J Math Chem (2008) 43:932–943
DOI 10.1007/s10910-007-9303-1

ORIGINAL PAPER

A new mapping rule for RNA secondary structures with its applications

Fenglan Bai · Dachao Li · Tianming Wang

Received: 16 July 2006 / Accepted: 6 July 2007 / Published online: 1 November 2007

Abstract According to the characterization of RNA secondary structures, the RNA secondary structures are transformed into elementary sequences, namely characteristic sequences of RNA secondary structures, by representing A, U, G, C in A-U/ G-C pairs, as A′, U′, G′, C′. Based on the representation, three recurrences for mapping RNA secondary structures into 1-D graph, 2-D graph and 3-D graph are given, respectively. Furthermore, a frequency-based method for RNA secondary structures is given in terms of 1-D graph.

Keywords RNA secondary structures · Graphical representation · Fourier transform

1 Introduction

Low-dimension graphical representations for complex high-dimension data have some advantages, for example, they are intuitive and easy to operate to extract the hided information. So they are applied widely in comparing biological sequences. Especially, graphical representations of DNA sequences, various distance matrices based on those representations, (such as E matrices, L/L matrices, M/M matrices, D/D matrices, and Q matrices), and more and more transformations of graphical representations into

F. Bai
Department of Mathematics, Dalian Jiaotong University, Dalian 116028, China
e-mail: baifenglan@djtu.edu.cn

D. Li
Department of Mathematics, Hainan Normal University, Haikou 571158, China
e-mail: dcli@hainnu.edu.cn

T. Wang (✉)
Department of Applied Mathematics, Dalian University of Technology, Dalian 116024, China
e-mail: wangtm@dlut.edu.cn

numerical representations are used [1–10]. In the available graphical representations of biological sequences, the interesting graphical representations have the following characters: restricted in a well defined plane region [11–19], or in a box in 3-dimension space [16], like Z-curve in 2-D or 3-D space, the spectrum-like representations in a segment of an axis [10–19]. From those graphical representations one can extract valuable mathematical invariants which are related to physichemical and biological properties, and can be used to describe biological sequences.

Ribonucleic Acid (RNA) is a biological macromolecular compound, and is composed of adenine (A), uracil (U), guanine G, and cytosine C, in which A pairs with U, and G pairs with C. The strand folding with itself forms single strand helix structures like DNA double helix structure, and some other complex structures in 3-D space.

RNA moleculars contain two kinds of construction information: RNA primary structures which are single strands composed of bases A,U,G and C, and third structures which are single helix structures formed by its primary structures through folding on themselves. RNA structures are very important for understanding the functions of them. RNA secondary structures are strict subsets of RNA third structures, and they play important role in conjecturing the third structures from RNA primary structures. Therefore, the research of RNA secondary structures is valuable to know the interaction between tRNA and proteins, and the stable process of mRNA. Representation methods of RNA secondary structures are valuable to the research of RNA secondary structures.

RNA secondary structures can be transformed to linear sequences like DNA sequences by suitable procedure [23–25]. Thus we can use the graphical representation methods dealing with DNA sequences to analyze some character of RNA secondary structures.

Recently, it becomes important in bioinformatics that a lot of methods processing signals have been sued to deal with nucleotide or protein sequences. Spectrum analysis as a popular signal processing method has been used to analyze DNA sequences. It can make locally potential periodic property become visual [20].

In DNA sequences analysis, the objective is a symbol strand consisting of characters A, C, G, and T. We can take it as a kind of discrete signals and use Fourier transform on them [21,22]. In this paper, we take RNA secondary structures as discrete signals and get spectrum-like curves to characterize them through Fourier transformation.

2 1-D, 2-D and 3-D representations of RNA secondary structures

Generally, RNA secondary structures are consisted of free bases A, C, G, U and base pairs A–U, C–G. We use A', U', G', C' to represent A, U, G, C in pairs. In this way, RNA secondary structures are transformed into elementary sequences, called characteristic sequences [17]. For example, the substructure of LRMV-3 (see Fig. 1) corresponds to the characteristic sequence $CC'U'C'C'AAAG'G'A'G'U$ (from $5'$ to $3'$).

Zupan and Randic proposed an algorithm [19] which is a simultaneous operation rule in 1-D, 2-D, and 3-D representations. These graphical representations are reversible, that is one can recover the original sequences by the coordinates of points in the curves.

