BIOINFORMATICS DISCOVERY NOTE Vol. 21 no. 7 2005, pages 841–845
doi:10.1093/bioinformatics/bti073

Sequence analysis

Profile-based detection of microRNA precursors in animal genomes

Matthieu Legendre[1], André Lambert[2] and Daniel Gautheret[1,*]

[1]INSERM ERM-206, Luminy Case 928, 13288 Marseille Cedex 09, France and [2]CNRS UMR6207, Luminy Case 907, 13288 Marseille Cedex 09, France

Received on August 12, 2004; revised on September 17, 2004; accepted on September 18, 2004
Advance Access publication October 27, 2004

ABSTRACT

Motivation: MicroRNAs (miRNA) are essential 21–22 nt regulatory RNAs produced from larger hairpin-like precursors. Local sequence alignment tools such as BLAST are able to identify new members of known miRNA families, but not all of them. We set out to estimate how many new miRNAs could be recovered using a profile-based strategy such as that implemented in the ERPIN program.

Results: We constructed alignments for 18 miRNA families and performed ERPIN searches on animal genomes. Results were compared to those of a WU-BLAST search at the same *E*-value cutoff. The two combined approaches produced 265 new miRNA candidates that were not found in miRNA databases. About 17% of hits were ERPIN specific. They showed better structural characteristics than BLAST-specific hits and included interesting candidates such as members of the miR-17 cluster in *Tetraodon*. Profile-based RNA detection will be an important complement of similarity search programs in the completion of miRNA collections.

Contact: gautheret@esil.univ-mrs.fr

INTRODUCTION

MicroRNAs (miRNAs) are small non-coding RNAs regulating gene expression through mRNA degradation or translation inhibition. miRNAs were first discovered in the nematode *Caenorhabditis elegans* and then identified in numerous plant and animal species. Experimental data have shown that they regulate functionally important pathways related to development (Reinhart *et al*., 2000), cell death (Brennecke *et al*., 2003), cancer (Calin *et al*., 2004) and neurological diseases (Dostie *et al*., 2003). The 21–22-nt-long mature miRNAs are synthesized from a longer 70–100-nt precursor (pre-miRNA) forming a long hairpin structure that contains the mature miRNA in either of its arms.

Although the first miRNAs were discovered using experimental methods (Lee and Ambros, 2001; Lagos-Quintana *et al*., 2001), many more were predicted through various computational screens, such as comparative genomics, that can detect entirely new RNA families (Lai *et al*., 2003; Lim *et al*., 2003a,b). Release 3.1 of the microRNA registry (Griffiths-Jones, 2004) contains 899 different miRNA precursors identified in Metazoa (*Drosophila melanogaster, C. elegans, Caenorhabditis briggsae, Homo sapiens, Mus musculus* and *Rattus norvegicus*) and plants (*Arabidopsis thaliana* and *Oryza sativa*). The microRNA registry contains experimentally validated miRNAs or close homologues identified using local alignment programs such as BLAST (Altschul *et al*., 1990). However, sequence alignment alone may fail to identify miRNAs that diverged too far apart in their primary sequence while retaining their base-paired structure. We expected that a more sensitive approach exploiting information from both miRNA structure and sequence could significantly improve detection of homologous miRNA genes. We here report the prediction of new members of known miRNA families using such a computational approach.

Two major computational techniques can exploit both RNA structure and sequence alignments for non-coding RNA (ncRNA) searches. Stochastic Context Free Grammars, or SCFGs (Sakakibara *et al*., 1994; Eddy and Durbin, 1994), provide a fully probabilistic description of RNA sequences and structure, albeit at a high computational cost which can be prohibitive at genomic scales. Alternatively, the ERPIN program represents RNA alignments as weight matrices or profiles (Gautheret and Lambert, 2001) and identifies matching sequences using a combined dynamic programming/profile scan algorithm. ERPIN profiles therefore captures both primary and secondary structure information, which is particularly well adapted to miRNA precursor identification. In its latest version, ERPIN also compensates for the incompleteness of training sets using pseudocounts and performs accurate *E*-value calculations. Here we use ERPIN to identify new microRNA precursors in animal genomes. We then compare ERPIN predictions to those made by the WU-BLAST program (Gish, 1996–2004, http://blast.wustl.edu) and show that current efforts to complete miRNA collections would benefit from using a profile-based approach.

MATERIALS AND METHODS

miRNA training sets

Precursor sequences were downloaded from the miRNA registry 2.2, containing 593 sequences (513 animal + 50 plant) (Griffiths-Jones, 2004). Due to the diversity of miRNAs, it is not possible to build a single profile describing all families. However, precursor sequences within a single family of related miRNAs (e.g. let-7) are reasonably conserved and can be aligned easily. To obtain families of homologous miRNAs, we started with a CLUSTALW (Thompson *et al*., 1994) alignment of all 513 animal miRNA precursors in the miRNA registry. From a visual inspection of the CLUSTAL-generated tree, we extracted the 18 most salient miRNA clusters, ranging in size from 6 to 27 sequences. Precursor sequences contained within each cluster were not necessarily homologous, but they were close enough in terms of sequence

*To whom correspondence should be addressed.

