W150–W153 *Nucleic Acids Research, 2004, Vol. 32, Web Server issue*
DOI: 10.1093/nar/gkh445

RDfolder: a web server for prediction of RNA secondary structure

Xiaomin Ying, Hong Luo[1], Jingchu Luo[1] and Wuju Li*

Beijing Institute of Basic Medical Sciences, PO Box 130(3), Beijing 100850, People's Republic of China and [1]Center of Bioinformatics, Peking University, Beijing 100871, People's Republic of China

Received February 15, 2004; Revised and Accepted April 20, 2004

ABSTRACT

Prediction of RNA secondary structure is important in the functional analysis of RNA molecules. The RDfolder web server described in this paper provides two methods for prediction of RNA secondary structure: random stacking of helical regions and helical regions distribution. The random stacking method predicts secondary structure by Monte Carlo simulations. The method of helical regions distribution predicts secondary structure based on the helices that appear most frequently in the set of structures, which are generated by the random stacking method. The RDfolder web server can be accessed at http://rna.cbi.pku.edu.cn.

INTRODUCTION

RNA molecules (tRNA, snRNA, etc.) participate in many essential and vital biological processes. Knowing their structure is very important in understanding their function. Owing to the difficulty of obtaining the real structures of RNA molecules *in vivo*, RNA secondary structure prediction methods are often used.

In the past thirty years, many prediction methods have been presented. Zuker's (1) minimum free energy method and its improved versions (2,3) are probably the most widely used. Several web servers for the prediction of RNA secondary structure have been developed by implementation of these algorithms. For example, the mfold web server (4) was developed to provide a universally available service for the prediction of RNA and DNA secondary structure using the minimum free energy approach. Pfold (5) was developed using stochastic context-free grammars, which can be used to find conservative secondary structures for a set of related RNA sequences. The Vienna RNA secondary structure server (6) provides three functions, namely, predicting RNA secondary structure for a single sequence, predicting consensus secondary structure for a set of aligned RNA sequences, and designing sequences which fold into a predefined structure. Tools for the analysis of RNA secondary structure are also provided on the Bielefeld bioinformatics server (7). In this paper, we describe a new web server for the prediction of RNA secondary structure based on the prediction methods described below.

Li and Wu (8) proposed a prediction method based on random stacking (RS) of helical regions. The structures from the RS method are calculated using Monte Carlo simulations. This method can find many possible secondary structures and calculate the corresponding occurrence frequencies. The structures with high occurrence frequencies are considered as the dominant structures. The validation of the mathematical model of high-level expression of foreign genes in the pBV220 vector (9) based on the RS method demonstrated that the RS method is efficient (10–12). However, it is our experience that, when the sequence is relatively long (>150 bases), the RS method cannot easily find the dominant structure. This weakness was overcome by the prediction method based on helical regions distribution (HD) (13). The HD method was developed to improve further the RS method, and it outperforms the RS method in that it can obtain stable secondary structure for longer sequences. The RDfolder web server provides both the RS and the HD methods via a web interface, to offer users more choices to predict RNA secondary structure. The RDfolder web server can be accessed at http://rna.cbi.pku.edu.cn.

DESCRIPTION AND APPLICATION

The RDfolder web server provides a succinct interface for users, which is shown in Figure 1. It contains eight steps for the prediction of the RNA secondary structure for a given sequence.

Input

Job name. A job name can be entered in the text box in the first step. Long names will be truncated to 40 characters. If no

*To whom correspondence should be addressed. Tel: +86 10 66931324; Fax: +86 10 68213039; Email: liwj@nic.bmi.ac.cn

