PLOS ONE

RESEARCH ARTICLE

A Dynamic 3D Graphical Representation for RNA Structure Analysis and Its Application in Non-Coding RNA Classification

Yi Zhang[1,2☯]*, Haiyun Huang[3☯], Xiaoqing Dong[1], Yiliang Fang[4], Kejing Wang[1], Lijuan Zhu[1], Ke Wang[1], Tao Huang[5]*, Jialiang Yang[1,6]*

1 Department of Mathematics, Hebei University of Science and Technology, Shijiazhuang, Hebei 050018, People's Republic of China, **2** Hebei Laboratory of Pharmaceutic Molecular Chemistry, Shijiazhuang, Hebei 050018, People's Republic of China, **3** Department of Information Retrieval of Library, Hebei University of Science and Technology, Shijiazhuang, Hebei 050018, People's Republic of China, **4** International Travel Healthcare Center, Fuzhou, Fujian 350001, People's Republic of China, **5** Institute of Health Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, People's Republic of China, **6** Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, United States of America

☯ These authors contributed equally to this work.
* Jialiang.yang@mssm.edu (JY); zhaqi1972@163.com (YZ); tohuangtao@126.com (TH)

Citation: Zhang Y, Huang H, Dong X, Fang Y, Wang K, Zhu L, et al. (2016) A Dynamic 3D Graphical Representation for RNA Structure Analysis and Its Application in Non-Coding RNA Classification. PLoS ONE 11(5): e0152238. doi:10.1371/journal.pone.0152238

Editor: Quan Zou, Tianjin University, CHINA

Received: December 22, 2015

Accepted: March 10, 2016

Published: May 23, 2016

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was partially supported by the National Science Foundation of China (No 11171088 to YZ), the Science and technology project of Hebei Province (No A2015208108 to YZ, No 1520341 to HH), the Science Fund of the Hebei University of Science and Technology Foundation (No 2014PT67 to YZ), the Hebei Province Foundation for Advanced Talents (No A201400121 to YZ), the Educational Commission of Hebei Province on of Humanities and Social Sciences (No SZ16180 to HH), the National Science and Technology Support Program (No

Abstract

With the development of new technologies in transcriptome and epigenetics, RNAs have been identified to play more and more important roles in life processes. Consequently, various methods have been proposed to assess the biological functions of RNAs and thus classify them functionally, among which comparative study of RNA structures is perhaps the most important one. To measure the structural similarity of RNAs and classify them, we propose a novel three dimensional (3D) graphical representation of RNA secondary structure, in which an RNA secondary structure is first transformed into a characteristic sequence based on chemical property of nucleic acids; a dynamic 3D graph is then constructed for the characteristic sequence; and lastly a numerical characterization of the 3D graph is used to represent the RNA secondary structure. We tested our algorithm on three datasets: (1) Dataset I consisting of nine RNA secondary structures of viruses, (2) Dataset II consisting of complex RNA secondary structures including pseudo-knots, and (3) Dataset III consisting of 18 non-coding RNA families. We also compare our method with other nine existing methods using Dataset II and III. The results demonstrate that our method is better than other methods in similarity measurement and classification of RNA secondary structures.

Introduction

As a bridge of information transmission, RNAs carry a wide variety of functions in biological systems, including performing catalytic function, regulating gene expression, and carrying genetic information [1]. In living cells, the single-stranded RNAs do not remain in a linear

2012BAK11B05 to YF), and Administration of Quality Supervision, Inspection, and Quarantine (AQSIQ) Support Program (No 2015IK037 to YF).

Competing Interests: The authors have declared that no competing interests exist.

form. Instead, their bases always fold into pairs that lead to the formation of RNA secondary structures [2]. Since the three dimensional (3D) structures and functions of RNAs are mostly determined by their secondary structures [3], it is important to account for both primary sequences and secondary structures in understanding the functional similarities among RNAs, especially for non-coding RNAs (ncRNAs) and RNAs with pseudo-knots.

Recently, ncRNAs have become a focus for both computational and experimental researches [4]. Though ncRNAs do not encode proteins, they are involved in a lot of cellular processes [5]. For example, ncRNAs play an important role in chromosome maintenance and segregation [6] and have also been implicated in neurological diseases and various cancers [7, 8]. Specifically, microRNAs are endogenous, short, non-coding RNA molecules that are directly involved in the posttranscriptional regulation of gene expression. Dysregulation of microRNAs is usually associated with diseases [9, 10]. Moreover, a few computational methods have been developed to detect causal genes of diseases at the whole-genome level [11, 12].

It is known that the functions of ncRNAs are mostly determined by their structures [5]. Though sequences can be well aligned by techniques like seed technique [13, 14], predicting the secondary structure of ncRNAs is very difficult [15]. Novel and effective methods to accurately evaluate the structural similarities of ncRNAs are highly demanded, especially for those with special structures like pseudo-knot. As a kind of RNA structures formed by stem nesting, pseudo-knot is responsible for a few important biological activities such as virus infiltration [16].

Historically, dynamic programming based algorithms with various scoring functions have been widely used to measure the similarities among RNA secondary structures [17–19]. For example, based on the alignment of a string representation of RNA secondary structure, a score function and a distance function were established to represent insertion, deletion, and substitution of bases in the compared structures [17, 18]. In addition, a tree representation of the RNA secondary structure elements [19] and base pairing probability matrices [20, 21] were also proposed to measure the similarities among RNA structures. However, methods based on dynamic programming are computationally inefficient, which makes it hard to predict RNAs with complex secondary structures like pseudo-knot.

In order to measure RNA similarity more efficiently, various alternative techniques have been tested. For example, a novel 2D graphical representation of RNA secondary structure was proposed in [22]. However, this method may cause the loss of information due to its non-uniqueness in representing an RNA. Jeffery introduced a chaos game representation of DNA sequences [23], based on which Li et al. proposed a non-degenerative 2D graphical representation of RNA secondary structure to solve the information loss problem [24]. On the ground of sequence and base chemical information, two similar 3D representation methods were proposed [23, 25]. However, they are space-demanding, especially for long RNAs. As a high dimension representation scheme, a 4D method was developed to resolve the problem of structure degeneracy and information loss, but it is not good for visualization [26]. In addition, a novel wavelet-based graphical representation method was used to classify non-coding RNA secondary structures [27]. However, the data obtained by this method is redundant because each base is characterized by three vectors.

In this paper, we proposed a novel dynamic 3D graphical representation of RNA secondary structures based on their sequences, chemical properties, and structural information. We evaluated our algorithm on three sets of RNA secondary structures including 9 viral RNAs, 33 RNAs with complex secondary structures, and 120 non-coding RNAs. A comparison study showed that our novel method outperformed other nine methods in deciphering RNA similarity.

Materials and Methods

Data

We adopted three RNA datasets in this study: (1) Dataset I consisting of 3'-terminus of RNAs from nine viruses reported by Bol (see Fig 1) [28]; (2) Dataset II containing two kinds of RNA secondary structures: 17 complex RNA secondary structures retrieved from RNase P database [29] and 16 RNA secondary structures with pseudo-knots retrieved from Pseud Base (Pseud Base++) [30] (see S1 Fig); (3) Dataset III including 60 non-coding RNA secondary structures from RNA STRAND database (RNA STRAND v2.0) [31, 32], and 60 non-coding RNA sequences randomly selected from 18 non-coding RNA families in Rfam database (see S2 Fig).

3D graphical representation of RNA secondary structures

The secondary structure of an RNA was constructed by four free bases A, G, C, and U, as well as base pairs formed by bonds between A-U, G-C, and G-U. For convenience, A, G, C, and U located in the base pairs were denoted by A', G', C', and U', respectively. In this way, we can use a sequence consisting of A, G, C, U and A', G', C', U' to represent an RNA secondary structure. The sequence is called the "characteristic sequence" of the RNA secondary structure. We provided a software to generate such sequence (see S1 Software) from any RNA secondary structure.

