Mireval: a web tool for simple microRNA prediction in genome sequences

William Ritchie1,2,*; François-Xavier Théodule1; and Daniel Gautheret3
1INSERM U928, Technologies Avancées pour le Génome et la Clinique, Luminy Case 906, 13288 Marseille, Cedex 09, France; 2Centenary Institute, Gene & Stem Cell Therapy Program, Newtown, 2042 Sydney, Australia; and 3Univ Paris-Sud 11, CNRS UMR 8621, Institut de Génétique et Microbiologie, Bat 400, 91405 Orsay Cedex, France

Received on February 25, 2008; revised and accepted on April 10, 2008
Advance Access publication May 3, 2008
Associate Editor: Dmitrij Frishman

1 INTRODUCTION

One of the reasons why microRNAs (miRNAs) have avoided our scrutiny for so long is because of their small size (20–22 nt) but also because they lack a common signature in their primary sequence. These two characteristics make miRNAs difficult to detect with classical gene finding algorithms. However, miRNAs can be detected by using characteristics such as the secondary structure and free-folding energy of their precursors, conservation of part of the miRNA sequence or similarity with other miRNAs. These characteristics have been exploited widely in the past years to improve the specificity and sensitivity of miRNA finding algorithms (Ritchie et al., 2005). However, scientists who have no prior knowledge of how these algorithms work or no computer programming skills may not be able to benefit from up-to-date miRNA detection capabilities.

We have developed a user-friendly online tool called Mireval that allows researchers with no bioinformatics skills to conduct a thorough analysis of an input sequence for novel miRNAs, using four different criteria and with no user-defined cutoff score. Mireval is a useful addition to existing miRNA resources such as miRBase (Griffith-Jones et al., 2007) that enable similarity searches with known miRNAs.

*To whom correspondence should be addressed.

2 IMPLEMENTATION

Mireval analyses a DNA sequence of up to 10 000 nt based on four criteria that we labeled ‘secondary structure analysis’, ‘conservation analysis’, ‘cluster analysis’ and ‘miRBase BLAST’. Users decide which of these analyses should be carried on and receive a full report that is easy to interpret. Our intent is to provide sufficient information so that users can make an informed choice on the most likely miRNA candidates without imposing an arbitrary scoring system.

2.1 Secondary structure analysis

The Dicer enzyme produces mature miRNAs from precursor hairpin structures. It is this specific hairpin shape that structure-based algorithms search for. Mireval offers two separate structural analysis algorithms that use two different, non-redundant approaches. Input sequences are analyzed with a sliding window of 80 nt at 10 nt steps. Each window is evaluated for stable secondary structures by RNA-fold (Hofacker et al., 1994) and each hairpin shape (a helix of at least 15 nt with no internal hairpin) is then analyzed as follows.

The first algorithm, ‘Triplet-SVM classifier’ (Xue et al., 2005), is based on support vector machines (SVM) and takes into account structural and primary sequence elements to classify candidates. It is able to distinguish pre-miRNAs from other hairpin shapes with 90% accuracy. In our implementation, the triplet-SVM classifier was trained on the human set of miRNA precursors taken from miRBase and a negative set extracted from coding regions of the human genome, as described in Xue et al. (2005).

The second algorithm was developed in our laboratory and is based solely on structural criteria (Ritchie et al., 2007). It is efficient in distinguishing miRNA precursors from hairpins that are formed by other ncRNA molecules. It works equally well on human sequences (Sensitivity = 0.8; Specificity = 0.76) and on sequences from more distantly related species (Sensitivity = 0.79; Specificity = 0.75).

Users may select to run either Triplet-SVM or both algorithms. The triplet-SVM classifier is used as the default algorithm because it is faster.
2.2 Conservation analysis

Many miRNAs are evolutionary conserved, showing a stronger conservation in the stem of the precursor hairpin than in the loop sequence. This pattern in which a stretch of ~60 nt is conserved with a drop in conservation towards the middle of the stretch is a characteristic of conserved miRNAs (Berezikov et al., 2005).

