miRQuest: integration of tools on a Web server for microRNA research

R.R. Aguiar[1], L.A. Ambrosio[1], G. Sepúlveda-Hermosilla[2], V. Maracaja-Coutinho[2,3,4] and A.R. Paschoal[1]

[1]Universidade Tecnológica Federal do Paraná, Cornélio Procópio, PR, Brasil
[2]Centro de Genómica y Bioinformática, Facultad de Ciencias, Universidad Mayor, Santiago, Chile
[3]Beagle Bioniformatics, Santiago, Chile
[4]Instituto Vandique, João Pessoa, PB, Brasil

Corresponding author: A.R. Paschoal
E-mail: paschoal@utfpr.edu.br

Genet. Mol. Res. 15 (1): gmr.15016861
Received May 27, 2015
Accepted November 26, 2015
Published March 24, 2016
DOI http://dx.doi.org/10.4238/gmr.15016861

ABSTRACT. This report describes the miRQuest - a novel middleware available in a Web server that allows the end user to do the miRNA research in a user-friendly way. It is known that there are many prediction tools for microRNA (miRNA) identification that use different programming languages and methods to realize this task. It is difficult to understand each tool and apply it to diverse datasets and organisms available for miRNA analysis. miRQuest can easily be used by biologists and researchers with limited experience with bioinformatics. We built it using the middleware architecture on a Web platform for miRNA research that performs two main functions: i) integration of different miRNA prediction tools for miRNA identification in a user-friendly environment; and ii) comparison of these prediction tools. In both cases, the user provides sequences (in FASTA format) as an input set for the analysis and comparisons. All the tools were selected on the basis of a survey of the literature on the available tools for miRNA prediction. As results, three different cases of use of the tools are also described, where one is the miRNA identification analysis in 30

different species. Finally, miRQuest seems to be a novel and useful tool; and it is freely available for both benchmarking and miRNA identification at http://mirquest.integrativebioinformatics.me/.

Key words: MicroRNA; RNA sequencing; Web server; Middleware; Identification; Bioinformatics

INTRODUCTION

MicroRNA (miRNA) is one of the most studied types of noncoding RNA. They are small noncoding RNA molecules that control gene expression of messenger RNAs via complementary interactions (Lai, 2003). The mature miRNA sequence ranges between 18 and 22 nucleotides (depending on the species); miRNAs are known to be involved in RNA silencing and post-transcriptional regulation of gene expression (Zhang et al., 2007). The interest of the scientific community in miRNA research is evident in the overwhelming number of biological databases available for this particular class of noncoding RNAs (Paschoal et al., 2012). This is especially due to the role of miRNAs as key regulators of important biological processes in different organisms (Oliveira et al., 2011) as well as their use as possible therapeutics and biomarkers (Saunders and Lim, 2009). According to the Non-coding RNA Databases Resource, http://www.ncrnadatabases.org/ (Paschoal et al., 2012), 51% of the indexed noncoding RNA databases (70 of 137, version 2.0) are specific for this particular class.

Nevertheless, one of the major issues in miRNA bioinformatics research is selection of the right tool for *in silico* prediction of novel microRNAs in a set of sequences of interest. In general, most of the software packages have not been designed for biologists with limited programming skills; in addition, these packages are normally developed with the focus on a particular group of organisms, i.e., plants or humans. In order to facilitate *in silico* miRNA analysis and to make it more accessible, we introduce here a novel integrative tool for miRNA analysis. miRQuest is a user-friendly Web server and middleware designed for two main functions: i) a miRNA prediction interface, which integrates four main tools; and ii) a benchmarking function, with which different sets of data can be used to identify which one of the four integrated tools is more suitable for the species of interest.

The Web server does not offer a novel computational method, but rather allows the users to simultaneously execute previously published methods using a sequence of interest and to evaluate and compare the results of these methods in a user-friendly environment. Three different cases of use of the tools are also described here. First, we perform various performance comparisons of the four integrated miRNA prediction tools on 30 species using sequences available in the miRBase database (Kozomara and Griffiths-Jones, 2014) by means of the benchmarking function; then, we use the tools for identification of different miRNAs that could be produced from long noncoding RNAs (lncRNAs) in the *Anopheles* mosquito; and finally, we predict miRNAs in a public dataset of a small-RNA sequencing run from the human neuroblastoma cell line SK-N-SH_RA. miRQuest is freely available for both benchmarking and miRNA identification at http://mirquest.integrativebioinformatics.me/.

