Available online at www.sciencedirect.com
ScienceDirect
Computational Biology and Chemistry
Computational Biology and Chemistry 31 (2007) 207–214
www.elsevier.com/locate/compbiolchem

Identification of specific sequence motifs in the upstream region of 242 human miRNA genes

Atsushi Inouchi [a], Shuichi Shinohara [b], Hiroshi Inoue [c], Kenji Kita [d], Mitsuo Itakura [c,*]

[a] *Graduate School of Engineering, The University of Tokushima, 2-1 Minamijosanjima, Tokushima City, Tokushima 770-8506, Japan*
[b] *Division of R&D Solution, Fujitsu Nagano Systems Engineering Ltd., Tsuruga Nabeyata 1403-3, Nagano City, Nagano 380-0813, Japan*
[c] *Division of Genetic Information, Institute for Genome Research, The University of Tokushima, 3-18-15, Kuramoto-cho, Tokushima City, Tokushima 770-8503, Japan*
[d] *Center for Advanced Information Technology, The University of Tokushima, 2-1 Minamijosanjima, Tokushima City, Tokushima 770-8506, Japan*

Received 22 October 2006; received in revised form 20 March 2007; accepted 27 March 2007

Abstract

We have identified novel over-represented and conserved motifs in the upstream regions of human and mouse miRNA stem-loop sequences by means of a new bioinformatic processing regimen. We observed sequence conservation −500 bp upstream in 189 human and mouse miRNAs declining with increasing distance from their putative miRNA stem-loop origin. We also found relatively GC-rich regions having more than 50% of guanine + cytosine (G + C) content at about −30 and −170 bp relative to human miRNA stem-loop sequence origin. To further identify specific sequence motifs that might be involved in the transcriptional regulation of miRNA precursors, we first searched 500 bp upstream sequences of 194 non-redundant human miRNA stem-loop sequences for frequently occurring motifs 5–15 bp long. We then found the comparable occurrences of the 20 most frequent motifs in the 2000 bp upstream regions of 242 human and 290 mouse miRNAs. The significantly reduced frequency of occurrence of all 20 motifs in the regions 2000 bp upstream of 23,570 human RefSeq genes demonstrated that these motifs were specific to the upstream miRNA sequences. The most frequently observed motif M1 (GTGCTTMTAGTGCAG), with a MEME *E*-value of 3.8e−57 was distributed within 500 bp upstream of stem-loop sequences and was also miRNA-specific. We suggest that these over-represented motif sites are good candidates for experimentally testing miRNA expression as well as possible interaction with regulatory factors.

Keywords: MicroRNA (miRNA); Motif search; Regulatory motif; Transcription factor binding sites; Promoter

1. Introduction

MicroRNAs (MiRNAs) are ∼22 nt long non-coding RNAs generated from a local hairpin structure of endogenous transcript. These small RNAs induce mRNA degradation or translational repression by binding to their target mRNAs (Bartel, 2004; Filipowicz et al., 2005; Sontheimer and Carthew, 2005). MiRNAs are now recognized as one of the major regulatory gene families, with a recent study suggesting that miRNAs regulate at least 20% of all human genes (X. Xie et al., 2005).

The details of miRNA genomics (Kim and Nam, 2006; Kim, 2005) disclosed that transcription of miRNA genes produces a long primary transcript (called pri-miRNA). A stem-loop precursor of ∼80 nt (called pre-miRNA, hairpin miRNA, or stem-loop precursor) is then cleaved from this pri-miRNA by the RNase III-type Drosha. Finally, the mature miRNA of ∼22 nt is produced from this stem-loop sequence by the RNase III-type enzyme Dicer.

MiRNAs are transcribed by RNA polymerase II (Pol II) (Cai et al., 2004; Lee et al., 2004). Pri-miRNAs contain cap structures and are polyadenylated, both unique properties of Pol II gene transcripts. Furthermore, human cells treated with α-amanitin, a Pol II inhibitor, decrease the level of miRNA transcription, and chromatin immunoprecipitation analysis reveals the physical association of Pol II with a miRNA promoter (Lee et al., 2004). Taken together, these reports indicate that the Pol II is the main RNA polymerase for miRNA gene transcription. Recent analysis of the miRNA promoter has identified canonical TATA-

* Corresponding author. Tel.: +81 88 633 9454; fax: +81 88 633 9455.
E-mail address: itakura@genome.tokushima-u.ac.jp (M. Itakura).

1476-9271/$ – see front matter © 2007 Published by Elsevier Ltd.
doi:10.1016/j.compbiolchem.2007.03.011

box motifs upstream of the transcription start site of miRNA (Z. Xie et al., 2005; Houbaviy et al., 2005).

To fully understand the mechanism for miRNA transcription it is necessary to analyze the promoters of various miRNA genes. Thus, identification of miRNA-specific elements could provide clues for understanding regulatory networks and miRNA gene prediction. Importantly then, bioinformatic searches for characteristic sequence motifs may well lead to identification of miRNA-specific promoter elements.

In a previous analysis, alignments of miRNA upstream sequences in orthologous *C. elegans*/*C. briggsae* regions showed a pronounced conserved sequence motif approximately 200 bp upstream of the miRNA stem-loop point (Ohler et al., 2004). Using this motif improved performance of the associated miRNA search algorithm. However, bioinformatic searches for miRNA-specific promoter elements in longer regions upstream of miRNA sequences have so far not been successful.

Our purpose here is to remedy this gap and to identify specific sequence motifs upstream of the miRNA stem-loop starting point that might be involved in miRNA expression. To discover over-represented sequence motifs, using computational motif-locating tools, we analyzed the set of human miRNA upstream sequences likely to contain transcriptional regulatory sequences.

