Journal of Molecular Structure: THEOCHEM 755 (2005) 131–136

THEO CHEM

www.elsevier.com/locate/theochem

A 2D graphical representation of RNA secondary structures and the analysis of similarity/dissimilarity based on it

Yu-hua Yao[a,*], Bo Liao[b], Tian-ming Wang[c]

[a]Department of Applied Mathematics, Dalian University of Technology, Dalian 116024, China
[b]School of Computer and Communication, Hunan University, Changsha Hunan 410082, China
[c]Department of Mathematics, Hainan Normal University, Haikou 571158, China

Received 22 March 2005; revised 8 August 2005; accepted 11 August 2005
Available online 10 October 2005

Abstract

A 2D graphical representation of RNA secondary structures is given in terms of classifications of bases of nucleic acids. The novel graphical representation can completely avoid loss of information associated with crossing and overlapping of the corresponding curve. Afterwards, we make quantitative analysis for a set of RNA secondary structures at the 3′-terminus of different viruses based on this graphical representation. The examination of similarities/dissimilarities illustrates the utility of the approach.

© 2005 Elsevier B.V. All rights reserved.

Keywords: RNA secondary structure; Graphical representation; Similarity; Virus

1. Introduction

Recently, several investigators have outlined graphical representations of DNA sequences. The advantage of graphical representations is that they allow visual inspection of data, helping in recognizing major differences among similar DNA sequences [1–13]. Another advantage of graphical representations of DNA sequences is the possibility to derive numerical characterization for DNA primary sequences. For example, to characterize DNA primary sequences, M. Randić introduced an approach to analyze the similarity among the coding sequences of the first exon of β-globin gene of 11 different species based on a 2D graphical representation presented by himself, rather than directly using string comparisons [1].

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. There are many algorithms for computing the similarity between RNA secondary structures [14–18]. Previously, almost all such comparisons are based on alignment of RNA structures: a distance function or a score function is used to represent insertion, deletion and substitution of letters in the compared structures. Using the distance function, one can compute similarity between RNA structures. It is well known that the alignments of RNA secondary structure are computer intensive. Sequences considered in alignment of RNA secondary structures are only string sequences. And there are a number of steps in such approaches that involve arbitrary decisions, e.g. decisions on the relative weight of different elementary string operations: deletion, insertion, substitution and penalties for unacceptable alignments.

The structural comparisons for RNA secondary structures based on the topological invariants of tree structures have been developed by Shapiro and Zhang [19], however the method does not directly use base paired nucleotides and unpaired nucleotides, and is not suitable to the RNA secondary structures with pseudoknots. Similar with the graphical representation of DNA sequences, Liao and Wang also outlined a 3D graphical representation of RNA secondary structures to compute the similarity of RNA

[*] Corresponding author. Tel.: +86 411 84701201; fax: +86 411 84706100.
E-mail address: yaoyuhua2288@163.com (Y.- Yao).

0166-1280/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.theochem.2005.08.009

