RESEARCH ARTICLE
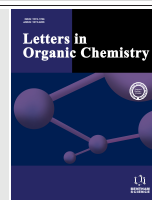
FledFold: A Novel Software for RNA Secondary Structure Prediction

Qi Zhao[1,2], Yuanning Liu[1], Yunna Duan[1], Tao Dai[1], Rui Xu[1], Hao Guo[2], Daiming Fan[2], Yongzhan Nie[2] and Hao Zhang[1,*]

[1]*Computer Science and Technology Department, Jilin University, Changchun, 130000, China;* [2]*State Key Laboratory of Cancer Biology, Fourth Military Medical University, Xi'an, 710000, China*

Abstract: ***Background***: RNA secondary structure is essential to understand the mechanism of RNAs.

Method: In this paper, fledFold, a novel software for RNA secondary structure prediction, is introduced. It combines both thermodynamic and kinetic factors of RNA secondary structures and can predict RNA secondary structures from their primary sequences with local personal computers.

Results: FledFold is implemented in C++ under Windows 7 and could run on windows 7 or later version with at least 2 GB of RAM. Fledfold is user friendly and could output results with multiple formats.

Conslusion: FledFold will be a valuable tool for RNA researches and it could be downloaded freely from http://www.jlucomputer.com/fledfold.php

1. INTRODUCTION

Recently, researches have discovered a large number of non-coding RNAs (ncRNA) [1, 2] which serve many different roles [3], such as modulating gene expression [4], catalyzing reactions [5], immunity [6] and development [7]. It has been well known that functions of ncRNAs are deeply related to their secondary structures (Fig. **1**) rather than their primary sequences. Therefore, the insight of RNA secondary structures has received increasing attention.

The concept of RNA secondary structure began with the work of Doty [9]. Generally, the RNA secondary structure could be defined as a set of canonical base pairs, including AU, GC and GU. Since it is often difficult to obtain X-ray diffraction [10] or nuclear magnetic resonance (NMR) spectroscopy data for RNA molecules to inspect their structures [11], predicting RNA structures from their primary sequences precisely is highly desirable. Mfold [12, 13] is the first practical programming algorithm which could predict the optimal secondary structure from a single RNA sequence. But the accuracy of mfold remains to be improved, especially when predicting long RNA sequences, such as full-length small subunit ribosomal RNA (rRNA) and large subunit rRNA [14]. Comparative analysis [15] is the most accurate method when a large number of homologous

sequences are available. However, this method needs both significant user inputs and a large number of well aligned homologous sequences.
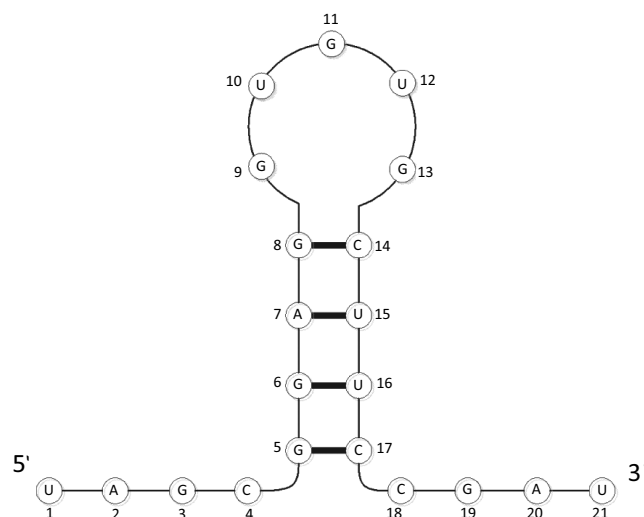
Fig. (1). RNA secondary structure diagram. Each numbered circle represents a base (A, C, G and U). The length of this RNA is 21. The thick lines connecting paired bases represent hydrogen bonds. (5, 8), (14,17) represent two regions which can be paired reversely, so that (5,17,4) is a helical region and the interval is 5 [8].

*Address correspondence to this author at the Computer Science and Technology Department, Jilin University, Changchun, 130000, China; Tel: 18604317609; Fax: 18604317609; E-mail: zhangh@jlu.edu.cn

In this contribution, an alternative software, fledFold, is described. FledFold combines both thermodynamic and kinetic factors of RNA secondary structure, and could predict RNA secondary structures from primary sequences. Our prior work [8] has shown that the accuracy of fledFold is higher than that of traditional methods, especially for RNAs without pseudoknots. Hence, it would be helpful to provide available software package for users.