Based on their chemical properties, the four bases A, C, G, and U can be divided into three groups: (i) amino M = {A,C} and keto K = {G,U}, (ii) purine R = {A,G} and pyrimidine Y = {C, U}, and (iii) weak H-bonds W = {A,U} and strong H-bonds S = {C,G}. According to the base classification scheme (i), (ii), and (iii), a characteristic sequence can be represented by three 3D graphs through three maps φ_1, φ_2, and φ_3 respectively (see Table 1), where A_i, G_i, C_i, U_i, A'_i, G'_i, C'_i, and U'_i are the cumulative occurrence numbers of A, G, C, U, A', G', C', and U', respectively in the characteristic sequence from the first base to the i-th base, $i = 1, 2, \ldots n$ with n being the length of the characteristic sequence.

Specifically, let $G = g_1g_2g_3\ldots g_n$ be the characteristic sequence of an RNA secondary structure, where n is the length of G. Each base g_i can be mapped into three dots $\varphi_1(g_i) = (x_{1i}, y_{1i}, z_{1i})$, $\varphi_2(g_i) = (x_{2i}, y_{2i}, z_{2i})$, and $\varphi_3(g_i) = (x_{3i}, y_{3i}, z_{3i})$ where $i = 1, 2, \ldots, n$. By connecting dots in $\varphi_1(g_i) = (x_{1i}, y_{1i}, z_{1i})$ $(i = 1, 2, \ldots, n)$ in order, we can obtain an M-K curve to represent the characteristic sequence G. Similarly, by connecting dots in $\varphi_2(g_i)$ $(i = 1, 2, \ldots, n)$ and $\varphi_3(g_i)$ $(i = 1, 2, \ldots, n)$ respectively, we can obtain R-Y and W-S curves. Taking the secondary structure of TSV-3 as an example, we plotted its M-K, R-Y, and W-S curves in Fig 2. Obviously, the x, y, and z coordinates for each of the three curves are all dynamic. The points projected onto the X-Y plane in our method are moving around the unit circle by a certain length and the z coordinate is changed by the content of bases.

Properties of the method

In this section, we introduced 3 properties showing 3 theoretical advantages of our method including non-degenerative (no information loss), easily reflecting the content of stem and loop, and the distribution of base frequencies. The proof of the properties were provided in S1 Text.

Property 1. The mapping between an RNA secondary structure and its dynamic 3D graphical representation is one to one, and the mapping on the X-Y plane is non-degenerative. Thus no information is lost.

Property 2. Our dynamic 3D graphical representation can easily reflect the content of bases and proportion of stem and loop structures.

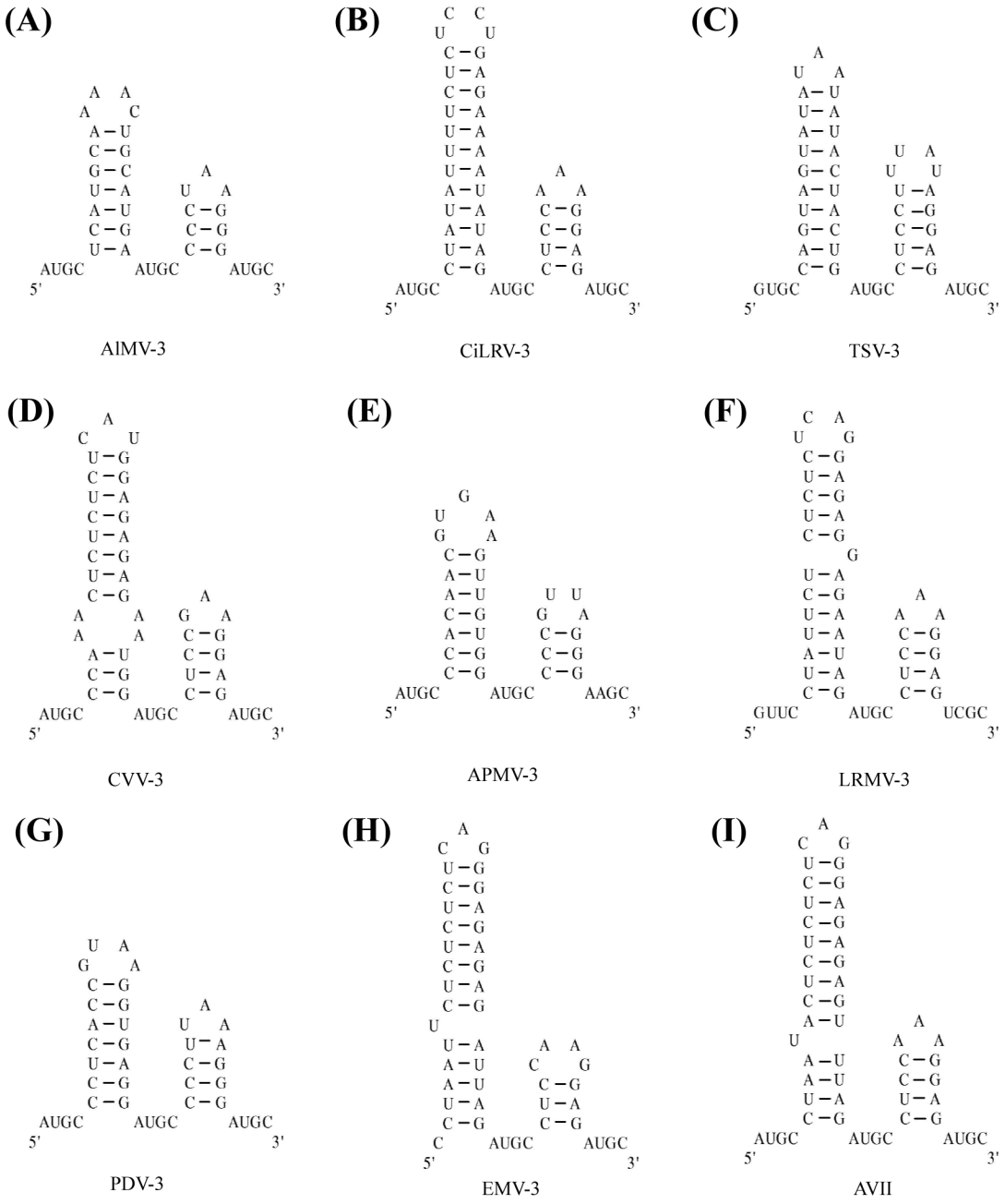

Fig 1. The RNA secondary structures at the 3′—terminus for 9 viruses. (A) alfalfa mosaic virus AlMV-3, (B) citrus leaf rugose virus CiLRV-3, (C) tobacco streak virus TSV-3, (D) citrus variegation virus CVV-3, (E) apple mosaic virus APMV-3, (F) lilac ring mottle virus LRMV-3, (G) prune dwarf ilarvirus PDV-3, (H) elm mottle virus EMV-3, and (I) asparagus virus 2 AVII.

doi:10.1371/journal.pone.0152238.g001

This property shows that a few information on the base distribution and compositions of RNA secondary structure, such as the proportion of stem and loop structures, can be intuitively reflected by the dynamic 3D graphical representation. It is of note that the proportion of stem and loop structures is extremely important for RNA secondary structure prediction [33].