2. Material and methods

2.1. Human and mouse data sets of upstream sequences

To collect the data on human and mouse miRNA, we used the miRNA database, miRBase Sequence Database Release 7.1 (http://microrna.sanger.ac.uk/; Griffiths-Jones et al., 2006). This database contains 326 human miRNA sequences of which 234 mature miRNA sequences have been experimentally verified. We also constructed sets of 242 human and 290 mouse miRNA upstream sequences likely to contain transcriptional regulatory sequences (Supplementary Table 1). This sequence set encompasses the 2000 bp region upstream of the miRNA annotated stem-loop sequences, including experimentally verified mature miRNA in the database.

To assess over-representation of the identified motifs, we also constructed another independent set of 136 human miRNA upstream sequences (Supplementary Table 1) that were collected from miRBase Sequence Database Release 8.2.

To validate the miRNA-specificity of identified motifs, we used the upstream sequences of the set of 23,570 human protein coding genes (downloadable from http://hgdownload.cse.ucsc.edu/goldenPath/hg18/bigZips/). This data set contained 2000 bp upstream sequences, measured from the annotated transcription start site of RefSeq genes. This data set included only the genes in which the transcription start is annotated separately from the translation start.

2.2. Analysis of sequence characteristics

To facilitate the examination of evolutionary conservation in the region upstream of human miRNA genes, we constructed an upstream data set of 189 miRNAs by extracting the relevant portions of human/mouse pairwise alignments. Out of 242 human miRNA upstream sequences, 189 were pairwise aligned with orthologous mouse sequences. From this we obtained 2000 bp upstream sequences of 189 human/mouse miRNAs. The assemblies used for this data set were the hg17/mm7 versions downloaded from the UCSC genome browser (http://genome.ucsc.edu/). To carry out the computation, we fixed a 10-base sliding window at the beginning of each putative alignment and then shifted the window by 1 bp from left to right to compute the conservation at each position.

In order to analyze the characteristics of human miRNA upstream sequences, we examined the G + C content of each window with the sliding window method as described above. We also screened human miRNA upstream sequences using a CpG island searcher (Takai and Jones, 2002; Takai and Jones, 2003). We set the threshold criteria for CpG islands as defined by Gardiner-Garden and Frommer (1987) (sequence length $\geq$ 200 bp, observed CpG/expected CpG $\geq$ 0.6, and G + C content $\geq$ 0.5).

2.3. Motif discovery

The overall process flowchart of our motif discovery method is displayed in Fig. 1. First, 194 human miRNA upstream sequences (500 bp) were submitted to the MEME program Version 3.5.3 (http://meme.sdsc.edu/; Bailey and Elkan, 1994). We used only 194 miRNA sequences. This is because of the set of 242 human miRNA upstream sequences including experimentally verified mature miRNA, 48 are members of the miRNA cluster on chromosome 19 located at positions 58,861,745–58,961,404 as well as the miRNA cluster on chromosome X located at positions 145,967,859–146,072,859 (HG17) (Bentwich et al., 2005). These 48 miRNAs have similar 500 bp upstream sequences, suggestive of gene duplication. Because including these duplicated sequences would strongly bias motif discovery, we excluded them from our data set processed by MEME (Supplementary Table 1). To detect the top 20 most frequent motifs, we used the MEME command: "meme <dataset> -mod anr -dna -minsites 4 -nmotifs 20 -minw 5 -maxw 15 -revcomp". We then evaluated the width of the resulting motifs of length 5–15 bp on both strands. The significance of a detected motif was reported as the *E*-value of the motif (the expected number of motifs in a random database of the same size), which is equal to the combined *p*-value of the sequence times the number of sequences in the database. All other MEME options were left at their default values. For these motifs found by MEME, we created sequence logos using WebLogo Version 2.8.2 (http://weblogo.berkeley.edu/; Crooks et al., 2004).

Next, the MEME motifs were individually aligned using the MAST program with default values (Bailey and Gribskov, 1998) to the data set of 242 human miRNA 2000 bp upstream sequences, the data set of 290 mouse miRNA 2000 bp upstream sequences, and the data set of 23,570 human 2000 bp upstream sequences of the annotated transcription start site of RefSeq genes. With these default parameters, MAST outputs motif results for sequences with *E*-values less than 10 and motif matches with *p*-values less than 0.0001.

