MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **57** (2007) 687-695

ISSN 0340 - 6253

RNA secondary structure mathematical representation without degeneracy

Bo Liao *, Wen Zhu, Jiawei Luo, Renfa Li
School of Computer and Communication, Hunan University
Changsha Hunan 410082,China

Received September 21, 2006

Abstract. A 4D representation of RNA secondary structures using a four cartesian coordinates system has been derived for mathematical denotation of RNA structure. The four-dimensional representation also avoids loss of information accompanying alternative 2D and 3D representation in which the curve standing for RNA structure overlaps and intersects itself, and resolves structures' degeneracy. The RNA pseudoknot also can be represented as four-dimensional representations. The examination of similarities/dissimilarities among the secondary structures belonging to different species illustrates the utility of our approach.

1. Introduction

Mathematical analysis of the 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[1-15]. Graphical techniques have emerged as a very powerful tool for the visualization and analysis of long DNA sequences. These techniques provide useful insights into local and global characteristics and the occurrences, variations and repetition of the nucleotides along a sequence which are not as easily obtainable by other methods[11,22-26].

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 representations of DNA sequences, we also can outline several graphical representations of RNA primary sequences based on 2-D and 3-D representation to compute the similarity of RNA primary sequences. Current RNA secondary structure comparison algorithms have focused exclusively on tree structures owing to their relative simplicity for quantitative analysis[16-18]. But tree structures refer to mathematical constructs for RNA secondary structures without pseudoknots. So we should present a new

*Corresponding author E-mail: dragonbw@163.com;Fax: 86-731-8821715

representation to analyze and to compare RNA secondary structures with pseudoknots. Recently, we have proposed 3D, 6D and 7D representation of RNA secondary structures[19-21,29], but the representation is not unique or the properties is not valuable.

Here, we present a four-dimensional representation of RNA secondary structures, which has no circuit or degeneracy, so that the correspondence between RNA secondary structures and RNA graphs is one to one. We make comparison among six RNA secondary structures(see Figure 1)which were reported by Bol [27] and T.Schlick[28] to illustrate the utility of our approach.

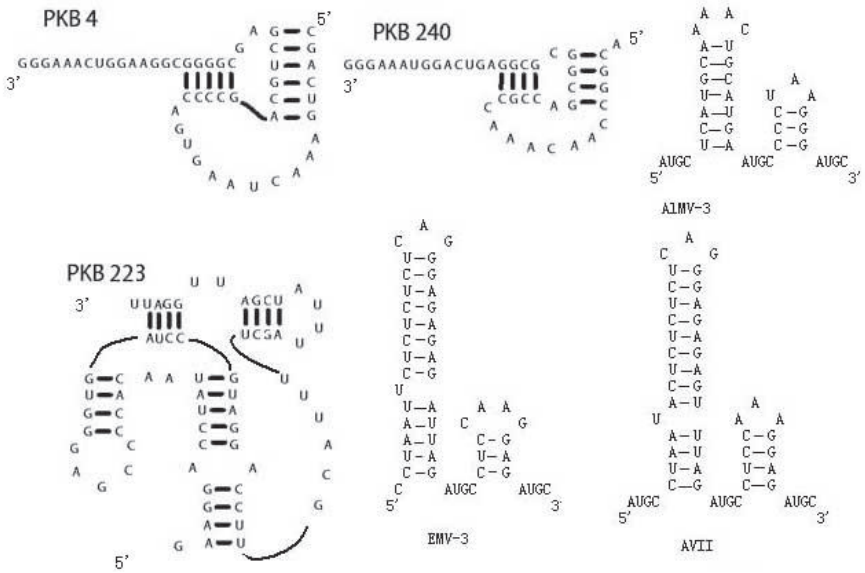

Figure 1: RNA secondary structures

2. 4-D 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. Let A', U', G', C' denote A, U, G, C in the base pair A-U and G-C, respectively. Then we can obtain a special sequence representation of the secondary structure. We call it characteristic sequence of the secondary structure. For example, pseudoknot B corresponds the characteristic sequence $C'U'G'G'C'G'AUUGCG'A'G'A'C'C'A'UGUC'G'C'$ $C'A'G'CUCU'G'G'U'C'U'C'CA$ (from $3'$ to $5'$)(see Figure 2).