Table 1. The definition of three maps φ_1, φ_2, and φ_3.

g_i	$\varphi_1(g_i) = (x_{1i}, y_{1i}, z_{1i})$			g_i	$\varphi_2(g_i) = (x_{2i}, y_{2i}, z_{2i})$			g_i	$\varphi_3(g_i) = (x_{3i}, y_{3i}, z_{3i})$		
	x_{1i}	y_{1i}	z_{1i}		x_{2i}	y_{2i}	z_{2i}		x_{3i}	y_{3i}	z_{3i}
{*AorC*}	$\frac{i}{n+1}$	$\sqrt{1-(\frac{i}{n+1})^2}$	A_i+C_i	{*AorG*}	$\frac{i}{n+1}$	$\sqrt{1-(\frac{i}{n+1})^2}$	A_i+G_i	{*AorU*}	$\frac{i}{n+1}$	$\sqrt{1-(\frac{i}{n+1})^2}$	A_i+U_i
{*GorU*}	$\frac{i}{n+1}$	$-\sqrt{1-(\frac{i}{n+1})^2}$	G_i+U_i	{*CorU*}	$\frac{i}{n+1}$	$-\sqrt{1-(\frac{i}{n+1})^2}$	C_i+U_i	{*CorG*}	$\frac{i}{n+1}$	$-\sqrt{1-(\frac{i}{n+1})^2}$	C_i+G_i
{*A′orC′*}	$\frac{-i}{n+1}$	$\sqrt{1-(\frac{i}{n+1})^2}$	$A'_i + C'_i$	{*A′orG′*}	$\frac{-i}{n+1}$	$\sqrt{1-(\frac{i}{n+1})^2}$	$A'_i + G'_i$	{*A′orU′*}	$\frac{-i}{n+1}$	$\sqrt{1-(\frac{i}{n+1})^2}$	$A'_i + U'_i$
{*G′orU′*}	$\frac{-i}{n+1}$	$-\sqrt{1-(\frac{i}{n+1})^2}$	$G'_i + U'_i$	{*C′orU′*}	$\frac{-i}{n+1}$	$-\sqrt{1-(\frac{i}{n+1})^2}$	$C'_i + U'_i$	{*C′orG′*}	$\frac{-i}{n+1}$	$-\sqrt{1-(\frac{i}{n+1})^2}$	$C'_i + G'_i$

doi:10.1371/journal.pone.0152238.t001

Property 3. For a characteristic sequence of the RNA secondary structure, let the frequencies of bases A', G', C', U', A, G, C, U be a', g', c', u', a, g, c, u, and $z_1^1 = \frac{1}{n}\sum_{i=1}^{n} z_{1i}^{A,C}$, $z_1^2 = \frac{1}{n}\sum_{i=1}^{n} z_{1i}^{G,U}$, $z_1^3 = \frac{1}{n}\sum_{i=1}^{n} z_{1i}^{A',C'}$, $z_1^4 = \frac{1}{n}\sum_{i=1}^{n} z_{1i}^{G',U'}$, $z_2^1 = \frac{1}{n}\sum_{i=1}^{n} z_{2i}^{A,G}$, $z_2^2 = \frac{1}{n}\sum_{i=1}^{n} z_{2i}^{C,U}$, $z_2^3 = \frac{1}{n}\sum_{i=1}^{n} z_{2i}^{A',G'}$, $z_2^4 = \frac{1}{n}\sum_{i=1}^{n} z_{2i}^{C',U'}$, $z_3^1 = \frac{1}{n}\sum_{i=1}^{n} z_{3i}^{A,U}$, $z_3^2 = \frac{1}{n}\sum_{i=1}^{n} z_{3i}^{G,C}$, $z_3^3 = \frac{1}{n}\sum_{i=1}^{n} z_{3i}^{A',U'}$, $z_3^4 = \frac{1}{n}\sum_{i=1}^{n} z_{3i}^{G',C'}$, $i = 1, 2, \ldots, n$, where n is the length of the characteristic sequence, then $z_1^1, z_1^2, z_1^3, z_1^4, z_2^1, z_2^2, z_2^3, z_2^4, z_3^1, z_3^2, z_3^3$, and z_3^4 indicate the distribution of base frequencies in the whole sequence.

Characterizing RNA secondary structures by 36-D vectors

For an RNA secondary structure, we have three sets of points (x_{1i}, y_{1i}, z_{1i}), (x_{2i}, y_{2i}, z_{2i}), and (x_{3i}, y_{3i}, z_{3i}), $i = 1, 2, \ldots, n$, where n is the length of the structure. We draw the geometrical center of the three curves to construct a 36-dimensional vector denoted by $[x_1^1, y_1^1, z_1^1, x_1^2, y_1^2, z_1^2, x_1^3, y_1^3, z_1^3, x_1^4, y_1^4, z_1^4, x_2^1, y_2^1, z_2^1, x_2^2, y_2^2, z_2^2, x_2^3, y_2^3, z_2^3, x_2^4, y_2^4, z_2^4, x_3^1, y_3^1, z_3^1, x_3^2, y_3^2, z_3^2, x_3^3, y_3^3, z_3^3, x_3^4, y_3^4, z_3^4]$ in which, $x_1^1 = \frac{1}{n}\sum_{i=1}^{n} x_{1i}^{A,C}$, $y_1^1 = \frac{1}{n}\sum_{i=1}^{n} y_{1i}^{A,C}$, $z_1^1 = \frac{1}{n}\sum_{i=1}^{n} z_{1i}^{A,C}$, $x_1^2 = \frac{1}{n}\sum_{i=1}^{n} x_{1i}^{G,U}$, $y_1^2 = \frac{1}{n}\sum_{i=1}^{n} y_{1i}^{G,U}$, $z_1^2 = \frac{1}{n}\sum_{i=1}^{n} z_{1i}^{G,U}$, $x_1^3 = \frac{1}{n}\sum_{i=1}^{n} x_{1i}^{A',C'}$, $y_1^3 = \frac{1}{n}\sum_{i=1}^{n} y_{1i}^{A',C'}$, $z_1^3 = \frac{1}{n}\sum_{i=1}^{n} z_{1i}^{A',C'}$, $x_1^4 = \frac{1}{n}\sum_{i=1}^{n} x_{1i}^{G',U'}$, $y_1^4 = \frac{1}{n}\sum_{i=1}^{n} y_{1i}^{G',U'}$, $z_1^4 = \frac{1}{n}\sum_{i=1}^{n} z_{1i}^{G',U'}$, $x_2^1 = \frac{1}{n}\sum_{i=1}^{n} x_{2i}^{A,G}$, $y_2^1 = \frac{1}{n}\sum_{i=1}^{n} y_{2i}^{A,G}$, $z_2^1 = \frac{1}{n}\sum_{i=1}^{n} z_{2i}^{A,G}$, $x_2^2 = \frac{1}{n}\sum_{i=1}^{n} x_{2i}^{C,U}$, $y_2^2 = \frac{1}{n}\sum_{i=1}^{n} y_{2i}^{C,U}$, $z_2^2 = \frac{1}{n}\sum_{i=1}^{n} z_{2i}^{C,U}$, $x_2^3 = \frac{1}{n}\sum_{i=1}^{n} x_{2i}^{A',G'}$, $y_2^3 = \frac{1}{n}\sum_{i=1}^{n} y_{2i}^{A',G'}$, $z_2^3 = \frac{1}{n}\sum_{i=1}^{n} z_{2i}^{A',G'}$, $x_2^4 = \frac{1}{n}\sum_{i=1}^{n} x_{2i}^{C',U'}$, $y_2^4 = \frac{1}{n}\sum_{i=1}^{n} y_{2i}^{C',U'}$, $z_2^4 = \frac{1}{n}\sum_{i=1}^{n} z_{2i}^{C',U'}$, $x_3^1 = \frac{1}{n}\sum_{i=1}^{n} x_{3i}^{A,U}$, $y_3^1 = \frac{1}{n}\sum_{i=1}^{n} y_{3i}^{A,U}$, $z_3^1 = \frac{1}{n}\sum_{i=1}^{n} z_{3i}^{A,U}$, $x_3^2 = \frac{1}{n}\sum_{i=1}^{n} x_{3i}^{G,C}$, $y_3^2 = \frac{1}{n}\sum_{i=1}^{n} y_{3i}^{G,C}$, $z_3^2 = \frac{1}{n}\sum_{i=1}^{n} z_{3i}^{G,C}$, $x_3^3 = \frac{1}{n}\sum_{i=1}^{n} x_{3i}^{A',U'}$, $y_3^3 = \frac{1}{n}\sum_{i=1}^{n} y_{3i}^{A',U'}$, $z_3^3 = \frac{1}{n}\sum_{i=1}^{n} z_{3i}^{A',U'}$, $x_3^4 = \frac{1}{n}\sum_{i=1}^{n} x_{3i}^{G',C'}$, $y_3^4 = \frac{1}{n}\sum_{i=1}^{n} y_{3i}^{G',C'}$, $z_3^4 = \frac{1}{n}\sum_{i=1}^{n} z_{3i}^{G',C'}$.

