Journal of Mathematical Chemistry, Vol. 42, No. 3, October 2007 (© 2006)
DOI: 10.1007/s10910-006-9135-4

On 3D graphical representation of RNA secondary structures and their applications

Liwei Liu*

Department of Applied Mathematics, Dalian University of Technology, Dalian 116024, China
E-mail: daliguowei@163.com

Tianming Wang

Department of Mathematics, Hainan Normal University, Haikou 571158, China

Received 28 March 2006; revised 13 April 2006

In this article, we proposed a 3D representation of RNA secondary structures. Based on this representation, we outline an approach by constructing a 3-component vector whose components are the normalized leading eigenvalues of the L/L matrices associated with RNA secondary structure. The examination of similarities/dissimilarities among the secondary structure at the 3′-terminus of different viruses illustrates the utility of the approach.

KEY WORDS: RNA secondary structure, similarity, virus, 3D graphical representation

1. Introduction

Mathematical analysis of large-volume genomic sequence or structure data is one of the challenges for bio-scientists. Graphical representation of DNA sequence provides a simple way of viewing, sorting and comparing various gene structures [1–7].

Ribonucleic acid (RNA) is an important molecule which performs a wide range of functions in the biological system. In particular, it is RNA(not DNA) that contains genetic information of virus such as HIV and therefore regulates the functions of such virus. RNA has recently become the center of much attention because of its catalytic properties, leading to an increased interest in obtaining structural information. Similar with the graphical representation of DNA sequences, we also can outline several graphical representation of RNA primary sequences based on 2D, 3D or 4D to compare the similarity of RNA primary

*Corresponding author.

sequences. Recently, Liao et al. have proposed 2D or 3D graphical representation of RNA secondary structures [8–11].

In this article we will make a comparison for the secondary structures at the 3′-terminus belonging to nine different species based on a 3D graphical representation. In figure 1, the secondary structures at the 3′-terminus belonging to nine different viruses are listed, which were reported by Reusken and Bol [12]. The similarities are computed by calculating the Euclidean distance between the end points of the vectors or calculating the correlation angle of the two vectors.

2. 3D representation of RNA secondary structures

The secondary structure of an RNA is a set of free bases and base pairs formed bonds between A–U and G–C. Following Zuker, we assume a model where there are no knots in the secondary structure. This means that for the secondary structure, the bonds are non-crossing. In this paper, we think of base pair G–U as free bases, although the pairing of G and U is frequently allowed. Let A', U', G', C' denote A, U, G, C in the base pair A–U and G–U, respectively. Then we can obtain a special sequence representation of the secondary structure. We call it characteristic sequence of the secondary structure. For example, the corresponding characteristic sequence of the substructure of AlMV-3 (figure 2) is $CGUAG'G'G'AAUC'C'C'CG$ (from 3′ to 5′).

We will illustrate the 3D characterization of RNA secondary structure. In 3D space points, vectors and directions have 3-components, and we will assign the following basic elementary directions to the four free bases and four pair bases.

Randic et al. proposed a 3D graphic representation of DNA primary sequences [1], where the coordinates of A, T, C, G are $(+1,-1,-1)$, $(+1,+1,+1)$, $(-1,-1,+1)$ and $(-1,+1,-1)$, respectively (see figure 3).

In this paper, we follow the idea proposed by Randic et al. U substituting T, the coordinates of four free base A, U, G, C are obtained. Graphically, four free bases are placed in to four vertices of a tetrahedron. We put four matched bases A', U', G', C' into the remaining four vertices of the cube side length 2. The rule of putting is that A' is put in an opposite corner of A in the edge parallel to x-axis, and U' is out in an opposite corner of U, C', G' are put in same way (see figure 4). The rule is called X-pattern. According this rule we get the coordinates of A', U', G', C' as follows

Free base	A $(+1,-1,-1)$,	U $(+1,+1,+1)$,
	C $(-1,-1,+1)$,	G $(-1,+1,-1)$.
Pair base	A' $(-1,-1,-1)$,	U' $(-1,+1,+1)$,
	C' $(+1,-1,+1)$,	G' $(+1,+1,-1)$.

