Journal of Mathematical Chemistry, Vol. 42, No. 4, November 2007 (© 2006)
DOI: 10.1007/s10910-006-9156-z

On a four-dimensional representation of RNA secondary structures

Bo Liao*, Wen Zhu and Pengcheng Li

School of Computer and Communication, Hunan University, Changsha Hunan 410082, China
E-mail: dragonbw@163.com

Received 13 April 2006; revised 10 May 2006

We propose a 4-D representation of RNA secondary structures. The four-dimensional representation resolves structures' degeneracy and avoids loss of information and the limitation that different structures correspond the same plot set (or presentation). The RNA pseudoknpts also can be represented as four-dimensional representations. Based on this representation, we outline an approach to compute the similarities between six RNA secondary structures for illustrating the utility of our approach.

KEY WORDS: RNA secondary structure, similarity, virus, 4D representation

1. Introduction

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. Mathematical analysis, structure prediction, visualizing RNA secondary structures and computing structural properties of the large volume RNA secondary structure data are the challenges for bio-scientists.

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. There are many algorithms for computing the similarity between RNA secondary structures [1–8]. Current RNA secondary structure comparison algorithms have focused exclusively on tree structures owing to their relative simplicity for quantitative analysis [3,6,8]. But tree structures refer to mathematical constructs

*Corresponding author.

for RNA secondary structures without pseudoknots. While graphical representation of RNA secondary structure provides a simple way of viewing, sorting and comparing various gene structures. Graphical techniques have emerged as a very powerful tool for the visualization and analysis of big RNA structures. These techniques provide useful insights into local and global characteristics and the occurrences, variations and repetitions of the bases and base pairs along a structure which are not as easily obtainable by other methods. Recently, we have proposed 3D, 6D and 7D graphical representation of RNA secondary structures [8–11,14], Bai proposed a random walk representation [12], Yao et al. [13] proposed a class of 2D graphical representations, but the representation is not unique. And, in these methods, different structures can correspond the same characteristic sequence so that different structures correspond the same representation. For example, there are two different RNA secondary structures (a) and (b) (see figure 1), one can obtain the same curve using the previous methods.

In this paper, we propose a 4-D representation of RNA secondary structures and outline an approach to make mathematical analysis and to compute the similarities between RNA secondary structures. The presented method avoids the limitation of the previous methods.

2. 4-D representation of RNA secondary structures

The secondary structure of an RNA is a set of single 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

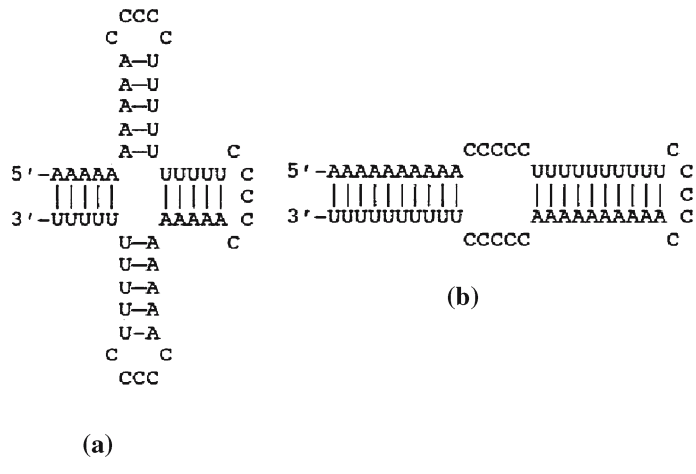

Figure 1. Two RNA secondary structures.