The 36-dimensional vector was adopted as a descriptor to characterize RNA secondary structures. It is of note that according to the 3 properties in the previous section and the

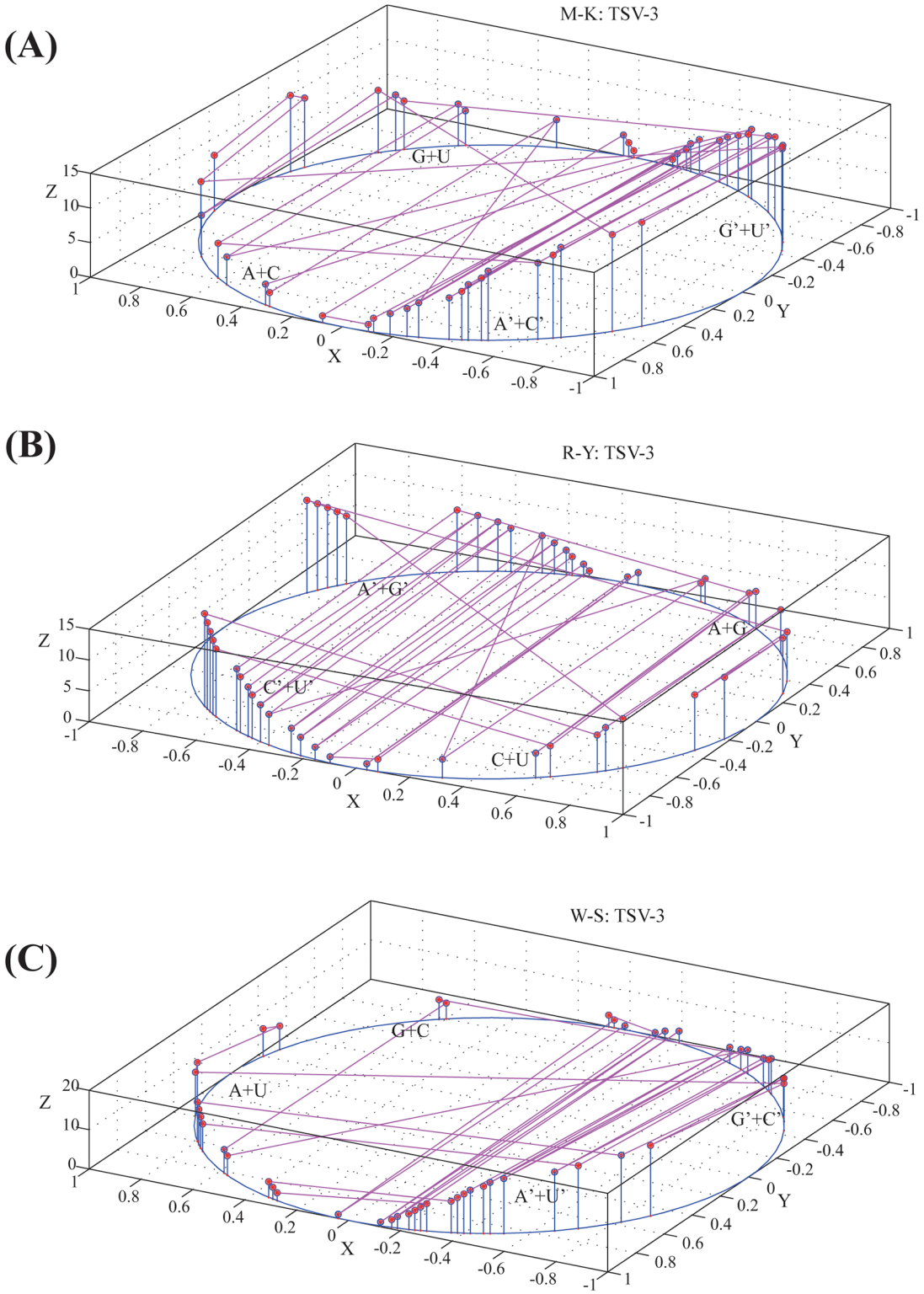

Fig 2. Three characteristic curves of TSV-3. (A) M-K curve, (B) R-Y curve, and (C) W-S curve. Let $G = g_1g_2g_3 \ldots g_n$ be the characteristic sequence of TSV-3, where n is the length of G. Each base g_i can be mapped into three dots $\varphi_1(g_i) = (x_{1i}, y_{1i}, z_{1i})$, $\varphi_2(g_i) = (x_{2i}, y_{2i}, z_{2i})$, and $\varphi_3(g_i) = (x_{3i}, y_{3i}, z_{3i})$, where $i = 1, 2, \ldots, n$. By connecting dots in $\varphi_1(g_i) = (x_{1i}, y_{1i}, z_{1i})$ $(i = 1, 2, \ldots, n)$ in order, we can obtain an M-K curve to represent the characteristic sequence G. Similarly, by connecting dots in $\varphi_2(g_i) = (x_{2i}, y_{2i}, z_{2i})$ and $\varphi_3(g_i) = (x_{3i}, y_{3i}, z_{3i})$ respectively, we can obtain R-Y and W-S curves.

doi:10.1371/journal.pone.0152238.g002

compactness and uniqueness of the three curves for a given RNA secondary structure, this dynamic graphical representation scheme can reflect the distribution of bases in the characteristic sequence and has small degeneracy. Thus, we adopted it to compute the similarity between RNA secondary structures. Specifically, for any two RNA secondary structures, we first constructed their 36D representative vectors, and then calculated the similarity between the two vectors by the quotient between (1) the Euclidean distance between their end-points (in graph) and (2) the cosine of the angle between the two vectors. Clearly, the smaller is the quotient, the more similar are the two RNA secondary structures.

Results and Discussion

Similarities between nine RNA secondary structures of virus

We drew in Fig 1 the RNA secondary structures for nine viruses in **Dataset I**, and listed their pairwise similarities in Table 2.

As Table 2 shows, the smallest entries are associated with pairs (AVII, LRMV-3), (LRMV-3, EMV-3), and (AVII, EMV-3), which indicates that AVII, LRMV-3, and EMV-3 are more similar to each other. On the other hand, APMV-3, AlMV-3, and PDV-3 show great dissimilarity with others. This is consistent with the results reported by Liao et al. [34, 35], Yao et al. [22], Li et al. [24], and Bai et al. [36]. For a better view of our results, we constructed a phylogenetic tree (see Fig 3) for the nine RNA structures based on the 9×9 similarity/dissimilarity matrix using UPGMA method in MEGA5.1. The results indicate that the 36D vectors can catch some intrinsic characteristics of RNA secondary structures.

Further test on Dataset II and III

To further evaluate the sensitivity of this algorithm in measuring the similarities among complex RNA secondary structures, we applied it to **Dataset II**, whose characteristic sequences were shown in S1 Table.

