DOI:10.3969/j.issn.1672-5565.2017.01.201608001

# 一种新的 RNA 二级结构三维图形表示及其应用

刘立伟*,刘铁晖
(大连交通大学理学院,辽宁 大连 116028)

**摘 要**:本研究提出了一种新的 RNA 二级结构的图形表示方法,这种方法不同于以往的表示方式。根据所提出的 RNA 二级结构的图形表示,将对 9 种病毒的 RNA 二级结构进行图形表示,构建系统进化树,进行序列间相似性的比较和分析。根据最终结果,可以很清晰地发现,AVII 与 LRMV 两种病毒是最为相似的,另外,较大的距离值出现在了 APMV 与 ALMV;PDV 与 AVII 中,这说明这几种 RNA 二级结构明显不相似。这一研究结果与前人相似性分析的结果是十分相似的,同时,所采取的方法更加简单易于区分观察且得到的结果又是十分可靠的,因此,这些更加证明了该方法是有效的。

**关键词**:RNA 二级结构;图形表示;系统进化树;相似性
**中图分类号**:Q522 **文献标志码**:A **文章编号**:1672-5565(2017)01-055-04

# A new 3-D graphical representation of RNA secondary structure and its application

LIU Liwei*, LIU Tiehui
(*Dalianjiaotong university school of science*, *Dalian Liaoning* 116028, *China*)

**Abstract**: Recently, we propose a new 3D graphical representation of RNA secondary structures. Based on this graph representation, we will construct the phylogenetic tree of the 9 viruses, and compare and analyze the similarity between the RNA secondary structures. According to the final results, we clearly find that Pair AVII and LRMV are the most similar. In addition, the larger distance values appear in the APMV and ALMV, PDV and AVII, indicating that these RNA secondary structure sequence has obvious difference. The results of this study are very similar to previous published results one. At the same time, the used method is more simple and easy to identify what we see, while the results is very reliable. Therefore, these results demonstrate the effectivity of our methed.

**Keywords**: RNA secondary structure; Graphical representation; Phylogenetic trees; Similarity

近期,随着生命科学和计算机科学的快速发展,生物信息学作为一个新兴的交叉学科非常活跃。它通过综合利用生物学,计算机科学,应用数学和信息技术而揭示大量而复杂的生物数据所赋有的生物学奥秘。RNA 在生命过程中起着非常重要的作用。许多实验已证实 RNA 的功能依赖于它本身的结构,从 RNA 结构的角度探索 RNA 的功能是一个十分重要的研究课题,因此 RNA 二级结构的相似性比较成为了这个课题的热点问题。随着基因组学和表观遗传学的发展,RNA 在生命活动中所扮演的角色更加被科学家重视。通过对 RNA 结构相似性的分析,进而能够帮助我们了解 RNA 一些新的生物功能。同时,廖波、张屹和曹志等[1-3]在 RNA 二级结构比较上都做出了相应的贡献。

Liao 等[1]给出了一种 RNA 二级结构的二元编码方法。文献[1]的做法是减少一个 RNA 二级结构分为三个二进制数字,并根据 RNA 二级结构的理化性质将编码的碱基分为三类,在提出编码规则的基础上进行操作,将 X1 异或 X2 用 X1 ⊕ X2 表示。结合所提出的编码规则,得到两个特征序列所对应的两个 RNA 公共子序列二级结构,最终得到最优的对齐方式,通过这种方式可以判断碱基之间或者碱基对和碱基对之间的突变,并容易进行序列比对。Zhang 等[2]比较 RNA 二级结构相似性并进行分类,

**收稿日期**:2016-08-13;**修回日期**:2016-09-25.
**基金项目**:辽宁省教育厅科学研究一般项目(No.L 2015093)。
* **通信作者**:刘立伟,男,副教授,研究方向:计算数学;E-mail:liutree80@163.com.

