mirtronDB: a mirtron knowledge base

Bruno Henrique Ribeiro Da Fonseca[1], Douglas Silva Domingues[1,2] and Alexandre Rossi Paschoal[1*]

[1]Department of Computer Science (DACOM), Bioinformatics Graduation Program (PPGBIOINFO), Federal University of Technology - Paraná, UTFPR, Cornélio Procópio, PR, Brazil.

[2]Department of Botany, Institute of Biosciences, São Paulo State University, UNESP, Rio Claro - SP, Brazil.

fonseca.adm@hotmail.com, doug@rc.unesp.br, *paschoal@utfpr.edu.br

Abstract. *Mirtrons arise from short introns with atypical cleavage by using the splicing mechanism. In the current literature, there is no repository centralizing and organizing the data available to the public. To fill this gap, we developed mirtronDB, the first knowledge database dedicated to mirtron, and it is available at http://mirtrondb.cp.utfpr.edu.br/. MirtronDB currently contains a total of 1,407 mirtron precursors and 2,426 mirtron mature sequences in 18 species. MirtronDB is a specialized resource that provides free and user- friendly access to knowledge on mirtron data; it is useful to explore mirtrons and their regulations. This pape was based the original publicaion in Bioinformatics IF: 5.481 (https://doi.org/10.1093/bioinformatics/btz153).*

Resumo. *Os mirtrons surgem de introns curtos com clivagem atípica usando o mecanismo de splicing. Até hoje não existe um repositório que organiza e centraliza os dados públicos de mitrons. De modo a preencher essa lacuna, apresenta-se o mirtronDB, o primeiro banco de dados dedicado aos mirtrons (http://mirtrondb.cp.utfpr.edu.br). MirtronDB contém um total de 1.407 precursores e 2.426 sequências maduras de mirtron em 18 espécies. MirtronDB é um recurso especializado que disponibiliza a comunidade científica uma plataforma amigável sobre oconhecimento dos mirtron. Este trabalho é baseado na publicação original na Bioinformatics IF: 5,481 (https://doi.org/10.1093/bioinformatics/btz153).*

1. Introduction

A Some studies in model organisms identified short hairpin introns displaying characteristics similar to miRNAs, which use the splicing mechanism as the first stage of the miRNA biogenesis cleavage [Ruby et al., 2007]. These noncanonical miRNAs, described as small introns, are collectively called "mirtrons" [Okamura et al., 2007]. Mirtron deregulation was identified as a potential source of several human pathologies [Qu and Adelson, 2012], whereas in plants, research suggests a feedback loop for the autoregulation of miRNA biogenesis (Budak and Akpinar, 2015). In the literature, although there are many databases devoted to miRNA, [i.e., Maracaja-Coutinho et al., 2019; Das et al., 2018; Lorenzetti et al., 2016; Szcześniak et al., 2014; Paschoal et al.,

2012; Pang et al., 2005] there is no repository for accessing knowledge on mirtron data. Not even miRBase [Griffiths-Jones et al., 2006], the miRNA state-of-the-art repository [Maracaja-Coutinho et al., 2019; Paschoal et al., 2012], has a specific analysis for mirtrons. Up until Nov. 2017, we identified 22 articles that had available public mirtron data. However, those datasets are dispersed and with neither standardization nor organization.

The organization of the data will facilitate research on mirtron characteristics, roles, and interactions in organisms, among other potential scientific endeavors. In this context and to fill this gap, we provide mirtronDB, a central mirtron knowledge data repository. For that, based on published available literature, we modeled a total of 1,407 mirtron precursors, and 2,426 mature mirtrons from 18 species (chordates, invertebrates and plants). MirtronDB has an online user- friendly interface for the user, who can search, browse, visualize, and download information. All datasets are publicly available in several formats. The user has access to (i) precursor mirtron similarity analysis; (ii) target gene predictions; and (iii) ceRNA predictions in plants. We expect that this resource will help increase the amount of research on mirtrons.

2. Materials and methods

MirtronDB was built using HTML 5, PHP 7.0, CSS 4.0, Bootstrap 3.3, Cytoscape.js and PostgreSQL in four steps: (1) Data collection; (2) Data modeling; (3) Data analysis; and (4) Website interface (Fig. S1). We collected the mirtron data available from June 2007 to November 2017 by searching the term "mirtron OR mirtrons" in the field “title/abstract” in NCBI PubMed (Supp. Table S1 in Fonseca et al., 2019) and in the papers thereby cited. The articles selected were manually analyzed and redundancies were removed. We created a standardized name: "organism name abbreviation + the word ‘mirtron’ + ID, and for mature we add the arm". We modeled a database and automatically imported the data (See Supp. Fig. S2 in Fonseca et al., 2019). The STATUS column in the search pages and details pages provides the mirtron functional information.