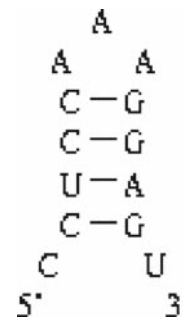

Fig. 1 The substructure of LRMV-3

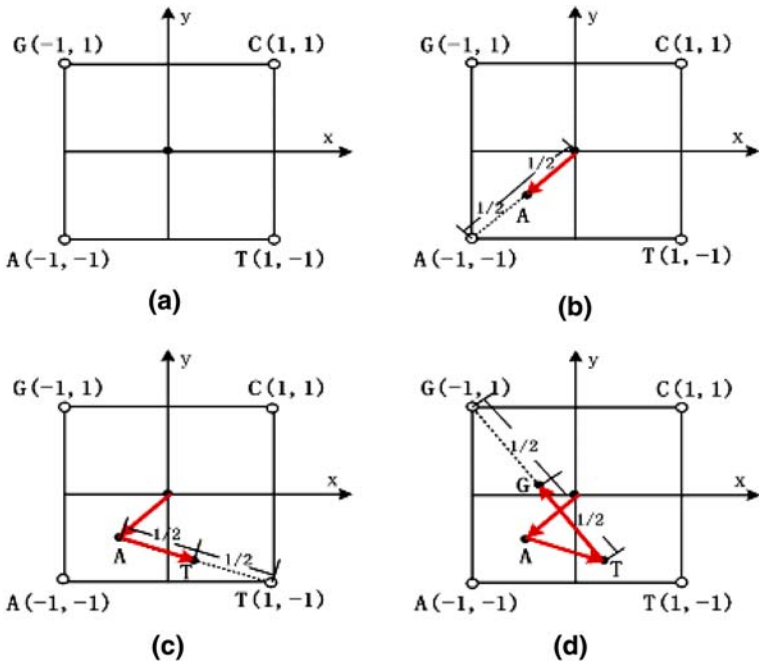

Fig. 2 **(a)** the coordinate of initial point of sequence is the origin **(b)** coordinate of the first base A of sequence ATG is (−0.5, −0.5): the middle point of corner A(−1, −1) and the origin. **(c)** the coordinate of the second base T is (0.25,−0.75): the middle point of A(−0.5, −0.5) and corner T(1,−1), **(d)** the coordinate of third base G is (−0.375, 0.125): the middle point of T(0.25, −0.75) and corner G(−1, 1)

They present 2-D "Z-curve" representation in a regular square in which the center of the square is at the origin of coordinate system and A, U, G, and C are at the four corners (−1, −1), (1, −1), (1,1), and (−1, 1). The Z-curve starting from the center of the square moves toward the corner where the corded base is located and stops at the middle point of the segment from origin to the corner, and continues moving to the next corner at which the next corded base is located, and so on. Here, we illustrate the first three steps to obtain the graphical representation from the DNA sequence ATG of the first exon of *β-globin gene* in the human genome, see Fig. 2.

In this paper, based on the algorithm proposed by Zupan and Randic [19], we give 1-D, 2-D, and 3-D graphical representations of RNA characteristic sequences [23–25], and illustrate the method through RNA secondary structures characteristic sequences of nine viruses(see Fig. 3).

2.1 1-D representation

First, we assign A, U, G, and C to $+x$-axis, and A′, U′, G′ and C′ to $-x$-axis, that is

$$\begin{array}{llll} \mathrm{A} \to x = +1, & \mathrm{U} \to x = +2, & \mathrm{G} \to x = +3, & \mathrm{C} \to x = +4, \\ \mathrm{A}' \to x = -1, & \mathrm{U}' \to x = -2, & \mathrm{G}' \to x = -3, & \mathrm{C}' \to x = -4. \end{array}$$

Certainly, the assignment is arbitrary, there are many other possible assignments. Let $S = s_1 s_2 \ldots$ be any characteristic sequence of an RNA secondary structure, where s_i is the i letter, and the length of the characteristic sequence is N. The recurrence for 1-D graphical representation is

$$R(x_{i+1}) = \frac{R(x_i) + S(x_{s_{i+1}})}{d}, \tag{1}$$

where, d is a real number not equal to zero, and $R\ (x_0) = 0$.

For example, take $d = 2$, we calculate the corresponding coordinates of subsequence C′U′C′CAAGG′ A′G′ of RNA characteristic sequence of virus EMV-3 as follows $S(x) = \{0, -2.0000, -2.0000, -3.0000, 0.5000, 0.7500, 0.8750, 1.9375, -0.5313, -0.7656, -1.8828\}$, and its graphical representation is shown in Fig. 4.