提出了一种三维(3D)的 RNA 二级结构的图形表示,基于核酸的化学性质,把其中一个 RNA 二级结构首先转化为一个特征序列,然后构造一个动态的三维图的特征序列,最后用三维图的数值特征化代表 RNA 二级结构。进行 RNA 二级结构相似性分析。还有 Cao 等[3]提出了一种有效的方法。在突变分析的基础上进行引入的表示,减少一个二级结构为三个二进制数字序列,根据所提出的三维立方体表示,将介绍一个扩展的二进制编码方法的 RNA 二级结构进行调整,通过转换的结构比对到序列比对。之前,Yao 等[4]在不同病毒图形表示的基础上,将 RNA 二级结构的相似性进行了一系列定量的比较。通过之前一些学者的经验理论,提出了一种新的表示方法。

本文主要介绍的是 RNA 二级结构的图形表示,并详细分析了 RNA 结构的表示方法,利用其特点提出了 RNA 结构的表示方法---距离矩阵表示法,在此基础上提出了基于距离矩阵表示法的相似性比对算法。主要内容包括如下几个方面:

(1)介绍了生物信息学中一些最基本和最热点的问题。初步对 RNA 二级结构相似性问题有个大概的了解,同时对现阶段生物信息学的研究进展进行简单的分析。

(2)提出一种新的 RNA 二级结构三维的图形表示方法。

(3)根据定义的图形表示的性质,提取了九维特征向量代表选取的 RNA 二级结构序列。然后将算法运用到 RNA 相似性分析上,同时进行进化树分析,比较其相似性。

## 1 RNA 二级结构的三维图形表示

RNA 二级结构是一组单碱基和碱基对通过氢键之间 A-U,G-C 的相互配对,相互作用形成一条 RNA 序列。根据 RNA 二级结构的特点,可以将一条 RNA 序列用碱基表示出来。图 1 所示为本研究选取的九种病毒的 RNA 二级结构[5]。

以其中一条为例,表示方法如下:

ALMV:AUGCUC'A'U'G'C'A'AAACU'G'C'A'U'G'A'AUGC'C'C'CUAAG'G'G'AUGC

从 5'开始,A 表示 5'端开始,U 为第二个碱基,依次类推,U'表示配对的碱基。根据这种规则可以得到一条用字母表示的 RNA 二级结构序列[6]。

RNA 序列片段用这种方式表示后,选择三种表示方法将每一个碱基放在三维结构里进行定义,将每个碱基给予它一个点坐标。

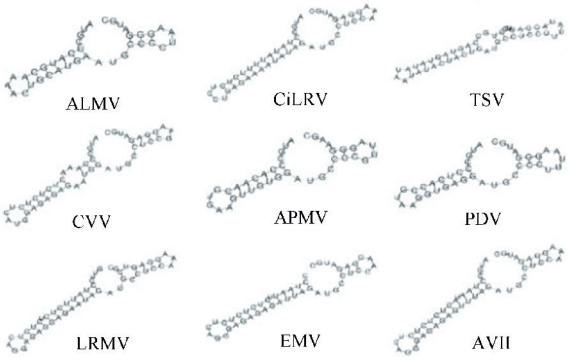



图 1 九种病毒的 RNA 二级结构

**Fig.1 RNA secondary structures of 9 viruses**

根据以往 Qi 等[7]所提出的 DNA 序列三维图形表示,也相类似的进行 RNA 序列三维图形表示:将 A,A', C,C', G,G', U,U'分别置于+$X$ 轴、-$X$ 轴、+$Y$ 轴和-$Y$ 轴上,而特征曲线也是沿着+$Z$ 轴延伸。因此,依照以上分类方法,每条 RNA 序列都会得到三种不同的表示形式,用数学形式表示如下:设 $Y=y_1y_2...y_n$ 为任意的 RNA 序列,则存在三个映射 $f_j, j=1,2,3$, $f_j(Y)=f_j(y_1)f_j(y_2)\cdots f_j(y_n)$ .因此:

$$(\mathrm{i})\ f_1(y_i)=\begin{cases}(1,0,i)\ ,若\ y_i=A/C\ ;\\(0,1,i)\ ,若\ y_i=A'/C'\ ;\\(-1,0,i)\ ,若\ y_i=U/G\ ;\\(0,-1,i)\ ,若\ y_i=U'/G'\ .\end{cases}\tag{1}$$

$$(\mathrm{ii})\ f_2(y_i)=\begin{cases}(1,0,i)\ ,若\ y_i=A/G\ ;\\(0,1,i)\ ,若\ y_i=A'/G'\ ;\\(-1,0,i)\ ,若\ y_i=C/U\ ;\\(0,-1,i)\ ,若\ y_i=C'/U'\ .\end{cases}\tag{2}$$

$$(\mathrm{iii})\ f_3(y_i)=\begin{cases}(1,0,i)\ ,若\ y_i=A/U\ ;\\(-1,0,i)\ ,若\ y_i=C/G\ ;\\(0,1,i)\ ,若\ y_i=A'/U'\ ;\\(0,-1,i)\ ,若\ y_i=C'/G'.\end{cases}\tag{3}$$

按照以上映射原则应用数学软件所画出的 ALMV 三维图(见图 2).

