Computational Detection of MicroRNAs in Animal Genomes

Matthieu LEGENDRE[1] and Daniel GAUTHERET[2]

INSERM ERM-206, Luminy Case 906, 13288 Marseille Cedex 09, France
(1) legendre@tagc.univ-mrs.fr
(2) gautheret@esil.univ-mrs.fr

Keywords. MicroRNA, Genomics, Gene annotation, non-coding RNA

Introduction

Micro RNAs (miRNA) are an important new class of non-coding RNA first discovered in nematodes [1] that regulate gene expression through partial hybridization to the 3'UTR of target mRNAs. The 21-22nt mature miRNA is produced from a 70-100 nt precursor adopting a stem-loop structure (**Fig. 1**). Precursors are found either as independent genes or in the form of clusters in a 1 to 5 kb genomic region, transcribed as a single unit. Although dozens of miRNA genes have been identified using experimental protocols [2], many others probably go undetected due to a tissue or time specific expression. In this context, there is a strong interest for computational strategies that would allow the identification of miRNAs based on sequence analysis alone.

We distinguish two types of computational strategies for miRNA discovery: (i) identification of miRNA genes homologous to known miRNAs, as for example new instances of lin-4-like or let-7-like miRNAs with the same target specificity, and (ii) identification of new classes of miRNAs with novel targets. Recently, there have been successful attempts to address this second issue, using a combination of phylogenetic footprinting (interspecies conservation of non-coding sequences) and RNA stem-loop prediction [3].

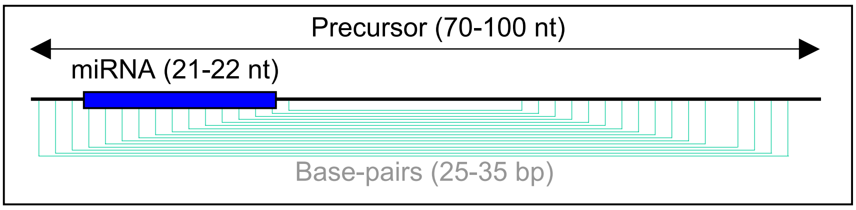

Fig. 1. Schematic representation of miRNA (box) and precursor stem-loop structure. The 21-22nt miRNA is generally embedded into one of the paired regions. Base pairing is imperfect, with multiple mismatches, bulges and internal loops.

Identifying new members of known classes may appear a simpler problem, as these RNAs share a conserved 21-22 nt fragment and complementary region (**Fig. 1**), supposedly an effective signature for standard similarity search programs. Indeed, simple BLAST [4] searches have been able to identify human and mouse orthologues of *C. elegans* miRNAs [5]. But is sequence comparison the best tool for such a task? Micro RNA sequences of the same class differ significantly not only in the precursor stem/loop sequence, but also in the 21-22 nt portion itself, which is not 100% conserved. Combined to the inherently low sensitivity of nucleic acid searches, this means that true miRNA genes can be missed. Profile searches are expected to yield much better results, provided that several related sequences are available. In this work, we use the ERPIN program [6] to conduct a search of let-7-like in animal genomes. ERPIN is a profile-based program that uses both primary and secondary structure information. We show that this program is able to identify several miRNAs that could not be identified using BLAST.

Results and Discussion

The ERPIN software constructs helix and single strand lod-score profiles from RNA alignments and scans sequence database for occurrences of these profiles using a modified dynamic programming algorithm. Recently, a method for Expect-value calculation was incorporated into the software, so that the relevance of predicted RNAs could be evaluated. Computed E-values were confirmed for various RNA motifs using randomized databases (A. Lambert & D. Gautheret, ms. in preparation). We used here a training set of 14 let-7 and let-7-like sequences from *C. elegans*, human and mouse, aligned using CLUSTALW and realigned manually to improve secondary structure superimposition. Instances of the let-7 motif were sought in 18 complete or partial animal genomes available as of May 2003. There were 91 ERPIN hits with an E value<0.01. Among these, eight (**Fig. 2**) could not be detected by the BLAST program at E<0.01, even after reducing word-length to 7 nt (default=11).

Fig. 2 also indicates the E-value threshold that would be required to detect the corresponding sequences with the BLAST program. Highly significant ERPIN hits at $E=10^{-7}$ or 10^{-8} would go undetected by BLAST at an E-value of 4 or higher, at word-length=7 (**Fig. 2**). In other words, BLAST would be able to detect these candidates only after search parameters are considerably relaxed and multiple false positives are tolerated.