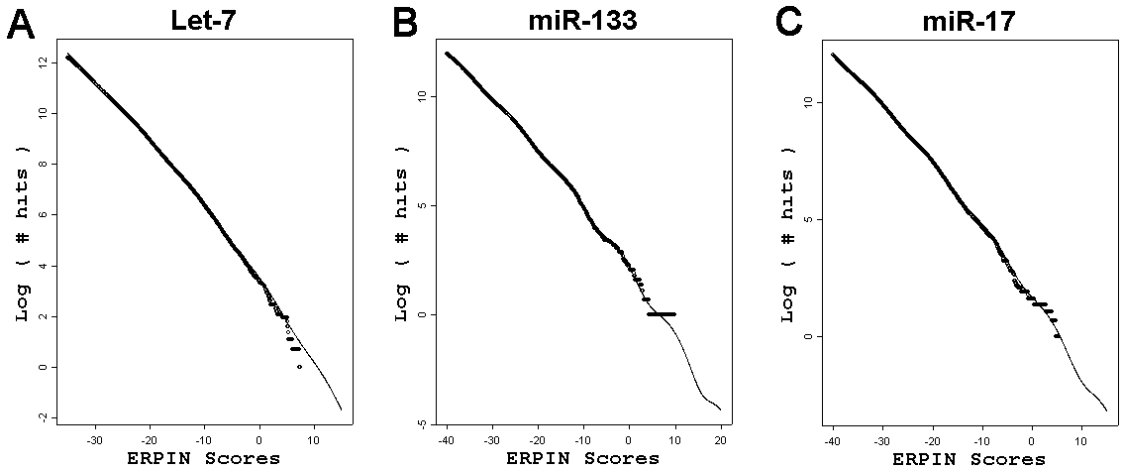

Fig. 1. Number of ERPIN hits in a 300 Mb random sequence, as evaluated from simulation (circles) and from *E*-values computed by ERPIN (line), for three different miRNA training sets: A, let-7 (27 sequences); B, miR-17 (10 sequences); C, miR-133 (6 sequences).

and structure to produce a useful signature for the subsequent profile search. Some known miRNA families are clearly grouped together in these clusters, such as the let-7 miRNAs, the miR-17 family or the miR-2/miR-13 family. Average identity within clusters was 77%. As no automated method is available to perform accurate structure-based RNA alignments, we realigned each cluster with CLUSTALW and deduced a consensus secondary structure using the ALIFOLD program (Hofacker *et al*., 2002) with default parameters and a covariance weight empirically set to 1 or 2. The 18 training sets are available through the ERPIN server Web site (Lambert *et al*., 2004, http://tagc.univ-mrs.fr/erpin/).

The novelty of miRNA candidates was assessed by comparing ERPIN or BLAST hits to the latest version of the miRNA registry (Griffiths-Jones, 2004) (at time of submission: version 3.1, containing 828 animal sequences) and to the 117 precursors identified by Tanzer and Stadler (2004) using a similarity search procedure.

Sequence databases

The search database contained the following genomic sequences: *C.elegans*, *C.briggsae*, *D.melanogaster*, *Human* (Goldenpath 8.33), *Mouse* (Goldenpath 30.16), *Rat* (Goldenpath 11.2) and *Zebrafish* (version 3) from Ensembl (ftp://ftp.emsembl.org); *Chimpanzee*, *Pig* and *Rabbit* genomes from Ensembl traces (ftp://ftp.ensembl.org/pub/traces/); *Chicken*, *Ciona intestinalis* (release 1.0), *Fugu rubripes* (version 3.0), *Sea urchin* and *Xenopus tropicalis* (assembly 1.0) from the JGI genome portal (ftp://ftp.jgi-psf.org/pub/JGI_data/); *Ciona savignyi* from the Broad Institute (ftp://ftp.broad.mit.edu/pub/annotation/ciona/assembly_4_25_2003/); *Bos taurus*, *Drosophila pseudoobscura* and *Rhesus macacus* from Baylor HGSC (ftp://ftp.hgsc.bcm.tmc.edu/pub/data); and *Tetraodon nigroviridis* from the NCBI Trace database (ftp://ftp.ncbi.nih.gov/pub/TraceDB/tetraodon_nigroviridis/).