Based on [37, 38], the 17 complex RNA secondary structures shown in S1(A) Fig of RNase P database can be divided into six groups: (1) Gamma Purple Bacteria RNase P structures (Klebsiella pneumoniae, Serratia marcescens, Escherichia coli and Chromatium vinosum), (2) Green sulfur Bacteria RNase P structures (Chlorobium limicola and Chlorobium tepidum), (3) Low G+C Gram positive RNase P structures (Bacillus subtilis and Enterococcus faecalis), (4) Cyanobacterial RNase P structures (Calothrix PCC7601, Anabaena PCC7120 and Synechocystis PCC6803), (5) Archaea Euryarchaeal RNase P structures (Thermococcus celer, Pyrococcus horikoshii and T. Litoralis), and (6) Nuclear RNase P structures (Pan troglodytes, Macaca mulatta, Pongo pygmaeus). In addition, according to pseudo-knot structures, the 16 pseudo-knot secondary structures shown in S1(B) Fig can be divided into five groups: (PKB44, PKB46, PKB4, PKB42, PKB43), (PKB94, PKB114, PKB84), (PKB131, PKB132), (PKB134, PKB135), and (PKB144, PKB140, PKB142, PKB143).

We plotted the similarities among the 33 RNA secondary structures (based on their 36D vector representations) in Fig 4. There are 11 branches corresponding to the six classes of RNase P structures and five classes of pseudo-knot structures respectively. Thus, our method perfectly separated the RNA classes in **Dataset II**.

By a similar process, we constructed a phylogenetic tree (Fig 5) for the secondary structures of 60 ncRNAs in **Dataset III**, whose characteristic sequences were shown in S2 Table. The phylogenetic tree presents clearly 18 branches corresponding to the 18 non-coding RNA families with only one mis-clustering, i.e., RF00001.Methanolobus.tindarius was classified into the cluster RF00374. The results show that our approach performs well in comparing the secondary structures of non-coding RNAs.

Table 2. The similarity/dissimilarity matrix for nine RNA secondary structures in Fig 1 based on RNA 36D vector representation.

species	AIMV-3	APMV-3	AVII	CiLRV-3	CVV-3	EMV-3	LRMV-3	PDV-3	TSV-3
AIMV-3	0	1.5387	4.4078	5.2166	2.6300	4.7192	3.9887	1.7040	4.4352
APMV-3		0	5.4632	7.0070	2.9442	6.2642	5.0473	1.5851	5.6373
AVII			0	2.0993	2.6788	1.0170	0.9822	4.1554	2.1266
CiLRV-3				0	4.6843	1.6580	1.9534	6.0340	1.2103
CVV-3					0	3.4462	2.8271	1.9629	4.1915
EMV-3						0	1.0041	4.7944	2.0004
LRMV-3							0	3.7523	1.7026
PDV-3								0	4.8894
TSV-3									0

doi:10.1371/journal.pone.0152238.t002

Comparison with other methods on Dataset II and III

We compared our similarity measure with other nine popular RNA comparison methods, including a similarity metric based on the wavelet decomposition of the TV-Curve of ncRNA [27], LZ complexity by Liu et al. [39], five 3D graphical representations of RNA secondary structures proposed by Liao et al. [34], Zhu et al. [25], Feng et al. [40], Liu et al. [41], and Luo et al. [42], and two different 2D graphical representations of RNA secondary structures provided by Li et al. [24] and Yao et al. [43] respectively. We applied the nine methods into **Dataset II** and **III**, and the RNA similarities measured by these methods were shown as UPGMA trees in S3–S20 Figs.

As can be seen, the method based on wavelet decomposition [27] failed in separating RNase P from others (see S3 Fig) and mis-classified RF00024.AF221913.1 and RF00001. Thermococcus.celer (see S4 Fig). Similarly, as shown in S5–S18 Figs, the five 3D and two 2D graphical representations performed weakly in comparing 33 RNA secondary structures in **Dataset II**. In addition, except for the method provided by Luo et al. [42], the other six methods also presented weak classifications for the 18 non-coding RNA families in **Dataset III**. Specifically, the method by Liu et al. [39] cannot distinguish RNAs with pseudo-knots from Pseud Base in **Dataset II** (see S19 and S20 Figs). An easy comparison in S3 Table showed that our method outperformed the above nine approaches. Our 36D geometrical center vector may capture some intrinsic characteristics of the non-coding RNA and pseudo-knot

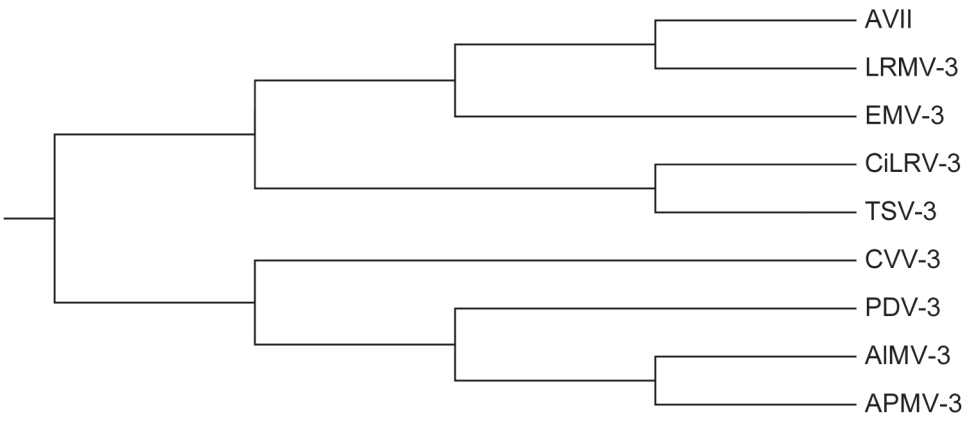

Fig 3. The phylogenetic tree for nine virus in Fig 1. The tree was constructed using UPGMA, in which the distance matrix is calculated by our RNA comparison method based on 36D vector representation.