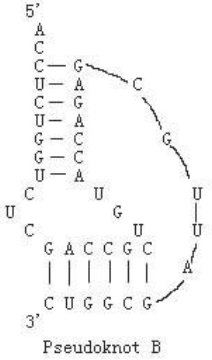

Figure 2: pseudoknot

We will illustrate the four-dimensional characterization of RNA secondary structure. We construct a map between the bases of characteristic sequences and plots in 4-D space, then we will obtain a 4-D representation of the corresponding RNA secondary structure. In 4-D space points, vectors and directions have four components, and we will assign the following basic elementary directions to the four free bases and two base pairs.

$(m, -\sqrt{n}, 0, 0) \longrightarrow A, (\sqrt{n}, -m, 0, 0) \longrightarrow G, (\sqrt{n}, m, 0, 0) \longrightarrow C, (m, \sqrt{n}, 0, 0) \longrightarrow U$

$(0, 0, s, -\sqrt{l}) \longrightarrow A', (0, 0, \sqrt{l}, -s) \longrightarrow G', (0, 0, \sqrt{l}, s) \longrightarrow C', (0, 0, s, \sqrt{l}) \longrightarrow U'$

where m, s is a real number, n, l is a positive real number but not a perfect square number. So that we will reduce a RNA secondary structure into a series of nodes $P_0, P_1, P_2, \ldots, P_N$, whose coordinates $x_i, y_i, z_i, v_i (i = 0, 1, 2, \ldots, N$, where N is the length of the RNA secondary structure being studied)satisfy

$$\begin{cases} x_i = a_i m + g_i\sqrt{n} + c_i\sqrt{n} + u_i m \\ y_i = -a_i\sqrt{n} - g_i m + c_i m + u_i\sqrt{n} \\ z_i = a'_i s + g'_i\sqrt{l} + c'_i\sqrt{l} + u'_i s \\ v_i = -a'_i\sqrt{l} - g'_i s + c'_i s + u'_i\sqrt{l} \end{cases}$$

where $a_i, c_i, g_i, u_i, a'_i, c'_i, g'_i$ and u'_i are the cumulative occurrence numbers of A, C, G , U, A', C', G', and U', respectively, in the subsequence from the 1st base to the i-th base in the sequence. We define $a_0 = c_0 = g_0 = u_0 = a'_0 = c'_0 = g'_0 = u'_0 = 0$.

3. Properties

Property 1 For a given RNA secondary structure there is a unique 4D representation corresponding to it.

Proof. Let (x_i, y_i, z_i, v_i) be the coordinates of the i-th base of RNA secondary structure, then we have

$a_i(m, -\sqrt{n}, 0, 0) + g_i(\sqrt{n}, -m, 0, 0) + c_i(\sqrt{n}, m, 0, 0) + u_i(m, \sqrt{n}, 0, 0) + a'_i(0, 0, s, -\sqrt{l}) + g'_i(0, 0, \sqrt{l}, -s) +$

$c'_i(0,0,\sqrt{l},s)+u'_i(0,0,s,\sqrt{l})=(x_i,y_i,z_i,v_i)$ i.e.

$$\begin{cases} a_i m+g_i\sqrt{n}+c_i\sqrt{n}+u_i m=x_i \\ -a_i\sqrt{n}-g_i m+c_i m+u_i\sqrt{n}=y_i \\ a'_i s+g'_i\sqrt{l}+c'_i\sqrt{l}+u'_i s=z_i \\ -a'_i\sqrt{l}-g'_i s+c'_i s+u'_i\sqrt{l}=v_i \end{cases} \tag{1}$$

