in the cotton life.

## 3. Results and discussion

### 3.1. Cotton miRNAs

*Arabidopsis thaliana* pre-miRNA sequences from the miRNA Registry (versionRfam 9.1, released February 2007 and release 11.0, April 2008) [33], were submitted to a basic local alignment search tool (BLAST) search against the known expressed sequence tags (ESTs) and genomic sequences of the cotton publicly available at http://cottondb.org/blast/blast.html/. The raw sequences (cotton ESTs and genomic sequences) with minimum *E*-values, as shown in Table 1, were aligned with corresponding *Arabidopsis thaliana* pre-miRNAs using the multiple sequence alignment tool Clustal W to obtain the candidate cotton pre-miRNA. The major steps are summarized in Fig. 1, and the process is described in detail in Section 2.

Twenty-two new cotton pre- miRNAs were identified after filtration and completion of the process. The 22 potential cotton miRNAs belong to 13 families (miR156, 157, 160, 164,166, 171, 390, 399, 827, 829, 836, 845 and 865) of miRNAs. We found seven miRNA families (miR160, 164, 827, 829, 836, 845 and 865) for the first time in cotton. The other miRNA families (miR156, 157, 166, 171, 390 and 399) are also reported in cotton by Zhang et al. and Qiu et al. [28,29]. All 22 novel cotton miRNAs were considered as valid candidates after satisfying the empirical formula for biogenesis and expression of the miRNAs, suggested by Ambros et al. [35]. Nine of the 22 cotton pre-miRNAs satisfied the criteria B, C and D; the remaining 13 pre-miRNAs confirmed the criteria C and D. According to Ambros et al. [35] only the criterion D is enough for homologous sequences to validate new miRNAs in different species.

Table 3
Comparison of cotton *Arabidopsis* and Li et al. Reference values of GC-content, core mfe, hairpin mfe and ch_ratio

| Reference | GC content | Core mfe | Hairpin mfe | ch_ratio |
|---|---|---|---|---|
| Li et al. [32] | 30–60 (93%) | −42 to −17 (99%) | −50.2 to −24.2 (99%) | 50–96 (99%) |
| Cotton miRNAs | 38.5–51.2 | −45 to −26.4 | −85.4 to −39.3 | 51–92 |
| *Arabidopsis* homologues | 36.4–51.1 | −54.2 to −23.5 | −79.4 to −48.3 | 42–93 |

The newly identified cotton pre-miRNAs have minimum folding free energies (mfe), with an average of about −40.0 kcal mol$^{-1}$, according to MFOLD [35,36], this is almost equal to the values of *Arabidopsis thaliana* precursor miRNAs and much lower than folding free energies of tRNA (−27.5 kcal mol$^{-1}$) and ribosomal RNA (rRNA) (−33 kcal mol$^{-1}$) [37]. All the mature sequences of cotton miRNAs are in the stem portion of the hairpin structures, as shown in Fig. 2. Our findings are same as those described by Zhang et al. [27]. Table 2 summarizes the minimum free folding energies (mfe), pre-miRNA length, mature miRNAs, and percent identity with *Arabidopsis thaliana* homologues and number of nucleotides difference.

The length of identified cotton pre-miRNAs ranges from 91 to 218 nt and mature sequences range from 20 to 24 nt. The six mature miRNA sequences in newly identified cotton miRNAs are perfectly (100%) matched with the corresponding homologues of *Arabidopsis thaliana*, whereas the remaining 16 mature cotton miRNA sequences differ by 1 to 4 nucleotides from their homologues. The result is same as those of Weber [4] and Zhang et al. [27], where the mature sequences have a difference of 4 nucleotides. Fifteen of the 22 cotton pre-miRNAs have mature sequences at the 5′ arm, while the remaining seven pre-miRNAs have 3′ arm sequences as illustrated in Fig. 2. This finding is similar to that of Reinhart et al. [17]. The predicted miRNA hairpin structures show that there are at least 12–21 nucleotides engaged in Watson–Crick or G/U base pairings between the mature miRNA and the opposite arms (miRNAs*) in the stem region, and the hairpin precursors do not contain large internal loops or bulges. These findings are the same as those of Zhang et al. [27].