doi:10.1371/journal.pone.0152238.g003

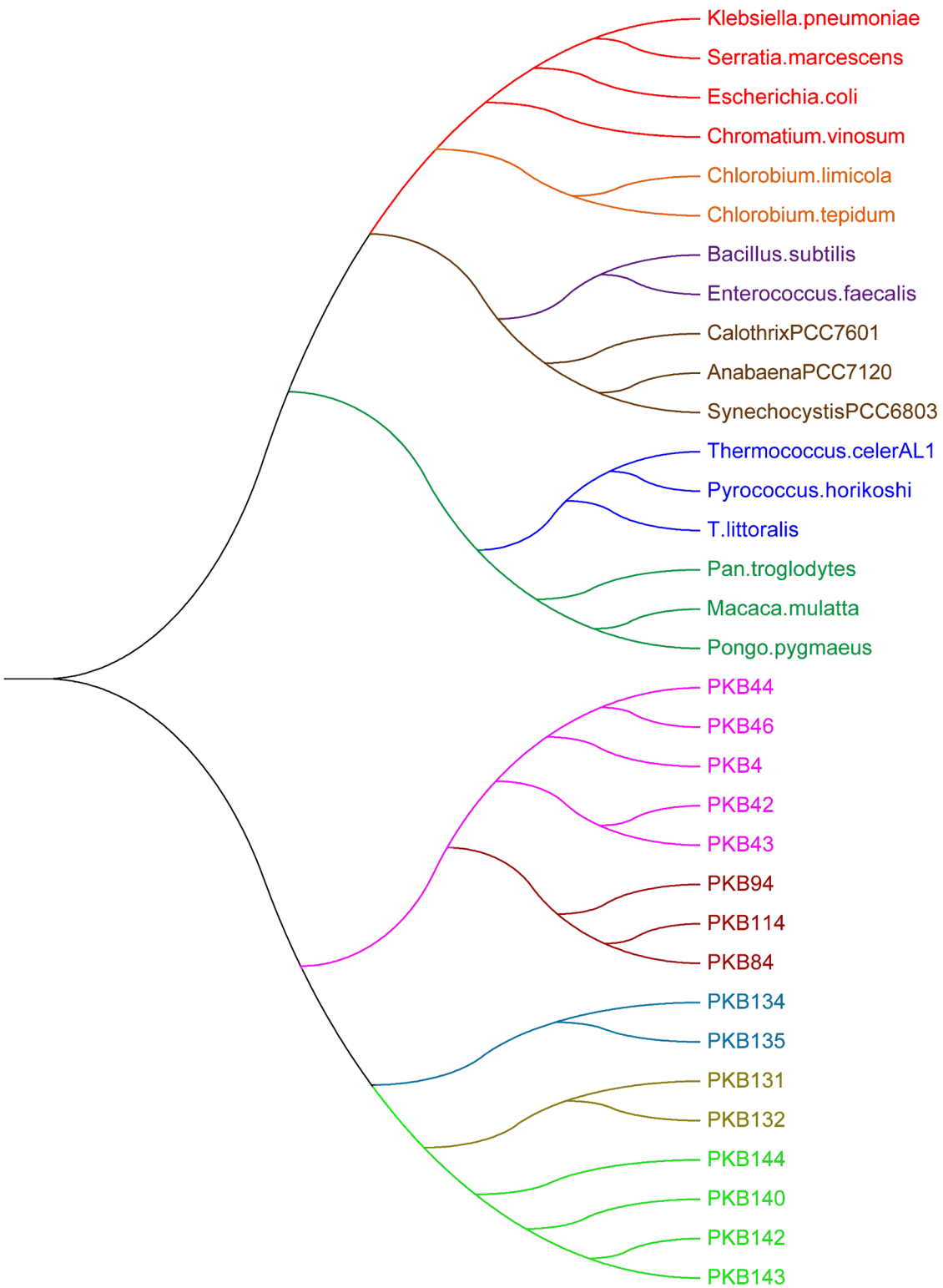

Fig 4. The phylogenetic tree for RNA secondary structures in S1 Fig. Branch colors indicate six clusters of RNase P structures from RNase P database and five clusters of pseudo-knot structures from Pseud Base. The six clusters of RNase P structures include: (1) Gamma Purple Bacteria RNase P structures, (2) Green sulfur Bacteria RNase P structures, (3) Low G+C Gram positive RNase P structures, (4) Cyanobacterial RNase P structures, (5) Archaea Euryarchaeal RNase P structures, and (6) Nuclear RNase P structures.

doi:10.1371/journal.pone.0152238.g004

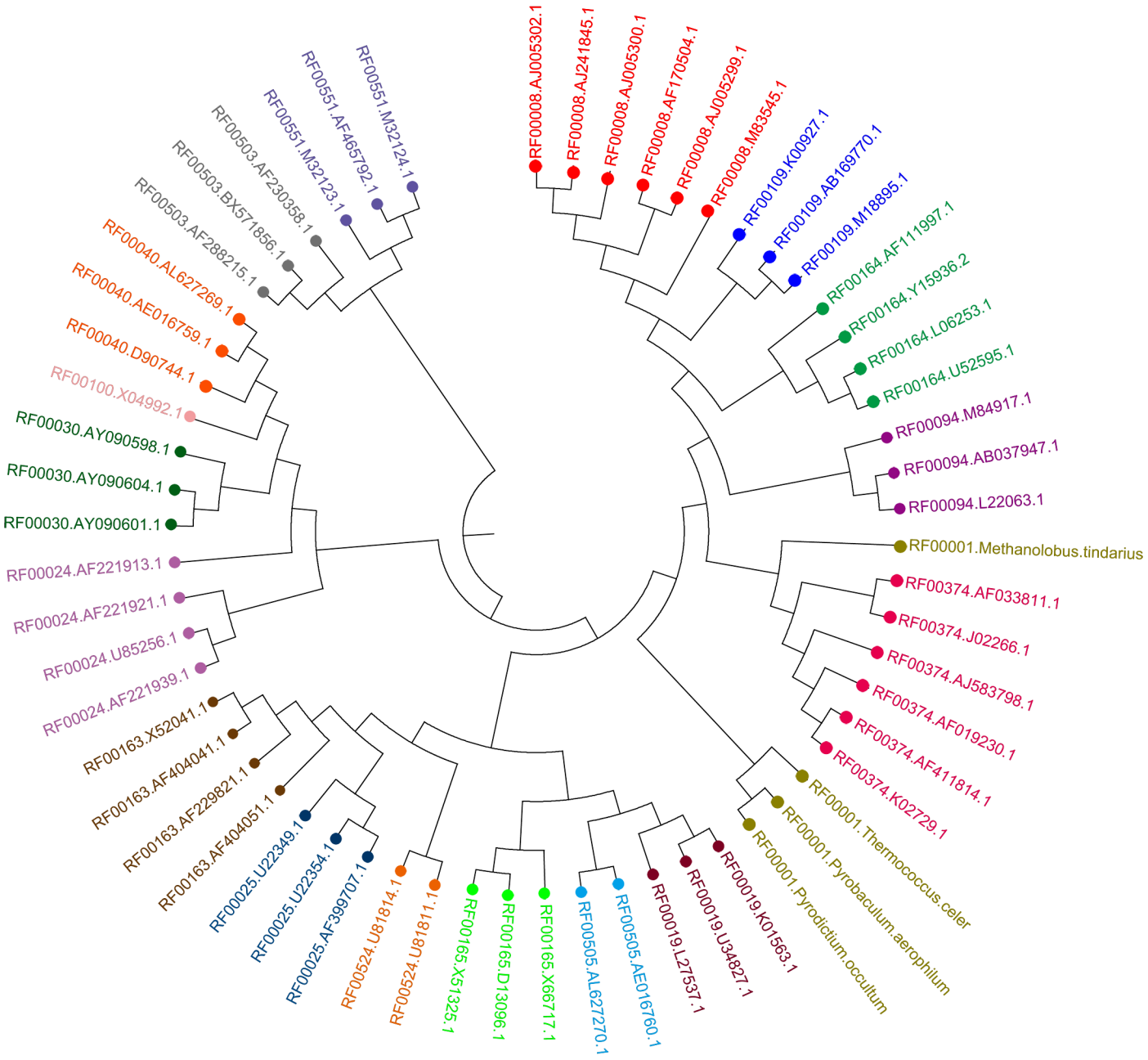

Fig 5. The phylogenetic tree for non-coding RNA secondary structures in S2 Fig. The 18 branch colors represent 18 non-coding RNA families respectively from RNAstrand database, including 5S rRNA (RF00001), Gammaretro_CES (RF00374), Hepatitis delta virus ribozyme (RF00094), Vimentin3 (RF00109), Corona_pk3 (RF00165), Y_RNA (RF00019), s2m (RF00164), Hammerhead ribozyme (type III) (RF00008), Ciliate telomerase RNA (RF00025), R2 RNA element (RF00524), Hammerhead ribozyme (type I) (RF00163), Vertebrate telomerase RNA (RF00024), rne5 (RF00040), RNase MRP (RF00030), 7SK RNA (RF00100), RNAIII (RF00503), RydC RNA (RF00505), and Bicoid 3 prime-UTR regulatory element (RF00551).

doi:10.1371/journal.pone.0152238.g005

structures. Moreover, the points projected onto the X-Y plane in our method are dynamic. This is different from Liao [34] and Zhu [25], in which they fixed x and y coordinates. These dynamic x and y coordinates may provide more "order" information along the secondary structure than the fixed ones.

Conclusion

Based on the chemical property and order of bases, we first proposed a dynamic 3D graphical visualization scheme for RNA secondary structures. We then extracted digital features of the dynamic 3D graphs to compute the similarity between two RNA secondary structures. The method was applied to ncRNAs from the Rfam database and achieved good classification results. In the end, we compared our method with other nine popular approaches and showed that our method outperformed them on RNAs with pseudo-knot and non-coding RNAs. In the future, it will be interesting to extend this method to capture more features of RNA secondary structure, and interpret the information carried by the graphical representation. The additional features and information will be useful in developing a more efficient method to measure RNA structure similarity.