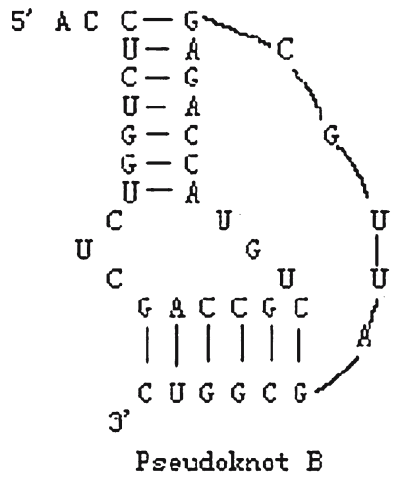

Figure 2. Pseudoknot.

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).

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.

$$\begin{cases} x_n = (\mathrm{A}_n + \mathrm{G}_n) - (\mathrm{C}_n + \mathrm{U}_n) \\ y_n = (\mathrm{A}_n + \mathrm{C}_n) - (\mathrm{G}_n + \mathrm{U}_n) \\ z_n = (\mathrm{A}_n + \mathrm{U}_n) - (\mathrm{C}_n + \mathrm{G}_n) \\ s_n = (\mathrm{A}'_n + \mathrm{U}'_n) - (\mathrm{C}'_n + \mathrm{G}'_n) \end{cases}$$

$\mathrm{A}'_n, \mathrm{U}'_n, \mathrm{C}'_n$ and G'_n satisfy

$$\begin{cases} \mathrm{A}'_n = \mathrm{A}^1_n + \sqrt{s_1}\mathrm{A}^2_n + \sqrt{s_2}\mathrm{A}^3_n + \sqrt{s_3}\mathrm{A}^4_n + \sqrt{s_4}\mathrm{A}^5_n \\ \mathrm{U}'_n = \mathrm{U}^1_n + \sqrt{s_1}\mathrm{U}^2_n + \sqrt{s_2}\mathrm{U}^3_n + \sqrt{s_3}\mathrm{U}^4_n + \sqrt{s_4}\mathrm{U}^5_n \\ \mathrm{G}'_n = \mathrm{G}^1_n + \sqrt{s_1}\mathrm{G}^2_n + \sqrt{s_2}\mathrm{G}^3_n + \sqrt{s_3}\mathrm{G}^4_n + \sqrt{s_4}\mathrm{G}^5_n \\ \mathrm{C}'_n = \mathrm{C}^1_n + \sqrt{s_1}\mathrm{C}^2_n + \sqrt{s_2}\mathrm{C}^3_n + \sqrt{s_3}\mathrm{C}^4_n + \sqrt{s_4}\mathrm{C}^5_n \end{cases},$$

where $\mathrm{A}_n, \mathrm{C}_n, \mathrm{G}_n$ and U_n are the cumulative occurrence numbers of A, C, G and U, respectively, in the subsequence from the 1st base to the nth base in the sequence. $\mathrm{A}^1_n, \mathrm{C}^1_n, \mathrm{G}^1_n$ and U^1_n are the cumulative occurrence numbers of $\mathrm{A}', \mathrm{C}', \mathrm{G}'$, and U' belonging to hairpin loop, respectively; $\mathrm{A}^2_n, \mathrm{C}^2_n, \mathrm{G}^2_n$ and U^2_n are the cumulative occurrence numbers of $\mathrm{A}', \mathrm{C}', \mathrm{G}'$, and U' belonging to interior loop, respectively. $\mathrm{A}^3_n, \mathrm{C}^3_n, \mathrm{G}^3_n$ and U^3_n are the cumulative occurrence numbers of $\mathrm{A}', \mathrm{C}', \mathrm{G}'$, and U' belonging to bulge, respectively; $\mathrm{A}^4_n, \mathrm{C}^4_n, \mathrm{G}^4_n$ and U^4_n are the

cumulative occurrence numbers of A', C', G', and U' belonging to multibranch loop, respectively; A_n^5, C_n^5, G_n^5 and U_n^5 are the cumulative occurrence numbers of A', C', G', and U' belonging to pseudoknot, respectively. s_k and $s_k u, k = 1, \ldots, 4$ are positive real number but not perfect square number, $s_i \neq s_j, i, j = 1, \ldots, 4$. We define $A_0 = C_0 = G_0 = U_0 = A_0^k = C_0^k = G_0^k = U_0^k = 0, k = 1, \ldots, 5$.