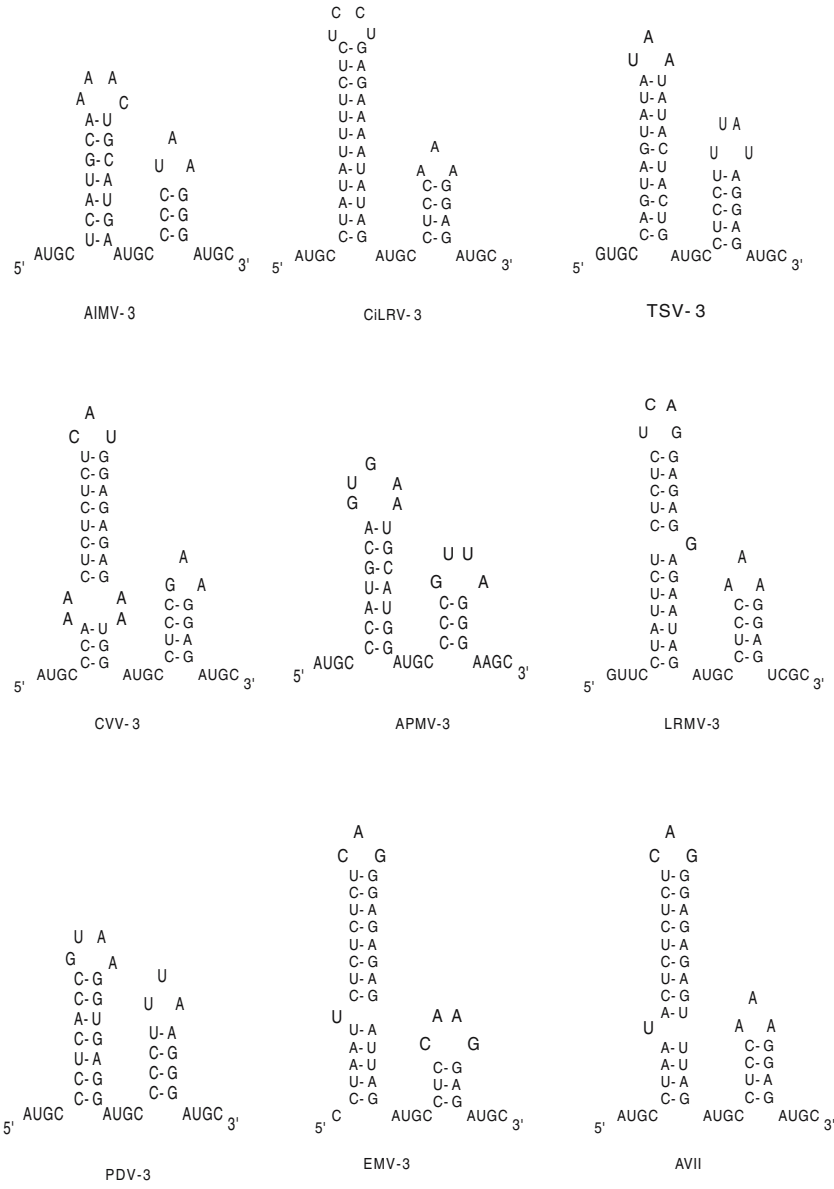

Figure 1. Secondary structure at the 3′-terminus of RNA 3 of alfalfa mosaic virus (AlMV-3 [13]), citrus leaf rugose virus (CiLRV-3 [14], tobacco streak virus (TSV-3 [15, 16]), citrus variegation virus (CVV-3 [14]), apple mosaic virus (APMV-3 [17]), prune dwarf ilarvirus (PDV-3 [18]), lilac ring mottle virus (LRMV-3 [19]), elm mottle virus (EMV-3 [20]) and asparagus virus II (AVII [21]). Numbering of nucleotides is from the 3′end of RNA 3.

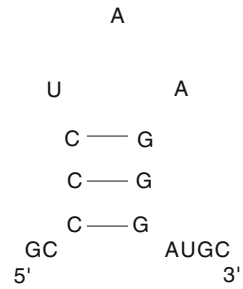

Figure 2. Substructure of AlMV-3.