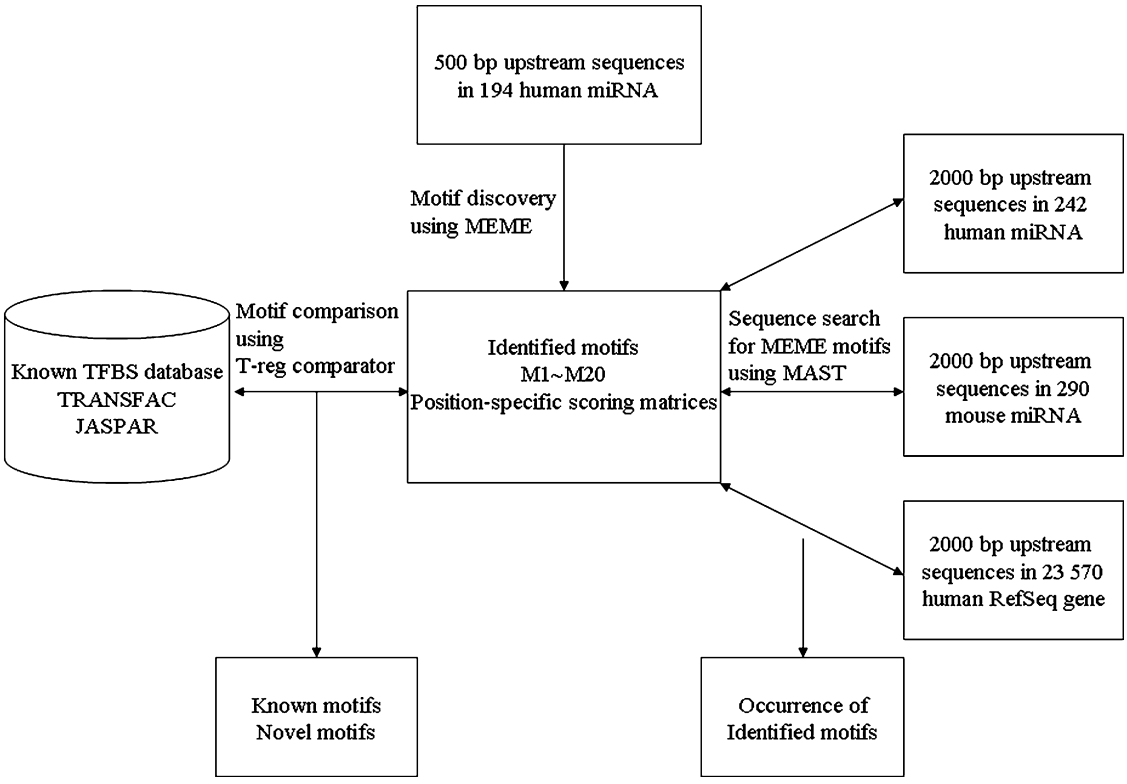

Fig. 1. Identification of specific sequence motifs in the upstream region of human miRNA genes. The discovery of the most frequent 20 motifs (M1–M20) (center) was followed by motif comparison (left) and the calculation of their frequency of occurrence in upstream sequences (right).

After this alignment, to determine the miRNA specificity of the MEME motifs, we took the proportion of sequences including each MEME motif and computed a statistic comparing this proportion between human miRNA upstream sequences and human RefSeq gene upstream sequences. For our statistic, we used a *Z*-score equation for a test of proportions as follows:

$$Z = \frac{|p_1 - p_2|}{\sqrt{p(1-p)(1/n_1 + 1/n_2)}}, \quad p = \frac{r_1 + r_2}{n_1 + n_2}$$

Here, n_1 is the number of human miRNA upstream sequences in the data set (242 sequences), n_2 the number of human RefSeq gene upstream sequences in the data set (23,570 sequences), r_1 the number of human miRNA upstream sequences including each MEME motif, and r_2 is the number of human RefSeq gene upstream sequences including each MEME motif.

For every MEME motif, we tested for the presence of a positional preference in the distance distribution between its instances and the miRNA stem-loop starting point. We divided the 2 kb upstream region of miRNA stem-loop into bins of 100 bp, and counted the number of sites located in each bin. For our statistic, we used a *Z*-score equation as follows:

$$Z = \frac{N - \mu}{\sigma}$$

Here, N is the number of sites in each bin and μ and σ are the mean and standard deviations on the distribution of the number of sites in different bins, respectively.

To compare the MEME motifs with known transcription factor binding sites (TFBS), we used the T-Reg Comparator software (http://treg.molgen.mpg.de/; Roepcke et al., 2005). The MEME program outputs position frequency matrices (PFM) (Stormo, 2000) for each sequence motif. We compared the PFMs for the MEME motifs with the ones in the TRANSFAC database (http://www.gene-regulation.com/; Matys et al., 2006) and the JASPAR database (http://mordor.cgb.ki.se/cgi-bin/jaspar2005/jaspar_db.pl; Sandelin et al., 2004). We chose the set of vertebrate matrices and a dissimilarity cutoff of 0.5 as recommended by Roepcke et al. (2005).

Alternatively, 194 human miRNA upstream sequences (500 bp) were submitted to the other motif discovery program, Weeder Version 1.3 (Pavesi et al., 2004). To detect the top 20 most frequent motifs, we used the Weeder command: "weederTFBS.out -f <dataset> -R 50 -O HS -W 12 -e 4 -M -S -T 20". We evaluated the resulting 12 bp motifs on both strands, because Weeder considers motifs ranging from 6 to 12 bp. Then we compared the Weeder motifs with the MEME motifs.

3. Results

3.1. Sequence characteristics in the upstream of miRNA stem-loop precursors

Conservation in the upstream region of 189 human/mouse miRNA stem-loop sequences is displayed in Fig. 2A. Positions (on the *X*-axis) are given relative to the beginning (5′-end) of the stem-loop sequences. The conservation measure (on the *Y*-axis)