ERPIN *E*-values

ERPIN *E*-values are computed based on a discrete convolution analysis of profile scores, as explained in detail in http://tagc.univ-mrs.fr/erpin/. Computed *E*-values were compared to simulations performed on random databases for a variety of training sets including tRNAs, snoRNAs, SECIS elements and miRNAs. In each case, a remarkable agreement was observed between computed *E*-value and simulation. Figure 1 shows simulated and computed numbers of hits at different score levels, for a search performed on 300 Mb random sequences of uniform nucleotide composition. Training sets corresponding to the let-7, miR-133 and miR-17 families are shown here, containing 27, 10 and 6 sequences respectively. In each case, computed *E*-values faithfully reproduce observed number of hits. Other miRNA training sets and other means of randomization based on order-1 or order-2 Markov models of actual genome sequence behave similarly (data not shown).

ERPIN and BLAST parameters

To improve the search when using underpopulated training sets, ERPIN uses pseudo-counts based on the Henikoff and Henikoff (1996) procedure. Implementation details are provided on the ERPIN Web site http://tagc.univ-mrs.fr/erpin/. A pseudo-count weight of 0.2 was used for all searches. Other ERPIN parameters were left at default values. Search regions included all helices and the conserved single strands shown in upper case in Table 1. Search parameters for all training sets, including those not shown in Table 1, are provided on the ERPIN Web site. All hits with an *E*-value below 0.01 were retained. BLAST searches were performed using the WU-BLAST package (Gish, 1996–2004, http://blast.wustl.edu), with a word-length parameter decreased from 11 to 7 to improve sensitivity.

Hits obtained from both programs were filtered using Dust (Tatusov, unpublished) and Repeat Masker (Smit, unpublished) to mask repeats and low complexity regions. All programs were run on a 10-CPU computer cluster, with an average runtime of 4.7 h to screen both strands of the complete 14.3 Gb database for a single miRNA family.

RESULTS AND DISCUSSION

To identify new miRNA precursors in animal genomes, we built training sets for 18 miRNA families, using a semi-automated procedure based on CLUSTALW (Thompson *et al*., 1994) for RNA alignment and ALIFOLD (Hofacker *et al*., 2002) for secondary structure annotation. These training sets were then fed into ERPIN to scan a 14.3 Gb database comprising a total of 20 complete or partial genomes. ERPIN detected a total of 553 hits with an *E*-value <0.01, of which 270 were not previously reported in the current miRNA databases. Most of these novel candidates have an even lower *E*-value (e.g. 216 hits at $E \leq 10^{-4}$), indicating that strong miRNA candidates remain unannotated in current databases. Only five out of the 270 candidates displayed a non-hairpin structure (e.g. V-shaped) when submitted to secondary structure prediction by the RNAfold program (Hofacker, 2003). The fact that most ERPIN predictions adopt a correct hairpin folding is comforting but expected,