2.1. Similarity analysis among organisms

We extracted the genomic information from several sources (See Supp. Table S2 in Fonseca et al., 2019). We performed a BLASTN alignment between all the precursor mirtrons against all other species genomes. We retained results that have above 95% query coverage and identity.

2.2 Mirtrons and miRNAs similarity analysis

The mature mirtrons were aligned to miRNAs from miRBase v22 [Griffiths-Jones et al., 2006] using the CD-HIT-EST-2D [Huang et al., 2010] and by using the alignment of 9 nucleotides (nt) at 0.98 of identity.

2.3 Target gene prediction and ceRNA prediction in plants

We predicted the targets gene for *H. sapiens* and plants. For human, we used TargetScan [Agarwal et al., 2015] with default parameters, and for plants, we used psRNATarget [Dai et al., 2017] with seed region parameter from 2 to 8 nt. For ceRNA, we used TAPIR [Bonnet et al., 2010] with default parameters to predict ceRNA in

plants. All mature mirtrons were compared against all lncRNAs from GreeNC database [Gallart et al., 2015].

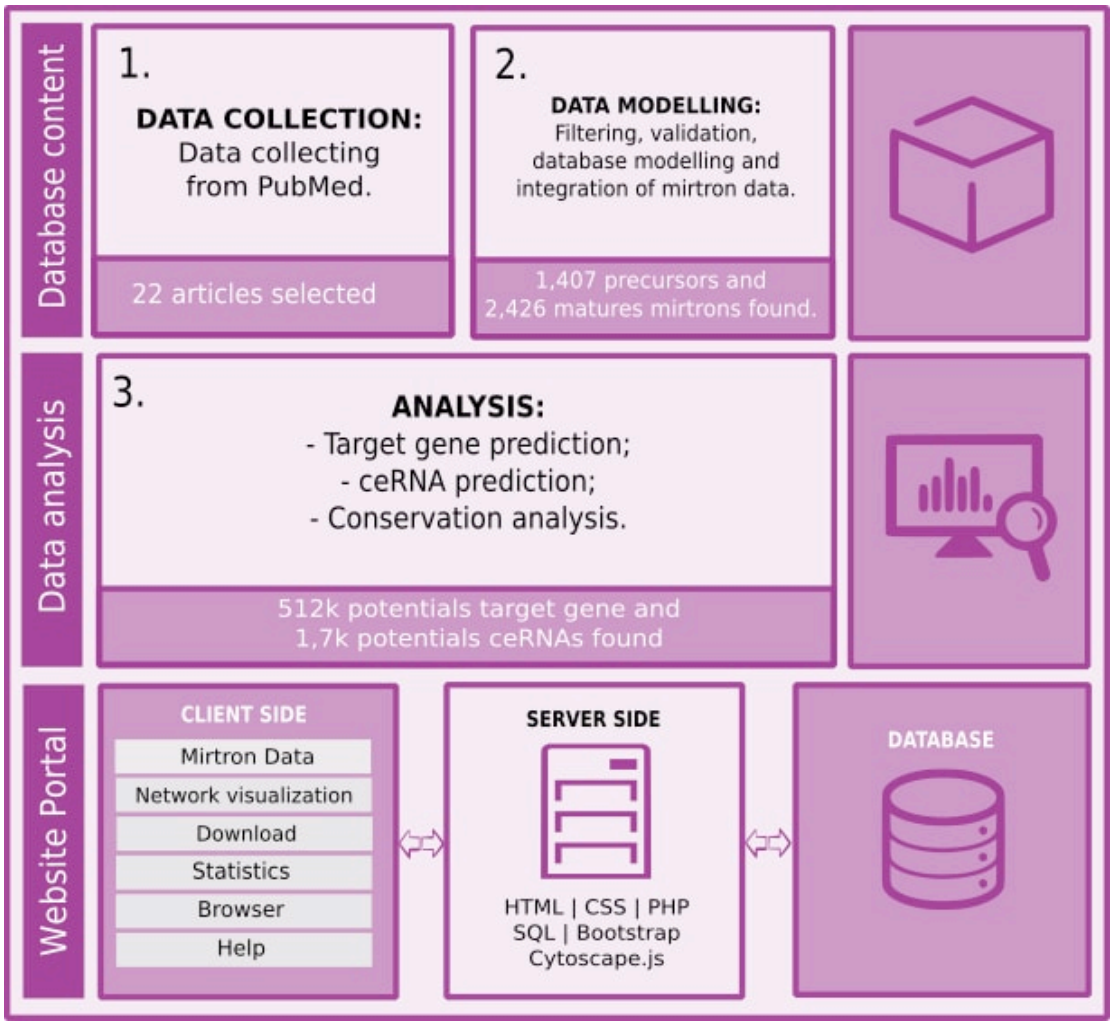

Figure 1. Schematic overview of steps to build mirtronDB.