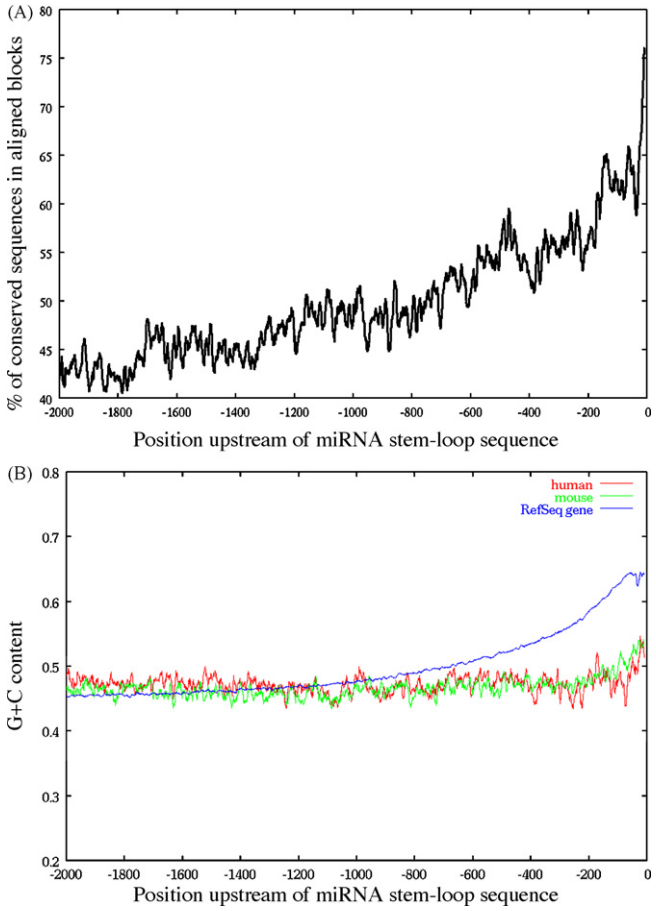

Fig. 2. (A) Conservation in the upstream region of 189 human miRNA genes. Percentage of conserved sequences in the upstream region of 189 aligned human miRNAs in a 10-base sliding window is plotted (*Y*-axis) against their position in the miRNA upstream sequences (*X*-axis, 0 = origin of stem-loop sequence). (B) The G + C content in the upstream region of human miRNA genes (red), mouse miRNA genes (green), and human RefSeq genes (blue). The G + C content in a 10-base sliding window is plotted (*Y*-axis) against their position in the each upstream sequence (*X*-axis, 0 = origin of stem-loop sequence).

is an average of the conservation of alignments of all human upstream sequences at each position. Conservation was at its highest level of about 75% at the immediately upstream region of stem-loop sequences. We noted that the degree of conservation declined with increasing distance from the stem-loop sequences. However, conservation remained relatively high, above 50%, even up to −500 bp from the stem-loop start point, with a small peak of conservation.

The G + C content in the upstream human miRNA stem-loop sequences, mouse miRNA stem-loop sequences and human RefSeq genes is shown in Fig. 2B. The G + C content (on the Y-axis) is the average of all sequences at each position. In the 2000 bp upstream human miRNA stem-loop sequences, the G + C content was in the 43.5–54.7% range (47.3% on average). At about −30 and −170 bp, relatively GC-rich regions having more than 50% of G + C content were found. Furthermore, 153 CpG islands were identified in the 2000 bp upstream regions of 106 in the total number of 242 human miRNAs by a CpG island searcher (see Supplementary Table 2).

3.2. Motif discovery

The top 20 motifs found by MEME are listed in Table 1, with their occurrence in the human and mouse 2000 bp upstream sequences of miRNA genes and also in the 2000 bp upstream sequences of human RefSeq genes. M1 to M20 showed comparable occurrence in human and mouse miRNA upstream sequences. All the top 20 motifs detected by conservation in 500 bp upstream of human miRNAs displayed significantly higher frequency of occurrence in human and mouse 2000 bp upstream sequences than in the comparable upstream RefSeq gene sequences. This difference between miRNA and RefSeq sequences was statistically significant at the 0.01 level as measured by a two-tailed test. In contrast, there is no occurrence of the M7 and M18 motifs in the upstream region of human RefSeq genes.

The occurrences of the top 20 motifs in the other data set of 136 human miRNA upstream sequences are shown in Supplementary Table 3. M1 to M20 showed the comparable occurrence in 242 and 136 human miRNA upstream sequences.

Sequence logos for the top 20 motifs are shown in Fig. 3, while position frequency matrices for these top 20 motifs are shown in Supplementary Table 4. Fig. 4 shows histograms of the location of the 3 most frequent MEME motifs (M1, M2, and M3) detected by MAST in the upstream sequences of 242 human and 290 mouse miRNAs, and 23,570 human RefSeq genes. (Histograms for M4 to M20 are shown in Supplementary Fig. 1.) Histograms of the locations for M1 to M20 appeared comparable in human and mouse miRNA upstream sequences. Fifty-four out of the 88 occurrences of M1 in the human data set are found in the 500 bp upstream region of miRNA stem-loop sequences (Fig. 4). The distribution of M1 has a peak at about 200 bp upstream from the stem-loop sequences. The positional preferences of the MEME motifs are given in Supplementary Table 5. Nine motifs (M1, 5, 8, 9, 11, 14, 16, 17, 20) out of 20 had a *Z*-score above 2.0 in one bin in each motif. The *Z*-score of positional preference of M1 was 2.4 in the bin from −300 to −200 bp. In Table 2, we display the results of comparing the MEME motifs with known transcription factor binding sites in the TRANSFAC and JASPAR databases. Fifteen motifs (M1, 4–9, 11, 13–16, 18–20) out of 20 had no known transcription factor binding sites, while the remaining five motifs had known transcription factor binding sites.

To examine the effect of repeat-masking on this analysis, we also analyzed the repeat-masked dataset of 242 human miRNA 2000 bp upstream sequences. The top 20 motifs found by MEME (MM1–20) are listed in Supplementary Table 6. MM1 had the same consensus as M1, GTGCTTMTAGTGCAG. The top 12 bp of MM4 (AAGTGCTTCCATGTT) was similar to the last 12 bp of M5 (AGTAAGTGCTTCCAT). MM15 (RGGGTGGGG) was similar to M7 (GGGRTGGGG). MM9 (CCNTGCAAAACTGAW) was similar to M11 (CCWTGCAAAACTGA). The rest of the MM motifs had no similarity to M1–20.

The top 20 motifs found by Weeder (W1–20) are listed in Supplementary Table 7, with the top 20 motifs of 12 bp found by MEME (M12-1–20). W1 (AAGTGCTTCCAT) and W14 (AAGTGCTTACAG) matched M12-3 (AAGTGCTTMCAK).