Table 1. ERPIN-specific microRNA precursor candidates

miRNA families	Sp[a]	Genomic sequence name		Position	E-value	New miRNA precursors sequences[b]
miR-9, miR-79	cin	scaffold_171	-	195229-195278	$7.57\,10^{-3}$	TTTGG t CATT t AGT ccggtgcagtgtgtgg---------- CAT --CAA ACT g GATA A CCAAA
miR-15,	gga	2369428_fasta.scree	+	32335-32392	$2.61\,10^{-3}$	AGCAGCACG gt CAGAGAGG cacaaaggaggtgatcctg---- CCTCTCTG ctc TGTGCTGGT
miR-16,	cel	Chromosome IV	-	4316433-4316492	$2.71\,10^{-3}$	TGCGGCACG ga AAAATGGG cgtggtttcgattgagaccacg- CCTATTTT tc- TGTGCCGCA
miR-195	fru	scaffold_1057	+	2403-2457	$4.77\,10^{-3}$	AGCAGCACG ga ATGGTTTG tgggttacactgagata------ CAGGCCAT ac- TGTGCTGCC
	fru	scaffold_1454	-	15258-15316	$2.66\,10^{-6}$	AGCAGCACA tc ATGTTTTG cagatttgtacagaatcactc-- CAAATCAT aa- TGTGCTGCC
	tni	C0AG1259cb07LP1	+	457-511	$7.10\,10^{-8}$	AGCAGCACG.ga.ATGGTTTG.tgagttacactgagata------.CAAGCCAT.gc-.TGTGCTGCC
	xtr	scaffold_1703	-	23636-23693	$7.74\,10^{-3}$	AGCGGCACA.gc.AGTGTGTA.acatgtaggaagccccgact---.CACACACT.at-.TGTGCTGCT
	xtr	scaffold_24135	+	2073-2133	$7.48\,10^{-3}$	AGCAGCACA ca GATATTGG cagtttcaaccctaaagctctg- CCAGTTCT gct TGTGCTGCT
	dre	Contigctg24119.3	-	177314-177374	$4.16\,10^{-5}$	AGCAGCATG.ta.AATATTGG.agttactccttggccaatgcct-.CCAATATT.gct.CGTGCTGCT
	dre	Contigctg30116.4	+	2382-2442	$2.09\,10^{-6}$	AGCAGCACA.tc.AATATTGG.cagctgccctctctctgggttg-.CCAGTATG.gtt.TGTGCTGCT
	dre	Contigctg30116.4	+	1857-1914	$4.38\,10^{-6}$	AGCAGCACA tc ATGGTTTG taagttataagggcaaattc--- CGAATCAT ga- TGTGCTGTC
	dre	ContigNA8583.1	-	1028-1083	$6.74\,10^{-5}$	AGCAGAACA.gc.AATACTGG.ataacaccacaatcacct------.CCAGTATT.ta-.TGTGCTGCT
	dre	ContigAL606727.3	-	61519-61574	$9.99\,10^{-5}$	AGCAGCACA.ga.ATGGTTTG.tgagttataacgggggtg------.CAGGCCGT.ac-.TGTGCTGCG
	dre	ContigNA30315.1	+	1092-1146	$1.03\,10^{-3}$	AGCAGCGCG tc ATGGTTTT caacattagagaaggtg------ CAAGCCAT ca- TTTGCTGCT
miR-17,18, 20, 93, 106	tni-1	C0AG162CG06LP1	+	770-838	$7.27\,10^{-6}$	GCGA aa AAGTGCTT t ca GT GCAG g TAG tgacgataacctggc- CTA CTGC ag tg t GAGCCCTT ctt TCCC
	tni-2	C0AG866CB02LP1	-	126-194	$5.49\,10^{-8}$	GCGA.ta.AAGTGCTT.aca..GT.GCAG.g.TAG.tgacgataacctggc-.CTA.CTGC.agtgt...GAGCACTT.ctt.TCCC
	tni-3	C0AG1220dd04LP1	-	714-778	$5.20\,10^{-7}$	TGTC.aa.AAGTGCTG.ttt..GT.GCAG.g.TAG.cgttcatccat-----.CTA.CTGC.aaaac...CAGCACTT.taa.GGCA
miR-34	xtr	scaffold_19278	-	11877-11934	$7.61\,10^{-4}$	GGCAGTGTA g TT AGC TGA t TG tgttaacataagacttg------ CA a TCA ct- AG c TAAACTACC
miR-133, miR-145	rno	Chromosome 9	+	13549277-13549341	$4.41\,10^{-7}$	TC t GGCTGGT caaac- GGA a CCAA gtccgtcttcctgagaggt--- TTGG TCC CCTTCA ACCAGCT a CA
	tni	C0AG1445dc08LP1	-	86-151	$1.06\,10^{-3}$	TG t GGCTGGT caaac- GGA a CCAA gtcaggtgttttctgtgaggt-- TTGG TCC CCTTCA ACCAGAC t AT
	tni	C0AG790DH08SP1	-	127-192	$2.39\,10^{-7}$	TG t GGCTGGT aaaac- GGA a CCAA gtcaggtgtttttgtgaggt-- TTGG TCC CCTTCA ACCAGCT a TG
	tni	C1AA007ZB10A1	-	207-272	$1.02\,10^{-5}$	TG.t.GGCTGGT.caaac-.GGA.a.CCAA.gtcaggtgtttctgtgaggt--.TTGG.TCC.CCTTCA.ACCAGCT.a.CT
	tni	C0AG890CE12SP1	-	670-735	$2.39\,10^{-7}$	TG.t.GGCTGGT.caaac-.GGA.a.CCAA.gtcaggtgtttctgtgaggt--.TTGG.TCC.CCTTCA.ACCAGCT.a.TG
	xtr	scaffold_7052	-	12019-12083	$2.64\,10^{-6}$	TG c GGCTGGT gaaaa- GGA a CCAC atcaacccagaaaaaggat--- TTGG TCC CCTTCA ACCAGCC g CA
	xtr	scaffold_60773	+	1544-1608	$1.90\,10^{-7}$	TA t GGCTGGT caaac- GGA a CCAA gtccgtcttccttagaggt--- TTGG TCC CCTTCA ACCAGCT a TT
	dre	ContigctgI0804.1	-	159936-160001	$2.00\,10^{-4}$	AG t TGCTGGT aaaac- GGA a CCAA gtcgggtgtttgcgagaggt-- TTGG TCC CTTTCA ACCAGCT a CT
