Detrended Fluctuation Analysis and Discrete Wavelet Transform for Similarity Comparison of RNA secondary structure

Lina Yang
Guangxi University
Yang Liu (1813302002@st.gxu.edu.cn)
Guangxi University https://orcid.org/0000-0003-0986-8466
Patrick Wang
Northeastern University
Xichun Li
Guangxi Normal University for Nationalities

Methodology article

Keywords: RNA secondary structure, DWT; fractal dimension, detrended uctuation analysis, sliding window

Posted Date: July 28th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-39527/v1

METHODOLOGY

Detrended Fluctuation Analysis and Discrete Wavelet Transform for Similarity Comparison of RNA Secondary Structure

Lina Yang[1], Yang Liu[1*†], Patrick Wang[2] and Xichun Li[3]

Abstract

Background: Ribonucleic acid (RNA) is an important biological macromolecule. Through in-depth studies of RNA, its function has been increasingly discovered. The function of RNA is mostly dependent on its secondary structure because of its conserved nature. The discovery of an approximate relationship between two RNA secondary structures can help to understand their functional relationship better. This discovery can also help in exploring many unknown functions. Currently, RNA secondary structural similarity analysis methods are mainly divided into alignment-based methods and alignment-free methods. Alignment-free methods can obtain similarities and differences among RNA secondary structures more quickly and more accurately than alignment-based methods.

Results: In this paper, a novel alignment-free method based on the triple vector curve representation of RNA is proposed. A combinational method involving a discrete wavelet transform and detrended fluctuation analysis (DFA) with a sliding window is used to generate the distance matrix. Finally, a phylogenetic tree is constructed using the distance matrix. Experiments are performed on RNA viruses and non-coding RNA datasets, and the phylogenetic trees generated by different methods are compared. The results show that our method yields more accurate results in the comparison of RNA secondary structures.

Conclusion: The method in this paper enables a more accurate analysis of the similarities between RNA secondary structures. This method has certain application value in the comparison of RNA secondary structures, especially in the analysis of longer RNA sequences.

Keywords: RNA secondary structure; DWT; fractal dimension; detrended fluctuation analysis; sliding window

*Correspondence:
1813302002@st.gxu.edu.cn
[1]Computer and Electronic Information, Guangxi University, 530004 Nanning, China
Full list of author information is available at the end of the article
†Equal contributor

Background

The secondary structure of RNA is a double-stranded structure constructed according to the principle of complementary base pairing. In RNA, U (uracil) is complementary to A (adenine), and G (guanine) is complementary to C (cytosine). RNA is usually transcribed from DNA and acts as a bridge between DNA and proteins [1]. The versatility of an RNA molecule depends on its secondary structure. RNAs with similar structures tend to have similar functions or properties, but the opposite does not necessarily hold true. Many RNA molecules are conserved at the structural level, but they have little sequence similarity. Therefore, the comparison of RNA secondary structures is key to elucidating their functional and evolutionary relationships. Most recent studies have focused on RNA secondary structure prediction [2–4], and comparisons of RNA secondary structures have not yet been sufficiently studied. At the present stage, the comparison methods for RNA secondary structures are mainly divided into two types: alignment-based methods and alignment-free methods. Alignment-based methods mainly rely on an RNA