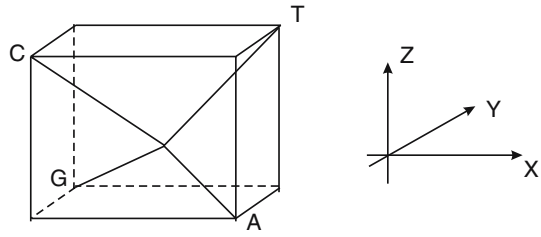

Figure 3. The tetrahedral directions assigned to A, G, C, T nucleic bases.

If A', U', G', C' are placed in the opposite vertices of A, U, G, C in edges parallel to y-axis, that is called Y-pattern. According this rule we get the coordinates of A', U', G', C' as follows

free base $A\ (+1, -1, -1)$, $U\ (+1, +1, +1)$,
$C\ (-1, -1, +1)$, $G\ (-1, +1, -1)$,
pair base $A'\ (+1, +1, -1)$, $U'\ (+1, -1, +1)$,
$C'\ (-1, +1, +1)$, $G'\ (-1, -1, -1)$.

If A', U', G', C' are placed in the opposite vertices of A, U, G, C in edges parallel to z-axis, that is called Z-pattern. According this rule we get the coordinates of A', U', G', C' as follows

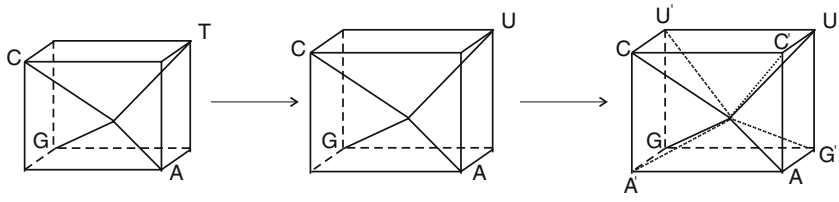

Figure 4. The square directions assigned to A, U, G, C free bases and A', U', G', C' pair bases based on x-pattern.

free base	A $(+1,-1,-1)$,	U $(+1,+1,+1)$,
	C $(-1,-1,+1)$,	G $(-1,+1,-1)$,
pair base	A' $(+1,-1,+1)$,	U' $(+1,+1,-1)$,
	C' $(-1,-1,-1)$,	G' $(-1,+1,+1)$.

Figures 5 and 6 correspond to the y-pattern and the z-pattern, respectively. Then, each RNA secondary structure corresponds to three curves.

Based on the x-pattern, we consider the beginning of AIMV-3:

The first point of the spatial curve is at point $(-1,-1,+1)$ which belongs to C, so directed from the origin. From here we move in the direction defined by $(-1,+1,-1)$ assigned to G telling that the first and the third coordinates have decreased while the second coordinate has increased. We arrive then at the point $(-2,0,0)$ as the location of G. From that point we move in the direction assigned to U, $(+1,+1,+1)$, which means that all the three coordinates of the position G $(-2,0,0)$, have to be increased by +1. This leads to point $(-1,+1,+1)$ as the location of U. The process continues, each time we algebraically add the (x, y, z) coordinates of the new point to that of the last point. This process is illustrated in table 1.

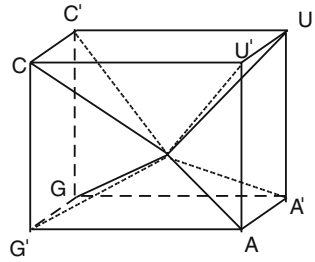

Figure 5. The square directions assigned to A, U, G, C free bases and A', U', G', C' pair bases based on the y-pattern.

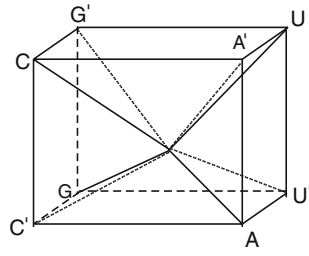

Figure 6. The square directions assigned to A, U, G, C free bases and A', U', G', C' pair bases based on the z-pattern.