	dre	ContigBX276101.6	-	114530-114595	$4.16\,10^{-6}$	TG t GGCTGGT caaat- GGA a CCAA gtcaggtgtttctgcgaggt-- TTGG TCC CCTTCA ACCAGCT a CT
miR-124	cin	scaffold_310	-	18721-18786	$7.72\,10^{-5}$	TGA gtc GTGT t TAC t GT gga-- CCTT g CTG tgtgacgtcac----- AAT - AAGG CAC GC g GTG a ATGC caa TCA
	cin	scaffold_310	-	18861-18926	$2.26\,10^{-3}$	TCG gta GTAT t TAT t GT gga-- CCTT g CTG tgtgacgtcac----- AAT - AAGG CAC GC g GTG a ATGC caa CGA
	csa	paired_scaffold_20	+	44747-44812	$9.83\,10^{-4}$	TCA tcc GCGT t TGC t GC gga-- CCTT c TTG tgtgacatca------ CAA t AAGG CAC GC g GTG a ATGC caa TGA
	csa	sep_scaffold_15403	+	33312-33377	$4.29\,10^{-3}$	TCA tcc GCGT t TGC t GT gga-- CCTT c TTG tgtgacatca------ CAA t AAGG CAC GC g GTG a ATGC caa TGA
let-7, miR-98	rno	Chromosome 14	-	14688839-14688894	$6.21\,10^{-3}$	GAG G TAG ta AGTTCTGT g GC ggaactccgagttgga---------- AC a GTAGGGCT tg CTA t TTT
	cin	scaffold_95	-	49117-49174	$3.27\,10^{-10}$	GAG G TAG tt GGTTGTAT t GT ttcccttgtgtaatgtta---------- AC t ATACAGCC cg CTA g CTT
	cin	scaffold_138	+	97023-97094	$7.56\,10^{-6}$	GAG G TAG ta GGTTATGC a GT tttgcgacattattctcgttgcgttggagata----- AC t GTGTAGCC ta CTG a CTC
	csa	paired_scaffo ld_355	-	222516-222586	$4.76\,10^{-5}$	GAG G TAG ta GGTTATGC a GT ttggcgacaatatttcgttgcaaaggagata------ AC t GTGTAACT tg CTG g CTC
	csa	paired_scaffold_355	-	225348-225420	$2.86\,10^{-6}$	GAG G TAG ta GGTTATGC a GT ttttaccttaaatattttagaagcacaagaact---- AC t GTGTAACC tt CTG c CTT
	csa	paired_scaffold_419	+	256598-256655	$5.20\,10^{-10}$	GAG G TAG tt GGTTGTAT t GT tttacaccttcggcgata---------- AC t ATACAGCC ag CTA a CTT
	csa	Sep_scaffold_30687	-	221296-221366	$1.13\,10^{-4}$	GAG.G.TAG.ta.GGTTATGC.a.GT.ttggcgacaatattccgttgcaaaggagata-------.AC.t.GTGTAACT.tg.CTG.g.CTC
	csa	Sep_scaffold_30687	-	224593-224666	$6.80\,10^{-6}$	GAG.G.TAG.ta.GGTTATGC.a.GT.ttttgccttgaatgtttttagaggcacaggagct---.AC.t.GTGTAACC.tt.CTG.c.CTT
	mmu	Chromosome 11	+	27707640-27707713	$7.20\,10^{-3}$	GTG G TAG at GATTATGT t GT aatcactacagaagagaggcttgcagcccttgaa--- AC a ACATAAAC ta CTG c CAT
miR-29, mir-285	dre	Contigctg21709.2	+	62578-62642	$1.19\,10^{-5}$	GA a GCTGAATTC aga TGGTG ccatagagtattttatggcatctag---------- CACCA ttt GAAATCAGT g TT
miR-35, 36, 37, 38, 41, 42, 140	cbr	c014000850.1.2317	+	2218-2286	$9.19\,10^{-3}$	CCTGCCCTGAT tc TTGCTG cag TGATATTT cgaacggcta------ GATTATCA cc GGGTGA ca ATTAGGACAGG
Lin-4, miR10, 99, 100, 125	fru	scaffold_2973	+	16269-16336	$7.48\,10^{-4}$	CTCG aac CC gt- AGA tc CG aa— CTTGTG ttaagtgactgcaagc-- CACAAG ct TG aa TCT ac AG gtc TGCG
	tni	C0AG1360AG05SP 1	-	723-790	$1.02\,10^{-3}$	CTCG aac CC gt- AGA tc CG aa— CTTGTG ttacctgactggaagc-- CACAAG ct CG ga TCT ac AG gtc TGCG
miR-92, 235, 310, 311, 312, 313	cin-1	scaffold_20	+	504448-504512	$9.86\,10^{-4}$	CGATAGGTCGTG tttggt- GTATCA tacttaagcaaacacag---- TATTGC ACCTGT CCCGGCCGATCG
	cin-2	scaffold_168	+	27057-27120	$5.20\,10^{-7}$	TTGCAGGTTTGG accgct- GCACCA gattatataaaatttc----- TATTGC ACTTGT CCCGGCCTTCAA
	csa-1	paired_scaffold_103	-	606271-606334	$2.46\,10^{-3}$	TTGTAGGTTTGG atcggt- ACATTA gattttataatatttc----- TATTGC ACTTGT CCCGGCCTTCGA
	csa-2	sep_scaffold_32025	-	237411-237474	$2.05\,10^{-4}$	TTGTAGGTTTGG.atcggt-.ACATTA.gattttataatatttc-----.TATTGC.ACTTGT.CCCGGCCTTCAA
miR-30, miR-259	dre	Contigctg14936.3	+	6971-7031	$3.08\,10^{-9}$	GTAAA --- CATC cccga CTG ga AGC tgtgctacgcggaaaacga-- GCT tt CAG ttg GATG TTTGC
	dre	Contigctg14936.3	+	7261-7317	$1.54\,10^{-3}$	GTAAA --- CATC ctaca CTC -- AGC tgtgagctgcagacgag---- GCT gg GCG gag GGTG TTTGC
	dre	Contigctg9441.1	-	94991-95047	$7.91\,10^{-4}$	GTAAA --- CATC ctaca CTC tc AGC tggagcgcagccgag------ GCC gg GAG tgg GATG TTTGC