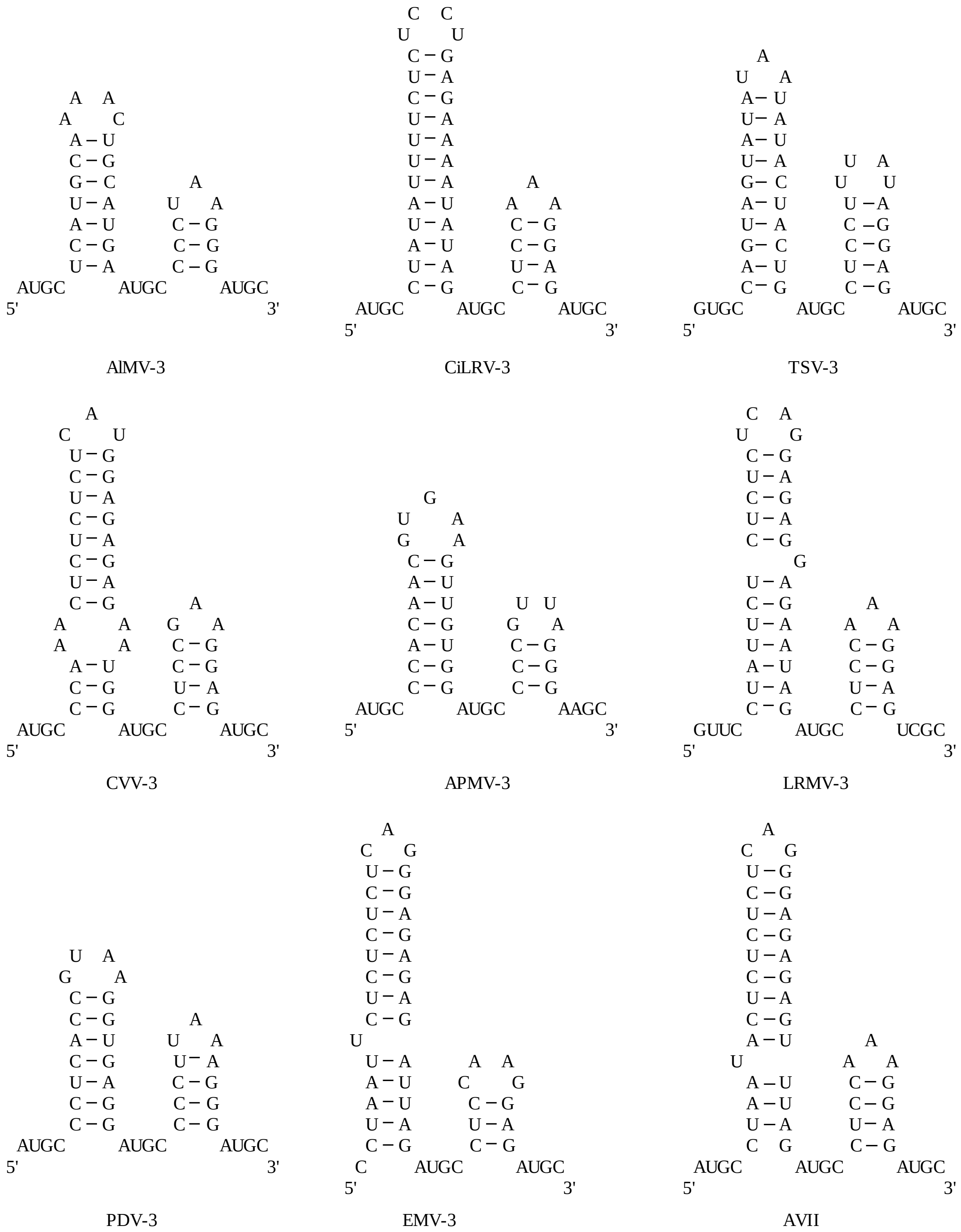

Fig. 1. Secondary structure at the 3′-terminus of RNA 3 of alfalfa mosaic virus (AlMV-3 [22]), citrus leaf rugose virus (CiLRV-3 [23]), tobacco streak virus (TSV-3 [24,25]), citrus variegation virus (CVV-3 [23]), apple mosaic virus (APMV-3 [26]), prune dwarf ilarvirus (PDV-3 [27]), lilac ring mottle virus (LRMV-3 [28]), elm mottle virus (EMV-3 [29]) and asparagus virus II (AVII [30]). Numbering of nucleotides is from the 3′end of RNA 3.

secondary structures [20]. Clearly, visualization of 3D graphical representation is not satisfying.

In this paper, a 2D graphical representation of RNA secondary structures is given in terms of classifications of bases of nucleic acids. Also, we will make a comparison for the secondary structures at the 3′-terminus belonging to nine different species based on this 2D graphical representation. In Fig. 1, the secondary structures at the 3′-terminus belonging to nine different viruses are listed, which were reported by Reusken and Bol [21]. The similarities are computed by calculating the Euclidean distance between the end point of the vectors or calculating the correlation

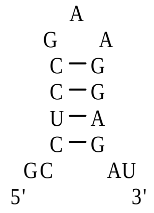

Fig. 2. Substructure of CVV-3.

angle of two vectors. Our method is suitable to the general RNA secondary structures (include the structures with pseudoknots).