Table 1
3D coordinates for the first five bases of AlMV-3 based on the x-pattern.

	AlMV-3			
no.	Base	x	y	z
1	C	−1	−1	1
2	G	−2	0	0
3	U	−1	+1	+1
4	A	0	0	0
5	G'	+1	+1	−1

Table 2
The leading eigenvalues of the L/L matrices associated with three essentially different patterns of the characteristic curves for the coding sequences of figure 1.

Patterns	AlMV-3	CiLRV-3	TSV-3	CVV-3	APMV-3	LRMV-3	PDV-3	EMV-3	AVII
X	8.2129	8.9379	8.5933	9.0861	8.4163	8.8831	8.5068	8.3774	8.7455
Y	7.6019	8.8317	8.8914	8.9873	8.0524	8.6754	7.8923	8.5079	8.7397
Z	7.5602	8.9093	8.7511	8.3225	7.9759	8.4234	7.6769	8.1222	8.4248

3. Similarities/dissimilarities among the RNA secondary structures of nine viruses

In order to find some of the invariants sensitive to the RNA secondary structure we will transform the 3D representation of the RNA secondary structure into another mathematical object, a matrix. Once we have a matrix representing a RNA secondary structure, we can use some of matrix invariants as descriptors of the sequences. One of the matrices is the L/L matrix, which are the same as Randic's [22]. Each curve corresponds to a L/L matrix, therefore one RNA secondary structure corresponds to the three L/L matrix.

We will characterize the coding sequences of the RNA secondary structure of nine species by means of the leading eigenvalue of the L/L matrix. In table 2 we give the leading eigenvalues of the L/L matrices associated with three essentially different patterns of the characteristic curves representing each of the coding sequences.

In table 3, we give the similarities and dissimilarities for the coding sequences of figure 1 based on the Euclidean distances between the end points of the 3-component vectors $(\lambda_1/n; \lambda_2/n; \lambda_3/n)$. The most similar are LRMV-3 and AVII with the lowest value 0.0068. The more similar are CVV-3 and EMV-3 with a value of 0.0069, CVV-3 and LRMV-3 with a value of 0.0074, LRMV-3 and EMV-3 with a value of 0.0078.

Table 3
The similarity/dissimilarity matrix for the coding sequences of figure 1 based on the Euclidean distances between the end points of the 3-component vectors of the normalized leading eigenvalues of the L/L matrices.

Species	AlMV-3	CiLRV-3	TSV-3	CVV-3	APMV-3	LRMV-3	PDV-3	EMV-3	AVII
AlMV-3	0	0.0457	0.0407	0.0541	0.0055	0.0582	0.0078	0.0530	0.0648
CiLRV-3		0	0.0092	0.0147	0.0428	0.0149	0.0396	0.0099	0.0205
TSV-3			0	0.0205	0.0370	0.0226	0.0350	0.0157	0.0277
CVV-3				0	0.0518	0.0074	0.0469	0.0069	0.0125
APMV-3					0	0.0559	0.0086	0.0502	0.0623
LRMV-3						0	0.0514	0.0078	0.0068
PDV-3							0	0.0463	0.0580
EMV-3								0	0.0126
AVII									0

Table 4
The similarity/dissimilarity matrix for the coding sequences of figure 1 based on the angle between the end points of the 3-component vectors of the normalized leading eigenvalues of the L/L matrices.

Species	AlMV-3	CiLRV-3	TSV-3	CVV-3	APMV-3	LRMV-3	PDV-3	EMV-3	AVII
AlMV-3	0	0.0351	0.0507	0.0324	0.0150	0.0223	0.0097	0.0385	0.0319
CiLRV-3		0	0.0188	0.0395	0.0210	0.0212	0.0414	0.0224	0.0192
TSV-3			0	0.0429	0.0357	0.0309	0.0549	0.0201	0.0228
CVV-3				0	0.0270	0.0196	0.0277	0.0229	0.0213
APMV-3					0	0.0095	0.0204	0.0248	0.0181
LRMV-3						0	0.0243	0.0162	0.0097
PDV-3							0	0.0397	0.0337
EMV-3								0	0.0067
AVII									0

In table 4, the similarities and dissimilarities for nine coding sequences that based on the correlation angle between two vectors. Observing table 4, we find the more similar species pairs are AVII-EMV-3, APMV-3-LRMV-3, PDV-3-AIMV-3 and AVII-LRMV-3.