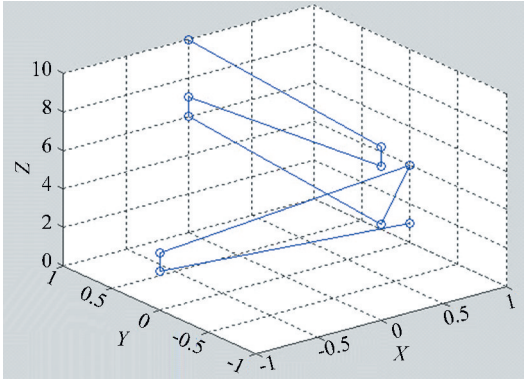



图 2 RNA 二级结构三维表示图(以 ALMV 的子结构为例)

**Fig.2 3-D graphical representation of RNA secondary structures (Substructure of ALMV.)**

## 2　构造系统进化树

将 RNA 二级结构图形表示结束以后,进行相似性比较。通过上述方法得到这些点坐标之后,下一步,将计算这些点之间的距离。同时,当在计算各点之间的距离时也选取了三种方法,分别是:*E* 矩阵,*M/M* 矩阵和 *L/L* 矩阵[8]。计算方法如下:

(1) *E* 矩阵: *E* 中的元素 $e_{ij}$ 即为曲线中的点 $i$ 与点 $j$ 之间的欧氏距离。

$$e_{ij}=\sqrt{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2+\cdots+(x_{in}-x_{jn})^2} \tag{4}$$

(2)*M/M* 矩阵:其中($i$, $j$)元由曲线上两个基对应点的欧式距离与它们之间存在的图论距离之比(即|$i$-$j$|)得到。

$$\text{矩阵元素}\begin{cases} m_{ij}=\dfrac{e_{ij}}{|i-j|}, & \text{若 } i\neq j; \\ m_{ij}=0, & \text{若 } i=j. \end{cases} \tag{5}$$

(3)*L/L* 矩阵:其中($i$, $j$)元由曲线上两个基对应点的欧式距离与两点之间的距离总和之比得到。

$$\text{矩阵元素}\begin{cases} l_{ij}=\dfrac{e_{ij}}{\sum_{k=i}^{j-1}e_{k(k+1)}}, & \text{若 } i\neq j; \\ l_{ij}=0, & \text{若 } i=j. \end{cases} \tag{6}$$

通过这三种方法,会得到三个最大特征值。因为开始进行 RNA 序列表示时就选取了三种表示方法,此时又选取了三种计算点之间距离的方法,所以,此刻会得到九个距离矩阵,因此就会有九个距离矩阵的最大特征值。

其中一个计算结果如下:

ALMV: *E* 矩阵向量(533.4072, 533.5630, 533.2587)

*M/M* 矩阵向量(40.2139,40.5043,40.2078)

*L/L* 矩阵向量(9.9845,8.6573,9.9938)

之后,将这些最大特征值组成一个向量,计算各向量间距离。向量之间的相似性,通过向量间的欧式距离进行计算,很明显,距离值越小,RNA 二级结构序列的相似性就越高。计算结果见表 1。