MATERIAL AND METHODS

Survey and selection of miRNA predictors

A systematic literature review was conducted on the basis of published studies indexed in

the following databases:, ACM, IEEE, and Google Scholar. The search was focused on research or review articles covering computational approaches or software packages for *in silico* miRNA identification, with the following search terms: "(miRNA or microRNA) and (tool or pipeline or computational or "*in silico*") and (identification or prediction)". In the search, we used "abstract title'' and in the search filter option.

Despite the vast quantity of available tools, many of them are outdated, no longer available, or do not work specifically as a stand-alone version. For these reasons, we used the following criteria for selection of the first four tools integrated in miRQuest: a) predictors with greater usage according to citations in the literature; b) high quality of results (e.g., accuracy) according to a literature review; c) availability, i.e., the tool was available for downloading; and finally, d) at least one predictor should have been designed for analysis of human or for plant RNAs, to avoid a bias to specific groups of organisms. As a result of all these criteria, we selected four main tools for this first version of miRQuest: Triplet-SVM (Xue et al., 2005); MiPred (Jiang et al., 2007); HHMMiR, version 1.2 (Kadri et al., 2009); and NOVOMIR, version 2011 (Teune and Steger, 2010).

miRQuest Web server

miRQuest is a Web application built in a middleware architecture that utilizes the concept of layers to facilitate the flow of information. The application was built on the Java platform, using Tomcat (version 8) as a Web server. The XML format was used as a standard for data extraction and exchange with Shell Scripts for execution of the processes related to RNA-fold (Brameier and Wiuf, 2007) and RNAshapes (Steffen et al., 2006). Both tools are commonly used for secondary-structure predictions; RNA-fold was used here with the microRNAs predicted by Triplet-SVM and HHMMiR, whereas RNAshapes with microRNAs predicted by NOVOMIR. miRQuest was implemented with Contexts and Dependency Injection to integrate the layers; Apache Shiro Framework was used for management of sessions, and Commons Email API for sending an e-mail with the results to the end user. Each predictor was installed in accordance with all implementation guidelines to ensure the correct version, packages, and subtools.

Case Study 1: Datasets for testing the performance of the implemented predictors on 30 species by means of the benchmark function

To test the performance of the selected predictors, we used distinct sets of positive and negative controls. In the positive control set, we used all miRNA precursors available in the miRBase database (version 19) for the following 30 species: animals (*Homo sapiens*, *Mus musculus*, *Macaca mulatta*, *Saguinis labiatus*, *Ovis aries*, *Bos taurus*, *Gallus gallus*, *Taeniopygia guttata*, *Danio rerio*, *Fugu rubripes*, *Xenopus tropicalis*, *Apis melífera*, *Drosophila melanogaster*, *Nasonia vitripennis*, *Caenorhabditis elegans*, and *Ciona intestinalis*), plants (*Physcomitrella patens*, *Chlamydomonas reinhardtii*, *Ectocarpus siliculosus*, *Pinus taeda*, *Glycine max*, *Solanum lycopersicum*, *Vitis vinífera*, *Arabidopsis thaliana*, *Sorghum bicolor*, and *Oryza sativa*), viruses (*HIV-1*, *HHV-4*, and *Rhesus lymphocryptovirus*), and a protist (*Dictyostelium discoideum*). We obtained 9573 miRNA precursor sequences. It should be noted that the training datasets that we used for the development of each prediction tool were removed from the positive dataset to avoid a bias.

As the negative control dataset, we used 8000 negative control sequences from publicly available data (Janssen et al., 2008). This negative control set is composed of four different groups

of sequences. The first one contains 2000 artificial random sequences uniformly distributed in the range from the shortest to the longest sequence, called the “random_uniform” dataset. The second group of negative controls, named “genes_uniform”, is composed of 2000 sequences extracted from protein-coding genes obtained from the National Center for Biotechnology Information (NCBI). These sequences were randomly cut to ensure similar lengths of miRNA precursors. The other two sets of negative controls, each composed of 2000 sequences, were generated by means of a nonuniform-length distribution and were named here as “genes_nonuniform” and “random_ nonuniform”.