There are several major features of the 4D representation as follows: First, the xyz-components of a 4D representation displays a z-curve of RNA primary sequence. That is to say, if $s_n \equiv 0$, then the 4D representation display a z-curve of RNA primary sequence. Second, for a given RNA secondary structure there is a unique 4D representation corresponding to it. Third, the terminal coordinate of a 4D representation for a given RNA secondary structure depends only on the nucleotide and the base pair composition of the structure, and is independent of the structure order of the RNA bases. Fourth, when appending one more free base to the RNA secondary structure studied, the increments of the x, y and z coordinates of the corresponding 4D representation must be equal either to $+1$ or -1, whatever the appending base is A,C,G or U. When appending one base pair to the RNA secondary structure studied, the increments of the s coordinate of the corresponding 4D representation must be equal either to $+2$ or -2, whatever the appending base pair is A–U or G–C. When appending one base matching a previous base to the RNA secondary structure studied, the increments of the x, y and z coordinates of the corresponding 4D representation must be equal either to $+1$ or -1, and the increments of the s coordinate of the corresponding 4D representation must be equal either to $+2$ or -2. Fifth, there are 13 transforms of RNA secondary bases and base pairs(see table 1).

Table 1
Thirteen transforms of RNA secondary structure and the corresponding transform table of the coordinates.

Transform	A	C	G	U	A–U	C–G	x	y	z	s
I	A	C	G	U	A–U	C–G	x	y	z	s
R_x	G	U	A	C	A–U	C–G	x	$-y$	$-z$	s
R_y	C	A	U	G	A–U	C–G	$-x$	y	$-z$	s
R_z	U	G	C	A	A–U	C–G	$-x$	$-y$	z	s
R_s	A	C	G	U	C–G	A–U	x	y	z	$-s$
R_{A}	A	U	C	G	A–U	C–G	z	s	y	s
R_{C}	G	C	U	A	A–U	C–G	z	$-x$	$-y$	s
R_{G}	U	A	G	C	A–U	C–G	$-z$	$-x$	y	s
R_{A}^2	A	G	U	C	A–U	C–G	y	z	x	s
R_{C}^2	U	C	A	G	A–U	C–G	$-y$	$-z$	x	s
R_{G}^2	C	U	G	A	A–U	C–G	$-y$	z	$-x$	s
R_{U}^2	G	A	C	U	A–U	C–G	y	$-z$	$-x$	s

Table 2
Properties of mutations.

	Δx_i	Δy_i	Δz_i	Δs_i	Direction
A → C	−2	0	−2	0	(−2, 0, −2, 0)
C → A	2	0	2	0	(2, 0, 2, 0)
A → G	0	−2	−2	0	(0, −2, −2, 0)
G → A	0	2	2	0	(0, 2, 2, 0)
A → U	−2	−2	0	0	(−2, −2, 0, 0)
U → A	2	2	0	0	(2, 2, 0, 0)
C → G	2	−2	0	0	(2, −2, 0, 0)
G → C	−2	2	0	0	(−2, 2, 0, 0)
C → U	0	−2	2	0	(0, −2, 2, 0)
U → C	0	2	−2	0	(0, 2, −2, 0)
G → U	−2	0	2	0	(−2, 0, 2, 0)
U → G	2	0	−2	0	(2, 0, −2, 0)
A–U → G–C	0	0	0	$-4\sqrt{s_k}$	$(0, 0, 0, -4\sqrt{s_k})$
G–C → A–U	0	0	0	$4\sqrt{s_k}$	$(0, 0, 0, 4\sqrt{s_k})$
A → Φ	−1	−1	−1	0	(−1, −1, −1, 0)
Φ → A	1	1	1	0	(1, 1, 1, 0)
C → Φ	1	−1	1	0	(1, −1, 1, 0)
Φ → C	−1	1	−1	0	(−1, 1, −1, 0)
G → Φ	−1	1	1	0	(−1, 1, 1, 0)
Φ → G	1	−1	−1	0	(1, −1, −1, 0)
U → Φ	1	1	−1	0	(1, 1, −1, 0)
Φ → U	−1	−1	1	0	(−1, −1, 1, 0)
G–C → Φ	0	0	0	$2\sqrt{s_k}$	$(0, 0, 0, 2\sqrt{s_k})$
Φ → G–C	0	0	0	$-2\sqrt{s_k}$	$(0, 0, 0, -2\sqrt{s_k})$
A–U → Φ	0	0	0	$-2\sqrt{s_k}$	$(0, 0, 0, -2\sqrt{s_k})$
Φ → A–U	0	0	0	$2\sqrt{s_k}$	$(0, 0, 0, 2\sqrt{s_k})$