Observing tables 3 and 4, we find that our result is similar to the results in [8, 11].

4. Conclusion

In this article we (1) outlined a construction of a 3D graphical representation of RNA secondary structures, illustrated on a portion of the nine viruses;

(2) described a particular scheme that allows 3D spatial representation of RNA to be transformed into a numerical matrix(L/L matrix) representation; and (3) based on the normalized leading eigenvalues from the L/L matrix, we computed similarities/dissimilarities among the RNA secondary structures of nine viruses. The advantage of our approach is that it allows visual inspection of data, helping in recognizing major similarities among different RNA structures or recognizing major differences among similar RNA structures. It is well-known that the alignments of RNA secondary structures are computer intensive that is direct comparison for RNA secondary structure. The structure invariant easily computed and compared is applied to compare RNA secondary structures, rather than strings' structure themselves.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China(Grant No.10571019).

References

[1] M. Randic, M. Vracko, A. Nandy and S.C. Basak, J. Chem. Inf. Comput. Sci. 40 (2000) 1235–1244.
[2] M. Randic, M. Vracko, N. Lers and D. plavsic, Chem. Phys. Lett. 371 (2003) 202–207.
[3] A. Nandy and P. Nandy, Chem. Phys. Lett. 368 (2003) 102–107.
[4] B. Liao and K. Ding, J. Comput. Chem. 26 (2005) 1519–1523.
[5] B. Liao, Chem. Phys. Lett. 401 (2005) 196–199.
[6] B. Liao, X. Shan, W. Zhu and R. Li, Chem. Phys. Lett. 422 (2006) 282–288.
[7] B. Liao, M. Tan and K. Ding, Chem. Phys. Lett. 414 (2005) 296–300.
[8] W. Zhu, B. Liao and K. Ding, J. Mol. Struct. THEOCHEM 757 (2005) 193–198.
[9] B. Liao, J. Luo, R. Li and W. Zhu, Int. J. Quantum. Chem. 106(8) (2006) 1749–1755.
[10] J. Luo, B. Liao, R. Li and W. Zhu, J. Math. Chem (online first).
[11] Y. Yao, X. Nan and T. Wang, J. Comput. Chem. 26 (2005) 1339–1346.
[12] Chantal B.E.M.Reusken and J.F. Bol, Nucl. Acids. Res. 14 (1996) 2660–2665.
[13] E.C. Koper-Zwarthoff, F.Th. Brederode, P. Walstra and J.F. Bol, Nucl. Acids. Res. 7 (1979) 1887–1900.
[14] S.W. Scott and X. Ge, J. Gen. Virol. 76 (1995) 957–963.
[15] E.C. Koper-Zwarthoff, F.Th. Brederode, P. Walstra and J.F. Bol, Nucl. Acids. Res. 8 (1980) 3307–3318.
[16] B.J. Cornelissen, H. Janssen, D. Zuidema and J.F. Bol, Nucl. Acids. Res. 12 (1984) 2427-2437.
[17] R.H. Alrefai, P. Shicl, L.L. Domier, C.J. D'Arcy, P.H. Berger and S.S. Korban, J. Gen. Virol. 75 (1994) 2847–2850.
[18] S.W. Scott and X. Ge, J. Gen. Virol. 76 (1995) 1801–1806.
[19] E.J. Bachman, S.W. Scott, G. Xin and V. Bowman Vance, Virology 201 (1994) 127–131.
[20] F. Houser-Scott, M.L. Baer, K.F. Liem, J.M. Cai and L. Gehrke, J. Virol 68 (1994) 2194–2205.
[21] EMBL/GenBank/DDBJ databases. Accession no.X86352.
[22] M. Randic, M. Vracko, N. Lers and D. Plavsic. Chem. Phys. Lett. 368 (2003) 1–6.