Criteria for the comparison of the prediction tools in the 30 species

For performance comparison among the predictors implemented in miRQuest, we analyzed the following metrics: 1) specificity, 2) sensitivity, 3) accuracy, 4) precision, and 5) the F1 score (Powers, 2011). Each criterion equation is described below. For each predictor, the same protocol was carried out: the same datasets were executed as positive and negative controls and the results were analyzed. We then calculated the true-positive, false-positive, true-negative, and false-negative values. Such data are necessary for calculation of the aforementioned metrics. Each tool was run using default configuration settings following manufacturer protocols.

$$\text{Sensitivity} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \quad \text{(Equation 1)}$$

$$\text{Specificity} = \frac{\text{TP}}{(\text{TN}+\text{FP})} \quad \text{(Equation 2)}$$

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \quad \text{(Equation 3)}$$

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{FP} + \text{FN} + \text{TN})} \quad \text{(Equation 4)}$$

$$\text{Recall} = \frac{TP}{(TN + FP)}$$
$$\text{F-Score} = 2 \times \frac{\left(\text{Precision} \times \text{Recall}\right)}{\left(\text{Precision} + \text{Recall}\right)} \quad \text{(Equation 5)}$$

Case Studies 2 and 3: Datasets used for identification of miRNAs in sequences from *Anopheles* and the human neuroblastoma cell line

The use of the miRNA identification function is illustrated here on the other two kinds of data: from the *Anopheles* mosquito and from humans. The first example is application of miRQuest

to identification of miRNAs derived from lncRNAs of the *Anopheles* mosquito (Jenkins et al., 2015). For this purpose, all 2949 lncRNAs identified by RNA deep sequencing (Jenkins et al., 2015) were downloaded. After that, the FASTA files were processed to obtain sequences shorter than 500 nucleotides (nt) - this is a requirement for some of the implemented predictors - and were used as input for the miRQuest Web server.

The second example of this functionality is the use of miRQuest to identify novel miRNAs in a dataset of an Illumina small-RNA sequencing run from the human neuroblastoma cell line SK-N-SH_RA. For this purpose, we visited the UCSC Genome Bioinformatics website (https://genome.ucsc.edu/) and downloaded the track table *"wgEncodeCshlShortRnaSeqSknshraCellShorttotalTapContigs*" from BED files related to the mapping of this small-RNA sequencing library (this dataset is from The Cold Spring Harbor Laboratory, the ENCODE Project, human genome version hg19). The data related to the contigs were filtered, and all sequences smaller than 17 nt were removed. The remaining transcripts were merged in BEDTools (Quinlan, 2014) to eliminate redundancies. For the final contigs shorter than 100 nt, we expanded the coordinates up to 100 nt to obtain structural looping for miRNA predictions. Finally, the sequences were extracted by means of BEDTools to convert these data into the FASTA format for use as input for miRQuest. Note that all steps described here can be performed by any biologist without programming skills in a user-friendly environment for analysis such as Galaxy (Giardine et al., 2005).

RESULTS AND DISCUSSION

Survey: miRNA prediction tools selected for implementation in miRQuest

To select the prediction tools for integration into miRQuest, we conducted a literature survey of miRNA identification tools in the major literature databases on the Internet (see Material and Methods for details). The survey allowed us to identify 28 prediction tools (**Table S1**); four of them were selected and implemented in this first version of miRQuest: Triplet-SVM, HHMMiR, NOVOMIR, and MiPred (Table 1). These tools were selected because they are stand-alone tools, highly cited in reviews or similar articles, and are freely available. In the following sections, we briefly describe each tool currently available in miRQuest.

Table 1. Detailed information on the four selected tools.

Name	Triplet-SVM	MiPred	HHMMiR	NOVOMIR
Organism	Humans	Humans	Humans	Plants
Training set	*Homo sapiens* and Pseudo pre-miRNA	*H. sapiens* and Pseudo pre-miRNA	*H. sapiens*	*A. thaliana*
Test set	*C. elegans*, *C. briggsae*, *D. melanogaster*, *D. pseudoobscura*, *D. rerio*, *G. gallus*, *M. musculus*, *R. norvegicus*	*H. sapiens* and Pseudo-microRNA	*A. thaliana*, *C. elegans*, *D. melanogaster*, *D. rerio*, *G. gallus*, *H. sapiens*, *M. musculus*, *O. sativa*	*A. thaliana pseudohairpin*
Database	Rfam 5.0	miRBase 10.1	miRBase 8.2	miRBase 10 and 14
	miRNA registry 5.0		Rfam 5.0	Rfam 7.0
Input file format	FASTA	FASTA	FASTA	FASTA
Number of output files	1	8	7	4
Feature info	Prediction result	Sequence name	Sequence name	Sequence name
		Sequence	Sequence	Prediction result
		Length	Secondary structure	Score
		Secondary structure	Positive likelihood	Secondary structure
		MFE	Negative likelihood	
		P value	Ratio of likelihoods	
		Prediction result Prediction confidence	Prediction result	