Supporting Information

S1 Fig. 33 RNA secondary structures in Dataset II. (A) 17 complicated RNA secondary structures from RNase P database: Klebsiella pneumoniae, Serratia marcescens, Escherichia coli K-12 W3110, Chromatium vinosum, Chlorobium limicola thiosulfatophilum, Chlorobium tepidum, Bacillus subtilis 168, Enterococcus (ex-Streptococcus) faecalis, Calothrix PCC7601, Anabaena PCC7120, Synechocystis PCC6803, Thermococcus celer AL-1, Pyrococcus horikoshii strain OT3, T. Litoralis, Pan troglodytes, Macaca mulatta, Pongo pygmaeus. (B) 16 RNA secondary structures with pseudo-knots from Pseud Base: PKB44, PKB46, PKB4, PKB42, PKB43, PKB94, PKB114, PKB84, PKB134, PKB135, PKB131, PKB132, PKB144, PKB140, PKB142, PKB143.
(DOC)

S2 Fig. 18 kinds of non-coding RNA secondary structures from RNAstrand database. (A) 5S rRNA (downloaded from Gutell Lab CRW Site in RNAstrand database and belonging to the RF00001 family of Rfam database). (B) Gammaretro_CES (RF00374). (C) Hepatitis delta virus ribozyme (RF00094). (D) Vimentin3 (RF00109). (E) Corona_pk3 (RF00165). (F) Y_RNA (RF00019). (G) s2m (RF00164). (H) Hammerhead ribozyme (type III) (RF00008). (I) Ciliate telomerase RNA (RF00025). (J) R2 RNA element (RF00524). (K) Hammerhead ribozyme (type I) (RF00163). (L) Vertebrate telomerase RNA (RF00024). (M) rne5 (RF00040). (N) RNase MRP (RF00030). (O) 7SK RNA (RF00100). (P) RNAIII (RF00503). (Q) RydC RNA (RF00505). (R) Bicoid 3 prime-UTR regulatory element (RF00551) (non-coding RNA secondary structures belonging to B-R are obtained from Rfam database in RNAstrand database).
(DOC)

S3 Fig. The Phylogenetic tree by multi-scale RNA comparison based on RNA triple vector curve representation for the secondary structures of RNAs in S1 Fig.
(DOC)

S4 Fig. The Phylogenetic tree by multi-scale RNA comparison based on RNA triple vector curve representation for the secondary structures of RNAs in S2 Fig.
(DOC)

S5 Fig. The two phylogenetic trees for the secondary structures of RNAs in S1 Fig based on the method by Liao *et al* [34]. (A) The phylogenetic tree based on the Euclidean distance. (B) The phylogenetic tree based on the Angle.
(DOC)

S6 Fig. The two phylogenetic trees for the secondary structures of RNAs in S2 Fig based on the method by Liao *et al* [34]. (A) The phylogenetic tree based on the Euclidean. (B) The

phylogenetic tree based on the Angle.
(DOC)

S7 Fig. The two phylogenetic trees for the secondary structures of RNAs in S1 Fig based on the method by Zhu *et al* [25]. (A) The phylogenetic tree based on the Euclidean distance. (B) The phylogenetic tree based on the Angle.
(DOC)

S8 Fig. The two phylogenetic trees for the secondary structures of RNAs in S2 Fig based on the method by Zhu *et al* [25]. (A) The phylogenetic tree based on the Euclidean distance. (B) The phylogenetic tree based on the Angle.
(DOC)

S9 Fig. The phylogenetic tree for the secondary structures of RNAs in S1 Fig based on the method by Feng *et al* [40].
(DOC)

S10 Fig. The phylogenetic tree for the secondary structures of RNAs in S2 Fig based on the method by Feng *et al* [40].
(DOC)

S11 Fig. The two phylogenetic trees for the secondary structures of RNAs in S1 Fig based on the method by Liu *et al* [41]. (A) The phylogenetic tree based on the Euclidean distance. (B) The phylogenetic tree based on the Angle.
(DOC)

S12 Fig. The two phylogenetic trees for the secondary structures of RNAs in S2 Fig based on the method by Liu *et al* [41]. (A) The phylogenetic tree based on the Euclidean distance. (B) The phylogenetic tree based on the Angle.
(DOC)

S13 Fig. The phylogenetic tree for the secondary structures of RNAs in S1 Fig based on the method by Luo *et al* [42].
(DOC)

S14 Fig. The phylogenetic tree for the secondary structures of RNAs in S2 Fig based on the method by Luo *et al* [42].
(DOC)

S15 Fig. The phylogenetic tree for the secondary structures of RNAs in S1 Fig based on the method by Li *et al* [24].
(DOC)

S16 Fig. The phylogenetic tree for the secondary structures of RNAs in S2 Fig based on the method by Li *et al* [24].
(DOC)

S17 Fig. The two phylogenetic trees for the secondary structures of RNAs in S1 Fig based on the method by Yao *et al* [43]. (A) The phylogenetic tree based on the Euclidean distance. (B) The phylogenetic tree based on the Angle.
(DOC)

S18 Fig. The two phylogenetic trees for the secondary structures of RNAs in S2 Fig based on the method by Yao *et al* [43]. (A) The phylogenetic tree based on the Euclidean distance.

(B) The phylogenetic tree based on the Angle.
(DOC)

S19 Fig. The three phylogenetic trees for the secondary structures of RNAs in S1 Fig based on the method by Liu *et al* [39]. (A) The phylogenetic tree based on non-$A(A')$ sequences. (B) The phylogenetic tree based on the non-$C(C')$ sequences. (C) The phylogenetic tree based on the non-$G(G')$ sequences.
(DOC)

S20 Fig. The three phylogenetic trees for the secondary structures of non-coding RNAs in S2 Fig based on the method by Liu *et al* [39]. (A) The phylogenetic tree based on the non-A (A') sequences. (B) The phylogenetic tree based on the non-$C(C')$ sequences. (C) The phylogenetic tree based on the non-$G(G')$ sequences.
(DOC)

S1 Software. RnaFeatureGenerator.
(ZIP)

S1 Table. The characteristic sequences for the secondary structures of RNAs in S1 Fig (A, G, C and U located in the base pairs are denoted as a, g, c and u).
(DOC)

S2 Table. The characteristic sequences for the secondary structures of RNAs in S2 Fig (A, G, C and U located in the base pairs are denoted as a, g, c and u).
(DOC)

S3 Table. The comparison between our method and the other nine algorithms.
(DOC)

S1 Text. The proof of three properties.
(DOC)

Author Contributions

Conceived and designed the experiments: YZ JY. Performed the experiments: HH XD KJW LZ KW. Analyzed the data: HH XD KJW LZ KW. Contributed reagents/materials/analysis tools: YZ JY TH. Wrote the paper: YZ JY HH YF.

References

1. Sato K, Kato Y, Hamada M, Akutsu T, Asai K. IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. Bioinformatics. 2011; 27(13):i85–93. PMID: 21685106. doi: 10.1093/bioinformatics/btr215
2. Zuker M. Calculating nucleic acid secondary structure. Curr Opin Struct Biol. 2000; 10(3):303–10. PMID: 10851192.
3. Tinoco I Jr., Bustamante C. How RNA folds. J Mol Biol. 1999; 293(2):271–81. PMID: 10550208.
4. Wang C, Wei L, Guo M, Zou Q. Computational approaches in detecting non- coding RNA. Curr Genomics. 2013; 14(6):371–7. doi: 10.2174/13892029113149990005 PMID: 24396270; PubMed Central PMCID: PMCPMC3861888.
5. Mattick JS, Makunin IV. Non-coding RNA. Hum Mol Genet. 2006; 15 Spec No 1:R17–29. PMID: 16651366.
6. Bernstein E, Allis CD. RNA meets chromatin. Genes Dev. 2005; 19(14):1635–55. PMID: 16024654.
7. Mattick JS. Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. Bioessays. 2003; 25(10):930–9. PMID: 14505360.