Obviously, 1-D representation method is reversible, that is, from the coordinate of the last point in the curve we can recover the RNA secondary structure sequence, whenever how long the sequence is. The 1-D graphical representations of RNA secondary structure sequences through recurrences are shown in Fig. 5.

In Fig. 5, the curves have no cycles, and the order of points in the curves is the same as the order of nucleotides in the RNA secondary structure sequences and the curve contains more information. In Fig. 5 we find that APMV-3 and PDV-3, EMV-3, LRMV-3 and AVII are similar.

2.2 2-D representations

Let

$$\begin{array}{llll} \mathrm{A} \to (-1, 0), & \mathrm{U} \to (1, 0), & \mathrm{G} \to (0, -1), & \mathrm{C} \to (0, 1) \\ \mathrm{A}' \to (-1, -1), & \mathrm{U}' \to (1, 1), & \mathrm{G}' \to (1, -1), & \mathrm{C}' \to (-1, 1) \end{array}$$

The other assignments and their permutations are allowed. We get the recurrence of $R\ (x_i,\ y_i)$ from which we can get graphical representation:

$$R(x_{i+1}, y_{i+1}) = \frac{R(x_i, y_i) + S(x_{s_{i+1}}, y_{s_{i+1}})}{d}, \tag{2}$$

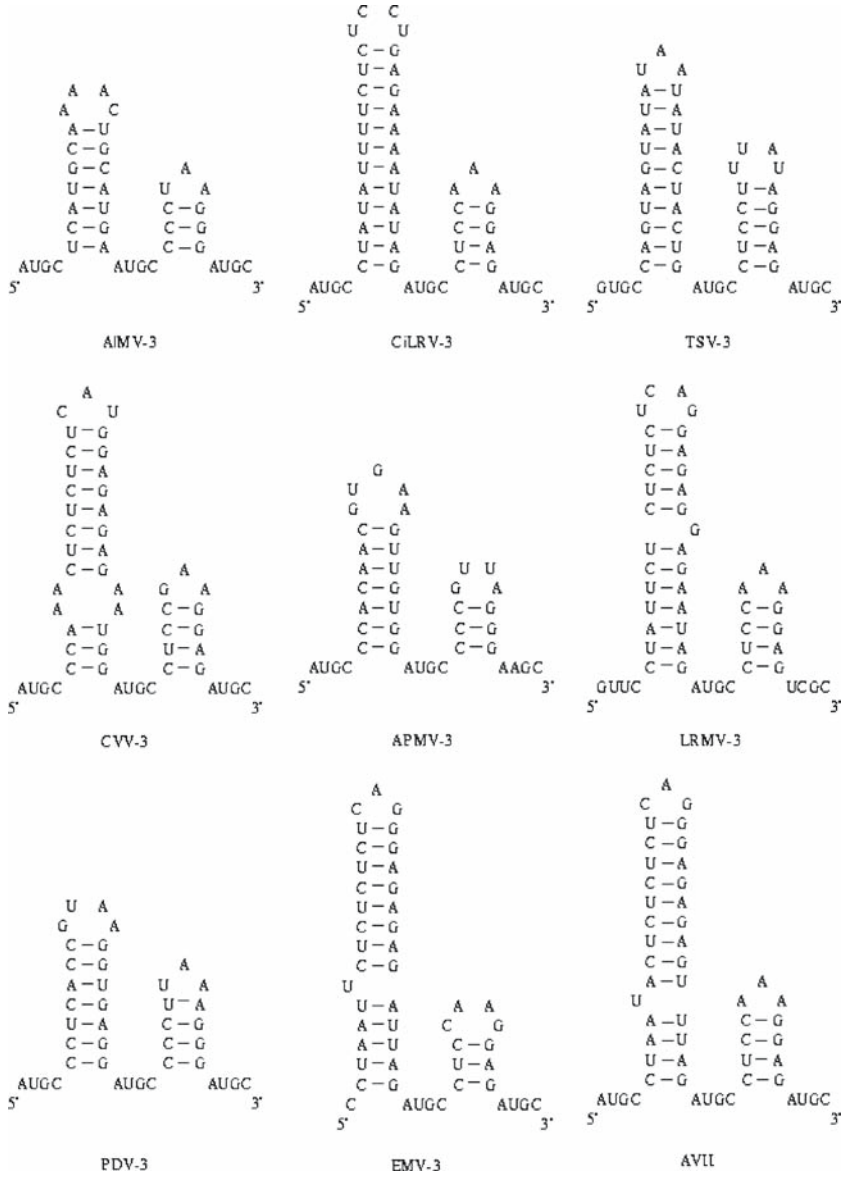

Fig. 3 Secondary structure at the 3′-terminus of RNA 3 of alfalfa mosaic virus(AlMV-3 [25]), citrus leaf rugose virus(CiLRV-3 [26]), tobacco streak virus (TSV-3 [28,29]), citrus variegation virus (CVV-3 [27]), apple mosaic virus (APMV-3 [29]), prune dwarf ilarvirus (PDV-3 [30]), lilac ring mottle virus (LRMV-3 [31]), elm mottle virus (EMV-3 [32]) and asparagus virus II (AVII[33])