[a] Species: cin (*Ciona intestinalis*), gga (*Gallus gallus*) cel (*Caenorhabditis elegans*), fru (*Fugu rubripes*), tni (*Tetraodon nigroviridis*), xtr (*Xenopus tropicalis*), dre (*Dario rerio*), rno (*Rattus norvegicus*), csa (*Ciona savignyi*), mmu (*Mus musculus*), cbr (*Caenorhabditis briggsae*.)

[b] Sequences are aligned as in training set; boxes indicate secondary structure helices.

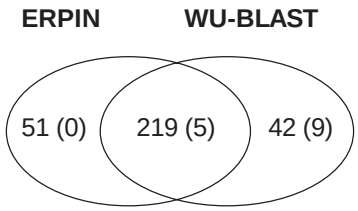

Fig. 2. Number of new miRNA candidates identified by ERPIN and WU-BLAST and not present in the miRNA registry release 3.1. Numbers in parenthesis indicate hits with an incorrect predicted folding.

Table 2. Number of correctly folded ERPIN and BLAST hits at different *E*-value cutoffs

E-value cutoff	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
ERPIN-specific	51	51	56	64	69
BLAST-specific	33	33	36	34	34
ERPIN and BLAST	214	214	185	152	121
Total	298	298	277	250	224

since ERPIN base-pair profiles generally entail strong penalties for base mismatches.

We next asked what proportion of the ERPIN hits would have been detected using a conventional similarity search program. We used the WU-BLAST program at a sensitive setting (BLASTN word length = 7), using each sequence in the 18 training sets in turn as a query. Running these queries against the same database produced 261 novel hits at an *E*-value cutoff of 0.01. Fourteen of these sequences were predicted to fold incorrectly by RNAfold.

Figure 2 shows a Venn diagram of ERPIN and WU-BLAST hits. There are 219 hits common to ERPIN and WU-BLAST, among which five are not folded correctly. Although most of the new precursors are detected by both programs, 51 candidates escaped sequence similarity search alone and were only detected by profile search. All of these were correctly folded into a single hairpin structure. There were also 42 BLAST-specific predictions, although nine of them were not predicted to fold correctly. ERPIN fails to detect some BLAST hits because it does not currently allow for insertions relative to the longest sequence in the training set, while BLAST does allow for insertions. However, ERPIN-specific hits are more abundant and better folded, especially at low *E*-values. If one only considers correctly folded candidates in Fig. 2, 72% are detected by both programs, 11% are BLAST specific and 17% are ERPIN specific. Furthermore, the proportion of ERPIN specific hits increases with *E*-value cutoff, as shown in Table 2 for cutoffs ranging from 10^{-1} to 10^{-5}. When using a high confidence level (*E*-values of 10^{-5} or better), ERPIN specific hits reach 30%. Candidate miRNA sequences identified by ERPIN, BLAST or both programs are available as supplemental data.

Table 1 shows the 51 ERPIN-specific hits identified in this study along with their consensus secondary structures. It appears that some single strands in our alignments could be paired together. This is a corollary of the ALIFOLD-based annotation procedure that tends to reject those base pairs not supported by covariation. Our goal here was to use an alignment/search procedure that was as automatic and straightforward as possible, but it is likely that alignments could be further refined, thus improving search performances.

An important miRNA annotation effort is currently being undertaken as part of the RFAM database (Griffiths-Jones *et al*., 2003). RFAM proposes both 'seed' alignments (validated miRNA precursors) and SCFG-based predictions. An accurate comparison of these and our predictions is not currently possible, as RFAM candidates have no associated *E*-value and are not published or described in detail. In any case, current RFAM predictions show no overlap with ERPIN candidates in Table 1.

