doi:10.3969/j.issn.1672-5565.2015.01.07

基于 PDB 数据库的三个 RNA 二级结构预测软件评估

刘 伟[1,3], 黄伊子[1,3], 李都悦[1,3], 向 妍[1,3], 周 玮[1,2,3*]

(1.湖南农业大学植物保护学院植物病虫害生物学与防控湖南省重点实验室, 长沙 410128;
2.湖南省烟草公司郴州市公司, 湖南 郴州 423000;
3.湖南农业大学湖南省生物农药与制剂加工工程技术研究中心, 长沙 410128)

摘 要:随着21世纪分子生物学研究的蓬勃发展,RNA二级结构预测成为其中一项重要内容。由于RNA二级结构预测的准确性最为关键,因此寻找高精度且易操作的二级结构预测工具显得非常重要。本文选取三种简单且易操作的二级结构预测软件,先基于PDB数据库收录的318个RNA发夹序列进行二级结构预测,进而通过比较预测结果与实验测定结果进行软件预测性能评估。比较结果显示,RNAstructure为三个软件中性能最优的RNA二级结构预测软件。

关键词:RNA二级结构;PDB数据库;二级结构预测;准确性

中图分类号:Q74 **文献标志码**:A **文章编号**:1672-5565(2015)-01-035-05

Evaluation of three RNA secondary structure prediction softwares based on PDB database

LIU Wei[1,3], HUANG Yizi[1,3], LI Douyue[1,3], XIANG Yan[1,3], ZHOU Wei[1,2,3*]

(1.*Hunan Provincial Key Laboratory for Biology and Control of Plant Diseases and Insect Pests*, *College of Plant Protection*, *Hunan Agricultural University*, *Changsha* 410128, *China*; 2.*Chenzhou Company of Hunan Tobacco Company*, *Chenzhou Hunan* 423000, *China*; 3.*Hunan Provincial Engineering & Technology Research Center for Biopesticide and Formulation Processing*, *Hunan Agricultural University*, *Changsha* 410128, *China*)

Abstract: With the development of molecular biology in the 21st century, the prediction of RNA secondary structure has become one of the most important contents in the field. Because the accuracy of RNA secondary structure prediction is crucial, it is very important to look for the secondary structure prediction tool with high precision and easy operation. In this article, three kinds of secondary structure prediction softwares were selected to evaluate their performances. Firstly, we predicted the RNA secondary structures of 318 RNA hairpins collected from PDB database, and then evaluated the performance of the softwares by comparing the predicted results with the experimental ones. Comparison results showed that RNA structure was superior to the other two kinds of softwares in predicting RNA secondary structure.

Keywords: RNA secondary structures; PDB database; Secondary structure prediction; Accuracy

RNA二级结构是指RNA分子在自然条件下盘绕、卷曲借助碱基间的氢键相互连接形成部分碱基配对和单链交替出现的茎环结构。RNA二级结构中碱基互补配对形成的双螺旋区成为茎区,而不形成互补配对的单链形成环。茎区主要按经典的Watson-Crick规则配对,即G和C配对,A和U配对。此外,在某些情况下也可形成G和U配对[1]。RNA的空间结构是识别RNA分子的重要依据和功能研究的基础和前提。虽然实验手段是获取二级结构的最可靠方法,但是由RNA分子难结晶而且降解

收稿日期:2014-09-15;**修回日期**:2014-11-26.

基金项目:国家自然科学基金青年项目(31301388);中国博士后面上项目(2014M562109);湖南省自然科学基金(14JJ3092);湖南省科学技术厅科技计划项目(2014GK3046);湖南农业大学大学生科技创新基金(团委)资助科研项目(18);湖南农业大学植物保护学院"大学生创新性实验计划项目(SAY1106)。

作者简介:刘伟,男,在读本科,研究方向:生物信息学;E-mail:liuwei_hnnd@163.com.

*通信作者:周玮,女,博士,副教授,研究方向:生物信息学;E-mail:mengrzhou@163.com.