The details of using fledFold for predicting RNA secondary structure are introduced in this paper. We believe that fledFold will be a valuable tool for RNA researches. Fledfold could be downloaded at http://www.jlucomputer.com/fledfold.php.

2. METHODS

The technical details of the fledFold can be found in our original publication [8], and here, we only highlight the pipeline of fledFold. FledFold combines both thermodynamics and kinetics, and was designed under the assumption that the RNA folding process from random coil state to full structure state is staged. In each folding stage, the final state of an RNA is determined by the optimal combination of helical regions which are most urgent to form under the current RNA state. FledFold utilizes the nearest neighbor (NN) model [16] to calculate the free energy of an RNA secondary structure, which assumes the free energy of an RNA secondary structure is the sum of the energy of its loops and helical regions. The thermodynamic parameters used in NN model is Turner 1999 [16]. FledFold predicts only the most likely secondary structure from a single RNA sequence, which makes it easy for the non-expert users. FledFold works in batch process pattern and the process pipeline for each RNA sequence is shown in Fig. (**2**).

3. RESULTS

The usage of fledFold is very simple, as shown in Fig. (**3**). There is no need to configure any parameters for fledFold or upload files to any websites, and the only input of fledFold is one or a group of FASTA files (need to be put under the path 'fledFold/sequences'). At present, only one RNA sequence is allowed in each FASTA file which is presented in characters 'A'-'Z' or 'a'-'z'. If a DNA sequence is input, it will be converted into the corresponding RNA sequence automatically. In addition, all the rare bases will be converted into the corresponding bases. FledFold will report errors if illegal characters are detected in the input FASTA files. The executable file 'fledFold.exe' under the path 'fledFold/' could be run either by simply double clicking or in command line. FledFold could process 1000 FASTA files at most at one time. When the processing is completed, the output files which describe RNA secondary structure in multiple formats will be generated and saved automatically under the path 'fledFled/sequences/'.

The results of fledFold are presented in dot-parenthesis format, Connectivity Table (CT) format and Scalable Vector Graphics (SVG) format with the same name as the corresponding input FASTA file (different suffixes). Dot-parenthesis files and CT files can be used to draw RNA structure figures conveniently and SVG files could give the highest print quality no matter how they are enlarged or shrunk, which are convenient for observing the details of predicted structures. Additional plug-in software is required to open SVG files in browser, such as Adobe SVG Viewer and Corel SVG Viewer.

FledFold is implemented in C++ under Windows 7 and could run on windows 7 or later version with at least 2 GB of RAM. A help document is provided in the package of fledFold to guide the users, which includes the introduction about the input and output of fledFold and some usage details.

Our prior work [8] suggested that the performance of fledFold is better than other algorithms especially for the RNA sequence without pseudoknots. FledFold takes only several seconds to predict the secondary structures of sequences with length shorter than 400 nt using our computers (Processer: i5, RAM: 4G, OS: Windows 7).
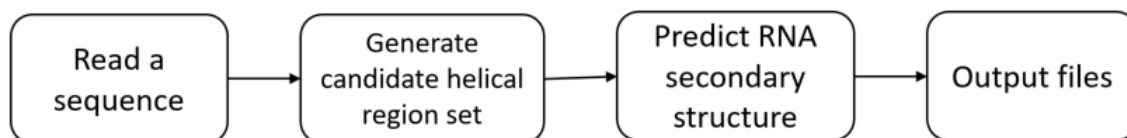

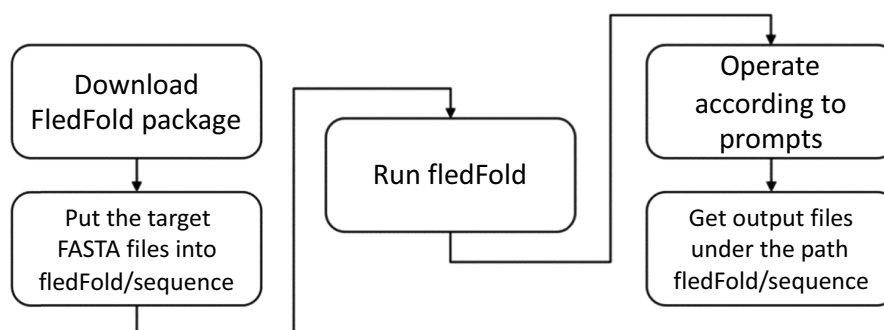**Fig. (2).** The overview of the processing procedure of fledFold.

Fig. (3). The procedure of using fledFold.

CONCLUSION

FledFold is convenient for users to predict RNA secondary structures from their primary sequences. Now, the version number of fledFold is 1.0, and several work are underway to improve it.