Table 1
The identified sequence motifs in the upstream regions of human miRNA stem-loop sequences

Motif	Width	*E*-value	Consensus sequence	Occurrence in 242 human miRNA upstream sequences			Occurrence in 290 mouse miRNA upstream sequences			Occurrence in 23,570 human RefSeq upstream sequences		
				No.	(%)	(forward/reverse)	No.	(%)	(forward/reverse)	No.	(%)	(forward/reverse)
M1	15	3.8e−057	GTGCTTMTAGTGCAG	31	(12.8)	(54/34)	29	(10.0)	(38/32)	27	(0.1)	(25/23)
M2	15	3.9e−056	WAAAAAAAAAAAAAA	151	(62.4)	(218/472)	151	(52.1)	(188/296)	8794	(37.3)	(20,316/19,135)
M3	12	1.5e−034	CCCCCWCCCCCC	145	(59.9)	(438/382)	148	(51.0)	(393/417)	1343	(5.7)	(4665/4370)
M4	15	3.3e−023	SSCCCCWGCCTGGCC	133	(55.0)	(289/376)	106	(36.6)	(249/250)	1329	(5.6)	(4500/4518)
M5	15	2.2e−020	AGTAAGTGCTTCCAT	41	(16.9)	(46/50)	25	(8.6)	(24/28)	7	(0.03)	(6/6)
M6	15	8.9e−026	GCYKGGACTCCTGGG	128	(52.9)	(369/253)	51	(17.6)	(74/78)	319	(1.4)	(769/795)
M7	9	2.5e+000	GGGRTGGGG	95	(39.3)	(223/153)	88	(30.3)	(158/169)	0	(0.0)	(0/0)
M8	15	9.0e−013	CCCYGCCCAGGCCCC	103	(42.6)	(257/277)	89	(30.7)	(204/200)	917	(3.9)	(2884/2919)
M9	12	4.1e−013	CTTTCTTTCTTCCTC	136	(56.2)	(338/245)	161	(55.5)	(446/352)	575	(2.4)	(1706/1694)
M10	15	1.0e−012	TTTWANTNTTTTTTT	140	(57.9)	(343/200)	113	(39.0)	(199/158)	544	(2.3)	(1167/1156)
M11	14	1.7e−008	CCWTGCAAAACTGA	20	(8.3)	(26/21)	15	(5.2)	(16/10)	11	(0.05)	(7/11)
M12	15	2.0e−005	TAGTGAAGCAGSTTA	29	(12.9)	(22/20)	36	(12.4)	(20/24)	6	(0.03)	(2/5)
M13	12	7.8e+000	CTCCTGCCTGGG	129	(53.3)	(335/327)	82	(28.3)	(148/115)	492	(2.1)	(1395/1394)
M14	15	5.9e−008	CTGCNGGCCCTGCTG	77	(31.8)	(154/125)	59	(20.3)	(116/83)	153	(0.6)	(284/339)
M15	12	2.5e+001	CNCTCCAGCTCC	93	(38.4)	(231/225)	116	(40.0)	(252/242)	136	(0.6)	(381/336)
M16	15	6.8e−002	GAGGCCGAGTTGGGC	108	(44.6)	(157/243)	51	(17.6)	(67/71)	2831	(12.0)	(5802/6160)
M17	12	2.1e+002	AGGTGTCTCCAA	19	(7.9)	(16/8)	11	(3.8)	(6/11)	21	(0.09)	(18/13)
M18	9	7.7e+002	TCCTGGAAG	62	(25.6)	(109/65)	31	(10.7)	(46/32)	0	(0.09)	(0/0)
M19	15	1.7e+002	TGGGGGTACCTGCKG	28	(11.6)	(31/28)	22	(7.6)	(26/16)	39	(0.2)	(40/39)
M20	15	3.3e−007	TGTTGGGAACAGAGG	46	(19.0)	(72/26)	25	(8.6)	(28/39)	35	(0.1)	(52/39)

In the right-most 3 columns representing "Occurrence", we indicate the number of MEME motifs aligned by MAST in each upstream data set, followed in parentheses, by the numbers of MEME motifs aligned by MAST in the forward/reverse strand in each data set. Each "Consensus sequence" is represented over an alphabet of 11 characters, consisting of A, C, G, T, S = C/G, W = A/T, Y = C/T, R = A/G, M = A/C, K = G/T, and N = A/C/G/T.

W3, W4, W8, W13, and W18 were partially similar to M12-3. Ten motifs (W2, 5, 6, 9–12, 15–17) out of 20 were similar to each other, including a triplet ATG. However, these W motifs had no similar MEME motif except M12-8 including the complementary sequence of CAT. TRANSFAC search for a transcription factor binding for M12-3 also having an exact match with W1 and W14 showed a partial similarity to HS$CD8A_03 (cccgccttggCCTCCCAAAgtgctgggat), a binding site of Ik-1.

3.3. Characteristics of the interesting MEME motifs

The most interesting MEME motif found, denoted M1, is also the most significantly conserved. This motif is represented by the consensus GTGCTTMTAGTGCAG, representing a sequence of the most frequent base at each position (its corresponding sequence logo is shown in Fig. 3). M1 has a MEME *E*-value of 3.8e−57 and is significantly over-represented in our data sets. In the human upstream data set, M1 is found by MAST 88 times in 31 sequences out of a total of 242 sequences (Table 1). Further, 21 out of the 31 sequences contain multiple M1 sites. Fifty-four out of these 88 occurrences in the human data set were found on the forward strand, a 1.6:1 preference. In the mouse upstream data set, M1 was also found in 29 sequences out of a total of 290 sequences. Thirty-eight out of these 70 occurrences in the mouse data set were found on the forward strand, a 1.2:1 preference. To verify whether M1 is simply a widely

