Published online 29 May 2016 *Nucleic Acids Research, 2016, Vol. 44, Web Server issue* ***W181–W184***
doi: 10.1093/nar/gkw459

miRNAFold: a web server for fast miRNA precursor prediction in genomes

Christophe Tav[1], Sébastien Tempel[1], Laurent Poligny[1] and Fariza Tahi[1,2,*]

[1]IBISC—IBGBI, Université Evry, Genopole, 23 Boulevard de France, 91037 Evry CEDEX, France and [2]Institute of Plant Sciences Paris Saclay (IPS2), CNRS, INRA, Université Paris-Sud, Université Evry, Université Paris-Saclay, Batiment 630, 91405 Orsay, France

Received February 12, 2016; Revised May 12, 2016; Accepted May 12, 2016

ABSTRACT

Computational methods are required for prediction of non-coding RNAs (ncRNAs), which are involved in many biological processes, especially at post-transcriptional level. Among these ncRNAs, miRNAs have been largely studied and biologists need efficient and fast tools for their identification. In particular, *ab initio* methods are usually required when predicting novel miRNAs. Here we present a web server dedicated for miRNA precursors identification at a large scale in genomes. It is based on an algorithm called miRNAFold that allows predicting miRNA hairpin structures quickly with high sensitivity. miRNAFold is implemented as a web server with an intuitive and user-friendly interface, as well as a standalone version. The web server is freely available at: http://EvryRNA.ibisc.univ-evry.fr/miRNAFold.

INTRODUCTION

MicroRNAs (miRNAs) play important roles in many biological processes and are known to be involved in many diseases, such as cancers and neurodegenerative diseases (1). Their identification is therefore an important task in biological and medical sciences. Especially, predicting new potential miRNAs is of high interest. With the availability of several sequenced genomes, homology-based methods are widely used for this purpose, yet they cannot detect miRNAs with low homology to the known sequences, e.g. species-specific miRNAs. *Ab initio* prediction methods can overcome this limitation as they are only based on the intrinsic properties of the sequences to predict.

Many tools exist in the literature for *ab initio* prediction of miRNA precursors (pre-miRNAs), which have a specific structure of hairpin. However, very few predictors allow searching for potential pre-miRNAs at a large scale in genomic sequences. Higashi *et al.* proposed very recently a tool called Mirinho (2), and showed its outperformance when compared to other tools like CSHMM (3) and miRPara (4). Mirinho is provided in a standalone version only, and in our knowledge, no web server for identifying possible pre-miRNAs in long genomic sequences exist in the literature.

We present in this paper a web server allowing biologists to identify pre-miRNAs in a fast way (about 15 seconds are needed to analyze a sequence of 1 Mb), and through a user-friendly and intuitive interface. It is based on miRNAFold algorithm, previously developed by our group (5), which allows predicting miRNA hairpin structures in a fast way. In this algorithm, an approximation of pre-miRNA hairpin structure is computed, followed by its refinement, allowing substantial decrease in the theoretical time complexity and practical running time, compared to the state of the art. miRNAFold has indeed a complexity in time of only $O(n^2)$, when all other existing algorithms, except Mirinho (2) (with the same complexity), are of complexities of at least $O(n^3)$. In (5), we have shown the outperformance of miRNAFold on different genomic sequences (of several species), compared to different existing tools in the literature, namely CID-miRNA (6), VMir (7) and miRPara (4). It indeed gave better prediction results in term of sensitivity and precision (selectivity), and was 60 times faster than the fastest tool. Since 2012, a web server has been freely available for miRNAFold exploitation and a standalone version for multi-platform usage has been freely provided upon request to many biologists and bioinformaticians around the world.