**表 1　*E* 矩阵的上三角矩阵**

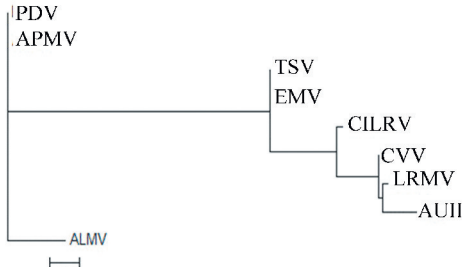
**Table 1　Upper triangular matrix of *E* matrix**

| 病毒种类 | ALMV | CiLRV | TSV | CVV | APMV | PDV | LRMV | EMV | AUII |
|---|---|---|---|---|---|---|---|---|---|
| ALMV | 0 | 650.229 6 | 530.218 1 | 712.429 0 | 96.183 5 | 95.816 0 | 712.485 1 | 530.290 1 | 775.814 6 |
| CiLRV | | 0 | 120.011 5 | 62.200 9 | 554.046 6 | 554.413 7 | 119.940 8 | 119.940 8 | 125.586 0 |
| TSV | | | 0 | 182.211 3 | 434.035 1 | 434.402 3 | 182.267 7 | 0.583 4 | 245.596 8 |
| CVV | | | | 0 | 616.245 9 | 616.613 2 | 0.134 5 | 182.139 1 | 63.386 0 |
| APMV | | | | | 0 | 0.466 3 | 616.302 2 | 434.107 2 | 679.631 5 |
| PDV | | | | | | 0 | 616.669 4 | 434.474 5 | 679.998 8 |
| LRMV | | | | | | | 0 | 182.195 1 | 63.330 3 |
| EMV | | | | | | | | 0 | 245.524 9 |
| AUII | | | | | | | | | 0 |

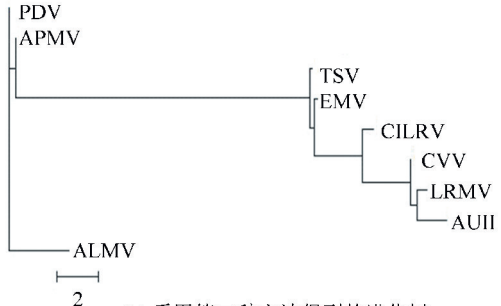
根据以上的结果,如果两条序列片段距离越小则越相似。最小距离法是在最小进化原理的基础上,构造一个距离矩阵来表示物种之间的进化距离。然后,通过这个距离矩阵,采用有效的方法将物种进行分类。然后进行系统进化树的构建,观察结果是否一致。在构建系统进化树时选取的是 Neighbor-Joining 方法。综上所述得出以下的结果,利用 Phylip 及 MEGA 软件描绘出系统进化树。三种方法得到三棵进化树如图 3 所示。

通过以上得到的三棵系统进化树之后,发现这三棵进化树并不完全一致,因此,有必要从这三棵进化树中提取它们的公共部分,也就是构建这三棵系统进化树的最大一致树(见图 4),这样能综合三种图形表示方法的信息。Jansson 等[9-10] 在这方面开发出了很多算法,在这里应用多数一致树(Majority consensus tree)建立最大一致树。

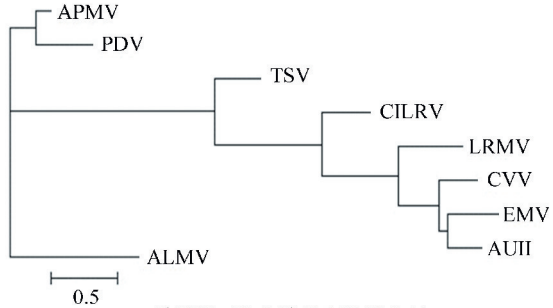
从这个系统进化树的图形表示中可以很清晰地看出,LRMV 与 AVII 距离最近,说明这 2 种病毒 RNA 二级结构最为相似;同理,AVII 与 CVV 相似性次之,等等。反之,APMV 与 ALMV 的距离最远,则这两条序列相似性最弱。由此可见这九种病毒 RNA 二级结构的相似性程度。所采用的方法更为简便且直观。且与以往 Liao 等[11-12] 的研究成果相似。文献[11-12]的研究结果表明:AVII, LRMV, EMV 是最为相似的;同时,APMV, PDV 与其他 RNA 二级结构之间是存在差异性的。由此可见,这一结果与本研究所得到的结果是相类似的。