3. Results

3.1 mirtronDB: database content

We found a total of 1,407 precursor mirtrons and 2,426 mature mirtrons in 18 species, and we extracted functional information, when available. All mirtrons collected are detailed in Supp. Table S3., and Supp. Fig, S3 presents the per year cumulative distribution of the mirtron data [Fonseca *et al*., 2019].

3.2 Precursor mirtron similarity analysis

We obtained 944 aligned precursors, where 896 were aligned in chordates (94.9%), 46 in invertebrates (4.9%) and 2 in plants (0.2%) (See Supp. Table S4 in Fonseca *et al*.,

2019). Four species had more than 3 mirtrons aligned in another genome: *H. sapiens*, *M. mulatta*, *P. troglodytes* and *D. melanogaster*.

3.3 Mature mirtron characterization

In chordates and invertebrates, most mature mirtrons have 22 nt (32.1%), and in plants, 28% of mature have 21 nt (Supp. Table S5 and Fig. S4 in Fonseca *et al*., 2019). We obtained logo sequences for mirtron arms, where chordates present more GC bases than invertebrates and plants (Fig. 2).

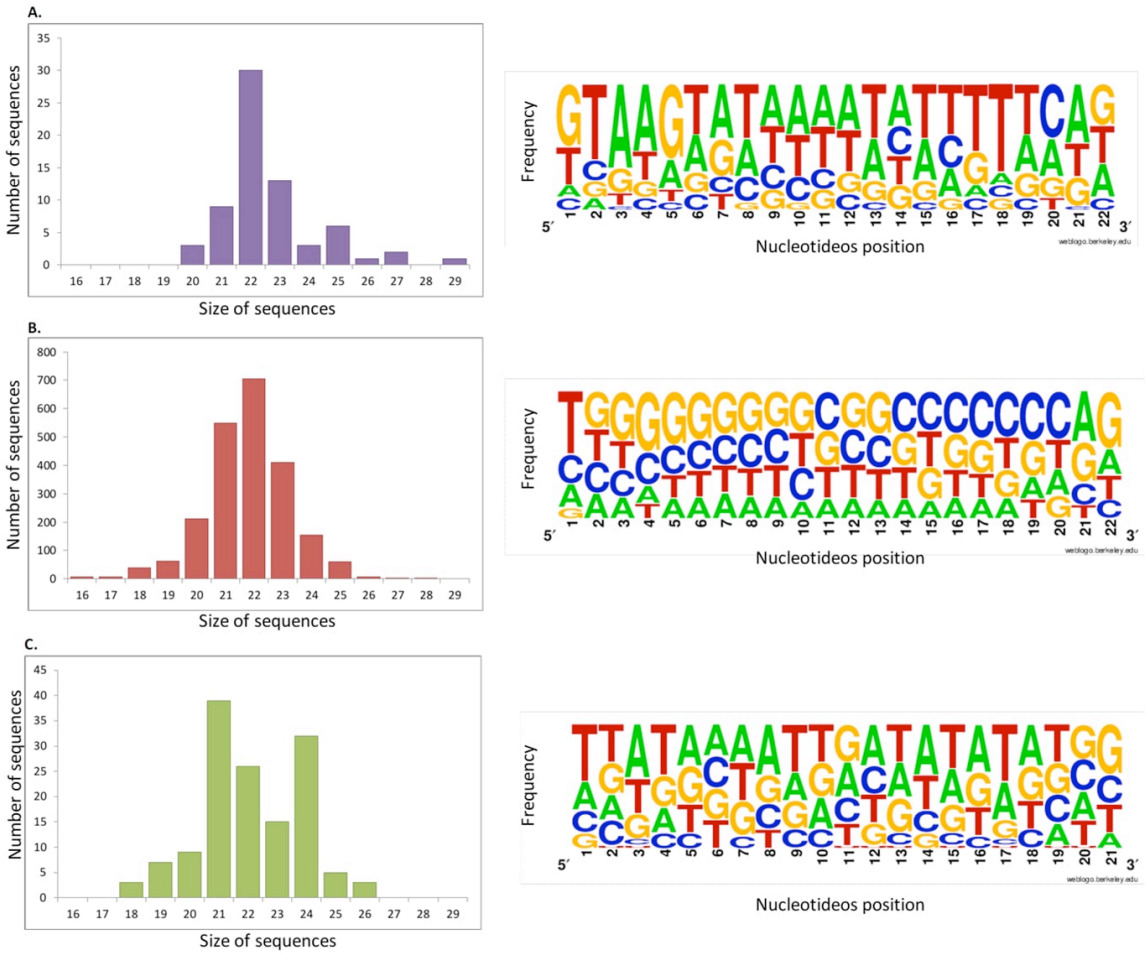

Figure 2. Mature mirtrons bases distribution by nucleotides number and frequency. A. Invertebrates, B. Chordates and C. Plants.