Obviously, x_i and y_i are irrational numbers of form $jm+k\sqrt{n}$, while z_i and v_i are irrational numbers of form $bs+d\sqrt{l}$, where j,k,b and d are integers. We suppose

$$x_i=j_x m+k_x\sqrt{n}$$

$$y_i=j_y m+k_y\sqrt{n}$$

$$z_i=b_z s+d_z\sqrt{l}$$

$$v_i=b_v s+d_v\sqrt{l}$$

then we have

$$\begin{cases} a_i+u_i=j_x \\ g_i+c_i=k_x \\ -g_i+c_i=j_y \\ -a_i+u_i=k_y \\ a'_i+u'_i=b_z \\ g'_i+c'_i=d_z \\ -g'_i+c'_i=b_v \\ -a'_i+u'_i=d_v \end{cases} \tag{2}$$

So, for given x-projection, y-projection ,z-projection and v-projection of any point $P=(x,y,z,v)$ on the structure, after uniquely determining $j_x,k_x,j_y,k_y,b_z,d_z,b_v,d_v$ from x,y,z and v, the number $a_p,g_p,c_p,u_p,a'_p,g'_p,c'_p$ and u'_p of A, G, C, U, A',G',C' and U' from the beginning of the sequence to the point P can be found by solving linear system(2). By successive x-projection, y-projection, z-projection and v-projection of points on the sequence, we can recover the original RNA secondary structure uniquely from the RNA graph.

The vector pointing to the point P_i from the origin O is denoted by r_i. The component of r_i,i.e. x_i,y_i,z_i and v_i are calculated by Eq.(1). Let $\Delta r_i=r_i-r_{i-1}$, then we have Property 2.

Property 2 For any $i=1,2,\ldots,N$, where N is the length of the studied RNA secondary structure, the vector Δr_i has only four possible direction. Furthermore, the length of Δr_i,i.e.,$|\Delta r_i|$, is always equal to $\sqrt{m^2+n}$ or $\sqrt{s^2+l}$, for any $i=1,2,\ldots,N$.

Proof. Actually, the components of Δr_i,i.e., $\Delta x_i,\Delta y_i,\Delta z_i$ and Δv_i can be calculated for each possible residue (A,G,C, U, $A',G',C'$and U') at the i-th position of the RNA secondary structure by using Eq.(1). For example, when the i-th residue is A, we find $\Delta x_i=m,\Delta y_i=-\sqrt{n},\Delta z_i=0$ and $\Delta v_i=0$. This result is independent of the conformation state of the (i-1)-th residue. The four numbers $(m,-\sqrt{n},0,0)$ are called the direction of Δr_i. The direction number and the length of Δr_i for each possible residue type at the i-th position are summarized as follows(Table 1).

Property 3 There is no circuit or degeneracy in our four-dimensional representation.

Proof. We assume that:(1) the number of nucleotide forming a circuit is e; (2) the number of A,G,C, U, A',G',C' and U' in a circuit is $a_e,g_e,c_e,u_e,a'_e,g'_e,c'_e$ and u'_e, respectively. So

Table 1: Eight possible direction

	Δx_i	Δy_i	Δz_i	Δv_i	$\lvert\Delta r_i\rvert$
A	m	$-\sqrt{n}$	0	0	$\sqrt{m^2+n}$
G	$\sqrt{n}$	$-m$	0	0	$\sqrt{m^2+n}$
C	$\sqrt{n}$	m	0	0	$\sqrt{m^2+n}$
U	m	$\sqrt{n}$	0	0	$\sqrt{m^2+n}$
A'	0	0	s	$-\sqrt{l}$	$\sqrt{s^2+l}$
G'	0	0	$\sqrt{l}$	$-s$	$\sqrt{s^2+l}$
C'	0	0	$\sqrt{l}$	s	$\sqrt{s^2+l}$
U'	0	0	s	$\sqrt{l}$	$\sqrt{s^2+l}$