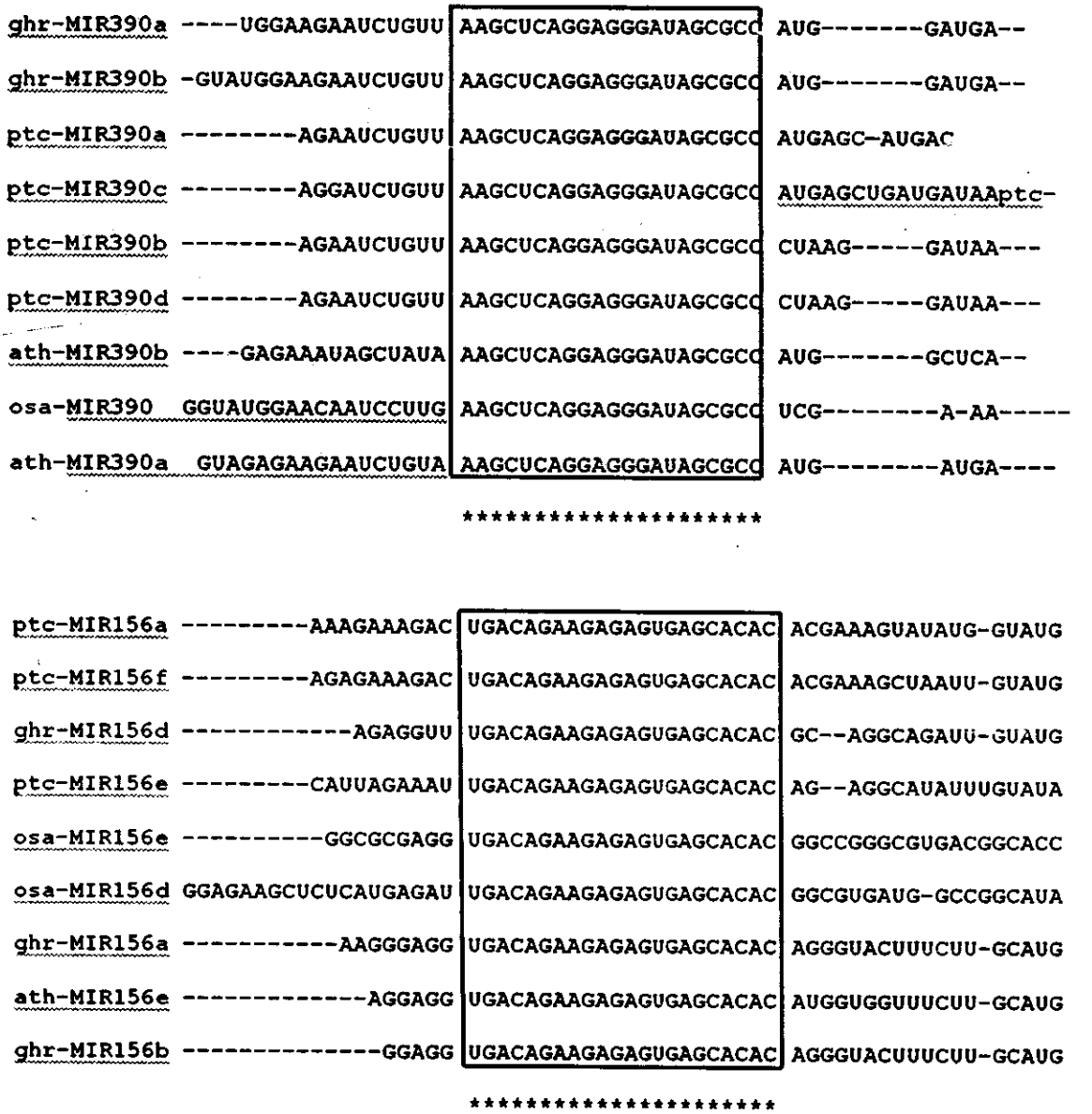


Fig. 3. Conservation of cotton miRNAs. Alignment of pre-miRNAs of *Populus trochorpa*, *Arabidopsis thaliana*, rice (*Oryza sativa*) and cotton (*Gossupium hirsutum*), using Clustal W (a multiple sequence alignment tool) showing conserved nature of mature sequence (underlined).

The relationship between predicted novel miRNAs and known protein is very crucial to validate them as strong candidates of miRNAs. The cotton pre-miRNAs BLAST against the protein database at the NCBI using blastx found no homology with known proteins. This result has confirmed our identified pre-miRNAs as strong candidates in cotton.

### 3.2. Sequence and structural features filter

The sequence and structural features filter was introduced by Li et al. [32]. It is useful to discriminate valid candidates from false positives. The filter is composed of four indices, namely GC content, core minimum free energy (mfe), hairpin mfe and the ratio of core mfe to hairpin mfe (ch_ratio).

As presented in Table 3, identified miRNAs of cotton have a range of GC content (~38.5–51.2), core mfe (~−45 to −26.4 kcal $mol^{-1}$), hairpin mfe (~−85.4 to −39.3 kcal $mol^{-1}$) and ch_ratio (~0.51 to 0.92). The GC content and ch_ratio are within the range given by Li et al. [32], but the lowest mfe ranges for core mfe and hairpin mfe exceed the values reported by Li et al. Basically the sequence and structure features filter developed by Li et al., based on the finding of Zeng et al. [38] that Drosha, a RNA-III like endonuclease, recognizes and cleaves the pri-miRNA at 22 nucleotides