3.4 Mirtrons availability in miRBase

We investigated if the mature mirtron sequences were represented in miRBase. Only 966 mirtrons (39.8%) appear in miRBase, reinforcing the novelty and provided by our mirtronDB (See Supp. Table S6 in Fonseca *et al*., 2019).

3.5 Target gene and ceRNA analysis

We identified a total of 512,298 and 3,884 potential targets, gene predictions, in human and in plants, respectively (See Supp. Table S3 in Fonseca *et al*., 2019). In plants, we also verified if the mirtrons could act as ceRNA candidates, where a total of 1,738 potential interactions were found (See Supp. Table S7 in in Fonseca *et al*., 2019).

3.6 mirtronDB: user interfaces and visualization

The mirtronDB portal provides a user-friendly web interface to access mirtron knowledge. With the "Search" function, the users can query mirtrons by organism, group, type, name, article, and use the JBrowser visualization. In the "Network" page, the users can build a mirtron network, and the results are displayed graphically.

4. Discussion

MirtronDB is a database that standardizes, integrates and provides mirtron data available from literature. We highlight that (i) all data collected is available in several formats; (ii) curated data make this repository a mirtron information reference; (iii) sequence, structure and conservation analysis are provided; and (iv) potential targets and ceRNA in mirtrons are also investigated. Data availability facilitates the development of new studies in biology. For example, we identified four mirtrons associated with diseases (See Supp. Material M1 in in Fonseca *et al*., 2019) in a cross-validation of mirtronDB with miRwayDB, which is a database with information of experimentally validated miRNA-pathway associations in pathophysiological conditions [Das et al., 2018]. Additionally, the mirtronDB data could promote novel standardized approaches to analyze the mirtrons in organisms that are not yet described.

6. Conclusion

MirtronDB is a comprehensive database about mirtrons that allows users to query data and download it. The analyses presented in this paper provide initial mirtron characterization and can be used as a guide about mirtrons' potential. This repository has the potential to promote advances in bioinformatics, such as what has been done by using data exploration and machine learning [Grzegorz et al., 2018]. We will update mirtronDB every year and the users can submit novel mirtrons to our website.

7. References

Agarwal, V. et al. (2015) Predicting effective microRNA target sites in mammalian mRNAs, *Elife*, 4.

Bonnet, E. et al. (2010) TAPIR, a web server for the prediction of plant microRNA targets, including target mimics, *Bioinformatics*, 26, 1566-1568.

Budak, H. and Akpinar, B. (2015) Plant miRNAs: biogenesis, organization and origins, *Functional & Integrative Genomics*, 15, 523-531.

Dai, X. et al. (2018) psRNATarget: a plant small RNA target analysis server (2017 release), *Nucleic Acids Research*.

Das, S. et al. (2018) miRwayDB: a database for experimentally validated microRNA-pathway associations in pathophysiological conditions. *Database* (Oxford). 2018: bay023.

Fonseca B., Domingues D., Paschoal AR. mirtronDB: a mirtron knowledge base. *Bioinformatics*. btz153. doi: 10.1093/bioinformatics/btz153.

Gallart, A. et al. (2015) GREENC: a Wiki-based database of plant lncRNAs. *Nucleic Acids Research*, 44, D1161-D1166.

Griffiths-Jones, S. et al. (2006) miRBase: microRNA sequences, targets and gene nomenclature, *Nucleic Acids Research*, 34, D140-D144.

Grzegorz, R. et al. (2018) Distinguishing mirtrons from canonical miRNAs with data exploration and machine learning methods, *Scientific Reports*, 8(1).

Huang, B. et al. (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences, *Bioinformatics*, 2, 680-682.

Lorenzetti, A. P. R. et al. (2016) PlanTE-MIR DB: a database for transposable element-related microRNAs in plant genomes. *Functional & Integrative Genomics*, 1-8.

Maracaja-Coutinho, V. et al. (2019), Chapter 10: Noncoding RNAs Databases: Current Status and Trend, Book: Computational Biology of Non-Coding RNA: Methods and Protocols, Springer New York, 1th edition.

Okamura, K. et al. (2007) The mirtron pathway generates microRNA-class regulatory RNAs in Drosophila, *Cell*, 89-100.

Paschoal, A. R. et al. (2012) Non-coding transcription characterization and annotation: A guide and web resource for non-coding RNA databases. *RNA Biology*, 274-282.

Qu, Z. and Adelson, D. (2012) Evolutionary conservation and functional roles of ncRNA, *Frontiers in Genetics*, 3, 205.

Ruby, J. G. et al. (2007) Intronic microRNA precursors that bypass Drosha processing, *Nature*, 448, 83.