快,采用实验方法测定分子结构很困难,并且代价高昂。近年来,采用计算机和数学模型预测 RNA 二级结构的方法被广泛采用,成为 RNA 结构和功能研究领域的热点问题[2]。RNA 二级结构作为决定 RNA 分子功能的重要环节,与许多重要生物学过程相联系。RNA 的二级结构广泛影响各类 RNA 的各种生物学过程,如影响 RNAi 的效率,也被广泛应用于寻找新的非编码 RNA[3-4]。因此,RNA 二级结构预测是进行 RNA 各项生物学功能研究的基础,RNA 二级结构预测的准确性直接关系到整个实验的进展,如何选取 RNA 二级结构预测软件就显得尤为重要。本文基于 PDB 实验数据对 RNAstructure、Centroidfold 和 RNAshapes 三个软件的二级结构预测功能进行比较,从中选取最优二级结构预测软件。

1 材料和方法

1.1 RNA 二级结构获取

RNA 结构的选取是本文研究的一个重要环节。供试 RNA 结构下载自 PDB 数据库。PDB(http://www.rcsb.org/pdb/home/home.do)是一个蛋白质、核酸等生物大分子的结构数据的数据库[5],由 Worldwide Protein Data Bank 监管。PDB 可以经由网络免费访问,是结构生物学研究中的重要资源。值得一提的是,虽然 PDB 的数据是由世界各地的科学家提交的,但每条提交的数据都会经过 PDB 工作人员的审核与注解,并检验数据是否合理。因此,在 PDB 数据库选取 RNA 数据是保证实验数据真实、可靠的基础。

因为该数据库数据量较大,且一直保持更新,所以选取 2006~2013 年期间收录的所有 RNA 结构。考虑到 RNA 结构的精确性,仅保留分辨率小于2.8Å 的 RNA 发夹。

1.2 预测方法及预测软件选取

RNA 二级结构预测方法的研究也比较多,比较经典方法有最小自由能法、动态规划算法和 Sankoff 算法。但最近也有些新的研究方法,如基于隐 markov 模型的 RNA 二级结构预测方法、基于进化神经的预测方法、基于半监督学习的随机文法模型方法等[6-8]。

软件选取是本文研究中的另一个重要环节。目前常用的 RNA 二级结构软件众多,综合考虑如实用性、操作难度和获取难度等各方面条件,我们选取了三个软件作为评估对象,分别是 RNAstructure、Centroidfold 和 RNAshapes,它们均是 RNA 二级结构预测中比较重要的软件。

RNAstructure (http://rna.urmc.rochester.edu/RNAstructure.html)是一款可在 Microsoft Windows 操作系统下免费使用的 RNA 结构预测和分析软件[1]。RNAstructure 使用 Zuker 算法预测 RNA 二级结构,预测一个结构分为两步。第一步是使用回归算法生成一个最优结构与一系列次优结构。生成次优结构的个数由用户输入的两个参数决定,第三个参数是重新排序最有可能的结构。使用公式重新计算每个结构的最小自由能,输出根据重新计算的最小自由能排序,这两步是连续进行的。该款软件的主要程序设计依赖于以下几个方面算法:1)最小自由能理论;2)碱基配对可能性原则;3)寡核苷酸与互补片段结合亲和力原则;4)共同序列保守结构分析原则。RNAstructure 具有操作界面友好、功能强大和给出良好图形界面输出的优点,它可以测序单一序列,也可以比较两个序列的结构,目前提供 Windows 和 Linux/UNIX 版本,不提供在线预测。

Centroidfold (http://www.ncrna.org/centroidfold/)是 RNA 二级结构预测中最精准的网络应用程序之一,它接受两种序列数据[9]:一个 RNA 序列和多个对齐的 RNA 序列。它的预测结果以碱基对符号和图形来表示,PDF 格式的图形也可接受。该服务器常用的应用是多序列对齐 RNA 二级结构预测,这个服务器的主要优点是用原始的 Centroidfold 软件作为预测引擎,从而在基准测试中获得最高的分数和最好的预测精确性,另外,使用这个软件进行 RNA 二级结构预测是免费的且不用登陆。CentroidFold 与 RNAfold、sfold 和 CONTRAfold 等相比,其性能相对来说比较好[10]。