$a_e+g_e+c_e+u_e+a'_e+g'_e+c'_e+u'_e=e$. Because $a_eA, g_eG, c_eC, u_eU, a'_eA', g'_eG', c'_eC'$ and u'_eU form a circuit, the following equation holds: $a_e(m,-\sqrt{n},0,0)+g_e(\sqrt{n},-m,0,0)+c_e(\sqrt{n},m,0,0)+u_e(m,\sqrt{n},0,0)+a'_e(0,0,s,-\sqrt{l})+g'_e(0,0,\sqrt{l},-s)+c'_e(0,0,\sqrt{l},s)+u'_e(0,0,s,\sqrt{l})=(0,0,0,0)$ i.e.

$$\begin{cases} a_em+g_e\sqrt{n}+c_e\sqrt{n}+u_em=0 \\ -a_e\sqrt{n}-g_em+c_em+u_e\sqrt{n}=0 \\ a'_es+g'_e\sqrt{l}+c'_e\sqrt{l}+u'_es=0 \\ -a'_e\sqrt{l}-g'_es+c'_es+u'_e\sqrt{l}=0 \end{cases} \tag{3}$$

Clearly Eq.(3) hold if , and only if $a_e=g_e=c_e=u_e=a'_e=g'_e=c'_e=u'_e=0$. Therefore, $e=0$, which means no circuit exists in this graphical representation.

Property 4 The 4D representation possesses the reflection symmetry.

Proof. usually the sequence is expressed in the order from $5'$ to $3'$. Suppose that the 4D representation for RNA secondary structure is described by $(x_i,y_i,z_i,v_i), i=0,1,2,\ldots,N$. Suppose again that the 4D representation for the reverse structure, i.e, the same sequence but from $3'$ to $5'$ is described by $(\hat{x_i},\hat{y_i},\hat{z_i},\hat{v_i})$, we find

$$\begin{cases} \hat{x_i}=x_N-x_{N-i} \\ \hat{y_i}=y_N-y_{N-i} \\ \hat{z_i}=z_N-z_{N-i} \\ \hat{v_i}=v_N-v_{N-i} \end{cases} \tag{4}$$

4. Similarities/Dissimilarities

In this section, we will make a comparison for the secondary structures belonging to six different species(see Figure1) based on our 4D representation.

A direct comparison of these RNA secondary structures using computer codes is somewhat less straightforward due to the fact that the RNA secondary structures have different lengths. We will construct a 4-component vector consisting of the normalized leading eigenvalue of the L/L matrix and M/M matrix of the 4D representation with different parameters. Similar with M.Randic's methods[7], we define the L/L matrix and M/M matrix as following: The elements of the L/L matrix are defined as the quotient of the Euclidean distance between a pair of vertices(dots) corresponding the bases of RNA secondary structures and the sum of distances between the same pair of vertices. In other words, $L/L=(l_{i,j})$, where $l_{i,j}=\frac{d_{i,j}}{\sum_{k=i}^{j-1}d_{k,k+1}}$, $d_{i,j}=$

$\sqrt{(x_i-x_j)^2+(y_i-y_j)^2+(z_i-z_j)^2+(v_i-v_j)^2}$. while $M/M=(m_{i,j})$, where $m_{i,j}=\frac{d_{i,j}}{|i-j|}$. For our application, we will use the following four 4D representation:

Case a. Letting $m=1/2; n=3/4; s=1/3$ and $l=4/5$, then we get

$$\begin{cases} x_i=a_i/2+g_i\sqrt{3/4}+c_i\sqrt{3/4}+u_i/2 \\ y_i=-a_i\sqrt{3/4}-g_i/2+c_i/2+u_i\sqrt{3/4} \\ z_i=a_i'/3+g_i'\sqrt{4/5}+c_i'\sqrt{4/5}+u_i'/3 \\ v_i=-a_i'\sqrt{4/5}-g_i'/3+c_i'/3+u_i'\sqrt{4/5} \end{cases}$$