Triplet-SVM

This is a tool that uses support vector machine (SVM) based on the statistical theory for

data classification (Xue et al., 2005). This predictor distinguishes real miRNAs from pseudohairpins based on the miRNA structure characteristics by using an SVM classifier.

MiPred

MiPred (Jiang et al., 2007) utilizes the random-forest prediction model for the miRNAs identification. This approach also considers 32 global and intrinsic hairpin folding attributes based on sequence, structure, statistical thermodynamics, and topology.

HHMMiR

HHMMiR is a predictor that uses a probabilistic model to identify miRNAs (Kadri et al., 2009). To this end, a hidden Markov hierarchical model (HHMM) is implemented in order to identify the characteristics of a hairpin loop (loop, extension, miRNA, and pri-extension). It also includes different algorithms for estimation of the parameters that the method uses to determine the sequence characteristics/distinctive structures in different regions of the precursor miRNA.

NOVOMIR

NOVOMIR is an algorithm for identifying specific miRNA genes in plant genomes (Teune and Steger, 2010). This is a Perl script that uses a series of filters and statistical models to discriminate a pre-miRNA from all other RNAs as well as to locate the miRNA in a supposed pre-miRNA sequence. This tool uses the paired hidden Markov model analogous to that described previously (Nam et al., 2005).

miRQuest: a strategic Web server for prediction of miRNAs

The development of a new computational approach to facilitate miRNA identification and benchmarking of the prediction tools is the main goal of this report. We integrated different miRNA predictors in a unique online platform. It should be noted that each predictive tool uses many languages and prediction techniques. For this reason, the development of the miRQuest Web server allowed us to offer a common user-friendly and standardized method (for use and evaluation of the results of different predictive software packages) that is most suitable for the organism under study (Figure 1). miRQuest can easily be used by biologists and researchers with limited bioinformatic or programming skills. The tool is freely available along with a tutorial at http://mirquest.integrativebioinformatics.me/.

The middleware architecture and workflow are described in detail in Figure 2A and B. The idea behind miRQuest is to create two main applications based on the selected (preexisting) miRNA prediction software: i) a tool for the comparison of miRNA *ab initio* predictors (benchmark option) and ii) a tool for miRNA identification. With the first tool, the user provides as input his or her own datasets (positive and negative controls) to test and compare the prediction tools under different scenarios. We identified a benchmark among these tools to demonstrate their application to 30 species (see next topic). With the second function, the user can use the sequences under study in the FASTA format as input to obtain a tabular report for annotation of the miRNAs. The user can also select one or more tools to perform the prediction and even change the parameters

based on each predictor. The results can be downloaded as either a CSV or XML file. Finally, for optimization of the hardware resources and for parallel processing, the system uses multiple queues and implements the first-in first-out concept (Figure 2B).

Figure 1. miRQuest Web server website where the end user can utilize a user-friendly and standardized method for evaluation of the results of different predictive software packages for microRNA research.

The system also has a function that allows for monitoring of all requests received by a particular user for processing. It defines a “weight” value that is implemented on the basis of the need for processing power for each tool and assigns the user request to its respective queue. The system also deploys the benchmark feature, meaning that the processing of the inputs is made basically for miRNA identification where the commands are executed depending on the predictor(s) selected. After this step, different formulas are used to determine the following performance metrics: specificity, sensitivity, accuracy, precision, and the F1 score. The information derived from

the processing of the data is stored in XML files and includes the group of classes responsible for the information provided.