(a) 采用第一种方法得到的进化树




(b) 采用第二种方法得到的进化树




(c) 采用第三种方法得到的进化树

图 **3** 根据进化距离所构建的系统进化树

**Fig.3 Phylogenetic trees constructed according to the distance**

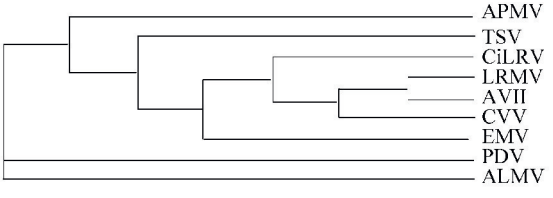



图 **4** 根据三棵系统进化树构建的最大一致树

**Fig.4 Maximum agreement tree is constructed by the three phylogenetic trees**

## 3 讨论及结论

对 RNA 二级结构与功能地研究是如今生物信息学一个十分重要的研究课题,但是对 RNA 结构相似性的预测分析仍然是很困难的。随着 RNA 结构相似性预测方法的日益发展,RNA 数据库的不断增多,RNA 结构预测的软件也日益增多。本文提出了一种新的 RNA 二级结构的图形表示方法。重点介绍了图形表示的构造,系统进化树的构建方法以及 RNA 二级结构序列间相似性的比较。根据所选取的图形表示方法,可以得到关于距离的特征值。随后,在这些距离特征值的基础上再利用预测软件构建系统进化树,基于这种方法,成功地提取了 RNA 二级结构相似性的一些基本信息。可见所选取的这种方法是可行的。

## 参考文献(References)

[1] LIAO B, CHEN W, SUN X, et al. A binary coding method of RNA secondary structure and its application[J]. Journal of Computational Chemistry, 2009, 30(14): 2205-2212. DOI: 10.1002/jcc.21227.

[2] ZHANG Y, HUANG H, DONG X, et al. A dynamic 3D graphical representation for RNA structure analysis and its application in non-coding RNA classification[J]. Plos One, 2016, 11(5): e0152238.DOI: 10.1371/journal.pone.0152238.

[3] CAO Z, LIAO B, LI R, et al. RNA secondary structure alignment based on an extended binary coding method[J]. International Journal of Quantum Chemistry, 2011, 111(5): 978-982.DOI: 10.1002/qua.22464.

[4] YAO Y, NAN X, WANG T. A class of 2D graphical representations of RNA secondary structures and the analysis of similarity based on them[J]. Journal of Computational Chemistry, 2005, 26(13): 1339-1346.DOI: 10.1002/jcc.20271.

[5] LI Ying, DUAN Ming, LIANG Yanchun. Multi-scale RNA comparison based on RNA triple vector curve representation [J]. BMC Bioinformatics, 2012, 13(1): 280. DOI: 10.1186/1471-2105-13-280.

[6] LIU Liwei, WANG Tianming. On 3D graphical representation of RNA secondary structures and their applications[J]. Journal of Mathematical Chemistry, 2007, 42(3): 595-602.DOI: 10.1007/s10910-006-9135-4.

[7] QI Zhaohui, FAN Tongrang. PN-curve: A 3D graphical representation of DNA sequences and their numerical characterization[J]. Chemical Physics Letters, 2007, 442(4-6): 434-440.DOI:10.1016/j.cplett.2007.06.029.

[8] 袁春欣. 核酸序列的图形表示理论及应用[D]. 大连:大连理工大学, 2007.
YUAN Chunxin. Theory and application of graphical representation of nucleic acid sequences[D]. Dalian: Dalian University of Technology, 2007.

[9] JANSSON J, SHEN C, SUNG W. Improved algorithms for constructing consensus trees[J]. Journal of the ACM, 2013, 63(3): 1800-1813.

[10] JANSSON J, SHEN C, SUNG W. Algorithms for the majority rule (+) consensus tree and the frequency difference consensus tree[J]. Algorithms in Bioinformatics. Springer Berlin Heidelberg, 2013(8126): 141-155.DOI:10.1007/978-3-642-40453-5_12.

[11] LIAO B, WANG T M. A 3D graphical representation of RNA secondary structures[J]. Journal of Biomolecular Structure & Dynamics, 2004, 21(6): 827-32. DOI: 10.1080/07391102.2004.10506972.

[12] LIAO B, WANG T, DING K. On a six-dimensional representation of RNA secondary structures[J]. Journal of Biomolecular Structure & Dynamics, 2005, 22(4): 1063-1071.DOI:10.1080/08927020500371332.