Case b. Letting $m=1/3; n=4/5; s=1/4$ and $l=1/3$, then we get

$$\begin{cases} x_i=a_i/3+g_i\sqrt{4/5}+c_i\sqrt{4/5}+u_i/3 \\ y_i=-a_i\sqrt{4/5}-g_i/3+c_i/3+u_i\sqrt{4/5} \\ z_i=a_i'/4+g_i'\sqrt{1/3}+c_i'\sqrt{1/3}+u_i'/4 \\ v_i=-a_i'\sqrt{1/3}-g_i'/4+c_i'1/4+u_i'\sqrt{1/3} \end{cases}$$

Case c. Letting $m=1/2; n=4/7; s=1/2$ and $l=5/9$, then we get

$$\begin{cases} x_i=a_i/2+g_i\sqrt{4/7}+c_i\sqrt{4/7}+u_i/2 \\ y_i=-a_i\sqrt{4/7}-g_i/2+c_i/2+u_i\sqrt{4/7} \\ z_i=a_i'/2+g_i'\sqrt{5/9}+c_i'\sqrt{5/9}+u_i'/2 \\ v_i=-a_i'\sqrt{5/9}-g_i'/2+c_i'/2+u_i'\sqrt{5/9} \end{cases}$$

Case d. Letting $m=1/4; n=9/11; s=3/4$ and $l=4/7$, then we get

$$\begin{cases} x_i=a_i/4+g_i\sqrt{9/11}+c_i\sqrt{9/11}+u_i/4 \\ y_i=-a_i\sqrt{9/11}-g_i/4+c_i/4+u_i\sqrt{9/11} \\ z_i=a_i'\times(\frac{3}{4})+g_i'\sqrt{4/7}+c_i'\sqrt{4/7}+u_i'\times(\frac{3}{4}) \\ v_i=-a_i'\sqrt{4/7}-g_i'\times(\frac{3}{4})+c_i'\times(\frac{3}{4})+u_i'\sqrt{4/7} \end{cases}$$

Table 2: The similarity/dissimilarity matrix for the six RNA secondary structures based on the Euclidean distances between the end points of the 4-component vectors of the normalized leading eigenvalues of the M/M matrices

Species	AlMV-3	pkb240	pkb 223	EMV-3	pkb 4	AVII
AlMV-3	0	0.1811	0.0929	0.1325	0.1762	0.1163
pkb240		0	0.1480	0.1532	0.0197	0.1645
pkb 223			0	0.0428	0.1351	0.0303
EMV-3				0	0.1370	0.0224
pkb 4					0	0.1496
AVII						0

In Table 2, we give the similarities and dissimilarities for the six RNA secondary structures based on the Euclidean distances between the end points of the 4-component vectors of the normalized leading eigenvalues of the M/M matrices. We believe that it is not accidental that the smallest entries in Table 2 are associated with the pairs(pkb240, pkb4), and (EMV-3, AVII). In Table 3, we give the similarities and dissimilarities for the six RNA secondary structures based on the Euclidean distances between the end points of the 4-component vectors of the normalized leading eigenvalues of the L/L matrices. The similarities of RNA secondary structures based

Table 3: The similarity/dissimilarity matrix for the six RNA secondary structures based on the Euclidean distances between the end points of the 4-component vectors of the normalized leading eigenvalues of the L/L matrices

Species	AlMV-3	pkb240	pkb 223	EMV-3	pkb 4	AVII
AlMV-3	0	0.1688	0.0461	0.0768	0.1638	0.0546
pkb240		0	0.1262	0.1310	0.0153	0.1441
pkb 223			0	0.0430	0.1198	0.0315
EMV-3				0	0.1208	0.0233
pkb 4					0	0.1352
AVII						0

on the normalized leading eigenvalues of the M/M matrices are compared with the similarities based on the normalized leading eigenvalues of the L/L matrices as illustrated in Figure 3. Entries that remain close to the line $y = x$ indicate one can obtain similar results using the two different methods.