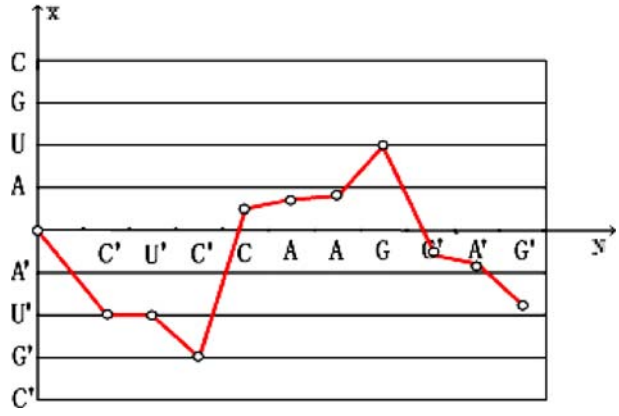

Fig. 4 1-D graphical representation of characteristic sequence of RNA secondary structure EMV-3

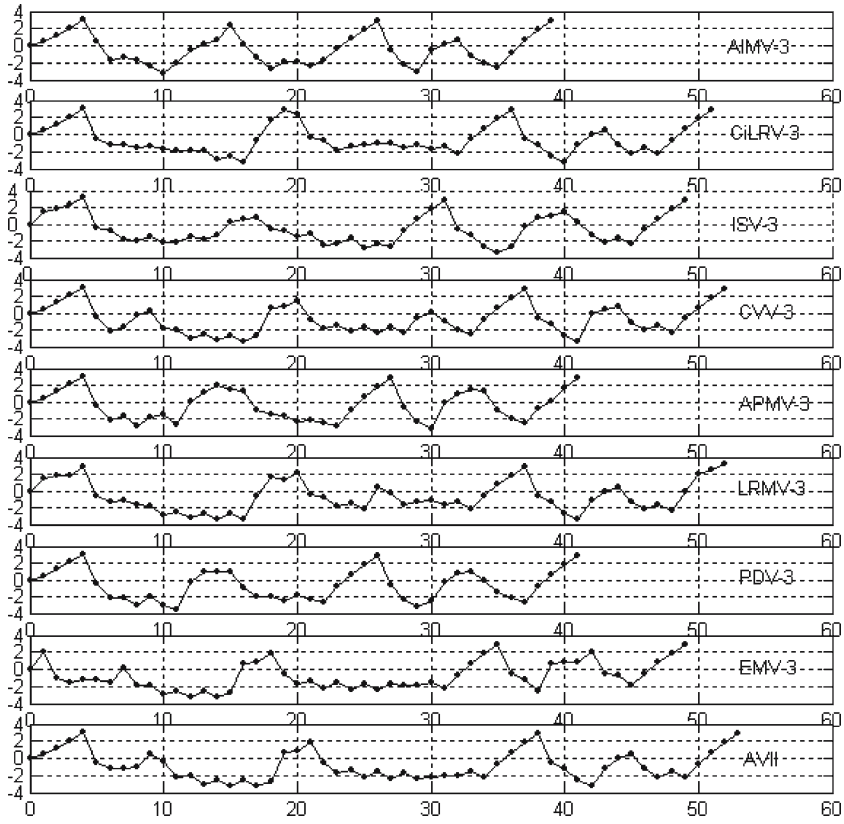

Fig. 5 1-D representations of nine characteristic sequences of RNA secondary structures in Fig. 3