8. Pang KC, Stephen S, Engstrom PG, Tajul-Arifin K, Chen W, Wahlestedt C, et al. RNAdb—a comprehensive mammalian noncoding RNA database. Nucleic Acids Res. 2005; 33(Database issue):D125–30. PMID: 15608161.
9. Wang QC, Wei LY, Guan XJ, Wu YF, Zou Q, Ji ZL. Briefing in family characteristics of microRNAs and their applications in cancer research. Bba-Proteins Proteom. 2014; 1844(1):191–7. doi: 10.1016/j.bbapap.2013.08.002 PubMed PMID: WOS:000330911500005.
10. Wei LY, Liao MH, Gao Y, Ji RR, He ZY, Zou Q. Improved and Promising Identification of Human MicroRNAs by Incorporating a High-Quality Negative Set. Ieee Acm T Comput Bi. 2014; 11(1):192–201. doi: 10.1109/Tcbb.2013.146 PubMed PMID: WOS:000336659800019.
11. Zou Q, Li JJ, Song L, Zeng XX, Wang GH. Similarity computation strategies in the microRNA-disease network: a survey. Brief Funct Genomics. 2016; 15(1):55–64. doi: 10.1093/bfgp/elv024 PubMed PMID: WOS:000370155900008. PMID: 26134276
12. Jiang R. Walking on multiple disease-gene networks to prioritize candidate genes. J Mol Cell Biol. 2015; 7(3):214–30. doi: 10.1093/jmcb/mjv008 PubMed PMID: WOS:000357857100004. PMID: 25681405
13. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990; 215(3):403–10. PMID: 2231712.
14. Yang J, Zhang L. Run probabilities of seed-like patterns and identifying good transition seeds. J Comput Biol. 2008; 15(10):1295–313. PMID: 19040365. doi: 10.1089/cmb.2007.0209
15. Lindgreen S, Gardner PP, Krogh A. MASTR: multiple alignment and structure prediction of non-coding RNAs using simulated annealing. Bioinformatics. 2007; 23(24):3304–11. PMID: 18006551.
16. Dam E, Pleij K, Draper D. Structural and functional aspects of RNA pseudoknots. Biochemistry. 1992; 31(47):11665–76. PMID: 1280160.
17. Bafna V, Muthukrishnan S, Ravi R. Computing similarity between RNA strings. DIMACS Technical Report. 1996; 96:30.
18. Dowell RD, Eddy SR. Efficient pairwise RNA structure prediction and alignment using sequence alignment constraints. BMC bioinformatics. 2006; 7:400. doi: 10.1186/1471-2105-7-400 PMID: 16952317; PubMed Central PMCID: PMC1579236.
19. Shapiro BA, Zhang K. Comparing multiple RNA secondary structures using tree comparisons. Bioinformatics. 1990; 6(4):309–18.
20. Hofacker IL, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P. Fast folding and comparison of RNA secondary structures. Monatshette Fur Chemie. 1994; 125:167–88.
21. McCaskill JS. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers. 1990; 29(6–7):1105–19. doi: 10.1002/bip.360290621 PMID: 1695107.
22. Yao YH, Nan XY, Wang TM. A class of 2D graphical representations of RNA secondary structures and the analysis of similarity based on them. Journal of computational chemistry. 2005; 26(13):1339–46. doi: 10.1002/jcc.20271 PMID: 16021599.
23. Jeffrey HJ. Chaos game representation of gene structure. Nucleic Acids Res. 1990; 18(8):2163–70. PMID: 2336393; PubMed Central PMCID: PMC330698.
24. Li C, Xing L, Wang X. Analysis of similarity of RNA secondary structures based on a 2D graphical representation. Chemical Physics Letters. 2008; 1–3:249–52.
25. Zhu W, Liao B, Ding K. A condensed 3D graphical representation of RNA secondary structures. Journal of Molecular Structure: THEOCHEM. 2005; 757(1–3):193–8.
26. Liao B, Zhu W, Li P. On a four-dimensional representation of RNA secondary structures. Journal of Mathematical Chemistry. 2007; 42(4):1015–22.
27. Li Y, Duan M, Liang Y. Multi-scale RNA comparison based on RNA triple vector curve representation. BMC bioinformatics. 2012; 13:280. doi: 10.1186/1471-2105-13-280 PMID: 23110635; PubMed Central PMCID: PMC3599440.
28. Reusken CB, Bol JF. Structural elements of the 3'-terminal coat protein binding site in alfalfa mosaic virus RNAs. Nucleic Acids Res. 1996; 24(14):2660–5. PMID: 8758992; PubMed Central PMCID: PMC145989.
29. Brown JW. The ribonuclease P database. Nucleic Acids Res. 1998; 26(1):351–2. PMID: 9399871; PubMed Central PMCID: PMC147188.
30. van Batenburg FH, Gultyaev AP, Pleij CW, Ng J, Oliehoek J. PseudoBase: a database with RNA pseudoknots. Nucleic Acids Res. 2000; 28(1):201–4. PMID: 10592225; PubMed Central PMCID: PMC102383.
31. Puton T, Kozlowski LP, Rother KM, Bujnicki JM. CompaRNA: a server for continuous benchmarking of automated methods for RNA secondary structure prediction. Nucleic Acids Res. 2013; 41(7):4307–23. doi: 10.1093/nar/gkt101 PMID: 23435231; PubMed Central PMCID: PMC3627593.

32. Andronescu M, Bereg V, Hoos HH, Condon A. RNA STRAND: the RNA secondary structure and statistical analysis database. BMC bioinformatics. 2008; 9:340. PMID: 18700982. doi: 10.1186/1471-2105-9-340

33. Xu X, Ji Y, Stormo GD. RNA Sampler: a new sampling based algorithm for common RNA secondary structure prediction and structural alignment. Bioinformatics. 2007; 23(15):1883–91. PMID: 17537756.

34. Liao B, Wang TM. A 3D graphical representation of RNA secondary structures. Journal of biomolecular structure & dynamics. 2004; 21(6):827–32. doi: 10.1080/07391102.2004.10506972 PMID: 15107004.

35. Liao B, Ding K, Wang TM. On a six-dimensional representation of RNA secondary structures. Journal of biomolecular structure & dynamics. 2005; 22(4):455–63. doi: 10.1080/07391102.2005.10507016 PMID: 15588108.

36. Bai FL, Zhu W, Wang TM. Analysis of similarity between RNA secondary structures. Chemical Physics Letters. 2005; 408(4–6):258–63.

37. Harris JK, Haas ES, Williams D, Frank DN, Brown JW. New insight into RNase P RNA structure from comparative analysis of the archaeal RNA. Rna. 2001; 7(2):220–32. PMID: 11233979.

38. Frank DN, Adamidi C, Ehringer MA, Pitulle C, Pace NR. Phylogenetic-comparative analysis of the eukaryal ribonuclease P RNA. Rna. 2000; 6(12):1895–904. PMID: 11142387.

39. Liu L, Bai F, Wang T. Comparing RNA Secondary Structures Based on LZ Complexity. Physics Procedia. 2012; 33:96–103.

40. Feng J, Wang T. A 3D graphical representation of RNA secondary structures based on chaos game representation. Chemical Physics Letters. 2008; 454(4–6):355–61.

41. Liu L, Wang T. On 3D graphical representation of RNA secondary structures and their applications. Journal of Mathematical Chemistry. 2007; 42(3). doi: 10.1007/s10910-006-9135-4

42. Luo J, Liao B, Li R, Zhu W. RNA secondary structure 3D graphical representation without degeneracy. Journal of Mathematical Chemistry. 2006; 39(3):629–36.

43. Yao Y, Liao B, Wang T. A 2D graphical representation of RNA secondary structures and the analysis of similarity/dissimilarity based on it. Journal of Molecular Structure: THEOCHEM. 2005; 755(1–3):131–6.