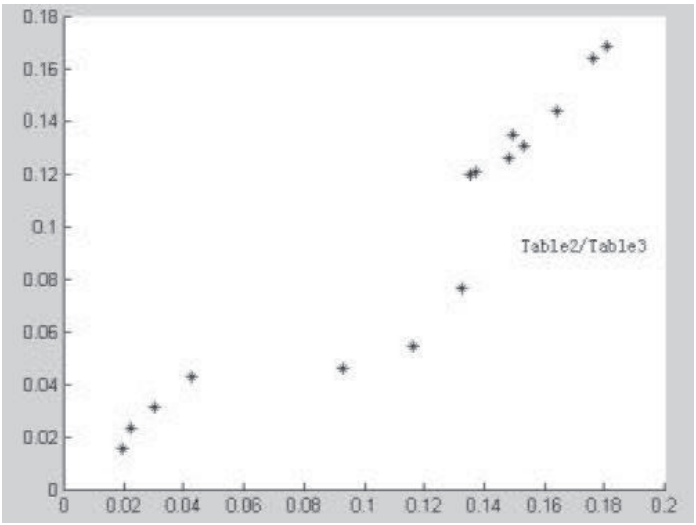

Figure 3: Comparison of Table 2 and Table 3

5. Conclusion

High complexity and degeneracy are major problems in previous RNA secondary structure representations. Our representation provides a direct plotting method to denote RNA secondary structures without degeneracy. From the RNA representation, the A,U,G,C,A-U and C-G usage as well as the original RNA structure can be recaptured mathematically without loss of textual information. The current four-dimensional representation of RNA secondary struc-

ture provides different approaches for both computational scientists and molecular biologists to analysis RNA secondary structures efficiently with different parameter n, m, s and l. A mathematical representation is presented and is applied to comparing the similarity between RNA secondary structures, the structure alignment is not deeded. In our future research, the current four-dimensional representation of RNA secondary structures and the similarity between RNA secondary structures will be applied to predict the biological function of RNA.

Acknowledgment

This work is supported by the National Natural Science Foundation of Hunan University.

References

[1] C. X. Yuan, B. Liao, T. M. Wang, New 3-D graphical representation of DNA sequences and their numerical characterization, Chem. Phys. Lett., 379(2003), 412-417.

[2] B. Liao, T. M. Wang, New 2D Graphical representation of DNA sequences, J. Comput. Chem, 25(2004), 1364-1368.

[3] B. Liao, T. M. Wang, 3-D graphical representation of DNA sequences and their numerical characterization, J. Mol. Struct.(THEOCHEM), 681(2004), 209-212

[4] B. Liao, T. M. Wang, Analysis of similarity of DNA sequences based on 3D graphical representation, Chem. Phys. Lett., 388(2004), 195-200.

[5] S. T. Yan, J. S. Wang, A. Niknejad, C. X. Lu, N. Jin, Y.K. Ho, DNA seuence representation without degeneracy, Nucl. Acids Res., 31(2003), 3078-3080.

[6] M. Randic, M. Vracko, A. Nandy, S. C. Basak, On 3-D graphical representation of DNA primary sequence and their numerical characterization, J. Chem. Inf. Comput. Sci., 40(2000), 1235-1244.

[7] M. Randic, M. Vracko, N. Lers, D. Plavsic, Novel 2-D graphical representation of DNA sequences and their numberical characterization, Chem. Phys. Lett., 368(2003), 1-6.