At present, fledFold can only run on a single machine, but in fact, many processes of fledFold can be executed in parallel. Therefore, the speed of fledFold will be improved significantly if distributed computation could be used. Providing distributed running capacity to fledFold is one of our future work. Only Windows version of fledFold is available at present and the development of Linux version, Mac version is ongoing. Once these versions of fledFold are completed, they will be uploaded to the same address as Windows version. In addition, fledFold cannot utilize prior knowledge, for example, the data from enzymatic cleavage [17], chemical mapping [18] or SHAPE [19]. These experiment data could significantly improve the accuracy of the prediction, hence, we plan to make use of prior knowledge to improve prediction results later version of fledFold.

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

AUTHOR CONTRIBUTIONS

Study conceived and designed: Y.L. and Q.Z. Performed the experiments: Q.Z., T.D. and R.X. Code generated: Q.Z. Analyzed the data: Q. Z. Wrote the manuscript: Q.Z., Y.L., H.Z., Y.D., H.G., D.F. and Y.N. contributed in revision.

REFERENCES

[1]   Mercer, T.R.; Dinger, M.E.; Mattick, J.S. Long non-coding RNAs: Insights into functions. *Nat. Rev. Genetics,* **2009***, 10*(3), 155-159.

[2]   Mattick, J.S.; Makunin, I.V. Non-coding RNA. *Human Mol. Genetics,* **2006***, 15*, R17-29.

[3]   Mattick, J.S. The functional genomics of noncoding RNA. *Science,* **2005***, 309*(5740), 1527-1528.

[4]   Wu, L.; Belasco, J.G. Let me count the ways: Mechanisms of gene regulation by miRNAs and siRNAs. *Mol. Cell,* **2008***, 29*, 1-7.

[5]   Rodnina, M.V.; Beringer, M.; Wintermeyer, W. How ribosomes make peptide bonds. *Trends Biochem. Sci.,* **2007***, 32*, 20-26.

[6]   Meister, G.; Tuschl, T. Mechanisms of gene silencing by double-stranded RNA. *Nature,* **2004***, 431*(7006), 343-349.

[7]   Mattick, J.S. Probing the phenomics of noncoding RNA. *eLife,* **2013***, 2*, e01968.

[8]   Liu, Y.; Zhao, Q.; Zhang, H.; Xu, R.; Li, Y.; Wei, L. A new method to predict RNA secondary structure based on RNA folding simulation. *IEEE/ACM Trans. Comput. Biol. Bioinform.,* **2016***, 13*(5), 990-995.

[9]   Pavesi, G.; Mauri, G.; Stefani, M.; Pesole, G. RNAProfile: An algorithm for finding conserved secondary structure motifs in unaligned RNA sequences. *Nucleic Acids Res.,* **2004***, 32*(10), 3258-3269.

[10]   Spencer, M. *X-ray diffraction studies of the secondary structure of RNA*, Cold Spring Harbor Laboratory Press: New York, **1963**.

[11]   Fürtig, B.; Richter, C.; Wöhnert, J.; Schwalbe, H. NMR spectroscopy of RNA. *Chembiochem,* **2003***, 4*(10), 936-962.

[12]   Zuker, M.; Stiegler, P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.,* **1981***, 9*, 133-148.

[13]   Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.,* **2003***, 31*(13), 3406-3415.

[14]   Bellaousov, S.; Mathews, D.H. ProbKnot: Fast prediction of RNA secondary structure including pseudoknots. *RNA,* **2010***, 16*(10), 1870-1880.

[15]   Pace, N.R.; Thomas, B.C.; Woese, C.R. Probing RNA structure, function, and history by comparative analysis. *Cold Spring Harbor Monograph Series,* **1999***, 37*, 113-142.

[16]   Turner, D.H.; Mathews, D.H. NNDB: The nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Res.,* **2010***, 38* (Database issue), D280-282.

[17]   Mathews, D.H.; Sabina, J.; Zuker, M.; Turner, D.H. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.,* **1999***, 288*(5), 911-940.

[18]   Mathews, D.H.; Disney, M.D.; Childs, J.L.; Schroeder, S.J.; Zuker, M.; Turner, D.H. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Nat. Acad. Sci. USA,* **2004***, 101*(19), 7287-7292.

[19]   Deigan, K.E.; Li, T.W.; Mathews, D.H.; Weeks, K.M. Accurate SHAPE-directed RNA structure determination. *Proc. Nat. Acad. Sci. USA,* **2009***, 106*, 97-102.