2. Outline of the 2D graphical representation of RNA secondary structures

The secondary structure of an RNA is a set of free bases and base pairs forming hydrogen bonds between A–U and G–C (base pair G–U is frequently allowed). Let A′, U′, G′, C′ denote A, U, G, C in the base pair A–U, G–C or 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 CVV-3 (Fig. 2) is GCC′U′C′C′GAAG′G′A′G′AU (from 5′ to 3′).

Nucleic acids are linear macromolecules. Analysis and research of RNA should consider their chemical property. In RNA primary sequences, the four bases A, U, G, C can be divided into two classes according to the strength of the hydrogen bond, i.e. weak H-bonds W = {A, U} and strong H-bonds S = {G, C}. The bases can be divided into another two classes, amino group M = {A, C} and keto group K = {G, U}. Besides, the division can be also made according to their chemical structures, i.e. purine R = {A, G} and pyrimidine Y = {C, U}.

According to the above three classifications of bases of nucleic acids, we present a 2D graphical representation of RNA secondary structure consisting of three characteristic curves based on the four horizontal lines system [31]. We draw four horizontal lines separated by unit distances, on which dots (rectangles) representing the bases constituting the considered sequence are placed. The representation requires first to associate the four bases (base pairs) with the four horizontal lines. The weak H-bonds W = {A, U} and the strong H-bonds S = {G, C} are labeled to the middle two lines, the paired weak H-bonds bases {A′, U′} and the paired strong H-bonds bases {G′, C′} are labeled to the upper line and the lower line, respectively. The consecutive bases along the horizontal axes are placed at unit displacement. Connecting adjacent dots, we obtain a zigzag like curve that better visually illustrates the substructure considered. The corresponding plot set is called characteristic plot set. The curve connecting all plots of the characteristic plot set in turn is called a characteristic curve. For example, in Fig. 3, we draw the W–S curve of the substructure of CVV-3.

In Fig. 4, we draw the W–S characteristic curves for the secondary structures at the 3′-terminus belonging to nine different viruses of Fig. 1. Observing Fig. 4, we can find that the characteristic curves of LRMV-3, EMV-3 and AVII are very similar to each other, and characteristic curve of CVV-3 is also more similar to the above three characteristic curves.

3. Numerical characterization of RNA secondary structure

In order to find some of the invariants sensitive to the form of the characteristic curve we will transform the graphical representation of the characteristic curve into another mathematical object, a matrix. Once we have a matrix representing a DNA sequence, we can use some of matrix invariants as descriptors of the sequence. We also associate the characteristic curve with some symmetric matrices: *E*, *M*/*M*, *L*/*L*, ${}^{k}L/{}^{k}L$, and ${}^{b}L/{}^{b}L$ matrix, which are introduced by M. Randić [31].

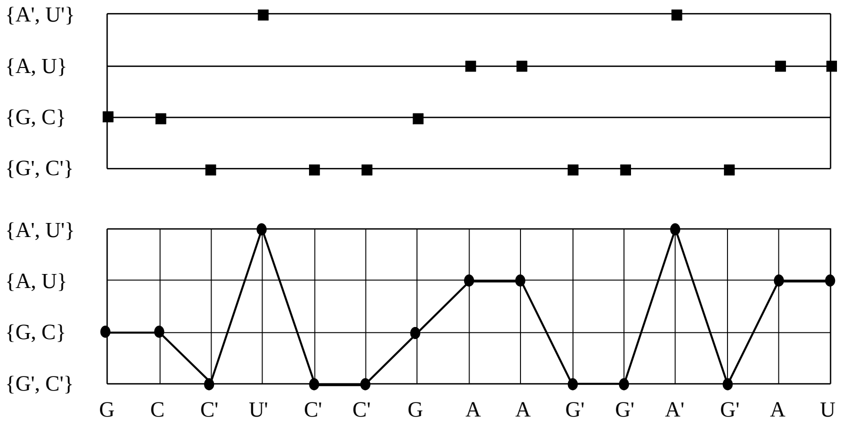

Fig. 3. The 2D graphical representation of the substructure of CVV-3. The rectangles (dots) denote the bases making up the sequence.

