Numerical characterization of RNA secondary structure

Jiaquan Zhan¹, Bo Liao,^{2,*} Yusen Zhang³

¹ Management Dept Huangdao District Qingdao Institute of Architecture & Engineering

Qingdao 266520, China

² Science 100, Graduate School of the Chinese Academy of Sciences

Beijing 100049, China

³ School of information Engineer, Shandong University at weihai Weihai 264209,China

Abstract: A secondary structure is a symbolic string composed of three kinds letters indicating the stack, external element and loop. A 2D graphical representation for this abstract symbolic sequence is proposed here. The curve is the unique representation for a given RNA secondary structure. Different geometrical properties of the curve are studies in details, which reflect the basic characteristics of the RNA secondary structure. Some characteristic matrices are derived from the definition of RNA secondary structure.

Key words: RNA secondary structure; 2D graphical representation; curve

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.

Almost all comparison of primary RNA structures are based on the comparison of strings. As is well-known, string comparisons are computer intensive, and despite the fact that practical schemes for sequence comparison have been outlined, there are a number of steeps in such approaches that involve arbitrary decisions e.g., decisions on the relative weights of different elementary string operations: deletion, insertions, substitution, and penalties for unacceptable alignments. The similarity between two structures have been formulated as problems of exact and approximate structure matching, finding a largest common substructure of the structures and computing optimal alignments under general scoring functions [1-8]. In order to find the numerical characterizations of structures, several author study the secondary structures using mathematical model approaches.[9-13]

In this paper, based on the special representation of three kinds of letters indicating the stack, external element and loop, we shall propose a 2-D graphical representation. Each RNA secondary structure corresponds to a unique curve representation and vice versa. In other words, each can be uniquely determined given the other. Therefore, the curve contains all the information that the secondary structure contains. It is found that the format of the curve can be of some advantages. Based on the mathematical definition of RNA secondary structure, some characteristic matrices are derived.

2 METHODS and RESULTS

^{*} Correspondence author; phone: 86-10-88256148; fax: 86-10-88256147; E-mail: dragonbw@163.com

2.1 Characteristic Matrices and the connectivity index

Definition (Waterman [1]): A secondary structure is a vertex-labeled graph on n vertices with an adjacency matrix **A** fulfilling

(i) $a_{i,i+1} = 1$ for $1 \le i \le n-1$

(ii) For each *i* there is at most a single $k \neq i-1, i+1$ such that $a_{i,k} = 1$

(iii) If $a_{i,j} = a_{k,l} = 1$ and i < k < j then i < l < j

Let $\mathbf{M} = \mathbf{A}\mathbf{D}^{-1}$, where **A** is the vertex-adjacency matrix, **D** is the diagonal matrix with the elements $\mathbf{D} = (d_{ii} = d_i)$ the number of vertex connecting i. Similar as D.J. Klein's approach[14], we introduce the following matrices:

$$H = D^{-1/2} M D^{1/2}$$

$$L = D - A = D^{1/2} (I - H) D^{1/2}$$

Where **I-H** is what sometimes called the normalized Laplacian matrix, **L** is call combinatorial Lapcian matrix. The wiener index **W** is also defined as $\mathbf{W} = \sum_{\substack{\lambda \neq 0}} 1/\lambda$, where λ is the eigenvalues of

L. The connectivity index is defined as $\chi = \frac{1}{2} \sum_{i \neq j} H_{ij}$.

For example,(i) a secondary structure of RNA



Figure 1: Substructure of AlMV-3

(ii) The adjacency matrix **A** and the diagonal matrix **D**(from 3' to 5')

	0		· /
	$(0\ 1\ 0\ 0\ 0\ 0\ 0\ 1)$		$(2\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0)$
	$1 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0$		030000000
	010100100		0 0 3 0 0 0 0 0 0
	001010000		0 0 0 2 0 0 0 0 0
A =	0 0 0 1 0 1 0 0 0	, D =	0 0 0 0 2 0 0 0 0
	000010100		0 0 0 0 0 2 0 0 0
	001001010		0 0 0 0 0 0 3 0 0
	010000101		00000030
	100000010		0000000002

(iii) Markov matrix M

	(0	1/3	0	0	0	0	0	0	1/2
	1/2	0	1/3	0	0	0	0	1/3	0
	0	1/3	0	1/2	0	0	1/3	0	0
	0	0	1/3	0	1/2	0	0	0	0
M =	0	0	0	1/2	0	1/2	0	0	0
	0	0	0	0	1/2	0	1/3	0	0
	0	0	1/3	0	0	1/2	0	1/3	0
	0	1/3	0	0	0	0	1/3	0	1/2
	(1/2)	0	0	0	0	0	0	1/3	0)