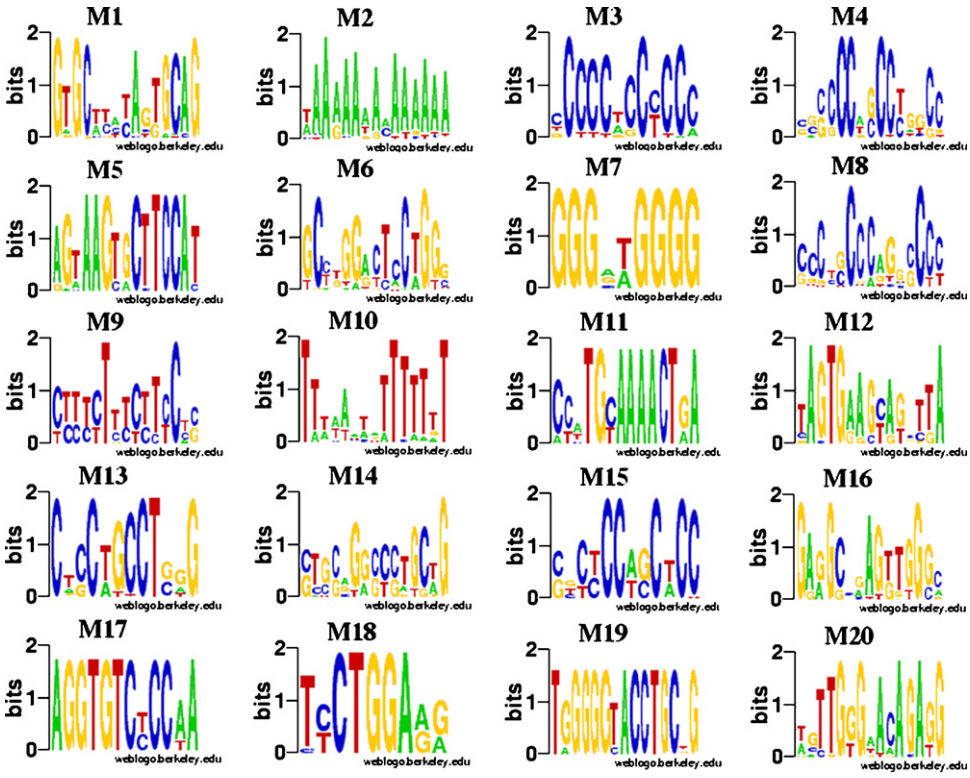

Fig. 3. Sequence logos for the 20 most frequent motifs identified by MEME. Each logo consists of stacks of nucleic acid symbols, one stack for each position in the sequence. The overall height of the stack indicates the sequence conservation at that position, while the height of the nucleic acid symbols within the stack indicates the relative frequency of each nucleic acid at that position.

occurring motif or instead a miRNA-specific one, we scanned all the 23,570 upstream sequences of RefSeq genes. In all, there were only 48 occurrences of the M1 motif in RefSeq in 27 sequences (25 occurrences on the forward strand and 23 occurrences on the reverse strand). Comparing these MEME motifs against known TFBS motifs by T-reg Comparator revealed no similar matrices for M1 in TRANSFAC and JASPAR databases (Table 2). Several TFBSs were detected with the higher dissimilarity score by the T-Reg Comparator shown in the parentheses; V$MYC_Q2 (1.153) and V$AP2REP_01 (1.226) in the TRANSFAC database, and Myc-Max (0.920) and c-MYB_1 (1.468) in the JASPAR database.

Another seemingly biologically interesting motif is M3. This 12 bp motif is represented by the consensus CCCCCWCCCCCC (the corresponding sequence logo is shown in Fig. 3). Previously, an investigation of the upstream sequences of 59 orthologous human/mouse miRNAs identified an over-represented motif, CCCWCCC, that was similar to the consensus CTCCGCCC present in conserved blocks in the upstream sequences of miRNAs in the nematode *Caenorhabditis elegans* (Ohler et al., 2004). For M3 we found a similar TFBS motif, ZNF42_1–4 (consensus NGGGGA) on the reverse strand (Table 2). This ZNF42_1–4 motif is also similar to M7, with the consensus sequence of GGGRTGGGG.

Table 2
Results of comparing the MEME-discovered motifs and known TFBS motifs

Motif	TFBS motif	Source DB	Orientation	Length	Offset	Overlap	Dissimilarity score	Consensus sequence
M2	Hunchback	JASPAR	Forward	10	−4	6	0.331	SMANAAAAAA
M3	ZNF42_1–4	JASPAR	Reverse	6	8	4	0.426	NGGGGA
M10	V$HOXA4_Q2	TRANSFAC	Reverse	8	−4	4	0.461	AWAATTRG
M12	V$GATA_Q6	TRANSFAC	Reverse	7	11	4	0.446	WGATAAN
	Snail	JASPAR	Forward	6	−1	5	0.111	CAGGTG
M17	deltaEF1	JASPAR	Reverse	6	−1	5	0.194	CACCTN
	V$E2A_Q6	TRANSFAC	Reverse	8	−3	5	0.387	CACCTGNC

The results include a list of matrices with divergence smaller than the dissimilarity cutoff of 0.5. For these matrices, the name, overlap, orientation, shift, actual dissimilarity score, and consensus sequence are provided. The T-Reg Comparator uses a database that currently contains data of JASPAR and TRANSFAC Version 8.4. The data from JASPAR and TRANSFAC Public are freely accessible on each web site. However, access to the data from the non-public part of TRANSFAC Version 8.4 is restricted to licensed users.