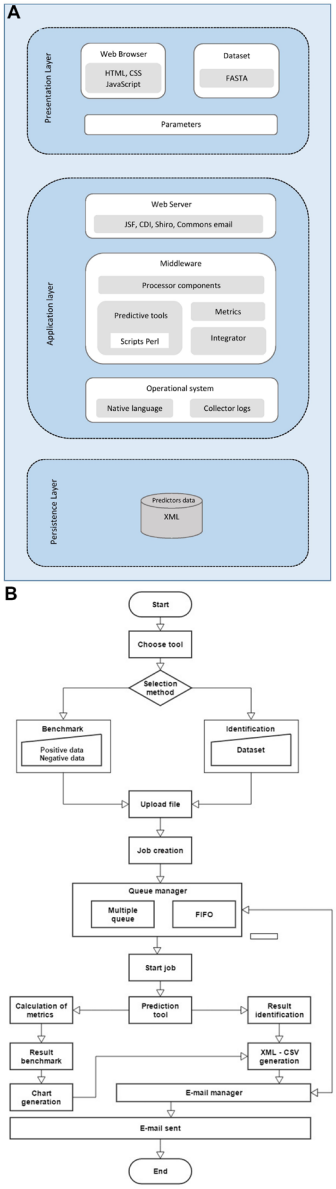

Figure 2. A. Software architecture. The partitions are organized in three main layers: Presentation Layer, Application Layer, and Persistence Layer. The Presentation Layer interacts directly with the user; here we have all the software-user interactions. The Application Layer is the main middleware layer; here all predictive tools were implemented as well all the rules, validation procedures, and various features. The Persistence Layer is responsible for organization and persistence of the results generated in the default XML format. **B.** Flowchart of the processes executed by the middleware. FIFO: first-in first-out.

miRQuest: cases of use

In this section, we introduce three different cases of miRQuest use. First, we present application of the benchmarking functionality to find the best miRNA predictor for use with 30 species. Then, we show two examples of miRNA identification using miRQuest with two kinds of data: i) identification of miRNAs possibly produced from lncRNAs of the *Anopheles* mosquito, and ii) identification of the repertoire of miRNAs in an Illumina small-RNA sequencing run from the human neuroblastoma cell line.

Benchmarking of available miRNA prediction tools on 30 species by means of miRQuest

To illustrate a case of application of the miRQuest benchmarking function, we compared the performance of each miRNA prediction tool on 30 species. We used different groups of positive and negative controls as described in the Material and Methods section. The positive-control dataset is composed of known miRNAs from diverse organisms (animals, plants, a protist, and viruses) and was obtained from the miRBase database (Kozomara and Griffiths-Jones, 2014). The negative control was based on the negative-control set described previously (Janssen et al., 2008). Note that the files containing the sequences used as the positive or negative control can be downloaded by clicking on the “Downloads” tab of our Web server.

The performance comparisons among the different prediction tools were carried out on the basis of the measured sensitivity, specificity, accuracy, precision, and the F1 score. A complete description of all the results is available in **File S1**. The specificity of HHMMiR, MiPred, Triplet-SVM, and NOVOMIR was 95, 98.78, 88.34, and 99.86%, respectively. In general, all the tools showed a balanced and good ability to correctly identify the negative control portion of the dataset. NOVOMiR and MiPred showed superior performance in comparison with HHMMiR and Triplet-SVM.

In terms of sensitivity, the results clearly showed that NOVOMIR can predict more miRNAs in plants, whereas MiPred is better for other organisms. These results are expected, given that NOVOMIR was developed for plant genomes and MiPred for human genomes (Figure 3). Even if we look only at model organisms (*H. sapiens*, *M. musculus*, *D. melanogaster*, *C. elegans*, and *A. thaliana*), the same results are observed (Figure 4). Of course, there are some exceptions. For viruses, Triplet-SVM and MiPred showed the best results. Nevertheless, the number of miRNAs described in the miRBase is too small for viruses (Kozomara and Griffiths-Jones, 2014), and our results may be biased. All the tests were performed following manufacturer protocols.

It is possible to notice that HHMMiR correctly identified the negative-control portion of the test set in specificity but showed bad results on identification of the positive-control set (other metrics). It should be noted that, as shown in Table 1, MiPred and NOVOMIR contain the most updated training datasets among the four tools. This advantage may result in the differences among the prediction tools.

All the tools showed high balanced accuracy (**File S1**). As for precision, NOVOMIR usually showed the best balance among the species with the exception of *A. mellifera*, whereas MiPred took the second place. Finally, it is worth mentioning the low-quality characteristics that we obtained for *C. intestinalis*. This result may be due to the poor annotation, or perhaps the tools are not prepared adequately in terms of data training for this organism.