(iv) H matrix

$$\mathbf{H} = \begin{pmatrix} 0 & 1/\sqrt{6} & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 \\ 1/\sqrt{6} & 0 & 1/3 & 0 & 0 & 0 & 0 & 1/3 & 0 \\ 0 & 1/3 & 0 & 1/\sqrt{6} & 0 & 0 & 1/3 & 0 & 0 \\ 0 & 0 & 1/3 & 0 & 1/2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 0 & 1/\sqrt{6} & 0 & 0 \\ 0 & 0 & 1/3 & 0 & 0 & 1/\sqrt{6} & 0 & 1/3 & 0 \\ 0 & 1/3 & 0 & 0 & 0 & 0 & 1/3 & 0 & 1/2 \\ 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 1/\sqrt{6} & 0 \end{pmatrix}$$

(v) The connectivity index
$$\chi = \frac{1}{2} \sum_{i \neq j} H_{ij} = 4.4663$$

(vi) The normalized Laplacian matrix I-H

$$\mathbf{I} - \mathbf{H} = \begin{pmatrix} 1 & -1/\sqrt{6} & 0 & 0 & 0 & 0 & 0 & 0 & -1/2 \\ -1/\sqrt{6} & 1 & -1/3 & 0 & 0 & 0 & 0 & -1/3 & 0 \\ 0 & -1/3 & 1 & -1/\sqrt{6} & 0 & 0 & -1/3 & 0 & 0 \\ 0 & 0 & -1/3 & 1 & -1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1/2 & 1 & -1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1/2 & 1 & -1/\sqrt{6} & 0 & 0 \\ 0 & 0 & -1/3 & 0 & 0 & 0 & -1/\sqrt{6} & 1 & -1/3 & 0 \\ 0 & -1/3 & 0 & 0 & 0 & 0 & 0 & -1/\sqrt{6} & 1 \end{pmatrix}$$

(vii) The combinatorial Laplacian matrix L

 $\mathbf{L} = \begin{pmatrix} 2 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ -1 & 3 & -1 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & -1 & 3 & -1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 2 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 2 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 2 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & -1 & 3 & -1 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & -1 & 3 & -1 \\ -1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 2 \end{pmatrix}$ (viii) The Wiener number index $W = \sum_{\lambda \neq 0} \frac{1}{\lambda} = 4.8827$

2.2 2-D graphical representation of RNA secondary structures

Lemma [10] Any secondary structure ψ can be uniquely decomposed into stacks, loops, and external elements.

Consider the RNA secondary structure with n residues first. Usually the sequence has the form $SEESLLSE \cdots$, where S, L, and E denote the stacks, loops, and external elements, respectively. Suppose that the cumulative numbers of stacks, loops, and external elements occurring in this sequence from the first residue to the nth residue are denoted by α_n , β_n and γ_n , respectively. Obviously, $\alpha_n + \beta_n + \gamma_n = n$.



Figure 2

The three integers α_n , β_n and γ_n can be mapped onto a point within the following regular triangle: considering the regular triangle ΔABC with the height equal to n, as shown in Figure 2, we find that the sum of the three sides is equal exactly to n. The point P to the sides BC, AC, and AB be equal to α_n , β_n and γ_n , respectively, as shown in Figure 2. The point P constitutes a mapping of the secondary structure content of the RNA concerned. This is a mapping of the one-to-one correspondence. A Cartesian coordinates system is set up as shown in Figure 2. The coordinates of the point $P_n(x, y)$ may be expressed in terms of α_n , β_n and γ_n as follows:

$$\begin{cases} x_n = (\beta_n - \alpha_n)/\sqrt{3} \\ y_n = \frac{2n}{3} - (\alpha_n + \beta_n) = \gamma_n - \frac{n}{3} \end{cases}$$
(1)

There are A, B, and C vertices in the triangle ΔABC . For convenience, the vector pointing to the A vertex from the origin O is said to be of an A direction. Any vector parallel to the A direction is said to be of the A direction, too. The definition of the B and C directions are completely similar. The vector pointing to the point P_n from the origin O is denoted by \mathbf{r}_n . The component of \mathbf{r}_i , i.e. x_n and y_n are calculated by Eq.(1). Let $\Delta \mathbf{r}_n = \mathbf{r}_n - \mathbf{r}_{n-1}$, then we have Property 1.

Property 1 For any $n = 1, 2, \dots, N$, here N is the length of the studied DNA sequence, the vector $\Delta \mathbf{r}_n$ has only three possible direction, i.e., either the direction or the B or the C direction, depending on the n-th residue being either S or L or E, in the RNA secondary structure inspected. Furthermore, the length of $\Delta \mathbf{r}_n$, i.e., $|\Delta \mathbf{r}_n|$, is always equal to $m^2 + n$, for any $n = 1, 2, \dots, N$.

Proof: Actually, the components of $\Delta \mathbf{r}_n$, i.e., Δx_n and Δy_n can be calculated for each possible residue (S L and E) at the n-th position of the DNA sequence by using Eq.(1). For example, when the n-th residue is S, we find $\Delta x_n = -\frac{1}{\sqrt{3}}$ and $\Delta y_n = -\frac{1}{3}$. This result is independent of the

conformation state of the (n-1)-th residue. The two numbers $\left(-\frac{1}{\sqrt{3}}, -\frac{1}{3}\right)$ are called the direction of

 $\Delta \mathbf{r}_n$. The direction number and the length of $\Delta \mathbf{r}_n$ for each possible residue type at the n-th position are summarized as follows.