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

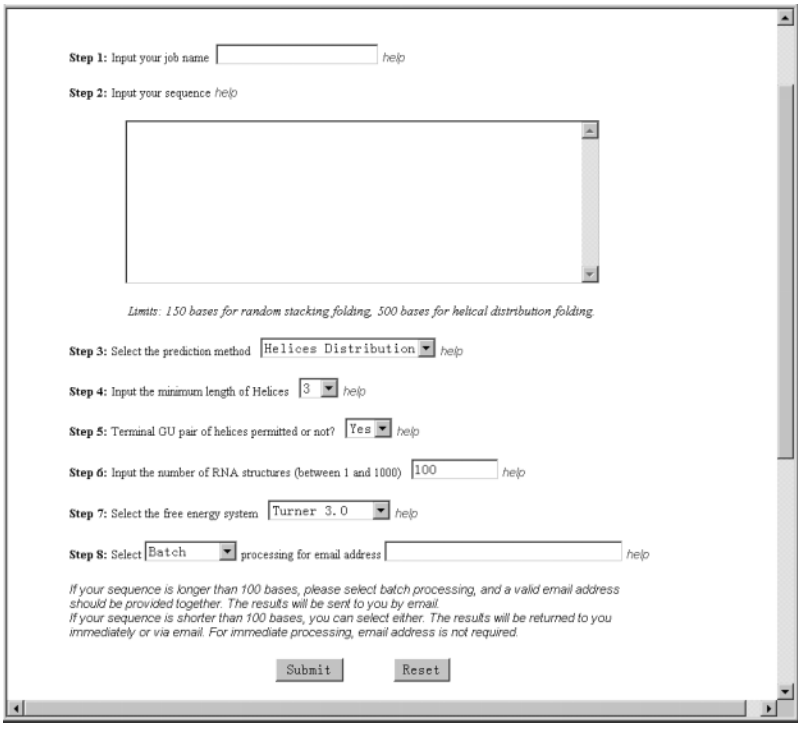

Figure 1. RDfolder user interface, illustrating the steps for the prediction of the secondary structure of a tRNA molecule (GenBank Accession no. X74085).

name is provided, the system clock time of the web server when the job is submitted will be taken as the job name.

Sequence. A sequence to be predicted can be either typed in or pasted into the sequence text box. All blanks and non-alphabetical characters are removed. Uppercase characters are converted to lowercase automatically in the program. If the sequence includes characters besides 'a', 'c', 'g', 't', 'u' and 'n', RDfolder will produce a notification page which informs users that unnecessary information or ambiguous characters are contained in the sequence. RDfolder does not support the IUPAC (International Union of Pure and Applied Chemistry) ambiguous DNA character convention except for 'n'. The letter 'n' is used for an unspecified base and is not allowed to base-pair. 't' is converted to 'u' in RDfolder. RDfolder does not recognize sequences in FASTA, GenBank or EMBL formats. It takes only raw sequences instead. The limit of sequence length is 150 bases for the RS method and 500 bases for the HD method, due to the current computing power of the web server.

Prediction method. The RDfolder server provides the RS and the HD methods. Users can select either by pulling down the list box. If the sequence length is >150 bases, the HD method should be selected. The default method is the HD.

The minimum length of helices. This parameter determines the minimum length of helices in the predicted RNA secondary structure. In both the RS and the HD methods, the first step is to search for all possible helical regions with lengths $\geqslant$ the selected parameter. Then, Monte Carlo simulations are used to predict the secondary structures. Therefore, the minimum length of helices will affect the final prediction results. Users can select from 2 to 10 by pulling down the list box. The default length is 3 bp.

Terminal GU pair of helices permitted or not. This parameter determines whether a terminal GU pair of helices is permitted or not. Users can choose 'Yes' or 'No' by pulling down the list box. A terminal GU pair of helices is permitted by default.

The number of RNA structures. This parameter controls the number of RNA secondary structures to be predicted. Users can input any number between 1 and 1000. For example, if 150 is entered, 150 structures will be calculated and listed. The default number is 100.

The free energy systems. RDfolder provides four free energy systems, i.e. 25°C, 37°C, 42°C (14) and the latest Turner 3.0 free energies (15). Users can select any of them by pulling down the list box. The default free energy system is the latest Turner 3.0.

Immediate and batch processing. If sequence lengths are <100 bases, either immediate or batch processing can be selected. For immediate processing, an email address is not required and the prediction results will be returned to the user immediately. If sequence lengths are >100 bases, batch processing must be selected and a valid email address should be provided. The prediction results will be sent to the user via email when the job is completed. Users can choose 'Immediate' or 'Batch' processing by pulling down the list box. Batch processing is the default selection.