upstream from the position of the first nucleotide (Pfn − 22) and the 24 nucleotides downstream the terminal nucleotide (Ptn + 24) to produce pre-miRNA. As the animal pre-miRNA length (60–70 nt) is shorter than plant pre-miRNA (90–140), in plant pre-miRNAs the mature sequence lies far away from the Pfn − 22 and Ptn + 24. In this regard, we made a modification in the Li et al. [32] calculation of core mfe by including the mature sequence in the core portion. The modification resulted in more base pairing, so the values of the lowest mfe exceeded the values of Li et al. We also implemented the filter with modification on the ath-miRNA homologues of the cotton and found a similar result.

### 3.3. Conservation of cotton miRNAs

The newly identified cotton miRNAs are conserved with *Arabidopsis thaliana*, rice (*Oryza sativa*) and *Populus trichocarpa* miRNAs. This finding suggests conservation of miRNAs in dicotyledon and monocotyledon plants. The ghr-miRNAs 156 and ghr-miRNAs 390 are more conserved than others, as illustrated in Fig. 3. The result is similar to that of Zhang et al. [27]. The conservation made them strong candidates of miRNAs in cotton and suggests their conserved physiological function.

The raw sequences of the cotton pre-miRNAs were BLAST against the NCBI database using blastn, which found that the majority of them have 100% homology with *Arabidopsis thaliana*, rice (*Oryza sativa*), *Populus trichocarpa* and maize (*Zea mays*) pre-miRNAs and no homology with protein coding sequences. This finding strongly validated our identified sequences as candidate miRNAs in cotton.