In this paper we present a new version of the web server, with a new interface and many new useful functionalities. The web server has been re-implemented in a new infrastructure that allows (i) input sequences of unlimited size (contrary to the limit of 1 Mb in the previous version), (ii) several simultaneous predictions without performance deterioration and (iii) a more simplified usage. The web server provides many new functionalities such as: (i) the input can be a FASTA file as well as a multi-FASTA file; (ii) many (more than 70) species-specific models and several examples of genomic sequences are provided; (iii) the structure

[*]To whom correspondence should be addressed. Tel: +33 1 64 85 35 07; Fax: +33 1 64 85 36 01; Email: fariza.tahi@ibisc.univ-evry.fr
Present address: Sébastien Tempel, LCB, CNRS UMR 7283, Aix Marseille Université, 31 Chemin Joseph Aiguier, 13009 Marseille, France.

of predicted pre-miRNAs are visualized in a user-friendly way thanks to the integration of Forna visualization tool (8); (iv) the results can be sent to the user if requested; and (v) the time running is displayed.

We compared miRNAFold to the very recent published tool Mirinho (2). We performed the benchmark on several sequences, including artificial sequences composed of pre-miRNAs and pseudo pre-miRNAs. miRNAFold gave better compromise between sensitivity and specificity compared to Mirinho. Mirinho returns many more pre-miRNA candidates than miRNAFold, yet the majority are false positives. Moreover, miRNAFold showed an outperformance in running time, which is of a substantial interest when the purpose is to analyze very large sequences. miRNAFold was about five times faster than Mirinho for a sequence of 30 500 pb and almost nine times faster on a sequence of ∼1 Mb.

The paper is organized as follows: we first describe succinctly miRNAFold algorithm (more details on the algorithm can be found in (5)), then the web server, including its input and output, the different parameters considered and some of its functionalities. Then we present some prediction and computing time results obtained on different sequences, before concluding.

ALGORITHM

We developed an original *ab initio* method called miRNAFold for pre-miRNA prediction in genomes (5). Our goal was to design an algorithm that is able to find efficiently pre-miRNA hairpin structures in whole genomes in a reasonable time. For this purpose, we adopted the following approach, which was motivated by different observations we made on pre-miRNAs deposited in miRBase (9).

We consider a sliding window of a given size L that is sufficiently long to contain a pre-miRNA, in which we try to detect pre-miRNA hairpin structures. In the first step, we search for long exact stems that verify a set of constraints. These are then considered as anchors of possible hairpins. In the second step, we extend the selected stems in order to obtain the longest non-exact stems satisfying another set of constraints. Each selected non-exact stem can be considered as a good approximation of a pre-miRNA hairpin, and thus provides the hairpin position. All possible hairpins are then built, considering the middle position of the non-exact stem as the middle position of the hairpin. Hairpins verifying some other defined constraints are then selected as pre-miRNA candidates. The search for hairpins in the window is of time and space complexity of $O(L^2)$. Therefore, the total time complexity of the algorithm is $O(L^2 \cdot n)$, where n is the sequence length.

Several selection criteria (constraints) have been defined for selecting the longest exact stems, then the longest non-exact stems, and finally the hairpins. There are respectively 12, 17 and 26 selection criteria, including sequence size and composition, structure free energy, size of bulges, internal and terminal loops, etc. The list of all considered criteria is provided as Supplementary Data in (5).

WEB SERVER

Input

The input of miRNAFold is a genomic sequence in a FASTA or multi-FASTA format.

There are several parameters that are provided with default values and can be modified by users. We have two types of parameters: (i) those relying on pre-miRNA general characteristics and (ii) those specific to the algorithm.

General parameters. The general parameters correspond to: the sliding window size (i.e. size of the longest pre-miRNA searched for, set by default to 150 bp), the hairpin minimal size and the maximal value of the hairpin free energy.

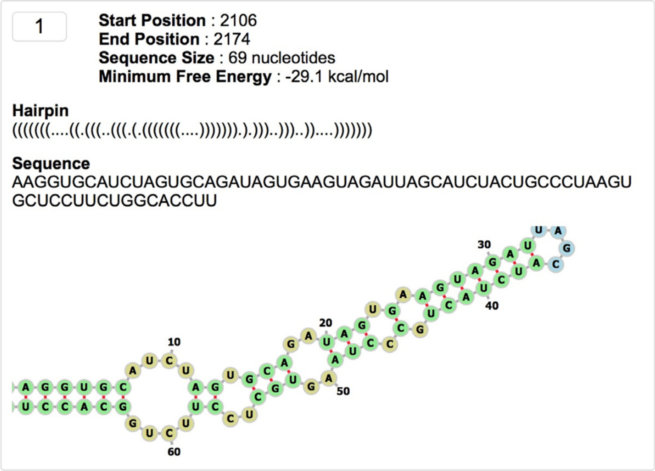
Specific parameters. As described above, several constraints or selection criteria are considered in the different steps of miRNAFold algorithm. We defined two parameters that depend on these constraints: (i) the percentage of validated criteria and (ii) the setting model to consider.