[8] E. Hamori, J. Ruskin, H curves, a novel method of representaion of nucleotides series especially suited for long DNA sequences, J. Biol. Chem., 258(1983), 1318-1327.

[9] E. Hamori, Novel DNA sequence representations, Nature, 314(1985),585-586.

[10] M. A. Gates, Simple DNA sequence representations, Nature, 316(1985), 219-219.

[11] A. Nandy, A new graphical representation and analysis of DNA sequence structure: Methodology and Application to Globin Genes, Curr. Sci., 66(1994), 309-314.

[12] A. Nandy, Two-dimensional graphical representation of DNA sequences and intron-exon discrimination in intron-rich sequences, Comput. Appl. Biosci., 12(1996), 55-62.

[13] B. Liao, A 2D graphical representation of DNA sequence, Chem. Phys. Lett., 401 (2005) 196-199.

[14] B. Liao, M. S. Tan, K. Q. Ding, A 4D representation of DNA sequences and its application, Chem. Phys. Lett., 402(2005), 380-383

[15] B. Liao, Y. S. Zhang, K. Q. Ding, T.M. Wang, Analysis of similarity/dissimilarity of DNA sequences based on a condensed curve representation, J. Mol. Struct.(THEOCHEM), 717(2005), 199-203.

[16] H. H. Gan, S. Pasquali, T. Schlick, Exploring the repertoire of RNA secondary motifs using graph theory:implications for RNA design, Nucl. Acids Res., 31(2003), 2926-2943.

[17] B. A. Shapiro, K. Z. Zhang, Comparing multiple RNA secondary structure using tree comparisons, Comput. Biomed. Res., 6(1990), 309-318.

[18] S. Y. Le, R. Nussinov, J. V. Maizel, Tree graphs of RNA secondary structures and their comparisons, Comput. Biomed. Res., 22(1989), 461-473.

[19] B. Liao, T. M. Wang, A 3D Graphical representation of RNA secondary structure, J. Biomol. Struc. Dyn. ,21(2004), 827-832.

[20] B. Liao, K. Q. Ding,T.M. Wang, On A Six-Dimensional Representation of RNA Secondary Structures, J. Biomol. Struc. Dyn., 22(2005), 455-464.

[21] B. Liao, T. M. Wang, K.Q. Ding, On A Seven-Dimensional Representation of RNA Secondary Structures, Mol. Simu., 31(14), 2005, 1063-1071.

[22] C. K. Peng, S. V. Buldyrev, A. L. Goldberger, Long-range correlations in nucleotide sequences, Nature, 356(1992),168-170.

[23] B. Liao , M. S. Tan, K. Q. Ding, Application of 2-D graphical representation of DNA sequence, Chem. Phys. Lett., 414(2005), 296-300.

[24] B. Liao , K. Q. Ding, Graphical Approach to Analyzing DNA Sequences, J. Comput. Chem., 14(26), 2005, 1519-1523.

[25] B. Liao, W. Zhu, Y. Lin, 3D graphical representation of DNA sequence without degeneracy and its applications in constructing philogenic tree, MATCH Commun. Math. Comput. Chem., 56(2006), 209-216.

[26] J. Gao, X. Zhang, Similarity analysys of RNA secondary structures based on 4D representation, MATCH Commun. Math. Comput. Chem., 56 (2006), 249-259.

[27] C. B. E. M. Reusken, J. F. Bol, Structural elements of the 30-terminal coat protein binding site in alfalfa mosaic virus RNAs, Nucl. Acids Res., 14(1996), 2660-2665.

[28] S. Pasquali, H. H. Gan, T. Schlick, Modular RNA architecture revealed by computational analysis of existing pseudoknots and ribosomal RNAs, Nucl. Acids Res., 33(2005), 1384-1398.

[29] J. W. Luo, B. Liao, R. F. Li, W. Zhu, RNA secondary structure 3D graphical representation without degeneracy, J. Math. Chem. 39(2006), 629-636.