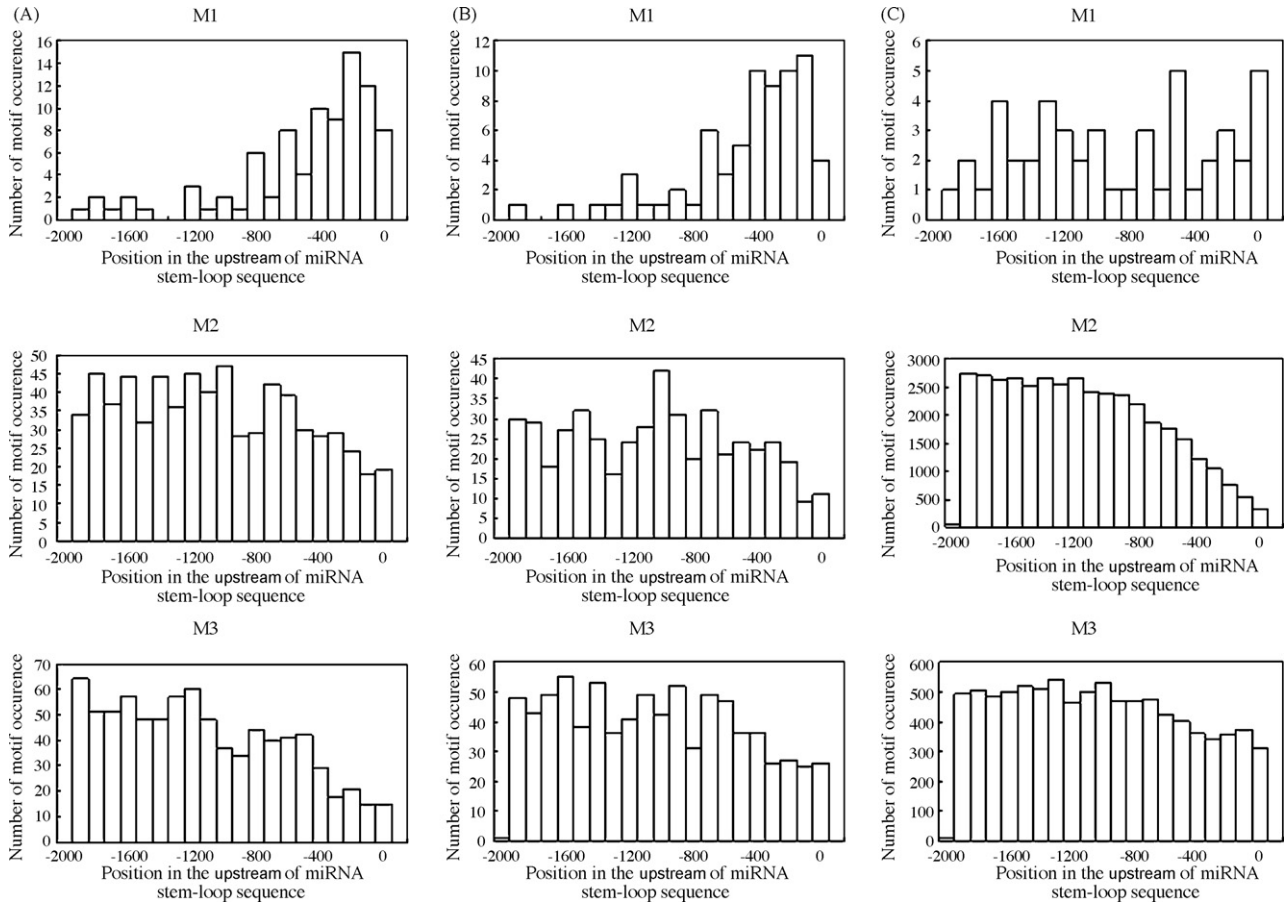

Fig. 4. Histograms of the location of the three most frequent MEME motifs M1, M2, and M3, plotted against their location from stem-loop origin, in bins of 100 bp. Histograms of the location of the MEME motifs in the upstream sequences of (A) human miRNA stem-loop sequences, (B) mouse miRNA stem-loop sequences, and (C) 23,570 human RefSeq genes.

4. Discussion

Based on the conservation within 500 bp human upstream of miRNA genes, we hypothesized that miRNA-specific expression motifs are likely to be present in these highly conserved regions in the upstream proximity of miRNA stem-loop sequence regions. The highest conservation at the immediate upstream region might not be related to expression due to its location. We therefore restricted our search by applying MEME to find conserved motifs in 500 bp upstream regions relative to annotated miRNA stem-loop sequences. The 2000 bp upstream human miRNA stem-loop sequences had a G + C content of 47.3% on average, relatively high compared to that in the whole human genome (about 41%) (Lander et al., 2001; Venter et al., 2001). Eighty-four out of the 106 upstream sequences including one or more CpG islands are located in intergenic regions. MiRNAs that have CpG-rich putative promoter upstream sequences may be independently transcribed and/or regulated by methylation.

The predominant occurrence of M1–M20 in the region upstream of miRNA genes compared to that of RefSeq genes suggests that M1 to M20 were specific to the miRNA upstream sequences. In addition, M1 to M20 had a comparable occurrence in the test data set of 136 human miRNA upstream sequences that was completely independent from the training data set for MEME. Therefore, we posit that M1 to M20 were specific not only to the training data set, but also to the miRNA upstream sequences.

Out of 20 motifs, M1 was the most conserved. The comparable occurrence of M1 in the human and mouse upstream sequences of miRNA genes suggests that M1 is evolutionarily conserved between human and mouse. The left flank of M1 is composed of the conserved sequence GTGC. The 5 bp sequence of its right flank is also highly conserved with consensus TGCAG. In contrast, the 6 bp sequence in the middle portion of M1 is not conserved, except for A at position 9. The apparent similarity of the last 4 bp of our M1 motif, GCAG to a cap signal, NCAN for transcription initiation, NCANHNNN (Bucher, 1990) may suggest the role of M1 for transcription initiation. Because several TFBSs were detected for M1 with higher dissimilarity scores, M1 might well bind with some of the known transcription factors.

Some MEME motifs such as M5, M7, M11, and M12 have a low frequency of occurrence. Few RefSeq upstream sequences contain these motifs, although both human and mouse miRNA upstream sequences have comparable frequencies of these motifs. In particular, M5, M7, and M11 motifs have no similar matrices in the TRANSFAC or JASPAR databases. These

three motifs might be candidates for novel, miRNA-specific functional elements.