For conservation analysis, mirEval relies on the genomic evolutionary rate profiling (GERP) algorithm (Cooper et al., 2005). GERP identifies conserved regions and estimates evolutionary rates for individual alignment columns. MirEval uses the pre-calculated GERP scores generated by ENSEMBL on a 10-way alignment of amniota vertebrates (Flicek et al., 2007). To access this data mirEval performs a Blast alignment of the input sequence on the NCBI ‘Genomes’ databank (Jenuth et al., 2000) to retrieve the genomic location of the sequence and then downloads the conservation track of this location from ENSEMBL by using the ENSEMBL Perl modules. Currently, the conservation analysis can only be conducted on 10 amniota vertebrates but will extend to other species when these are included in ENSEMBL.

2.3 Cluster analysis

miRNAs are often found in genomic clusters. Using this property, Sewer and colleagues (2005) discovered up to 100 novel miRNAs in human, rat and mouse by searching 20 kb genomic regions flanking known miRNAs.

To report on such clustered regions, mirEval first retrieves the genomic position that corresponds to the middle of the input sequence (as described in Section 2.2) and retrieves all miRNA elements from the ENSEMBL ncRNA track that lie within a 20 kb flanking region of this position. If three or more known miRNAs are found at this position, MirEval alerts the user by a text message on the output screen.

The ncRNA track at ENSEMBL is directly fed by miRBase (Griffiths-Jones et al., 2007) through the DAS server system and is therefore always up to date.

2.4 miRBase BLAST

Many miRNAs belong to a family of homologous genes (Tanzer and Stadler, 2006). Therefore, sequence similarity search against a set of known miRNAs is a necessary step in miRNA identification and has been implemented in MirEval.

MirEval reports on sequence similarities between the input sequence and all miRNAs in miRBase by directly submitting Blast requests to the miRBase server. Sequences larger than 1000 nt are first cut up into stretches of 1000 nt with a 100 nt overlap and submitted with a 1 s pause between each submission to reduce load on the miRBase server.

2.5 Output and computational issues

MirEval’s report is condensed into an easy to read, color-coded output (Fig. 1). This display is essential in evaluating stretches that are likely to be novel miRNAs as overlapping features such as high levels of conservation and a hairpin shaped secondary structure gain more importance when evaluated simultaneously.

MirEval is the first miRNA search tool that allows a thorough multi-criteria analysis of input sequences and delivers an unbiased, clear report. One of our main concerns was to ensure that this tool was easy to maintain, long lasting and impervious to version changes. To this end we favored data retrieval methods through connections to stable databases rather than maintaining data locally. This is the case for the ‘conservation analysis’, ‘miRBase BLAST’ and ‘cluster analysis’. The structural analyzes, however, are performed locally but we chose two algorithms that perform well on a wide array of species and that do not need to be retrained on new datasets.

Because MirEval connects to remote databases, runtimes may depend on the accessibility of these databases as well as input sequence length. Longer sequences will require longer Blast searches and larger segments to be retrieved from the ENSEMBL server, therefore adding to the runtime. Runtime for structural analyses varies with sequence length and the number of hairpin shapes in the sequence. We measured runtimes on sets of 100 human sequences with lengths varying between 100 and 10 000 nt (Supplementary Table 1). Times include client-server transfer, queuing and sequence verification time and are therefore close to what users will experience when using MirEval. Because each analysis uses a different server (two of our own servers for structural analysis and external servers for conservation and clustering), the total runtime...
is approximately equal to the analysis with the longest runtime, i.e. about 3.5 min for 5 kb.

ACKNOWLEDGEMENTS

Funding: This work was supported in part by INCa grant PL0079, ‘microRNA, stem cells and cancer’.

Conflict of Interest: none declared.

REFERENCES


Ritchie, W. et al. (2007) RNA stem-loops: to be or not to be cleaved by RNase III. RNA, 13, 457–462.