RNAshapes (http://bibiserv.techfak.uni-bielefeld.de/rnashapes/)使树状域与结构映射,保持了邻接与嵌套的结构特点,但无视螺旋长度[11-12]。它与动态规划算法紧密结合,因此可在此间用于二级结构预测,这不但避免了指数爆炸,还给了我们一个充分和完整的 RNA 分子折叠空间。RNAshapes 提供了三个强大的 RNA 分析工具:1)计算不同形状中的一组代表结构,并从中选取最相符的[13];2)计算形状累计概率[14];3)与共识结构进行比较预测,并作为 Sankoff 算法的选择方案[15]。

1.3 结果评估

ViewerLite 是一款操作简单、界面友好的结构示图软件。将从 PDB 数据库下载的包含有 RNA 结构的 PDB 文件载入 ViewerLite 软件中,软件将显示出相应的 RNA 二级结构图,保存好图像并记录好相关结果,作为预测结果评估的标准。

图 1 是用软件 Viewerlite 显示出的 RNA 发夹立

体结构(以 1YN1 为例,其序列为 GCGAGUUGAC-UACUCGC),其结构可以旋转和缩放,因此可以方便而准确地验证软件预测结构是否与实验结构相匹配。

图 1　Viewerlite 呈现的 RNA 发夹(1YN1)结构图

Fig.1　The spatial graph of RNA hairpin (1YN1) using ViewerLite software

2　结果与分析

从 PDB 数据库中选择了 318 个实验测定的 RNA 发夹,然后分别用三种不同软件对其发夹序列进行二级结构预测,比较后记录各个软件预测结果与实验结果的匹配、不匹配和难以识别的数量。匹配是指所选预测软件所呈现出来的图形与 viewerlite 的空间图形在碱基配对上是没有差异的(见图 2);不匹配即碱基配对不一致(见图 3)。除了匹配和不匹配外,还有其他情况,包括无法识别和输入 RNA 序列后无其二级结构呈现,无法识别是指由于二级结构的复杂性,无法辨别出预测出来的结构是否与 viewerlite 上的碱基配对一致(见图 4、图 5)。

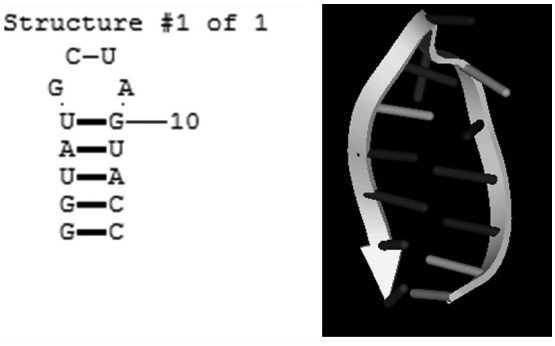

图 2　RNA 二级结构匹配情况展示

Fig.2　RNA secondary structure matching condition

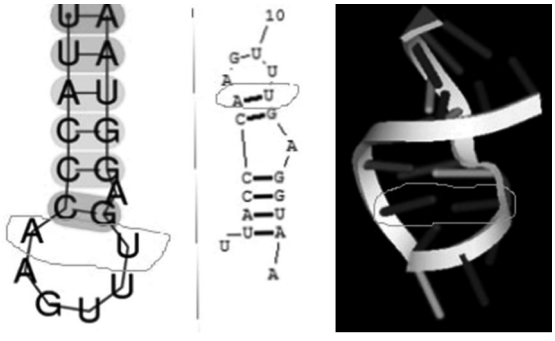

图 3　RNA 二级结构不匹配情况展示(1SLO)

Fig.3　RNA secondary structure mismatching condition

注:由立体图(右图)可知,RNA(1SLO)形成的是四环发夹,即框里面 A 和 U 是配对的,RNAstructure 的预测结果与之一致,但 Centroidfold 的预测结果(左图)是六环发夹,即 A 与 U 没有配对。因此,Centroidfold 的结果是不匹配的,而 RNAstructure 是匹配的。

Notes:The A and G is paired in box from the space diagram which has four ring hairpin and it's consistent with the prediction results of RNAstructure, but the result from Centroidfold has six ring hairpin and A and G is unpair.So it is paired for RNAstructure and unpaired for Centroidfold.