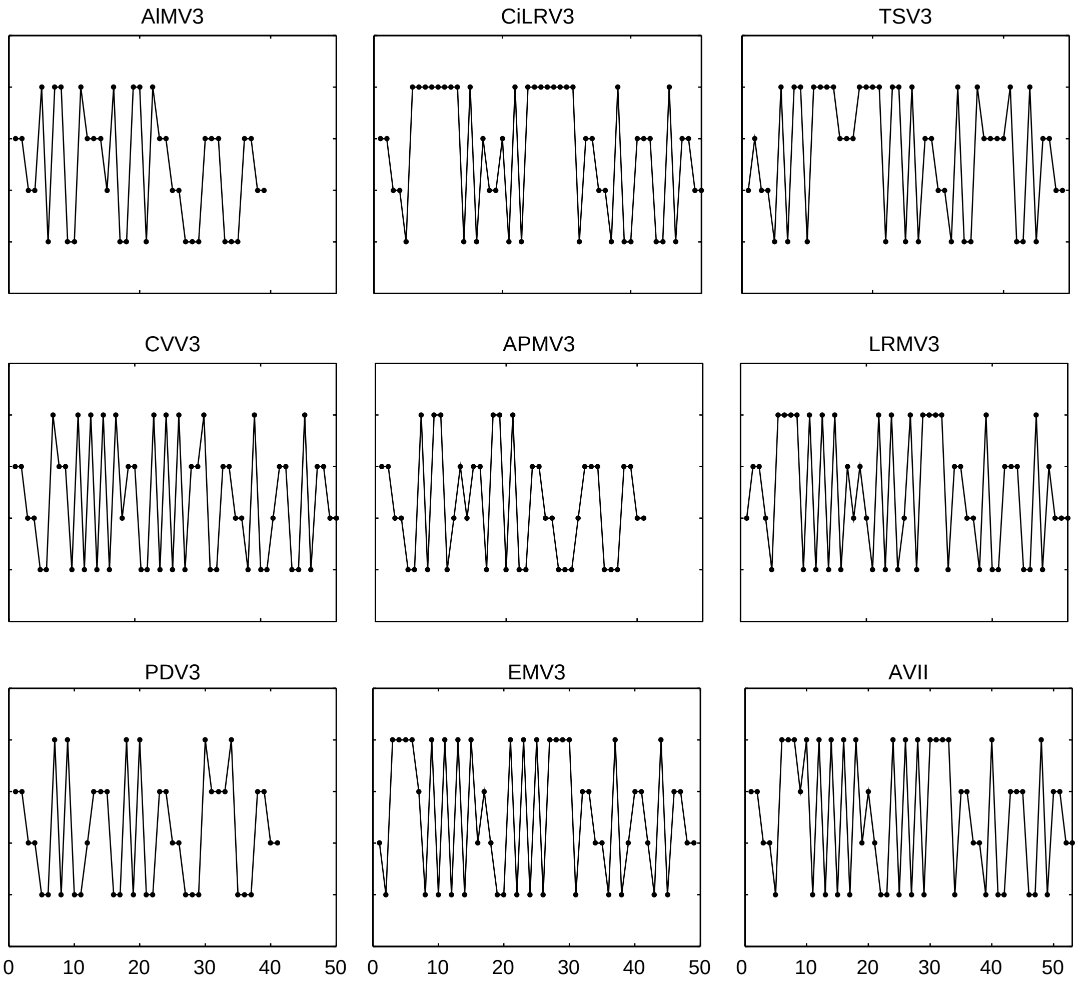

Fig. 4. The W–S characteristic curves for the secondary structures at the $3'$-terminus belonging to nine different viruses are illustrated in Fig. 1.