Because a pre-miRNA verifies very rarely all the defined criteria, we set a percentage of constraints that a pre-miRNA candidate must satisfy at each step of the algorithm in order to be selected. This percentage can vary between 0 and 100%. If it is set to zero, all hairpins are selected, but not only those with pre-miRNA characteristics. When it is set to 100%, the constraints become quite restricted and, hence, we did not obtain any candidate in our benchmark. The lower this percentage, the more candidates we can obtain, thus the higher the sensitivity and the lower the precision, and reversely. Obviously, the running time is longer with a lower percentage. In the different tests performed, we observed that the best compromise between the sensitivity and the precision is obtained when setting this percentage to 70% (results shown in (5)). We therefore consider 70% as default value for this parameter.

The criteria considered here have been observed on known pre-miRNAs. The thresholds associated to these criteria are thus extracted from miRBase data. We provide the model corresponding to all miRBase pre-miRNAs (*All genomes* model), which is set by default, as well as other optional species-specific models. In general, the species-specific models yield better results than the *All genomes* model does. On the current version of the web server, the models of 78 species, including human, are available. These models correspond to the species that contain more than 100 pre-miRNAs in miRBase.

Output

The web server gives as output the set of predicted pre-miRNAs (see Figure 1). The sequence, position, size, thermodynamic property and hairpin structure for each pre-miRNA candidate are provided. The structure is represented in a bracket format and is visualized with Forna (8). This tool, available in the ViennaRNA Web Services, has several advantages: (i) it can be integrated in an easy way in a web server, (ii) it allows zooming the figure using the cursor and (iii) it provides well-represented structures with sequence position of each 10 nt and different colors for stems



1

Start Position : 2106
End Position : 2174
Sequence Size : 69 nucleotides
Minimum Free Energy : -29.1 kcal/mol

Hairpin
(((((((....((.(((..(((.(.(((((((....))))))).).)))..)))..))....)))))))

Sequence
AAGGUGCAUCUAGUGCAGAUAGUGAAGUAGAUUAGCAUCUACUGCCCUAAGU
GCUCCUUCUGGCACCUU

Figure 1. An example of a predicted pre-miRNA in the output page of miRNAFold web server.

(in green), internal loops and bulges (in brown) and terminal loops (in blue).

Predicted pre-miRNAs are given in several pages, with five per page. The running time is also provided, as well as the computer specificities.

Additional functionalities

In order to have a useful and user-friendly web server, we provide several additional functionalities (that were not available in the initial version of the web server), such as:

(i) Several examples of genomic sequences are provided in FASTA format.
(ii) A multi-FASTA file (containing several sequences) can also be given at input (miRNAFold will analyze iteratively each of the sequences).
(iii) The results can be sent by email to the user if requested.
(iv) The results (with or without the secondary structure plot) can be downloaded.
(v) The execution time is given.

RESULTS

In (5), we showed results obtained by miRNAFold on an artificial sequence and on real sequences of different species, as well as comparisons with several existing tools, namely: CID-miRNA (6), VMir (7) and miRPara (4). The different tests have shown the outperformance of miRNAFold in prediction results, as well as in computing time. Note that the constraint values, i.e. the setting model files, were built from miRBase 17.

Here we present prediction results obtained on new genomic sequences, considering the latest version of miRBase (V21). We compare these results to Mirinho (2), the most recent published tool for pre-miRNA prediction in genomic sequences. Mirinho has the same principle as miRNAFold since it also searches first for the principal stem that is called stem-arm. It uses a sliding window of length 25 nt in order to find putative stem-arms. From the defined stem-arms, the Needleman-Wunsh alignment algorithm using the thermodynamic nearest-neighbor model is applied to search for stable structure pre-miRNAs. Mirinho is available in a standalone version.