### 3.4. miRNAs and ESTs

In 22 identified miRNAs of cotton, nine are from the ESTs. Zhang et al. have already reported miRNAs from the ESTs of various plant species [27]. A similar finding was given by Li et al. [32] from human ESTs and Qiu et al. from cotton ESTs [29]. The EST base identification of miRNAs is the confirmation of their expression. It also shows a link between the miRNAs and the tissues, organs or developmental stage of the organisms to which the ESTs belong. On basis of EST expression in our identified miRNAs, the miRNAs 156e, 390b and 390c are expressed in the meristematic region of fiber, root and stem. miRNAs 157a and 157b are expressed at the seedling stage of plant and 390a is expressed in the root of the cotton plant.

ATTATTTCATTGTCACCCAAAAAGGCCAAAACCCCACCACTCTCCCCA

TCTCTCTCCCCCCTCCCTTGAACAGAGGCTAAACTCTTGGGGGGGATA

AAGAGATAAGAGAGAATCCAAGCAACGCACATACACAAAACCCTTAGC

TGCCACCAAAGCCTCTTTTGAAGATATA[illegible]TTTTAGGGGAGCATTC

ATTCTGAGGGGAATGTTGTCTGGACCGGGGA[illegible]ACTTGTGTTTCTAATC

CGTTTTCCCTGCAATATTGTTCTATTAACG[illegible]GGGATCTTTATATTGC

TTATTTTTGTAATGGATTAAGGTTGTACTA[illegible]TGTCGTCGGACCAGGCT

TCATTCCCCCCAATCATTGCTCCCATATGTAACTCCCATGGACTGCAG

CATGTAAAAAAGGTTTGGGTTGTTTTTTTGGATCATTGGGGTGTTTTA

TTTACTGAAATTAAGGTTGTTTGTTACTTTGCAG[illegible]TGCAATGAAAAG

Fig. 4. Illustration of intron in the mRNA. The red sequence represents cotton miRNA-1666, green represents consensus sequences at the 5′ splice site and the 3′ splice site in the pre-mRNA (GU-AG rule), and blue represents the branch site.

Table 4
Putative miRNA target genes in cotton

| miRNA Family | Target function in cotton | Cotton TC putative targets | Site | Conserved with *Arabidopsis* |
|---|---|---|---|---|
| 157 | | TC41514 | 1131–1151 | Yes |
| | Squamosa promoter binding protein-like | TC34708 | 1219–1239 | Yes |
| | | TC35884 | 576–596 | Yes |
| | SBP-domain protein | TC36356 | 510–530 | Yes |
| | | TC38424 | 901–921 | Yes |
| | | CO092899 | 778–798 | Yes |
| | | TC34910 | 555–575 | |
| | | TC31499 | 191–211 | |
| | | TC34032 | 1146–1166 | |
| | | TC37796 | 3101–1321 | |
| | Transcription factor RAU1 | TC39015 | 262–282 | |
| | | BF270515 | 86–106 | |
| | | CO120687 | 110–130 | |
| | | TC32276 | 1–20 | |
| 164 | GrpE like protein | TC39893 | 20–40 | |
| | DTDP-glucose 4,6-dehydratase | TC37141 | 706–726 | |
| | | CO109521 | 690–710 | |
| | | CO108738 | 214–233 | |
| 166 | PHAVOLUTA-like HD-ZIPIII protein | TC31648 | 99–119 | Yes |
| | | CO111138 | 453–473 | |
| | | CO079323 | 497–517 | |
| 160 | Auxin response factor | TC37975 | 214–234 | Yes |
| | Auxin response factor | TC29029 | 1023–1043 | Yes |
| | Auxin response factor | TC35005 | 406–426 | Yes |
| | | BF275852 | 41–61 | |
| 845 | | CO083727 | 102–122 | |
| | | CO076175 | 416–436 | |
| | | CO076013 | 354–374 | |
| | | CO089849 | 244–264 | |
| 390 | | CO115013 | 404–424 | |
| 865 | Alpha-L-arabinofuranosidase | TC28161 | 1809–1829 | |
| | | CA993632 | 17–37 | |
| | phospholipase D beta | NP527405 | 2466–2485 | |
| | | TC29506 | 320–340 | |
| | Phospholipase D beta | TC33576 | 2839–2858 | |
| | | TC29893 | 103–123 | |
| | *S*-Adenosylmethionine decarboxylase proenzyme 2 | TC31564 | 978–998 | |
| | | CO121595 | 433–453 | |
| | TMV response-related gene | TC28611 | 1358–1378 | |
| 399 | | TC27809 | 561–581 | |
| | NHL-repeat containing like | TC30044 | 491–511 | |
| | AMP-binding protein | TC30011 | 497–517 | |
| | CER3 protein | TC41229 | 635–655 | |
| | Cell division protein ftsH homologue | TC37093 | 2417–2437 | |
| 529 | | CO076013 | 350–373 | |
| | | CO083727 | 98–121 | |
| | | CO075095 | 118–141 | |
| | | AW187526 | 244–267 | |
| | | CO089849 | 240–263 | |
| | Beta tubulin | TC37321 | 1265–1288 | |
| 836 | | CO122741 | 689–712 | |
| | | CO128688 | 690–713 | Yes |
| | Minor allergen | TC40204 | 830–853 | |
| 827 | | CO088990 | 599–619 | |
| | U1snRNP-specific protein | TC35354 | 912–932 | |
| | Myosin-like protein | TC39799 | 992–1012 | |
| | | AI731815 | 490–510 | |
| | P23 co-chaperone | TC28167 | 917–937 | |