Recently, Tanzer and Stadler (2004) studied the evolution of 'miR-17' gene clusters located on chromosomes 7, 13 and X in the human genome. This cluster, composed of miR-17, miR-18, miR19a, miR-19b, miR-20, miR-25, miR-92, miR-93, miR-106a and miR-106b, is proposed to result from ancient miRNA duplication events followed by duplications of the entire cluster. Our initial clustering procedure grouped the majority of these miRNAs into a single training set (Fig. 3A), containing miR-17, miR-18, miR-20, miR-93, miR-106a and miR-106b, whilst miR-92 fell into another training set (Fig. 3B). The phylogenetic trees in Figure 3 show the relationships between training set sequences (light grey) and new hits identified from these training sets by both ERPIN and BLAST (noted 'EB') or by ERPIN only (noted 'E'). No BLAST-specific hit was obtained in these runs. It is noteworthy that ERPIN performed better than BLAST in identifying distant homologues in these examples. For instance, BLAST did not identify any new member of the miR-106a/miR-17 family, while ERPIN identified two homologues in *Tetraodon* (Fig. 3A, black arrows). Likewise, the miR-93 homologue identified in *Tetraodon* is ERPIN specific (Fig. 3A, white arrow); and the miR-92 and miR-92a homologues identified in *Ciona* are all ERPIN-specific (Fig. 3B, arrows). In these examples, BLAST identified orthologues between two mammalian or two *Drosophila* species, but profile search detected miRNAs that diverged earlier in the metazoan lineage, such as *Ciona* hits identified from their vertebrate homologues (e.g. miR-92, Fig. 3B).

To determine whether a miR-17-like cluster was indeed present in *Tetraodon*, we mapped the three ERPIN-specific *Tetraodon* hits (Fig. 3A and Table 1) onto the current genomic assembly available at Genoscope (http://www.genoscope.cns.fr/externe/tetraodon). Although the *Tetraodon* hits tni-1 and tni-2 are from different scaffolds, they both map to the same location on chromosome 1 (position 5477786–5777854) with one mutation for tni-2. Either one of these hits comes from an as yet unassembled part of the *Tetraodon* genome, or they both come from the same miRNA with a sequence error or variation. The tni-3 miRNA maps at position 5477580–5477644 of chromosome 1, in the vicinity of tni-1/tni-2. This suggests that a cluster containing both miR-93 and miR-106/miR-17 homologues is present on chromosome 1 of *Tetraodon*, which is probably related to the miR-17 cluster in mammals. To our knowledge, these putative members of the *Tetraodon* miR-17 cluster were not previously reported.

RNA genes do not produce as noticeable phylogenetic footprints as their protein-coding counterparts, and their detection by similarity search program is often restricted to relatively close homologues. In the case of miRNA precursors, a program such as BLAST or WU-BLAST failed in our examples to detect some divergent orthologues such as between vertebrates and tunicates, or even between mammals and fishes. This tendency is not verified for all miRNA clusters: in some cases (data not shown), BLAST hits were more divergent from training set sequences than ERPIN hits. Therefore, both programs may fail to detect 'interesting' distant homologues. In any case, using

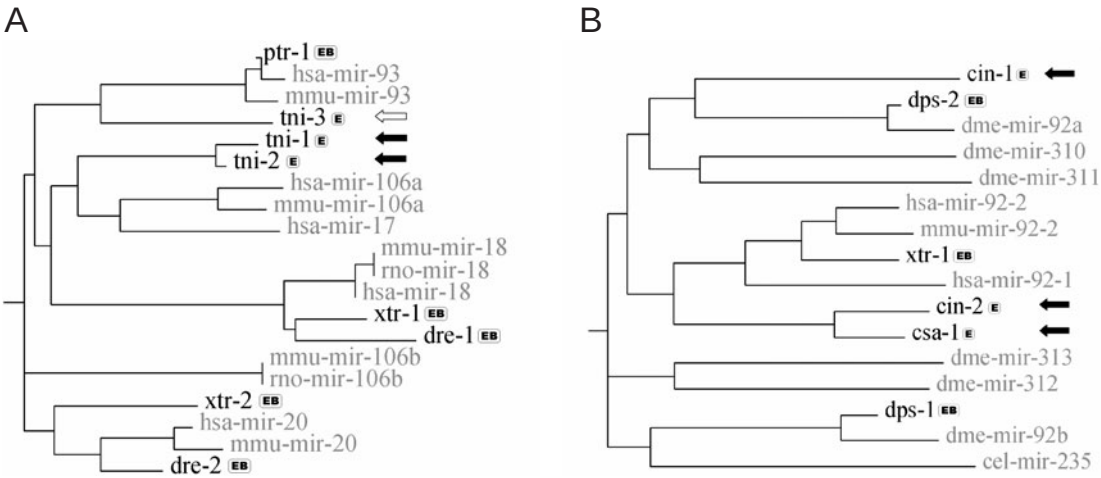

Fig. 3. Neighbor-joining tree of miRNA clusters obtained from two different training sets. A, 'miR-17' cluster; B, 'miR-92' cluster. Training set sequences are shown in grey; new candidates identified in this study are shown in black. Refer to Table 1 for sequences of ERPIN candidates and to supplemental data for sequences of other candidates. The symbol 'E' indicates hits identified by ERPIN only, and the symbol 'EB' indicates hits identified by both ERPIN and BLAST. Organisms: ptr (*Pan troglodytes*), hsa (*Homo sapiens*), mmu (*Mus musculus*), tni (*Tetraodon nigroviridis*), rno (*Rattus norvegicus*), xtr (*Xenopus tropicalis*), dre (*Dario rerio*), cin (*Ciona intestinalis*), dps (*Drosophila pseudoobscura*), dme (*Drosophila melanogaster*) and cel (*Caenorhabditis elegans*). Arrows refer to miRNAs discussed in text.