RESULTS

In order to demonstrate the RDfolder server, a tRNA sequence (GenBank Accession no. X74085) was submitted as a test

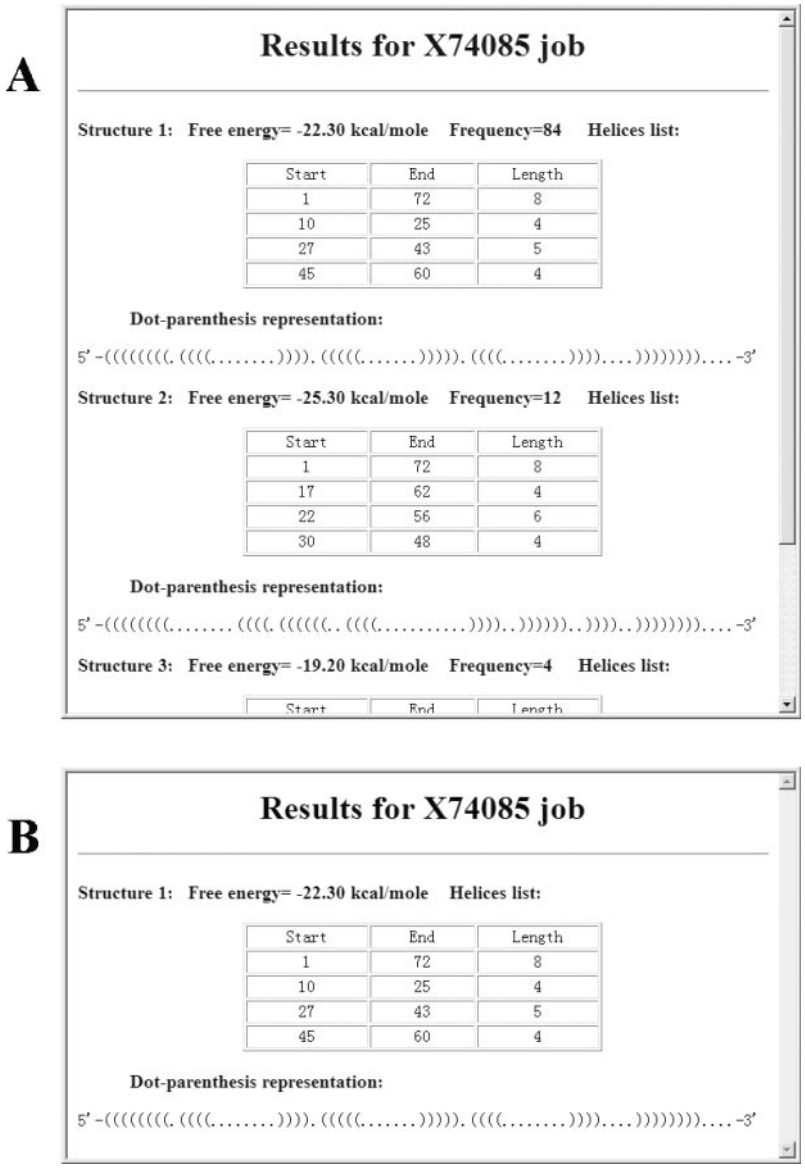

Figure 2. Prediction results from RDfolder server. (**A**) The results of the RS method. All possible secondary structures with different occurrence frequencies are sorted. The first structure, with the highest frequency, is considered as the dominant structure, which is a typical cloverleaf structure. (**B**) The result of the HD method. The predicted structure is also a typical cloverleaf structure.

sequence. The prediction results of the RS method and the HD method are shown in Figure 2. The predicted secondary structure is represented as a list of helices and in dot–parenthesis notation. Each helix consists of three parameters, namely, the start, the end and the length of the helix. The Dot–parenthesis representation is a string of dots and parentheses of the same length as the predicted sequence. A dot in the string indicates that the corresponding nucleotide is unpaired. If nucleotides i and j are paired, where $i < j$, a left parenthesis '(' at position i and a right parenthesis ')' at position j are shown instead.

Owing to the use of Monte Carlo simulations, the RS and the HD methods are relatively time-consuming compared with Zuker's minimum free energy method. For this reason, we provide immediate processing and batch processing for users. For RNA sequences of $\leqslant$100 bases, the computation time is <40 CPU s each for the RS method and <15 CPU s each for the HD method on our current server (for information on our server see Equipment). In addition, a large-scale test on 1161 tRNA sequences from the Rfam family (16) has been carried out. Assuming the cloverleaf structures to be the correct structures, we find that the best prediction accuracies for the RS and the HD method are 0.39 (456/1161) and 0.33 (388/1161) respectively, which are higher than the 0.23 (270/1161) accuracy of Zuker's minimum free energy method.