$\Omega \to \Phi$ corresponds a deletion, while $\Phi \to \Omega$ corresponds a insertion, $\Omega \in \{A, C, G, U, A–U, G–C\}$; $k = 0, 1, \ldots, 4$; $s_0 = 1$, which corresponds a hairpin loop.

For example, the operation R_s is also called exchange of the base pair, in which

$$\text{A–U} \Longleftrightarrow \text{C–G}$$
$$x \Longleftrightarrow x,\ y \Longleftrightarrow y,\ z \Longleftrightarrow z,\ s \Longleftrightarrow -s.$$

As a general rule, there are four basic types of changes in DNA or RNA. They are substitution of a nucleotide(or base pair) for another nucleotide (or base pair), deletion of nucleotides (or base pair), insertion of nucleotides (or base pair), and inversion of nucleotides. We shall consider the properties of mutations based on this 4D representation of RNA secondary structure. We assume the mutation appear on the ith base. Let (x_i, y_i), (x_i', y_i') be the coordinates of the primal base and mutational base, respectively. $\Delta x_i = x_i' - x_i$, $\Delta y_i = y_i' - y_i$. The two numbers $(\Delta x_i, \Delta y_i)$ are called the direction of the mutation. In table 2, we list the properties of mutations.

3. Similarities/dissimilarities

For any RNA secondary structures, we have a set of points (x_i, y_i, z_i, s_i), $i = 1, 2, 3, \ldots, N$, where N is the length of the structure. The coordinates of the geometrical center of the points, denoted by x^0, y^0, z^0 and s^0, may be calculated as follows:

$$x^0 = \frac{1}{N}\sum_{i=1}^{N} x_i,\ y^0 = \frac{1}{N}\sum_{i=1}^{N} y_i,\ z^0 = \frac{1}{N}\sum_{i=1}^{N} z_i,\ s^0 = \frac{1}{N}\sum_{i=1}^{N} s_i. \quad (1)$$

The covariance matrix CM of the points are defined:

$$\mathrm{CM} = \begin{pmatrix} \mathrm{CM}_{xx} & \mathrm{CM}_{xy} & \mathrm{CM}_{xz} & \mathrm{CM}_{xs} \\ \mathrm{CM}_{yx} & \mathrm{CM}_{yy} & \mathrm{CM}_{yz} & \mathrm{CM}_{ys} \\ \mathrm{CM}_{zx} & \mathrm{CM}_{zy} & \mathrm{CM}_{zz} & \mathrm{CM}_{zs} \\ \mathrm{CM}_{sx} & \mathrm{CM}_{sy} & \mathrm{CM}_{sz} & \mathrm{CM}_{ss} \end{pmatrix}.$$