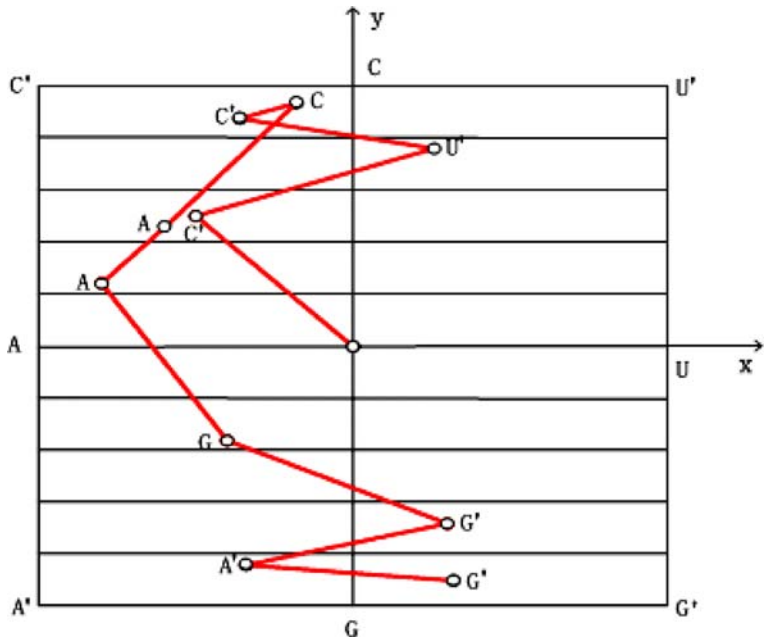

Fig. 6 2-D graphical representation of characteristic sequence of RNA secondary structure EMV-3

where, d is a real number not equal to zero, $R\ (x_0,\ y_0) = 0$.

For example, take $d = 2$, the characteristic sequence C′U′C′CAAGG′A′G′ for the substructure of virus EMV-3 can be transformed into a set of coordinate points, $R(x, y) = \{(0, 0),\ (-0.5000,\ 0.5000),\ (0.2500,\ 0.7500), (-0.3750,\ 0.8750), (-0.1875,\ 0.9375), (-0.5938,\ 0.4688), (-0.7969,\ 0.2344), (-0.3984,\ -0.3828), (0.3008,\ -0.6914), (-0.3496,\ -0.8457),\ (0.3252,\ -0.9229)\}$, and the graphical representation can be seen in Fig. 6.

2.3 3-D representations

Let

$$\begin{array}{llll} \mathrm{A} \to (-1,-1,1), & \mathrm{U} \to (1,-1,-1), & \mathrm{G} \to (1,1,1), & \mathrm{C} \to (-1,1,-1) \\ \mathrm{A}' \to (-1,-1-1), & \mathrm{U}' \to (1,-1,1), & \mathrm{G}' \to (1,1,-1), & \mathrm{C}' \to (-1,1,1) \end{array}$$

Alternate assignments of A, U, G, C, A′, U′, G′, C′ to the corner points are allowed, as before. We can get the recurrence formula $R(x_i, y_i, z_i)$ of 3-D representations

$$R(x_{i+1}, y_{i+1}, z_{i+1}) = \frac{R(x_i, y_i, z_i) + S(x_{s_{i+1}}, y_{s_{i+1}}, z_{s_{i+1}})}{d}, \tag{3}$$

where, d is a non-negative real number, and $R(x_0,\ y_0,\ z_0) = 0$.

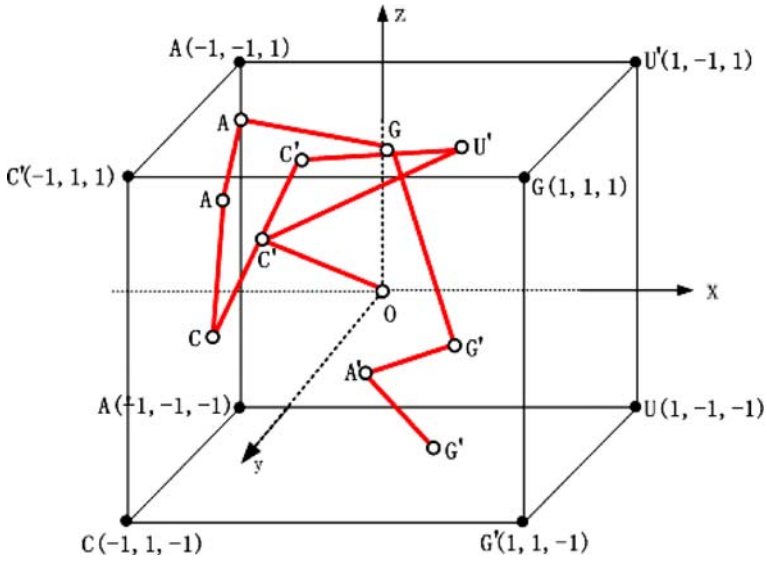

Fig. 7 3-D graphical representation of characteristic sequence of RNA secondary structure EMV-3