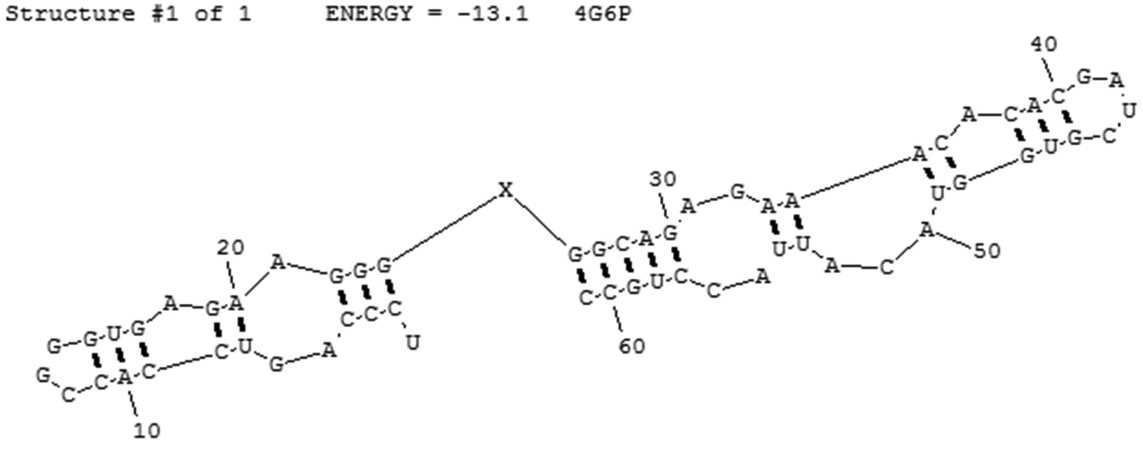

图 4　RNA(4G6P)预测二级结构

Fig.4　The predicted RNA (4G6P) secondary structure

注"图 4 为 RNAstructure 软件的预测结构,但 centroidfold 和 RNAshapes 无法对其进行预测。

Notes:The results can be predicted by RNAsructure and it is diffcult for centroidfold and RNAshapes to do it.

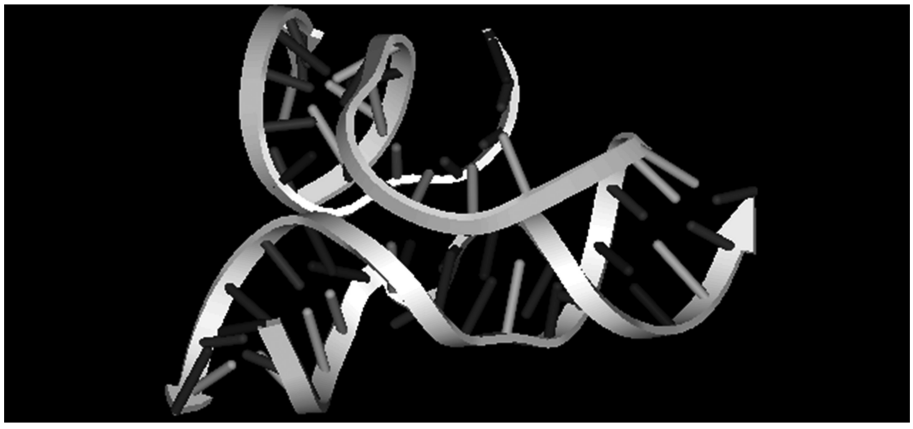

图 5 viewerlite 显示的 RNA(4G6P)空间图

Fig.5 The spatial graph of RNA (4G6P) using viewerlite

注:与图 4 相比较难以识别。

Notes:It is difficult to compare with Fig.4.