The ESTs usually represent a portion of the messenger RNA (mRNA). Here an important question arises as to how we can predict a non-coding miRNA from a coding mRNA. We have three hypotheses with which to answer the question.

The first hypothesis is that the miRNA is in the intronic region of the mRNA, as already described by Li et al. [32] and Weber [4]. In the mRNA processing the poly A tail is attached earlier than the intron removal [39], so there is a chance that the oligo dT primed mRNA may have intronic regions. ghr-miRNA 166 reported by Qiu et al. [29] is an example of the intronic region miRNA as illustrated in Fig. 4, according to the GU–AG rule [40].

The second hypothesis suggests the presence of the miRNA on the opposite strand of the mRNA coding (sense) strand of DNA. Our ghr-miRNA156e and gab-miRNA160 are the reverse complementary sequences of the known reported mRNAs sequences.

The third hypothesis suggests that pre-miRNAs like mRNAs have a poly A tail, so they pool up with oligo dT-primed RNAs.

### 3.5. Prediction of cotton miRNA targets

The link between miRNAs and their targets is an important step for validation of miRNAs identified on basis of bioinformatics. The conserved nature of miRNAs in different organisms suggests their conserved function [27]. We used the miRU software [30], an automated on-line search of miRNA targets in plants, to identify the cotton miRNA targets using *Arabidopsis thaliana* as the reference organism for the gene function conservation. The conserved and non-conserved targets were obtained, as shown in Table 4. The conserved targets between cotton and *Arabidopsis thaliana* confirm conserved function for orthologues of miRNAs. We found that the miRNA 157 and 160 families are more conserved in cotton and *Arabidopsis thaliana*. The non-conserved putative targets may be involved in processes that are species specific or tissue specific [22]. Most of the putative targets in cotton function as transcription factors.

## 4. Conclusions

We have identified 22 novel candidate miRNAs belonging to 13 families in cotton from ESTs and genomic sequences based on bioinformatics. Seven families are reported for the first time. This is the first step toward a description of the miRNAs in cotton and will enlighten the future pathway leading to understand the function and processing of miRNAs in cotton. It is a bioinformatics approach for new pre-miRNAs identification from plant species whose genome is not yet sequenced. The use of the pre-miRNAs in homology search, the Ambros et al. empirical formula long with the modified sequence and structural features filter is a suitable combination for plant miRNA discovery in the future. The EST-based identification is a confirmation of the miRNA expression. Our results indicate the evolutionary conserved nature of miRNAs in plants.