The results on accuracy seem to be close and good for all the predictors, whereas

NOVOMIR, in general, seems to be the best tool in terms of accuracy in this set of 30 species. The results on the F1 score (**File S1**) were similar to the sensitivity results of NOVOMIR in plants and MiPred in nonplant species, with the exception of *S. labiatus* and viruses.

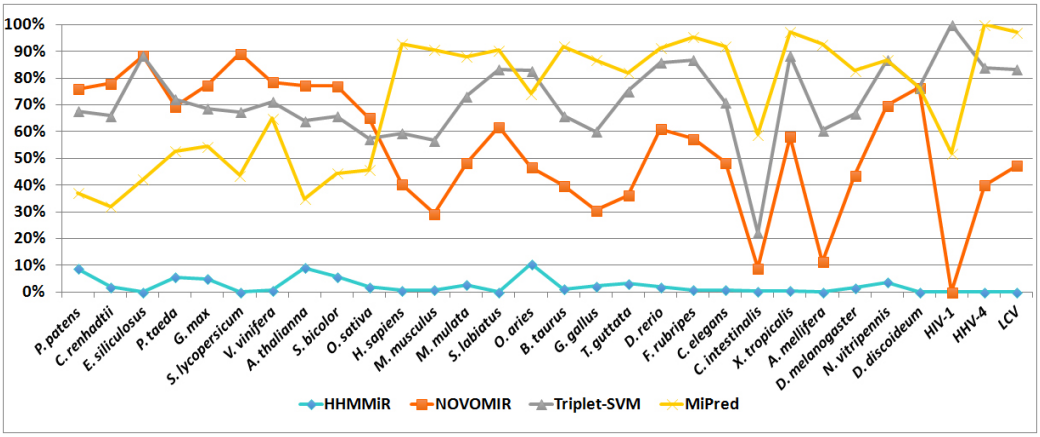

Figure 3. Results of comparison of sensitivity in the 30 species by means of the benchmark functionality.

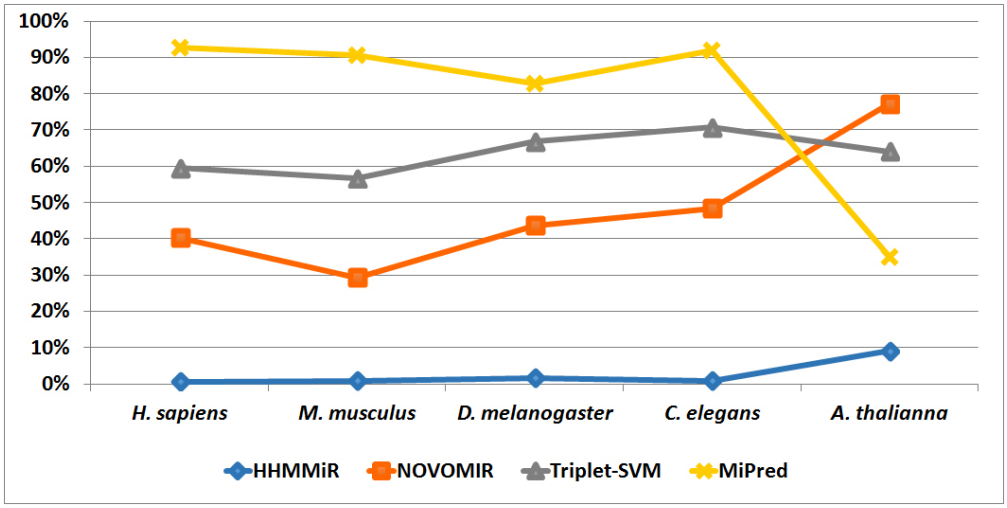

Figure 4. Results of comparison of sensitivity in model organisms by means of the benchmark functionality.

Identification of miRNAs derived from lncRNAs in Anopheles gambiae by means of miRQuest

Anopheles gambiae is the most important vector of malaria. Jenkins et al. (2015) used RNA deep sequencing and identified 2949 lncRNAs in this insect. We used miRQuest to identify the group of miRNAs (Jenkins et al., 2015) that could be generated by the processing of lncRNAs.