The eigenvalues of the E, ${}^{k}L/{}^{k}L$ ($k=1$, 2, 5, 10, 50) and ${}^{b}L/{}^{b}L$ matrix for the W–S curve of the substructure of CVV-3 are given in Table 1. There is some parallelism among the computed eigenvalues of the matrices, as could have been expected. The eigenvalues, and in particular the leading eigenvalues, can be used as descriptors of a RNA secondary structure. We choose the leading eigenvalues of L/L matrices as RNA secondary structure descriptors. Since

Table 1
The eigenvalues, λ_i ($i=1, 2,\ldots,15$), of the E, ${}^{k}L/{}^{k}L$ ($k=1$, 2, 5, 10, 50), and ${}^{b}L/{}^{b}L$ matrices of the substructure of CVV-3

Eigenvalue	Matrix						
	E	L/L	${}^{2}L/{}^{2}L$	${}^{5}L/{}^{5}L$	${}^{10}L/{}^{10}L$	${}^{50}L/{}^{50}L$	${}^{b}L/{}^{b}L$
λ_1	82.1448	9.1090	6.6043	4.1813	3.3268	2.4294	2.3827
λ_2	−0.6866	1.0121	1.7965	2.0241	1.9529	1.8457	1.8478
λ_3	−0.7527	0.5686	1.2630	1.8702	1.8596	1.7160	1.7144
λ_4	−0.8021	0.0078	0.6449	1.3433	1.4263	1.4065	1.4142
λ_5	−0.8220	−0.3462	0.0494	0.5514	0.7778	1.0316	1.0477
λ_6	−0.8752	−0.5592	−0.2871	0.1265	0.4153	0.7535	0.7654
λ_7	−1.2304	−0.8416	−0.7639	−0.5807	−0.3255	0.1858	0.2118
λ_8	−1.7329	−0.8868	−0.8232	−0.6556	−0.4194	−0.0119	0.0000
λ_9	−1.8440	−0.9996	−1.0026	−0.9992	−0.9169	−0.5934	−0.5859
λ_{10}	−2.4625	−1.0085	−1.0129	−1.0362	−0.9809	−0.7707	−0.7654
λ_{11}	−2.9327	−1.0209	−1.0398	−1.0812	−1.0985	−1.1723	−1.1728
λ_{12}	−4.8307	−1.0357	−1.0661	−1.1367	−1.2072	−1.4069	−1.4142
λ_{13}	−5.8350	−1.0935	−1.1596	−1.2949	−1.4225	−1.6325	−1.6414
λ_{14}	−13.5476	−1.3429	−1.4690	−1.5971	−1.6779	−1.8431	−1.8478
λ_{15}	−43.7905	−1.5628	−1.7337	−1.7153	−1.7099	−1.9375	−1.9566

Table 2
The leading eigenvalues of the *L*/*L* matrices associated with three essentially different patterns of the characteristic curves for the secondary structures at the 3′-terminus belonging to nine viruses of Fig. 1

Patterns	AlMV-3	CiLRV-3	TSV-3	CVV-3	APMV-3	LRMV-3	PDV-3	EMV-3	AVII
W–S	24.1920	30.6472	27.9983	26.1455	25.6403	26.4264	24.7416	24.5014	26.4307
M–K	25.3327	25.8672	24.9480	26.9156	32.6359	24.8512	26.9803	22.8753	26.2761
R–Y	23.7325	35.5733	28.8833	40.2375	25.7379	40.2400	29.5445	38.7424	40.6211

the characteristic curve does not represent the genuine molecular geometry, we are not interested in the interpretation of the leading eigenvalues of these matrices, but are interested in them as numerical parameters that may facilitate the comparisons of RNA secondary structure.

We will characterize the RNA secondary structures at the 3′-terminus of nine viruses in Fig. 1, by means of the leading eigenvalue of the *L*/*L* matrix. In Table 2, we show the leading eigenvalues of the *L*/*L* matrices associated with three essentially different patterns of the characteristic curves representing each of the RNA secondary structures. Observing Table 2, we can find that the largest leading eigenvalue occurs for pattern R–Y except for AlMV-3 and APMV-3.