表 1 是各个软件的二级结构预测比较结果,由该表格可以看出,RNAstructure 的匹配率最高,其后依次是 RNAshapes 和 Centroidfold。

表 1 三种 RNA 二级结构预测软件预测结果比较

Table 1 Predicted results comparison of three RNA secondary structure prediction softwares

软件	匹配	不匹配	其他	匹配率	不匹配率
RNAstructure	176	132	10	55.35%	41.51%
Centroidfold	108	152	58	33.96%	47.80%
RNAshapes	145	173	0	45.60%	54.40%

图 6 是分别用软件 RNAstructure、Centroidfold 和 RNAshapes 基于 1YN1 发夹序列预测出的二级结构。左图为 RNAstructure 预测结果,上面信息比较详细,图中对结构的名称(可自己命名)、能量值以及碱基对的排序都有明显的注解。从中图看 Centridfold 预测结构颜色分明,有色彩填充,比较美观,Centroidfold 预测出的 RNA 二级结构是这三个预测软件中在视觉效果上是比较好的选择。但是 Centroidfold 是一种在线软件,在没有网络的情况下不能进行二级结构预测。RNAshapes 预测结果相对于其它软件来说(右图)比较简约,其操作起来相对于另两个软件难度要高,但其绘图功能还是比较强大。由图 6 预测结果的结构比较图可知,这三个软件对 RNA(1YN1)的预测结果都是符合要求的,但这只是相对于简单的二级结构来说,遇到复杂的 RNA 结构,他们的差异性较明显。

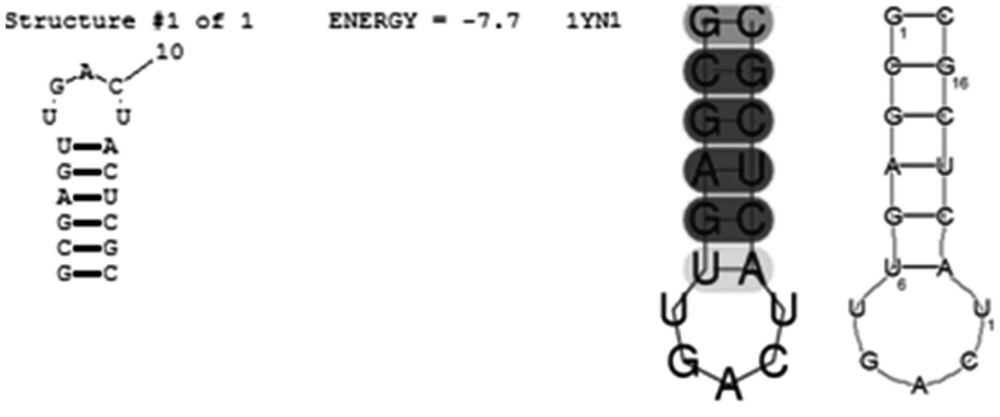

图 6 基于 1YN1 发夹序列预测出的二级结构

Fig.6 The predicted results of 1YN1

注:(1):RNAstructure 预测;(2):Centroidfold 预测;(3):RNAshapes 预测。

Notes:(1):Prediction by RNAstructure; (2):Prediction by Centroidfold;(3):Prediction byRNAshapes.

3　讨　论

通过上述实验数据和预测结果,本文所选的三个预测软件的优劣性很明显。对于 RNA 二级结构的预测,RNAstructure 的性能是其中最好的,其在匹配率以及结构信息方面都较其他软件有优势。同时本次实验也存在很多改进之处,比如,本文下载的 RNA 结构量受年限和分辨率制约,后期工作可考虑覆盖到整个时期且加入分辨率更低但分子更大的 RNA 结构,本文评估软件仅选取三个常用软件,可考虑扩大 RNA 二级结构预测软件的规模。

参考文献(References)

[1] 吴建祖.生物信息学分析实践[M]. 北京:科学出版社,2010.
WU Jianzu. The analysis and practice of bioinformatics [M]. Beijing:Science Press,2010.