The subsequence of M5 (AAGTGCTTMCAK) was found by Weeder as W1 and W14. This motif partially matched the left flank of M1 (GTGCTTMTAGTGCAG). These results lend weight to our hypothesis of a biological function for M1 and M5. We should be able to pursue this hypothesis further, because numerous computational programs have become available for motif discovery. These algorithms are different from each other, and output different motifs. Using different algorithms and comparing the results should improve the reliability of motif discovery methods.

With the repeat masked data set, MEME found motifs that were in part distinct from those found with the unmasked data set. General motif discovery algorithms identify sequence motifs that are statistically over-represented in a data set. Some repeat sequences might thus be identified as the MEME motifs. As a result, conventionally repetitive DNA elements are removed from the sequences input to the motif discovery program. However, the biological functions of repetitive DNA elements remain to be understood. Therefore, in order to avoid missing any possible conserved sequences, repeat masking should not be applied.

More than 500 different miRNAs have already been identified in animals and plants, with the number of known miRNA genes expected to increase to 500–1000 per species (Bartel, 2004). With data sets from these additional species, it should be possible to construct a comprehensive list of functional elements in the upstream region of miRNA genes. Identification of miRNA-specific motifs should also help to discover new miRNA genes.

In conclusion, we have identified over-represented and conserved motifs in the upstream regions of human and mouse miRNA stem-loop sequences. Our analyses on upstream sequence characteristics of human miRNAs provide an explicit list of candidates for human miRNA-specific regulatory motifs. In particular, we propose that our identified M1 sites may well be promising candidates for experimental verification of possible regulatory functions.

Acknowledgments

This study was supported by a grant from the Ministry of Education, Science, and Technology (Knowledge Cluster Initiative) of Japan. We thank members of the Institute for Genome Research for helpful advice and discussions.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.compbiolchem.2007.03.011.

References

Bailey, T.L., Elkan, C., 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc. Int. Conf. Intell. Syst. Mol. Biol. 2, 28–36.

Bailey, T.L., Gribskov, M., 1998. Combining evidence using *p*-values: application to sequence homology searches. Bioinformatics 14, 48–54.

Bartel, D.P., 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 116, 281–297.

Bentwich, I., Avniel, A., Karov, Y., Aharonov, R., Gilad, S., Barad, O., Barzilai, A., Einat, P., Einav, U., Meiri, E., Sharon, E., Spector, Y., Bentwich, Z., 2005. Identification of hundreds of conserved and nonconserved human microRNAs. Nat. Genet. 37, 766–770.

Bucher, P., 1990. Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. J. Mol. Biol. 212, 563–578.

Cai, X., Hagedorn, C.H., Cullen, B.R., 2004. Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. RNA 10, 1957–1966.

Crooks, G.E., Hon, G., Chandonia, J.M., Brenner, S.E., 2004. WebLogo: a sequence logo generator. Genome Res. 14, 1188–1190.

Filipowicz, W., Jaskiewicz, L., Kolb, F.A., Pillai, R.S., 2005. Post-transcriptional gene silencing by siRNAs and miRNAs. Curr. Opin. Struct. Biol. 15, 331–341.

Gardiner-Garden, M., Frommer, M., 1987. CpG islands in vertebrate genomes. J. Mol. Biol. 196, 261–282.

Griffiths-Jones, S., Grocock, R.J., van Dongen, S., Bateman, A., Enright, A.J., 2006. miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res. 34, D140–D144.

Houbaviy, H.B., Dennis, L., Jaenisch, R., Sharp, P.A., 2005. Characterization of a highly variable eutherian microRNA gene. RNA 11, 1245–1257.

Kim, V.N., 2005. MicroRNA biogenesis: coordinated cropping and dicing. Nat. Rev. Mol. Cell. Biol. 6, 376–385.

Kim, V.N., Nam, J.W., 2006. Genomics of microRNA. Trends Genet. 22, 165–173.

Lander, E.S., et al., 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921.

Lee, Y., Kim, M., Han, J., Yeom, K.H., Lee, S., Baek, S.H., Kim, V.N., 2004. MicroRNA genes are transcribed by RNA polymerase II. EMBO J. 23, 4051–4060.

Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., Voss, N., Stegmaier, P., Lewicki-Potapov, B., Saxel, H., Kel, A.E., Wingender, E., 2006. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res. 34, D108–D110.

Ohler, U., Yekta, S., Lim, L.P., Bartel, D.P., Burge, C.B., 2004. Patterns of flanking sequence conservation and a characteristic upstream motif for microRNA gene identification. RNA 10, 1309–1322.

Pavesi, G., Mereghetti, P., Mauri, G., Pesole, G., 2004. Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. Nucleic Acids Res. 32, W199–W203.

Roepcke, S., Grossmann, S., Rahmann, S., Vingron, M., 2005. T-Reg Comparator: an analysis tool for the comparison of position weight matrices. Nucleic Acids Res. 33, W438–W441.

Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W.W., Lenhard, B., 2004. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res. 32, D91–D94.

Sontheimer, E.J., Carthew, R.W., 2005. Silence from within: endogenous siRNAs and miRNAs. Cell 122, 9–12.

Stormo, G.D., 2000. DNA binding sites: representation and discovery. Bioinformatics 16, 16–23.

Takai, D., Jones, P.A., 2002. Comprehensive analysis of CpG islands in human chromosome 21 and 22. PNAS 99, 3740–3745.

Takai, D., Jones, P.A., 2003. The CpG island searcher: a new WWW resource. In Silico Biol. 3, 235–240.

Venter, J.C., et al., 2001. The sequence of the human genome. Science 291, 1304–1351.

Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., Lander, E.S., Kellis, M., 2005. Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. Nature 434, 338–345.

Xie, Z., Allen, E., Fahlgren, N., Calamar, A., Givan, S.A., Carrington, J.C., 2005. Expression of Arabidopsis MIRNA genes. Plant Physiol. 138, 2145–2154.