For example, take $d = 2$, the subsequence C′U′C′CAAGG′AG′ of characteristic sequence for the RNA secondary structure of virus EMV-3 can be transformed into a set of coordinate points as follows:

$R(x, y, z) = \{(0, 0, 0), (-0.5000, 0.5000, 0.5000), (0.2500, -0.2500, 0.7500), (-0.3750, 0.3750, 0.8750), (-0.6875, 0.6875, -0.0625), (-0.8438, -0.1563, 0.4688), (-0.9219, -0.5781, 0.7344), (0.0391, 0.2109, 0.8672), (0.5195, 0.6055, -0.0664), (-0.2402, -0.1973, -0.5332), (0.3799, 0.4014, -0.7666)\}$, and the 3-D graphical representation can be seen in Fig. 7.

We notice that whenever the lengths are, all the points are positioned in a square/cube in the 2-D/3-D representation, i.e., this method can put the graphical representations of any sequences in a suitable region, and can display the distributions of nucleotides.

For example, in 3-D graphical representations, all nucleotides A appear in a cube, one of whose vertices is origin and another is at point $(-1, -1, 1)$. The edge length of cube is 1. In fact, the initial point is origin, if the first letter of a sequence is A, according to recurrence Eq. (3), its coordinate is $(-1/2, -1/2, 1/2)$. Obviously, the coordinates of all points in the cube satisfy the conditions $-1 < x < 0, -1 < y < 0, 0 < z < 1$. If the ith letter is A, the coordinate of A is $R(x_i, y_i, z_i) = \frac{(R(x_{i-1}, y_{i-1}, z_{i-1}) + (-1, -1, 1))}{2}$ by recurrence Eq. (3), whenever $R(x_{i-1}, y_{i-1}, z_{i-1})$ is, whose coordinates satisfy the $-1 < x_{i-1} < 1, -1 < y_{i-1} < 1, -1 < z_{i-1} < 1$, so $-1 < (x - 1)/2 < 0, -1 < (y - 1)/2 < 0, 0 < (z + 1)/2 < 1$, $R(x_i, y_i, z_i)$ surely is in the cube. For other nucleotides the situation is the same. From the above claim, we can judge the content of each nucleotide by the graphical representations. Obviously, for each assignment, the representation is unique and reversible.

Table 1 The matrix of correlation coefficients of RNA secondary structures based 2-D representations

Base	AIMV-3	CiLRV-3	TSV-3	CVV-3	APMV-3	LRMV-3	PDV-3	EMV-3	AVII
AIMV-3	1.0000	0.9935	0.9942	0.9959	0.9944	0.9961	0.9940	0.9982	0.9980
CiLRV-3		1.0000	0.9953	0.9942	0.9934	0.9974	0.9933	0.9961	0.9955
TSV-3			1.0000	0.9958	0.9972	0.9973	0.9975	0.9964	0.9977
CVV-3				1.0000	0.9973	0.9969	0.9980	0.9975	0.9976
APMV-3					1.0000	0.9975	0.9987	0.9968	0.9975
LRMV-3						1.0000	0.9980	0.9988	0.9988
PDV-3							1.0000	0.9977	0.9982
EMV-3								1.0000	0.9989
AVII									1.0000

2.4 Applications of 2-D and 3-D representations

Based on 2-D and 3-D representations, RNA secondary structures can be transformed to numerical sequences to calculate corresponding L/L matrices. Because the invariants of matrices can characterize the oroperties of structures [8,9], we use the invariants of L/L matrix as a measure to compare the similarities between the structures. But there exist many invariants, such as leading eigenvalue, determinant, traces of matrices, sums of lines, and so on. In this paper, we use the first 20 eigenvalues of L/L matrix as components of a vector to describe RNA secondary structures [8], and use correlation coefficients of vectors to characterize the similarities of sequences [26].

Let $r_{\alpha\beta}$ denote correlation coefficients between $x_\alpha(n)$ and $x_\beta(n)$, we have:

(1) $|r_{\alpha\beta}| \leq 1, r_{\alpha\alpha} = 1$;
(2) $r_{\alpha\beta} = r_{\beta\alpha}$.

When the absolute value of $r_{\alpha\beta}$ approaches to 1, vectors $x_\alpha(n)$ and $x_\beta(n)$ are closer, otherwise they are less similar, where,

$$r_{\alpha\beta} = \frac{\sum_{n=0}^{N-1} x_\alpha(n) x_\beta(n)}{\left[\sum_{n=0}^{N-1} |x_\alpha(n)|^2 \sum_{n=0}^{N-1} \left|x_\beta(n)\right|^2\right]^{1/2}}.$$

We computed the correlation coefficients of RNA secondary structures for nine viruses listed in Fig. 3, see Tables 1 and 2.