[2] 夏飞,朱强华,金国庆,等.基于 CPU-GPU 混合计算平台的 RNA 二级结构预测算法并行化研究[J].国防科技大学学报,2013,(6):138-146.
XIA Fei, ZHU Qianghua, JIN Guoqing, et al. RNA secondary structure prediction parallel algorithm based on CPU-GPU hybrid computing platform [J]. Journal of National University of Defense Technology, 2013, (6): 138-146.

[3] 张浩文, 杨禹丞, 鲁志. 非编码 RNA 的生物信息学研究方法: RNA 结构预测及其应用[J]. 生命科学, 2014, 26(003): 219-227.
ZHANG Haowen, YANG Yucheng, LU Zhi. Noncoding RNA of bioinformatics methods:RNA structure prediction and its application[J]. Life Science, 2014, 26(003): 219-227.

[4] 桂坚斌,孙迎,高武,等.RNA 二级结构在 siRNA 设计中的应用[J].北京生物医学工程,2012,31(6):652-656.
GUI Jianbin, SUN Ying, GAO Wu, et al. Application of RNA secondary structure in siRNA design[J]. Beijing Biomedical Engineering, 2012, 31(6):652-656.

[5] BERMAN H M. The protein data bank: a historical perspective [J]. Acta Crystallographica Section A: Foundations of Crystallography, 2007, 64(1): 88-95.

[6] 董浩,刘元宁,张浩,等.基于隐 Markov 模型的 RNA 二级结构预测新方法[J].计算机研究与发展,2012,49(4):812-817.
DONG Hao, LIU Yuanning, ZHANG Hao, et al. A method of RNA secondary structure prediction based on hidden markov model[J]. Research and Development of Computer, 2012, 49(4):812-817.

[7] 牟超,何静媛,石杨,等.基于进化神经网络的 RNA 二级结构预测方法[J].四川大学学报(自然科学版),2014,51(1):64-68.
MOU Chao, HE Jingyuan, SHI Yang, et al. An evolutionary neural network approach to predict RNA secondary structure[J]. Journal of Sichuan University, 2014, 51(1):64-68.

[8] 唐四薪,赵辉煌,周勇等.RNA 二级结构预测:基于半监督学习的随机文法模型方法[J].计算机与应用化学,2013,(9):1038-1042.
TANG Sixin, ZHAO Huihuang, ZHOU Yong, et al. Prediction of RNA secondary structure: stochastic grammar model based on semi supervised learning method [J].Computers and Applied Chemistry, 2013, (9):1038-1042.

[9] SATO K, HAMADA M, ASAI K, et al. Centroidfold: a web server for RNA secondary structure prediction[J]. Nucleic Acids Research, 2009, 37(suppl 2): W277-W280.

[10] HAMADA M, KIRYU H, SATO K, et al. Prediction of RNA secondary structure using generalized centroid estimators[J]. Bioinformatics, 2009, 25(4): 465-473.

[11] VOSS B, GIEGERICH R, REHMSMEIER M. Complete probabilistic analysis of RNA shapes[J]. BMC Biology, 2006, 4(1): 5.

[12] STEFFEN P, VOSS B, REHMSMEIER M, et al. RNAshapes: an integrated RNA analysis package based on abstract shapes[J]. Bioinformatics, 2006, 22(4): 500-503.

[13] GIEGERICH R, VOSS B, REHMSMEIER M. Abstract shapes of RNA[J]. Nucleic Acids Research, 2004, 32(16): 4843-4851.

[14] JANSSEN S, GIEGERICH R. Faster computation of exact RNA shape probabilities[J]. Bioinformatics, 2010, 26(5): 632-639.

[15] REEDER J, GIEGERICH R. Consensus shapes: an alternative to the sank off algorithm for RNA consensus structure prediction[J]. Bioinformatics, 2005, 21(17): 3516-3523.