4. Similarities/dissimilarities among the RNA secondary structures of nine virus

We shall illustrate the use of the 3D quantitative characterization of RNA secondary structures with an examination of similarities/dissimilarities among the nine virus of Fig. 1. We construct a three-component vectors consisting of the normalized leading eigenvalue λ/N, where λ is the leading eigenvalue of the matrix *L*/*L* and *N* is the number of bases making up the corresponding RNA secondary structure. The underlying assumption is that if two vectors point to a similar direction in three-dimensional space, then the two RNA secondary structures represented by the three-component vectors are similar.

The similarities among such vectors can be computed in three ways: (1) we calculate the Euclidean distance between the end point of the vectors; (2) we calculate the correlation angle of two vectors; and (3) we calculate the cosine of the correlation angle of two vectors. When one calculates the correlation angle of two vectors, the cosine of the correlation angle of two vectors is easily obtained. The smaller is the Euclidean distance between the end points of two vectors, the more similar are the RNA secondary structures. The smaller is the correlation angle between two vectors, the more similar are the RNA secondary structures. On the other hand, the larger is the cosine of the correlation angle between two vectors, the more similar are the RNA secondary structures.

In Table 3, we present the similarity/dissimilarity matrix for the secondary structures at the 3′-terminus belonging to nine viruses of Fig. 1 based on the Euclidean distances between the end points of the three-component vectors of the normalized leading eigenvalues of the *L*/*L* matrices. We believe that it is not accidental that the smallest entries in Table 3 are associated with the pairs (LRMV-3, AVII), (LRMV-3, EMV-3) and (CVV-3, AVII). On the other hand, the larger entries in the similarity/dissimilarity matrix appear in the rows belonging to APMV-3.

In Table 4, the similarity/dissimilarity matrix for the secondary structures at the 3′-terminus belonging to nine viruses of Fig. 1 based on the angle between the three-component vectors of the normalized leading eigenvalues of the *L*/*L* matrices are showed. Observing Table 4, we find that APMV-3 is very dissimilar to others among the nine species because its corresponding rows have larger entries. On the other hand, the more similar species pairs are (CVV-3, AVII), (LRMV-3, EMV-3), and (LRMV-3, AVII).

Table 3
The similarity/dissimilarity matrix for the secondary structures at the 3′-terminus belonging to nine viruses of Fig. 1 based on the Euclidean distances between the end points of the three-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	0.1690	0.1499	0.2419	0.1478	0.2634	0.1137	0.2846	0.2517
CiLRV-3		0	0.1120	0.1247	0.2981	0.1236	0.1526	0.1431	0.1238
TSV-3			0	0.1969	0.2944	0.1974	0.2010	0.2176	0.1918
CVV-3				0	0.3374	0.0401	0.1808	0.0536	0.0234
APMV-3					0	0.3691	0.1677	0.3881	0.3541
LRMV-3						0	0.2106	0.0217	0.0216
PDV-3							0	0.2284	0.1985
EMV-3								0	0.0378
AVII									0

Table 4
The similarity/dissimilarity matrix for the secondary structures at the $3'$-terminus belonging to nine viruses of Fig. 1 based on the angle between the three-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	0.1552	0.0892	0.2249	0.0904	0.2450	0.0851	0.2655	0.2337
CiLRV-3		0	0.0723	0.1182	0.2358	0.1179	0.1079	0.1365	0.1180
TSV-3			0	0.1724	0.1771	0.1819	0.0823	0.2019	0.1767
CVV-3				0	0.2771	0.0349	0.1428	0.0495	0.0144
APMV-3					0	0.3042	0.1386	0.3237	0.2888
LRMV-3						0	0.1669	0.0205	0.0206
PDV-3							0	0.1869	0.1531
EMV-3								0	0.0358
AVII									0

5. Conclusion

It is well known that the alignments of RNA secondary structures are computer intensive that is direct comparison for RNA secondary structures. Structure considered in alignments of RNA secondary structures is only string structures. Here, We represent the RNA secondary structures as 2D curves and make similarity analysis between RNA secondary structures. In our approach, the characteristic curves of RNA secondary structures are very simple, and the similarity can be computed easily. Also, our approach allows visual inspection of data, helping in recognizing major similarities among different RNA secondary structures. The mathematical invariants, normalized leading eigenvalues, are applied to compare RNA secondary structures, rather than string structure.