We found that at least 23.9% (705 of 2949) of the *Anopheles* lncRNAs could be miRNA precursors. The numbers of predictions per tool were as follows: Triplet-SVM predicted 1370 miRNAs, whereas HHMMiR predicted 679 lncRNAs hosting 887 miRNAs. On the other hand, MiPred and NOVOMIR predicted a much smaller number of miRNAs in the input dataset: only 23 miRNAs were predicted by NOVOMIR, and 18 by MiPred. Due to the lack of information on *Anopheles* miRNAs, it is important to consider this information for selection of the most reliably predicted miRNAs. According to the analysis above (Figures 3 and 4), the best tool for prediction of miRNAs in insects is MiPred. On the other hand, only a small number of insects was used in that analysis, and the miRNAs predicted by other tools can also be functional. Thus, further experimental analysis is necessary to validate this finding.

Identification of miRNAs in a small-RNA sequencing run from the human neuroblastoma cell line

For this purpose, we used miRQuest to predict the set of miRNAs in a small-RNA sequencing library available from the ENCODE Project for the neuroblastoma cell line SK-N-SH_RA (data downloaded directly from the UCSC Genome Bioinformatics website). The data were processed (see Methods), and we obtained a final dataset of 34,986 transcripts, which we used as input for miRQuest. Our results revealed that 12.08% (4227 of 34,986) of the transcripts were possible miRNAs. The results for each tool were as follows: MiPred identified 3512 miRNAs, Triplet-SVM 1564, HHMMiR 696, and NOVOMIR identified 228 miRNAs. Just as in *Anopheles*, according to our benchmark comparative analysis (Figures 3 and 4), the best tool for prediction of human miRNAs is MiPred. In this analysis, the majority of predicted miRNAs were obtained using this tool (3512 potential miRNAs). The benchmark analysis suggests that these MiPred results on miRNAs are more reliable than the prediction results of the other three tools; this finding should be validated experimentally.

CONCLUSIONS

Here, we propose a new Web server and middleware system called miRQuest for miRNA research. This middleware integrates four main miRNA *ab initio* tools within two main functions. The first function is used for various comparative performance analyses among these tools on any datasets of interest. The proposed system produces a comparative report in different formats (CSV and XML). The second function serves for identification of miRNAs proper and is performed in a streamlined way by means of each of the four integrated predictors. The user provides input (the sequences in FASTA format), and the Web server returns a report of the prediction results. The following are main benefits for the end user: i) he or she does not need to worry about installation and management of the tools, ii) the end user can perform a miRNA prediction analysis without programming skills. Furthermore, this study presents a literature survey of the available miRNA *ab initio* tools for selection of the most appropriate ones for use in miRQuest. In addition, we discuss the performance of each of the four selected predictors according to the benchmark functionality of miRQuest for prediction of miRNAs in 30 species from different groups: animals, plants, viruses, and protists. In addition, the proposed Web server was used to identify the set of miRNAs that could be produced from lncRNAs in the *Anopheles* mosquito and to identify miRNAs in a dataset of small-RNA sequencing from the ENCODE project (in a neuroblastoma cell line). miRQuest can be easily accessed and used by the scientific community at http://mirquest.integrativebioinformatics.me.

This is a Web server and it does not offer a novel computational method but rather it allows the users to execute previously published methods on a sequence of interest and to evaluate and compare the results of these methods on their sequences in a user-friendly environment. It should be noted that the integration of a large number of miRNA prediction tools is not a trivial task; for this reason and because of some issues related to availability, we choose to implement only four main tools for miRNA analysis in the current version of miRQuest. We believe that this analysis is an important starting point for miRNA research, and our Web server should make bioinformatics tools and analysis more accessible to nonprogrammers. Finally, this seems to be the first system that integrates tools for miRNA analysis on an open-access Web server.

Conflicts of interest

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

This study summarizes the Master's thesis (of R.R. Aguiar) in Informatics from Programa de Pós-Graduação em Informática: Mestrado Profissional at the Federal University of Technology at Paraná (UTFPR), Cornelio Procópio, PR, Brazil. A.R. Paschoal received funding from CNPq for the Project "Chamada: MCTI/CNPQ/Universal 14/2014 - Faixa A - Process #454505/2014-0" and is enrolled in the Master's Program in Bioinformatics (PPGBIOINFO) at UTFPR. V. Maracaja-Coutinho received funding from Centro de Genómica y Bioinformática, Facultad de Ciencias, Universidad Mayor, and from the Start-Up Chile Program (#2012-13791) from Corporación de Fomento a la Producción for Beagle Bioinformatics, both in Chile.