The elements of covariance matrix CM satisfy

$$\begin{cases} \mathrm{CM}_{xx} = \frac{1}{N}\sum_1^N (x_i - x^0)(x_i - x^0) \\ \mathrm{CM}_{xy} = \frac{1}{N}\sum_1^N (x_i - x^0)(y_i - y^0) = \mathrm{CM}_{yx} \\ \mathrm{CM}_{xz} = \frac{1}{N}\sum_1^N (x_i - x^0)(z_i - z^0) = \mathrm{CM}_{zx} \\ \mathrm{CM}_{xs} = \frac{1}{N}\sum_1^N (x_i - x^0)(s_i - s^0) = \mathrm{CM}_{sx} \\ \mathrm{CM}_{yy} = \frac{1}{N}\sum_1^N (y_i - y^0)(y_i - y^0) \\ \mathrm{CM}_{yz} = \frac{1}{N}\sum_1^N (z_i - z^0)(y_i - y^0) = \mathrm{CM}_{zy} \\ \mathrm{CM}_{ys} = \frac{1}{N}\sum_1^N (s_i - s^0)(y_i - y^0) = \mathrm{CM}_{sy} \\ \mathrm{CM}_{zz} = \frac{1}{N}\sum_1^N (z_i - z^0)(z_i - z^0) \\ \mathrm{CM}_{zs} = \frac{1}{N}\sum_1^N (z_i - z^0)(s_i - s^0) = \mathrm{CM}_{sz} \\ \mathrm{CM}_{ss} = \frac{1}{N}\sum_1^N (s_i - s^0)(s_i - s^0) \end{cases} \quad (2)$$

The above 10 numbers give a quantitative description of a set of point (x_i, y_i, z_i, s_i), $i = 1, 2, \ldots, N$, scattering in a four-dimensional space. Obviously, the matrix is a real symmetric 4×4 one. There is a leading eigenvalue for a matrix CM. So that there are four eigenvalues corresponding a RNA secondary structure. We will construct a four-component vector consisting of the four eigenvalues to compute the similarity between RNA secondary structures. In order to facilitate the quantitative comparison of different species in terms of their collective parameters, we introduce a distance scale as defined below. Suppose that there are two species i and j, the parameters are $\lambda_1^i, \lambda_2^i, \lambda_3^i, \lambda_4^i, \lambda_1^j, \lambda_2^j, \lambda_3^j, \lambda_4^j$, respectively, where $\lambda_1^i, \lambda_2^i, \lambda_3^i, \lambda_4^i$ are the four eigenvalues of matrix CM_i corresponding to species i. The distance d_{ij} between the two points is:

$$d_{ij} = \sqrt{(\lambda_1^i - \lambda_1^j)^2 + (\lambda_2^i - \lambda_2^j)^2 + (\lambda_3^i - \lambda_3^j)^2 + (\lambda_4^i - \lambda_4^j)^2}, \quad i, j = 1, 2, \ldots, M \quad (3)$$

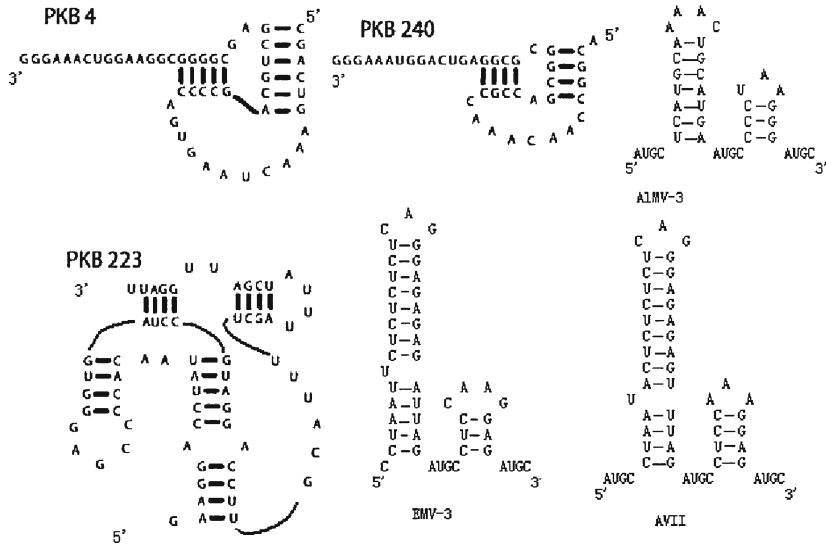

Figure 3. RNA secondary structures.