## Acknowledgments

We wish to acknowledge Sam Griffiths-Jones (miRNA Registry, The Wellcome Trust Sanger Institute, UK) for suggestions and correction. This work is supported by a PhD scholarship awarded by the Higher Education Commission, Pakistan.

## References

[1] Human Genome Sequencing Consortium, Initial sequencing and analysis of the human genome, Nature 420 (2002) 520–562.
[2] E. Mica, L. Gianfranceschi, M.E. Pe, Characterization of five microRNA families in maize, J. Expl. Bot 57 (11) (2006) 2601–2612.
[3] E. Bonnet, J. Wuyts, P. Rouzé, Y. Van de Peer, Detection of 91 potential conserved plant microRNAs in Arabidopsis thaliana and *Oryza sativa* identifies important target genes, Proc. Natl. Acad. Sci. USA 101 (2004) 11511–11516.
[4] M.J. Weber, New human and mouse microRNA genes found by homology search, FEBS Lett. 272 (2005) 59–73.
[5] D.P. Bartel, MicroRNAs: genomics, biogenesis, mechanism, and function, Cell 116 (2004) 281–297.
[6] J.C. Carrington, V. Ambros, Role of microRNAs in plant and animal development, Science 301 (2003) 336–338.
[7] Y. Lee, C. Ahn, J. Han, et al., The nuclear RNase III Drosha initiates microRNA processing, Nature 425 (2003) 415–419.
[8] S.C. Hammond, E. Bernstein, D. Beach, G.J. Hannon, An RNA-directed nuclease mediates posttranscriptional gene silencing in Drosophila cells, Nature 404 (2000) 293–296.
[9] E. Bernstein, A.A. Caudy, S.M. Hammond, G.J. Hannon, Role for a bidentate ribonuclease in the initiation step of RNA interference. Nature 409 (2001) 363–366.
[10] Y. Kurihara, Y. Watanabe, Arabidopsis micro-RNA biogenesis through Dicer-like 1 protein functions, Proc. Natl. Acad. Sci. USA 101 (2004) 12753–12758.
[11] M.J. Aukerman, H. Sakai, Regulation of flowering time and floral organ identity by a microRNA and its APETALA2-Like target genes, Plant Cell 15 (2003) 2730–2741.
[12] G. Tang, B.J. Reinhart, D.P. Bartel, P.D. Zamore, A biochemical framework for RNA silencing in plants, Genes Dev. 17 (2003) 49–63.
[13] C.D. Novina, P.A. Sharp, The RNAi revolution, Nature 430 (2004) 161–164.
[14] C.A. Kidner, R.A. Martienssen, The developmental role of microRNA in plants, Curr. Opin. Plant Biol. 8 (2005) 38–44.
[15] R.C. Lee, R.L. Feinbaum, V. Ambros, C. The, elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14, Cell 75 (1993) 843–854.
[16] B.J. Reinhart, F.J. Slack, M. Basson, A.E. Pasquinelli, J.C. Bettinger, A.E. Rougvie, H.R. Horvitz, G. Ruvkun, The 21- nucleotide let-7 RNA regulates developmental timing in *Caenorhabditis elegans*. Nature 403 (2000) 901–906.
[17] B.J. Reinhart, E.G. Weinstein, M.W. Rhoades, B. Bartel, D.P. Bartel, MicroRNAs in plants, Genes Dev. 16 (2002) 1616–1626.
[18] M.Q. Martindale, J. Baguna, Expression of the 22 nucleotide let-7 heterochronic RNA throughout the Metazoa: A role in life history evolution? Evol. Dev. 5 (2003) 372–378.
[19] X. Chen, A microRNA as a translational repressor of APETALA2 in Arabidopsis flower development, Science 303 (2003) 2022–2025.
[20] M. Yoshikawa, A. Peragine, M.Y. Park, R.S. Poethig, A pathway for the biogenesis of trans-acting siRNAs in Arabidopsis, Genes Dev. 19 (2005) 2164–2175.
[21] E. Allen, et al., microRNA-directed phasing during transacting siRNA biogenesis in plants, Cell 121 (2005) 207–221.
[22] S. Lu, Y.H. Sun, R. Shi, C. Clark, L. Li, V.L. Chiang, Novel and mechanical stress responsive microRNAs in Populus trichocarpa that are absent from Arabidopsis, Plant Cell 17 (2005) 2186–2203.
[23] R. Sunkar, J.K. Zhu, Novel and stress-regulated microRNAs and other small RNAs from Arabidopsis, Plant Cell 16 (2004) 2001–2019.