ERPIN increased the number of novel miRNA candidates by 17% at an E-value of 10^{-2}, or by 25% at an E-value of 10^{-4}, when compared to a BLAST search. Therefore, a comprehensive annotation of miRNA in genomes should involve both sequence similarity search and an RNA-specific profile or SCFG search. The detection of miRNA candidates for experimental validation will be significantly improved by combining these computational tools.

ACKNOWLEDGEMENTS

The authors thank Dr. Pascal Hingamp and Dr. Rémi Houlgatte for their careful reading of the manuscript and useful suggestions.

SUPPLEMENTARY DATA

Supplementary data for this paper are available on *Bioinformatics* online.

REFERENCES

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Brennecke,J., Hipfner,D.R., Stark,A., Russell,R.B. and Cohen,S.M. (2003) Bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene hid in *Drosophila*. *Cell*, **113**, 25–36.

Calin,G.A., Sevignani,C., Dumitru,C.D., Hyslop,T., Noch,E., Yendamuri,S., Shimizu,M., Rattan,S., Bullrich,F., Negrini,M. *et al*. (2004) Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers. *Proc. Natl Acad. Sci. USA*, **101**, 2999–3004.

Dostie,J., Mourelatos,Z., Yang,M., Sharma,A. and Dreyfuss,G. (2003) Numerous microRNPs in neuronal cells containing novel microRNAs. *RNA*, **9**, 180–186.

Eddy,S.R. and Durbin,R. (1994) RNA sequence analysis using covariance models. *Nucleic Acids Res.*, **22**, 2079–2088.

Gautheret,D. and Lambert,A. (2001) Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles. *J. Mol. Biol.*, **313**, 1003–1011.

Griffiths-Jones,S. (2004) The microRNA Registry. *Nucleic Acids Res.*, **32**, D109–D111 (database issue).

Griffiths-Jones,S., Bateman,A., Marshall,M., Khanna,A. and Eddy,S.R. (2003) Rfam: an RNA family database. *Nucleic Acids Res.*, **31**, 439–441.

Henikoff,J.G. and Henikoff,S. (1996) Using substitution probabilities to improve position-specific scoring matrices. *Comput. Appl. Biosci.*, **12**, 135–143.

Hofacker,I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res.*, **31**, 3429–3431.

Hofacker,I.L., Fekete,M. and Stadler,P.F. (2002) Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.*, **319**, 1059–1066.

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

Lai,E.C., Tomancak,P., Williams,R.W. and Rubin,G.M. (2003) Computational identification of *Drosophila* microRNA genes. *Genome Biol.*, **4**, R42.

Lambert,A., Fontaine,J.F., Legendre,M., Leclerc,F., Permal,E., Major,F., Putzer,H., Delfour,O., Michot,B. and Gautheret,D. (2004) The ERPIN server: an interface to profile-based RNA motif identification. *Nucleic Acids Res.*, **32**, W160–W165.

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

Lim,L.P., Glasner,M.E., Yekta,S., Burge,C.B. and Bartel,D.P. (2003a) Vertebrate microRNA genes. *Science*, **299**, 1540.

Lim,L.P., Lau,N.C., Weinstein,E.G., Abdelhakim,A., Yekta,S., Rhoades,M.W., Burge,C.B. and Bartel,D.P. (2003b) The microRNAs of *Caenorhabditis elegans*. *Genes Dev.*, **17**, 991–1008.

Reinhart,B.J., Slack,F.J., Basson,M., Pasquinelli,A.E., Bettinger,J.C., Rougvie,A.E., Horvitz,H.R. and Ruvkun,G. (2000) The 21-nucleotide let-7 RNA regulates developmental timing in *Caenorhabditis elegans*. *Nature*, **403**, 901–906.

Sakakibara,Y., Brown,M., Hughey,R., Mian,I.S., Sjolander,K., Underwood,R.C. and Haussler,D. (1994) Stochastic context-free grammars for tRNA modeling. *Nucleic Acids Res.*, **22**, 5112–5120.

Tanzer,A. and Stadler,P.F. (2004) Molecular evolution of a microRNA cluster. *J. Mol. Biol.*, **339**, 327–335.

Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.