where d_{ij} denotes the distance between the end points of the vectors representing the ith and the jth genomes, and M is the total number of all genomes. Then we obtain a real $M \times M$ symmetric matrix whose elements are d_{ij}. As an example, we compute the similarities between six RNA secondary structures (see figure 3) to illustrate the utility of our approach.

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 four-component vectors of the eigenvalues of the CM matrices. We believe that it is not accidental that the smallest entries in table 3 are associated with the pairs (pkb240, pkb4),(AlMV-3-AVII) and (EMV-3, AVII).

Table 3
The similarity/dissimilarity matrix for the six RNA secondary structures based on the Euclidean distances between the end points of the four-component vectors of the eigenvalues of the CM matrices.

Species	AlMV-3	pkb240	pkb 4	pkb 223	EMV-3	AVII
AlMV-3	0	1.2774e+003	1.3884e+003	331.4504	302.2486	145.7325
pkb240		0	127.2905	1.2289e+003	1.5751e+003	1.4184e+003
pkb 223			0	1.3226e+003	1.6877e+003	1.5305e+003
EMV-3				0	533.5348	413.7371
pkb 4					0	159.2839
AVII						0

4. Conclusion

We have presented a presentation of RNA secondary structure and outlined an approach to compute the similarities between RNA secondary structures. The advantage of our approach is that it allows one to construct a numerical characterization and use it to judge the mutation. It is well-known that the alignments of RNA secondary structures are computer intensive. In alignments of RNA structures are considers only string's structures. Here we use an intensive approach which considers not only sequence structures but also chemical structures for RNA secondary structures.

Acknowledgments

This work is supported in part by the China Postdoctoral Science Foundation and the National Natural Science Foundation of Hunan University.

References

[1] V. Bafna, S. Muthukrisnan and R. Ravi, Comput. Sci. 937 (1995) 1
[2] F. Corpet and B. Michot, Comput. Appl. Biosci. 10 (1995) 389
[3] S.Y. Le, R. Nussinov and J.V. Mazel, Comput. Biomed. Res. 22 (1989) 461
[4] S.Y. Le, J. Onens, R. Nussinov, J.H. Chen, B. Shapiro and J.R. Mazel, Comput. Biomed. 5 (1989) 205
[5] B. Shapiro, Comput. Appl. Biosci. 4(3) (1998) 387
[6] B. Shapiro and K. Zhang, Comput. Appl. Biosci. 6(4) (1990) 309
[7] K. Zhang, in: *Pro. IEEE. Internat. Joint Symp On Intelligence and Sytems Rockviue, Maryland*, May, (1998) 126–132
[8] H.H. Gan, S. Pasquali and T. Schlick, Nuclei Acids Res. 31 (2003) 2926
[9] B. Liao and T. Wang, J. Biomol. struct. Dyn. 21 (2004) 827
[10] B. Liao, K.Q. Ding and T.M. Wang, J. Biomol. Struct. Dyn. 22 (2005) 455
[11] B. Liao and T.M. Wang, Mol. Simulat. 14 (2005) 1063
[12] F.L. Bai, W. Zhu and T.M. Wang, Chem. Phys. Lett. 408 (2005) 258
[13] Y.H. Yao, X.Y. Nan and T.M. Wang, J. Comput. Chem. 26 (2005) 1339
[14] W. Zhu, B. Liao and K.Q. Ding, J. Mole. Struct. THEOCHEM. 757 (2005) 193