Acknowledgements

The author thank the anonymous referees for many valuable suggestions that have improved this manuscript.

References

[1] M. Randić, M. Vračko, N. Lerš, D. Plavšić, Chem. Phys. Lett. 371 (2003) 202.
[2] M. Randić, M. Vračko, J. Chem. Inf. Comput. Sci. 40 (2000) 599.
[3] A. Nandy, Curr. Sci. 66 (1994) 309.
[4] M. Randić, A.T. Balaban, J. Chem. Inf. Comput. Sci. 43 (2003) 532.
[5] B. Liao, T.M. Wang, J. Comput. Chem. 25 (11) (2004) 1364.
[6] B. Liao, T.M. Wang, Chem. Phys. Lett. 388 (2004) 195.
[7] C.X. Yuan, B. Liao, T.M. Wang, Chem. Phys. Lett. 379 (2003) 412.
[8] B. Liao, T.M. Wang, J. Comput. Chem. 25 (2004) 1364.
[9] B. Liao, T.M. Wang, Journal of Molecular Structure: THEOCHEM 681 (2004) 209.
[10] B. Liao, Chem. Phys. Lett. 401 (2005) 196.
[11] Y.H. Yao, T.M. Wang, Chem. Phys. Lett. 398 (2004) 318.
[12] B. Liao, Y.S. Zhang, K.Q. Ding, T.M. Wang, Journal of Molecular Structure: THEOCHEM 717 (2005) 199.
[13] B. Liao, M.S. Tan, K.Q. Ding, Chem. Phys. Lett. 402 (2005) 380.
[14] V. Bafna, S. Muthukrisnan, R. Ravi, Comput. Sci. 937 (1995) 1.
[15] F. Corpet, B. Michot, Comput. Appl. Biosci. 10 (4) (1995) 389.
[16] S.Y. Le, R. Nussinov, J.V. Mazel, Comput. Biomed. Res. 22 (1989) 461.
[17] S.Y. Le, J. Onens, R. Nussinov, J.H. Chen, B. Shapiro, J.R. Mazel, Comput. Biomed. Res. 5 (1989) 205.
[18] B. Shapiro, Comput. Appl. Biosci. 4 (3) (1998) 387.
[19] B. Shapiro, K. Zhang, Comput. Appl. Biosci. 6 (4) (1990) 309.
[20] B. Liao, T.M. Wang, J. Biomol. Str. Dyn. 21 (2004) 827.
[21] B.E.M. Chantal, J.F. Reusken, Bol, Nucl. Acids Res. 14 (1996) 2660.
[22] E.C. Koper-Zwarthoff, F.T. Brederode, P. Walstra, J.F. Bol, Nucl. Acids Res. 7 (1979) 1887.
[23] S.W. Scott, X. Ge, J. Gen. Virol. 76 (1995) 957.
[24] E.C. Koper-Zwarthoff, F.T. Brederode, P. Walstra, J.F. Bol, Nucl. Acids Res. 8 (1980) 3307.
[25] B.J. Cornelissen, H. Janssen, D. Zuidema, J.F. Bol, Nucl. Acids Res. 12 (1984) 2427.
[26] R.H. Alrefai, P.J. Shicl, L.L. Domier, C.J. D'Arcy, P.H. Berger, S.S. Korban, J. Gen. Virol. 75 (1994) 2847.
[27] S.W. Scott, X. Ge, J. Gen. Virol. 76 (1995) 1801.
[28] E.J. Bachman, S.W. Scott, G. Xin, V.B. Vance, Virology 201 (1994) 127.
[29] F. Houser-Scott, M.L. Baer, K.F. Liem, J.M. Cai, L. Gehrke, J. Virol. 68 (1994) 2194.
[30] EMBL/GenBank/DDBJ databases. Accession No. X86352.
[31] M. Randić, M. Vračko, N. Lerš, D. Plavšić, Chem. Phys. Lett. 368 (2003) 1.