Species	S	E-value	Sequence	Blast E
Ciona	45.3	1.2e-07	TGGctg-AGGTAGttGGTtGTATTGTTtccct----------TGTgta--ATGtt-------------AACTATACaGCCcgCTAGCTtaa-CCA	4.1
Ciona	31.7	0.00057	CACaatgAGGTAGtaGGTtATGCAGTTatagcctccctttaTACtga--GTGgagcacaggaga---TACTGTGTaACCttCTGCCTtt--GTG	1.5
Ciona	34.7	0.00015	TTAtg--AGGTAGtaGGTtATGCAGTTttgcgacattattc-TCGttg--CGTtggagat--------AACTGTGTaGCCtaCTGACTc---TAA	1.8
Human	14.0	0.0044	TAGgga-AGGTAGgcGTAgGTAGGGTAggaagaaactttgtgCTGgat--TTGctcag----------CACTGAACaAGCg-CTCTCTg---CCA	>1000
C.elega	45.5	1.8e-08	CGGtg--AGGTAGtaGGTtGTATAGTTtg-------------GAAtatt-ACCaccggtg--------AACTATGCaATTttCTACCTta--CCG	>1000
	27.8	0.00031	GCAtctgAGGTAGtaTGTaATATTGTAgac------------TGTctataATGtc-------------CACAATGTtTCAa-CTAACTcggcTGT	>1000
Drosoph	16.3	0.0014	CAAattgAGGTAGtaGGTtGTATAGTAgt-------------AATtac--ACAtca------------TACTATACaATGtgCTAGCTttctTTG	0.4
Nematod	16.0	0.0047	GTCcg--AGGTAGaaGTAgATGAAGTGct-------------TCGtct--CGTaagtcacgt------AACTGCGTaAGTtaCGGCCGac--GAT	>1000

Fig. 2. ERPIN let-7 hits with E-value < 0.01 not found by BLAST at E<0.01 and word-length=7, using any of the 14 sequences in the training set. The WU-BLAST program [7] was used for all tests. Columns indicate species, ERPIN score, ERPIN E-value, hit sequence, and BLAST E-value at which this sequence would be identified. Regions of the let-7 motif used for ERPIN search are in uppercase. Scanned genome sequences: Bovine, C. briggsae, C. elegans, Chicken, Ciona, Drosophila, Xenopus, Human, Mouse, Pan troglodytes, Nematode subdivision, Parasite subdivision, Rabbit, Sea urchin, Tetraodon, Fugu, Zebrafish. The "Nematode" hit is from *Parastrongyloides trichosuri.*

There are several causes for this interesting behavior of profile searches *vs*. simple sequence similarity. One of them is the presence of large variations in the apical loop separating the two conserved regions, splitting the alignment into two discontinuous, low scoring fragments. In addition, conserved regions are interrupted by several point mutations that may delude BLAST at the stage of lookup table search. Most importantly perhaps, point mutations in the miRNA and elsewhere in double stranded regions generally occur in a concerted fashion that retains base pairing. For instance, G and C at two positions mutates into C and G. Such covariations are captured into the specific type of profile used by the ERPIN program (see [6]), resulting in an improved discrimination of covarying *vs.* non-covarying hits.

To further illustrate the importance of secondary structure constraints, we raised the BLAST E-value cutoff to E=10 in order to capture the three *Ciona* hits in Fig. 2. At this level, BLAST finds 14 other solutions in the *Ciona* genome. We submitted all candidates to the Mfold secondary structure prediction program [8]. While the three ERPIN hits fold into the typical miRNA stem loop structure, 12 out of the 14 BLAST-specific hits are predicted to form other secondary structures, a strong indication that these are not true miRNAs.

We do not expect that ERPIN will perform better than sequence alignment programs in all cases. False negatives may arise for example when RNA structures have new bulge or internal loop patterns not observed in the training set. However, there are absolute benefits of lod-score profiles over sequence comparisons that justify using ERPIN for the exhaustive screening of any given miRNA gene family. This is well illustrated here by the *Ciona intestinalis* hits. The presence of let-7 in *Ciona* was shown by Northern blot analysis [9] but exact gene sequences were not previously reported.

References

[1] Lee RC et al. (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75:843-854.

[2] Lau NC et al. (2001) An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science*. 294:858-62.

[3] Lim LP et al. (2003) The microRNAs of Caenorhabditis elegans. *Genes Dev*. 17:991-1008.

[4] Altschul S. et al. (1997) Gapped Blast and Psi-Blast : a new generation of protein database search programs. *Nucl. Acids. Res*. 25, 3389-3402.

[5] Lagos-Quintana M. et al. (2001): Identification of novel genes coding for small expressed RNAs. *Science*, 294:853-858.

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

[7] Gish, W. (1996-2003) http://blast.wustl.edu.

[8] Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res*. 31: 1-10.

[9] Pasquinelli AE et al. (2000) Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature*. 408:86-9.