Datasets

We compared miRNAFold with Mirinho on several sequences, including three artificial sequences:

(i) The artificial sequence considered in (5) and composed of 101 human pre-miRNAs and 100 human pseudo pre-miRNAs (Artificial1).
(ii) A sequence composed of 863 human pre-miRNAs and 7422 human pseudo pre-miRNAs (dataset from (10)) (Artificial2).
(iii) A sequence composed of 1677 cross-species pre-miRNAs and 8266 cross-species pseudo pre-miRNAs (dataset from (10)) (Artificial3).

For the two first sequences, which are composed of human data, we consider both *All genomes* model and *Homo-Sapiens* model (see above). For the third sequence, composed of cross species sequences, we consider only the *All genomes* model.

Prediction results

Figure 2 displays sensitivity and specificity values obtained by miRNAFold and Mirinho on the three artificial sequences. This outcome suggests that miRNAFold offers a much more favorable compromise between sensitivity and specificity as compared to Mirinho for most users. More results are provided in Supplementary Data: Supplementary Table S1 shows different measures on the results obtained on the artificial sequences and Supplementary Table S2 shows results obtained on real genomic sequences of different species. As shown in Supplementary Table S2, Mirinho identifies real precursors within genomic sequences at the cost of vastly more predicted precursors than miRNAFold, many of which are likely to be false positives.

We recently proposed a tool called miRBoost (10), based on machine learning approach, for classifying pre-miRNA candidates into true and pseudo pre-miRNAs. This tool, available on our EvryRNA platform (http://EvryRNA.ibisc.univ-evry.fr), can therefore be used to reduce the number of wrongly predicted pre-miRNAs. Another tool we developed, called ncRNAclassifier (11), also available on EvryRNA, can help to reduce the number of false positives by identifying the ones that correspond to transposable elements. The use of these provided tools can therefore allow biologists to obtain a significant set of new pre-miRNAs that thus could be validated (or invalidated) with experimental techniques.

Time computing

Table 1 presents the running time of miRNAFold and Mirinho on the artificial sequences described above, using an Intel Core i7 3720QN, 8 processors 2.6 Ghz and 8Gb of memory. As we can see, miRNAFold is much faster than

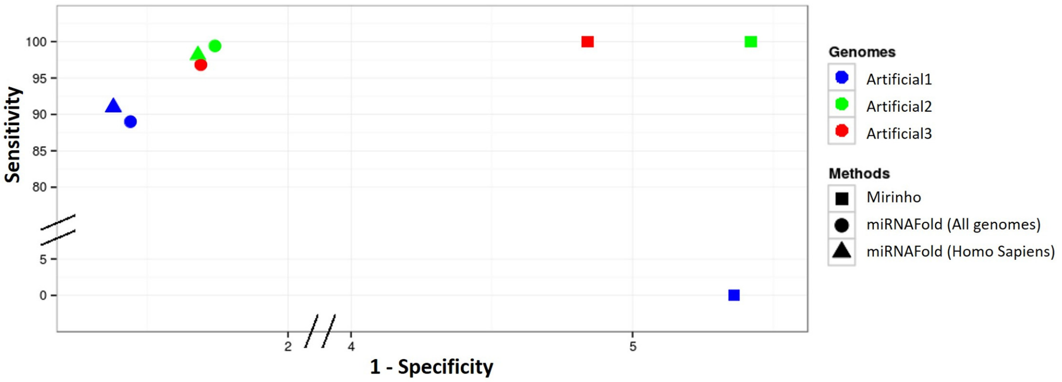


Figure 2. Plot of sensitivity versus specificity of miRNAFold and Mirinho obtained on three artificial sequences. miRNAFold was run with *All genomes* model for the three sequences and with *Homo-Sapiens* model for Artificial1 and Artificial2 sequences that are composed of human sequences. Sensitivity = TP/(TP + FN) and Specificity = TN/(TN + FP), where TP: True Positive; TN: True Negative; FP: False Positive and FN: False Negative.