In Table 1 we can see that EMV-3 and AVII, LRMV-3 and AVII, and LRMV-3 and AVII are more similar. In Table 2, EMV-3 and CVV-3 are the most similar pair, then PDV-3 and APMV-3, EMV-3 and AVII, LRMV-3 and AVII, LRMV-3 and EMV-3, and CVV-3 and AVII are similar, that is more coincident with the real structure and the results in [23–25].

Table 2 The matrix of correlation coefficients of RNA secondary structures based 3-D representations

Base	AIMV-3	CiLRV-3	TSV-3	CVV-3	APMV-3	LRMV-3	PDV-3	EMV-3	AVII
AIMV-3	1.0000	0.9940	0.9982	0.9947	0.9976	0.9961	0.9988	0.9959	0.9955
CiLRV-3		1.0000	0.9975	0.9969	0.9960	0.9980	0.9957	0.9976	0.9982
TSV-3			1.0000	0.9956	0.9981	0.9979	0.9983	0.9967	0.9971
CVV-3				1.0000	0.9981	0.9983	0.9973	0.9995	0.9987
APMV-3					1.0000	0.9979	0.9992	0.9986	0.9981
LRMV-3						1.0000	0.9980	0.9986	0.9988
PDV-3							1.0000	0.9984	0.9978
EMV-3								1.0000	0.9990
AVII									1.0000

The advantages of the method are (1) it can be used for large RNA secondary structures, (2) the invariants are easy to calculate, and (3) the computation is easy and fast, and the result is better.

3 Spectrum analysis of RNA secondary structure sequences

The characteristic sequences of RNA secondary structure sequences are mapped to numerical sequences $x(n)(x(n) = R(x_n)), n = 1, 2, \ldots, N$ (the length of the characteristic sequence) by recurrence (1), $x(n)$ is the mapping value of the n-th character.

The discrete fourier transform formula [26] for sequence $x(n)$ is

$$DFTx(n) = X(k) = \sum_{n=1}^{N-1} x(n)e^{-j2\pi kn/N}, \quad k = 0, 1, 2, \ldots, N-1 \tag{4}$$

where N is the length of the sequence.

We will analyze RNA secondary structures of nine viruses in Fig. 1 by our method. Fig. 8 is spectrum-like curves of numerical sequences $x(n)$ by DFT.

From Fig. 8 we can see the periodical properties in some regions of each graph, such as the spectrum graph of CVV-3 , who obviously periodically changes in region 10–18 and 24–29. The spectrum distributions of sequences are very clear. We can see that in spectrum graphs of APMV-3 and AIMV-3, the peaks locate near 3 and they are nearly equal, and peaks locate near 1 and are near equal in spectrum graphs EMV-3 and AVII. We can judge from spectrum graphs that APMV-3 and AIMV-3, EMV-3 and AVII are similar, LRMV-3, EMV-3 and AVII are more similar, so do the CVV-3 and AVII. These results coincide with the results in [23–25] and basically are the same as the analysis in time domain (see Fig. 5).

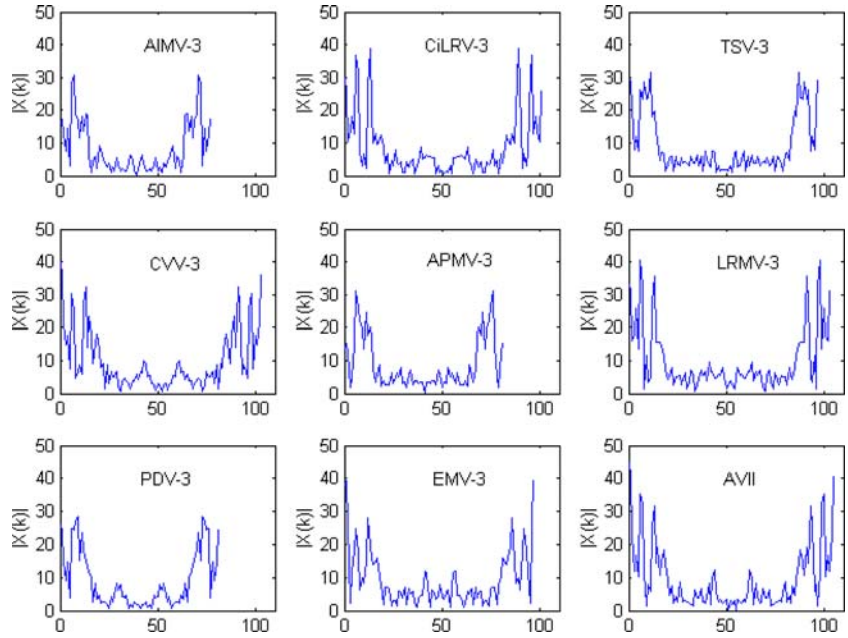

Fig. 8 spectrum graphs for RNA characteristic sequences of nine viruses in Fig. 1