[24] S.M. Johnson, H. Grosshans, J. Shingara, M. Byrom, R. Jarvis, A. Cheng, E. Labourier, K.L. Reinert, D. Brown, F.J. Slack, RAS is regulated by the let-7 microRNA family, Cell 120 (5) (2005) 635–647.
[25] Y. Bennasser, S.Y. Le, M.L. Yeung, K.T. Jeang, HIV-1 encoded candidate micro-RNAs and their cellular targets, Retroviro l (1) (2004) 43.
[26] M. Arteaga-Vázquez, J. Caballero-Pérez, J.-P. Vielle-Calzada, A family of MmcroRNAs present in plants and animals, Plant Cell 18 (2006) 3355–3369.
[27] B. Zhang, X. Pan, C.H. Cannon, G.P. Cobb, T.A. Anderson, Conservation and divergence of plant microRNA genes, Plant J 46 (2006) 243–259.
[28] B. Zhang, Q. Wang, K. Wang, X. Pan, F. Liu, T. Guo, G.P. Cobb, T.A. Anderson, Identification of cotton microRNAs and their targets, Gene 397 (1-2) (2007) 26–37.
[29] C.X. Qiu, F.L. Xie, Y.Y. Zhu, K. Guo, S.Q. Huang, L. Nie, Z.M. Yang, Computational identification of microRNAs and their targets in *Gossypium hirsutum* expressed sequence tags, Gene 395 (1-2) (2007) 49–61.
[30] F.A. Stephen, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST, A new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402.
[31] M. Zuker, Mfold web server for nucleic acid folding and hybridization prediction, Nucleic Acids Res. 31 (2003) 3406–3415.
[32] S.C. Li, C.U. Pan, W.C. Lin, Bioinformatic discovery of microRNA precursors from human ESTs and introns, BMC Genet 7 (2006) 164.
[33] S. Griffiths-Jones, The microRNA Registry, Nucleic Acids Res. 32D (2004) 109–111.
[34] Y. Zhang, miRU: an automated plant miRNA target prediction server, Nucleic Acids Res. 33 (web server issue) (2005) W701–W704.
[35] V. Ambros, B. Bartel, D.P. Bartel, et al., A uniform system for microRNA annotation, RNA 9 (2003) 277–279.
[36] D.H. Mathews, J. Sabina, M. Zuker, D.H. Turner, Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure, J. Mol. Biol. 288 (1999) 911–940.
[37] E. Bonnet, J. Wuyts, P. Rouze, Y. Van de Peer, Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences, Bioinformatics 20 (2004) 2911–2917.
[38] Y. Zeng, P. Yi, B.R. Cullen, Recognition and cleavage of primary microRNA precursors by the nuclear processing enzyme Drosha, EMBO J 24 (1) (2005) 138–148.
[39] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, P. Walter, How cells read the genome: from DNA to protein, in: Molecular Biology of The Cell, 4th Edition (2002), pp. 231–280.
[40] M.G. Mayer, L.M. Floeter-Winter, Pre-mRNA trans-splicing: from kinetoplastids to mammals, an easy language for life diversity, Mem. Inst. Oswaldo Cruz 100 (2005) 501–513.