Table 1. Comparison of miRNAFold and Mirinho running time

	Sequence size	Mirinho time	miRNAFold time
Artificial1	30 500 pb	2.023 s	0.417 s
Artificial2	782 579 bp	101.612 s	11.802 s
Artificial3	971 358 bp	126.381 s	14.680 s

Mirinho. Besides, the longer the sequence, the higher the difference between the two running times, which shows the outperformance of miRNAFold for sequence analysis at a large scale.

CONCLUSION

In this paper, we have presented a web server called miRNAFold for pre-miRNA identification, which is very useful for biologists and bioinformaticians working in (or interested by) the field of ncRNAs. In our knowledge, no equivalent web servers, in rapidity, provided information, functionalities and prediction results, as well as in ergonomic aspects, exist elsewhere.

In order to reduce the number of false positives, we aim to integrate in an optimized manner miRNAFold with miRBoost (10) (described above) in order to obtain a new version of miRNAFold that will give better prediction results in an equivalent computing time. We will also integrate on the web server the possibility to check, thanks to ncRNAclassifier tool (11), whether a predicted pre-miRNA is derived from a transposable element or is a false positive candidate because being a transposable element.

Another ongoing work is the development of miRNAFold version for HPC and GPU, in order to gain more in computation.

Finally, miRNAFold web server is a part of a platform called EvryRNA developed by our team, and dedicated for RNA analysis, identification and prediction, including structure prediction. EvryRNA platform is part of Genopole biocluster. (EvryRNA server is accessible with any standard web browser (Mozilla Firefox, Google Chrome, Apple Safari and Internet Explorer).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

From UEVE and Council of Essonne Region.

Conflict of interest statement. None declared.

REFERENCES

1. Esteller,M. (2011) Non-coding RNAs in human disease. *Nat. Rev. Genet.*, **12**, 861–874.
2. Higashi,S., Fournier,C., Gautier,C., Gaspin,C. and Sagot,M. (2015) Mirinho: An efficient and general plant and animal pre-miRNA predictor for genomic and deep sequencing data. *BMC Bioinformatics*, **16**, 179.
3. Agarwal,S., Vaz,C., Bhattacharya,A. and Srinivasan,A. (2010) Prediction of novel precursor miRNAs using a context-sensitive hidden Markov model (CSHMM). *BMC Bioinformatics*, **11**(Suppl. 1), S29.
4. Wu,Y., Wei,B., Liu,H., Li,T. and Rayner,S. (2011) MiRPara: a SVM-based software tool for prediction of most probable microRNA coding regions in genome scale sequences. *BMC Bioinformatics*, **12**, 107.
5. Tempel,S. and Tahi,F. (2012) A fast ab-initio method for predicting miRNA precursors in genomes. *Nucleic Acids Res.*, **40**, e80.
6. Tyagi,S., Vaz,C., Gupta,V., Bhatia,R., Maheshwari,S., Srinivasan,A. and Bhattacharya,A. (2008) CID-miRNA: a web server for prediction of novel miRNA precursors in human genome. *Biochem. Biophys. Res. Commun.*, **372**, 831–834.
7. Grundhoff,A., Sullivan,C. and Ganem,D. (2006) A combined computational and microarray-based approach identifies novel microRNAs encoded by human gamma-herpesviruses. *RNA*, **12**, 733–750.
8. Kerpedjiev,P., Hammer,S. and Hofacker,I. (2015) Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams. *Bioinformatics*, **31**, 3377–3379.
9. Kozomara,A. and Griffiths-Jones,S. (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.*, **39**(Suppl. 1), D152–D157.
10. Tran,V. D. T., Tempel,S. and Zerath,B. (2015) miRBoost: boosting support vector machines for microRNA precursor classification. *RNA*, **21**, 775–785.
11. Tempel,S., Pollet,N. and Tahi,F. (2012) ncRNAclassifier: a tool for the detection of transposable element sequences in RNA hairpins and their classification. *BMC Bioinformatics*, **13**, 246.