4 Conclusion

In this paper, RNA secondary structures are transformed into 1-D, 2-D, or 3-D graphical representations by certain mapping rules (recurrence formulae). In this way, we can tell the difference among the sequences intuitively. Then numerical sequences are obtained based on the representations and numerical characterizations are extracted from the representations to analyze the similarities of the RNA secondary structures.

The first advantage of recurrence formulae 1–3 are that they are reversible, that is the original characteristic sequences of RNA secondary structures can be recovered through numerical sequences from the coordinate of the last point to the first point one by one.

The second advantage is that any sequences (such that protein sequences consisted of 20 amino acids) can be transformed into 1-D, 2-D,or 3-D graphical representations.

In the spectrum analysis of nucleotide sequences, the periodical property is reflected by the discrete Fourier transformation. It needs further study on whether other properties of the spectrums can be reflected by the DFT. In fact, the discrete Fourier transformation only reflects local properties of the sequences, systematic and synthetic methods are more effective.

The method proposed in this paper is used to analyze the similarities of RNA secondary structures. In fact, RNA secondary structures contain more biological information. How to establish better mathematical model is a further target.

Acknowledgments It is supported by the National Natural Science Foundation of China (10571019).

References

1. M. Randic, M. Vracko, J. Chem. Inf. Comput. **40**, 599 (2000)
2. M. Randic, M. Vracko, A. Nandy, S.C. Basak, J. Inf. Comput. **40**, 1235 (2000)
3. M. Randic, M. Vracko, L. Nella, P. Dejan, Chem. Phys. Lett. **368**, 1 (2003)
4. M. Randic, M. Vracko, N. Lers, D. Plavsic, Chem. Phy. Lett. **371**, 202 (2003)
5. A. Nandy, P. Nandy, Chem. Phys. Lett. **368**, 102 (2003)
6. A. Nandy, Comput. Appl. Biosci. **12**, 55 (1996)
7. A. Nandy, Curr. Sci. **66**, 821 (1994)
8. B. Liao, T.M. Wang, J. Comput. Chem. **11**, 1364 (2005)
9. B. Liao, T.M. Wang, J. Chem. Inf. Comput. Sci. **44**, 1666 (2004)
10. F.L. Bai, T.M. Wang, J. Biomol. Struc. Dyn. **23**, 537 (2006)
11. H.I. Jeffrey, Nucleic Acid Res. **18**, 2163 (1990)
12. N. Goldman, Nucleic Acid Res. **21**, 2487 (1993)
13. S. Basu, A. Pan, C. Dutta, J. Das, J. Mol. Graphs. Modell. **15**, 279 (1997)
14. M. Randic, SAR QSAR Environ. Res. **15**(3), 147 (2004)
15. M. Randic, J. Zupan, SAR QSAR Environ. Res. **15**(3), 191 (2004)
16. M. Randic, J. Zupan, A.T. Balaban, Chem. Phys. Lett. **397**, 247 (2004)
17. M. Randic, Chem. Phys. Lett. **386**, 468 (2004)
18. M. Randic, J. Chem. Inf. Comput. Sci. **41**, 1330 (2001)
19. J. Zupan, M. Randic, J. Chem. Inf. Model. **45**, 309 (2005)
20. M. Randic, D. Butina, J. Zupan, Chem. Phys. Lett. **419**, 528 (2006)
21. S.V. Buldyrev, A.L. Goldberger, S. Havlin, Phys. Rev. E. **51**, 5084 (1995)
22. D. Anastassiou, Bioinfor matics. **16**, 1073 (2000)
23. F.L. Bai, W. Zhu, T.M. Wang, Chem. Phys. Lett. **408**, 258 (2005)
24. B. Liao, T.M. Wang, J. Biomol. Struct. Dyn. **21**, 827 (2004)
25. B. Liao, K.Q. Ding, T.M. Wang, J. Biomol. Struct. Dyn. **22**, 455 (2005)
26. K.D. Zong, G.S. Hu, *Digital Signals Disposal* (Springer, Bei Jing, 1998)