EQUIPMENT

The current web service is running an Apache web server on a PC Linux box with dual Intel Xeon 2 GHz processors and 4 GB RAM. The operating system is Redhat Linux, version 8.0.

FUTURE PLANS

The RDfolder server presented here provides only two methods for prediction of RNA secondary structure. The length of the sequences is limited to 150 bases for the RS method and 500 bases for the HD method due to the computational constraints of the server. Moreover, for long sequences, prediction of optimal or sub-optimal structures is not so reliable (17). Therefore, we plan to offer prediction methods for single-stranded regions in RNA secondary structure, and to provide users with an easy-to-use method to understand RNA–RNA and RNA–protein interactions.

In addition, as the RDfolder currently provides helix lists and dot–parenthesis representations of secondary structures, work is under way to provide graphical depictions of secondary structures to further improve the visualization of the results.

ACKNOWLEDGEMENTS

We would like to thank Xiaochen Bo for his great help in answering questions about Linux and programming. We thank T.W. Tan, T. Littlejohn and W.T. Li for their help in English improvement. We also thank the anonymous reviewers for their valuable suggestions. This work is supported by grant #30270315 from the National Natural Science Foundation of China.

REFERENCES

1. Zuker,M. and Stiegler,P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, **9**, 133–148.
2. Jaeger,J.A., Turner,D.H. and Zuker,M. (1989) Improved predictions of secondary structures for RNA. *Proc. Natl Acad. Sci. USA*, **86**, 7706–7710.
3. Jaeger,J.A., Turner,D.H. and Zuker,M. (1990) Predicting optimal and suboptimal secondary structure for RNA. *Methods Enzymol.*, **183**, 281–306.
4. Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–3415.
5. Knudsen,B. and Hein,J. (2003) Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res.*, **31**, 3423–3428.
6. Hofacker,I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res.*, **31**, 3429–3431.
7. Sczyrba,A., Kruger,J., Mersch,H., Kurtz,S. and Giegerich,R. (2003) RNA-related tools on the Bielefeld Bioinformatics Server. *Nucleic Acids Res.*, **31**, 3767–3770.
8. Li,W.J. and Wu,J.J. (1996) Prediction of RNA secondary structure based on random stacking of helical regions. *Acta Biophys. Sinica*, **12**, 213–218.
9. Zhang,Z.Q., Yao,L.H. and Hou,Y.D. (1990) Construction and application of a high level expression vector containing P_RP_L promoter. *Chin. J. Viol.*, **6**, 111–116.
10. Li,W.J., Lei,H.X., Pei,W.H. and Wu,J.J. (1998) GeneDn: for high-level expression design of heterologous genes in a prokaryotic system. *Bioinformatics*, **14**, 884–885.
11. Pei,W.H., Shen,B.F. and Li,W.J. (1998) Computer-aided design in high-expression of recombinant ricin-A chain in *E.Coli*. *J. Cell. Mol. Immunol.*, **14**, 33–36.
12. Pei,W.H., Hu,M.R., Li,W.J. and Shen,B.F. (2000) The gene cloning and bioactivity of the expression product of the human FKBP12. *Chin. J. Biochem. Mol. Biol.*, **16**, 322–325.
13. Li,W.J. and Wu,J.J. (1998) Prediction of RNA secondary structure based on helical regions distribution. *Bioinformatics*, **14**, 700–706.
14. Turner,D.H., Sugimoto,N., Jaeger,J.A., Longfellow,C.E., Freier,S.M. and Kierzek,R. (1987) Improved parameters for prediction of RNA structure. *Cold Spring Harbor Symp. Quant. Biol.*, **52**, 123–133.
15. Mathews,D.H., Sabina,J., Zuker,M. and Turner,D.H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, **288**, 911–940.
16. Griffiths,J.S., Bateman,A., Marshall,M., Khanna,A. and Eddy,S.R. (2003) Rfam: an RNA family database. *Nucleic Acids Res.*, **31**, 439–441.
17. Higgs,P.G. (2000) RNA secondary structure: physical and computational aspects. *Q. Rev. Biophys.*, **33**, 199–253.