REFERENCES

Brameier M and Wiuf C (2007). Ab initio identification of human microRNAs based on structure motifs. *BMC Bioinformatics* 8: 478.http://dx.doi.org/10.1186/1471-2105-8-478

Giardine B, Riemer C, Hardison RC, Burhans R, et al. (2005). Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.* 15: 1451-1455.http://dx.doi.org/10.1101/gr.4086505

Janssen S, Reeder J and Giegerich R (2008). Shape based indexing for faster search of RNA family databases. *BMC Bioinformatics* 9: 131.http://dx.doi.org/10.1186/1471-2105-9-131

Jenkins AM, Waterhouse RM and Muskavitch MA (2015). Long non-coding RNA discovery across the genus *anopheles* reveals conserved secondary structures within and beyond the Gambiae complex. *BMC Genomics* 16: 337.http://dx.doi.org/10.1186/s12864-015-1507-3

Jiang P, Wu H, Wang W, Ma W, et al. (2007). MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic Acids Res.* 35: W339-344.http://dx.doi.org/10.1093/nar/gkm368

Kadri S, Hinman V and Benos PV (2009). HHMMiR: efficient de novo prediction of microRNAs using hierarchical hidden Markov models. *BMC Bioinformatics* 10 (Suppl 1): S35.http://dx.doi.org/10.1186/1471-2105-10-S1-S35

Kozomara A and Griffiths-Jones S (2014). miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* 42: D68-D73.http://dx.doi.org/10.1093/nar/gkt1181

Lai EC (2003). microRNAs: runts of the genome assert themselves. *Curr. Biol.* 13: R925-R936.http://dx.doi.org/10.1016/j.cub.2003.11.017

Nam JW, Shin KR, Han J, Lee Y, et al. (2005). Human microRNA prediction through a probabilistic co-learning model of sequence and structure. *Nucleic Acids Res.* 33: 3570-3581.http://dx.doi.org/10.1093/nar/gki668

Oliveira KC, Carvalho ML, Maracaja-Coutinho V, Kitajima JP, et al. (2011). Non-coding RNAs in schistosomes: an unexplored world. *An. Acad. Bras. Cienc.* 83: 673-694.http://dx.doi.org/10.1590/S0001-37652011000200026

Paschoal AR, Maracaja-Coutinho V, Setubal JC, Simões ZL, et al. (2012). Non-coding transcription characterization and

annotation: a guide and web resource for non-coding RNA databases. *RNA Biol.* 9: 274-282.http://dx.doi.org/10.4161/rna.19352

Powers DMW (2011). Evaluation: from precision, recall and f-measure to ROC, Informedness, markedness and Correlation. *J. Mach. Learn. Tech* 2: 37-63.

Quinlan AR (2014). BEDTools: The swiss-army tool for genome feature analysis. *Curr. Protoc. Bioinformatics* 47: 11.12.1-11.12.34.

Saunders MA and Lim LP (2009). (micro)Genomic medicine: microRNAs as therapeutics and biomarkers. *RNA Biol.* 6: 324-328.http://dx.doi.org/10.4161/rna.6.3.8871

Steffen P, Voss B, Rehmsmeier M, Reeder J, et al. (2006). RNAshapes: an integrated RNA analysis package based on abstract shapes. *Bioinformatics* 22: 500-503.http://dx.doi.org/10.1093/bioinformatics/btk010

Teune JH and Steger G (2010). NOVOMIR: De Novo Prediction of MicroRNA-Coding Regions in a Single Plant-Genome. *J. Nucleic Acids* 2010: 495904.http://dx.doi.org/10.4061/2010/495904

Xue C, Li F, He T, Liu GP, et al. (2005). Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinformatics* 6: 310.http://dx.doi.org/10.1186/1471-2105-6-310

Zhang B, Wang Q and Pan X (2007). MicroRNAs and their regulatory roles in animals and plants. *J. Cell. Physiol.* 210: 279-289. http://dx.doi.org/10.1002/jcp.20869

Supplementary material

Table S1. List of all miRNA *ab initio* tool found in the literature revision.

http://www.geneticsmr.com/year2016/vol15-1/pdf/gmr6861_supplementary.pdf

File S1. Performance comparisons among the different prediction tools.

http://www.geneticsmr.com/year2016/vol15-1/pdf/gmr6861_files1.xls