	Δx_n	Δy_n	$ \Delta \mathbf{r}_n $
S	$-\frac{1}{\sqrt{2}}$	$-\frac{1}{3}$	$\frac{2}{3}$
L	$-\frac{1}{\sqrt{2}}$	$-\frac{1}{2}$	$\frac{2}{2}$
E	$\frac{\sqrt{3}}{0}$	3	3
		$\overline{3}$	$\overline{3}$

Property 2 For any positive integers n, m and n>m, the vector equation $\mathbf{r}_n = \mathbf{r}_m$ is valid if only if

Where α_{n-m} , β_{n-m} , and γ_{n-m} are the cumulative numbers of the residues S,L, and E occurring in the subsequence from the m-th to the nth residue in the sequence inspected.

Proof: Obviously, $\mathbf{r}_n = \mathbf{r}_m$ implies $\mathbf{r}_n - \mathbf{r}_m = 0$ that or

$$\begin{cases} x_n - x_m = (\beta_{n-m} - \alpha_{n-m})/\sqrt{3} = 0\\ y_n - y_m = \frac{2(n-m)}{3} - (\alpha_{n-m} + \beta_{n-m}) = 0 \end{cases}$$

This leads to Eq.(2) immediately. Eq.(2) describes the loop property of the 2D graphical representation of RNA secondary structure.

Property 3 The 2D representation possesses the reflection symmetry.

Proof: Usually the sequence is expressed in the order from 5' to 3'. Suppose that the 2D representation for DNA sequence is described by $(x_n, y_n), n = 0, 1, 2, \dots, N$. Suppose again

that the 2D representation for the reverse sequence, i.e, the same sequence but from 3' to 5' is described by (\hat{x}_n, \hat{y}_n) , I find

$$\begin{cases} \hat{x}_n = x_N - x_{N-n} \\ \hat{y}_n = y_N - y_{N-n} \end{cases}$$

for $n = 0, 1, 2, \dots, N$.

3 CONCLUSIONS

We have presented a 2D graphical representation for the abstract symbolic sequence of RNA secondary structure. The curve is the unique representation for a given RNA secondary structure. Different geometrical properties of the curve are studies in details, which reflect the basic characteristics of the RNA secondary structure. Some characteristic matrices are derived from the definition of RNA secondary structure. Different structures have different characteristic matrices and different graphical representation. We also can apply these numerical characterizations to make comparisons between RNA secondary structures.

4 REFERENCES

[1] M. S. Waterman, Introduction to Computational Biology: Maps, Sequences and Genomes. Chapman & Hall, London, 1995.

[2] Jason T. L. Wang, Kaizhong Zhang, Identifying approximately common substructures in tree based on a restricted edit distance, Information Science, 2000,126,165-189.

- [3] D. Angluin ,Finding patterns common a set of strings, Journal of computer and system sciences ,1980, 21, 46-62.
- [4] William J. Masek, A Faster Algorithm Computing string Edit Distances, Journal of computer and system sciences ,1980, 20 ,18-31.
- [5] Chratal, V. and Sankoff, D. Longest Common subsequences of two random sequences. J.Appl.Probab, 1975, 12 306-315
- [6] Hirschberg, D. S. A linear space algorithm for computing maximal common subsequences, Common, ACM ,18,341-343.
- [7] Christian N. S. Pedersen Algorithms in Computational Biology , in: BRICS Dissertation Series 200
- [8] Bo Liao, Tian-ming Wang, Largest common substructure of RNA structure, Internet Journal of Molecula Design, 3(2004),361-367.
- [9] X. G Vienmt, M. v. de chaumont, Enumeration of RNA's secondary structures by complexity, in: V. Capasso, E. Grosso, S. L. Paver-Fontana(Eds), Mathematics in Medicine and Biology, Lect. Notes in Biomath, Vol. 57 Springer, Berlin, 1985, 360-365.
- [10] Ivo .L. Hofacker, Peter Schuster, Peter F. Stadler, Combinatorics of RNA secondary structures. Discr. Appl. Math, 1998,88 ,207-237.
- [11] Bo Liao, Tian-ming Wang, General combinatorics of RNA hairpins and cloverleaves, J.Chem.Inf.Comput.Sci, 43(4)2003,1138-1142.
- [12] Bo Liao, Tian-ming Wang, General combinatorics of RNA secondary structures, Mathematical Biosciences, 191(2004), 69-81.
- [13] Bo Liao, Tianming Wang, A 3D Graphical Representation of RNA Secondary structures, Journal of Biomolecular Structure & Dynamics, 21(6), 2004, 827-832.
- [14] Douglas J. Klein, Jose Luis Palacios, Milan Randic, Nenad Trinajstic, Random Walks and Chemical Graph Theory, J.Chem.Inf.Comput.Sci, 44(2004), 1521-1525.