secondary structure represented by a string or tree [5–10]. The Sankoff algorithm uses the free energy minimization method to synchronize the folding and alignment of two or more RNA sequences. However, it is difficult to achieve. Therefore, to be more effective in RNA secondary structure alignment, a number of improved algorithms were subsequently developed , such as Consan and, Dynalign [11], PMcomp [12], Stemloc [13], Foldalign [14], locARNA [15], SPARSE [16], MARNA [17], FoldAlignM [18], Murlet [19], CARNA [20] and RAF [21]. In response to the unsolved problems of the Sankoff algorithm, Consan developed a good unified scoring algorithm, a pair-SCFG structural alignment algorithm, for paired alignment and folding. Dynalign obtained a common structure with low free energy by comparing two sequences that lack sequence identity and free energy minimization. PMcomp uses McCaskill′s method to calculate the basic paired probability matrix from which the paired and progressive multiple alignments are calculated. Due to the nature of the locARNA method being limited to high complexity, SPARSE introduced the algorithm "sparsified prediction and alignment of RNAs based on their structure ensembles (SPARSE)", which has lower time complexity and improved accuracy. Alignment-free methods divide the process into two parts: folding and alignment. Among them, some tree-based methods, such as RNAforester [22], RNAdistance [8], RNAStrAt [23] and RNApdist [24], are widely used. The RNAforester tree comparison algorithm computes the local similarity in RNA secondary structures. RNAdistance calculates the similarity of the RNA secondary structures by measuring the editing distance of the tree. RNApdist compares the RNA secondary structures based on base-pairing probabilities. The Vienna RNA package can now be used to implement both the RNAdistance and RNApdist methods [25, 26]. Notably, these comparison algorithms have high time complexity. General alignment-free methods are usually based on the numerical representations of RNA secondary structures, followed by the development of the many graphical representations of RNA secondary structures. The visualization of an RNA secondary structure using a graphical representation is more intuitive, thus providing a new way of thinking about the comparison of RNA secondary structures. Previously, researchers used eight symbols to represent RNA, but this representation was accompanied by a loss of information [27]. A feature sequence may correspond to diverse RNA secondary structures. That is, the feature sequences obtained by such a method are not unique. To reduce the loss of information, based on a 4×4 matrix of RNA sequences proposed by Randic [28], Yao used four horizontal lines and eight symbols to represent the RNA secondary structure [29]. The method avoids the loss of information due to the crossover and overlap of curves, but the information loss caused by the limitations of feature invariant extraction still exists. Then, a method for RNA secondary structure visualization with no information loss was proposed by Randic [30]. In [31], Li proposed a novel representation of RNA secondary structures (TV-Curve) and introduced a wavelet decomposition to compare RNA secondary structures.

Wavelet analysis is a synthesis of ideas in pure mathematics, physics and engineering. A commonly used windowed Fourier transform, such as the short-time Fourier transform (STFT), analyzes the signal with a fixed sliding window. Obviously, fixed-length sliding window processing is not suitable for all signals. The linear time-frequency analysis for non-stationary signals should have different resolutions at different positions in the time-frequency plain; in other words, it should

be a multi-resolution analysis method. Wavelet transform is a multi-resolution analysis method. Furthermore, wavelets are a fairly simple and effective mathematical tool that have been widely used. [32] combined a wavelet transform with a neural

Table 1 Eight Viruses from the RFAM Database

Virus	Accession ID	Length
ALMV	NC_001495.1	3644
APMV	NC_003465.1	3476
CiLRV	JX237459.1	3404
CVV	NC_009538.1	3433
EMV	NC_003569.1	2874
LRMV	NC_038776.1	2287
PDV	KY883326.1	3320
TSV	NC_003842.1	2926

network to achieve highly accurate machine fault diagnosis. In [33], wavelet analysis was used to identify membrane protein types. It was also used to realize RNA secondary structure similarity analysis. In practical applications, especially when implementing the wavelet transform on a computer, the signal must be discretized and analyzed by the discrete wavelet transform. The discrete wavelet transform was used for the feature extraction of skin lesions to obtain the best combination

[width=7cm]Fig.1(a).eps

(a)

[width=7cm]Fig.1(b).eps

(b)

[width=6cm]Fig.1(c).eps

(c)

Figure 1 (a) Feature extraction of RNA secondary structure. (b) The third-level discrete wavelet transform. (c) The DFA fractal dimension of (b) obtained with a sliding window width of 17.

of features [34]. By detecting and analyzing protein secondary structures by discrete wavelet transforms, the correlation of the amino acid sequence and secondary structure was tested by the hydrophobic value of the amino acids [35].

The fractal dimension (FD) reflects the validity of the space occupied by a complex form, which is a measure of the irregularity of a complex form [36]. FD acknowledges that various parts of the world may be similar in some way to the whole region under certain conditions or processes, and it recognizes that changes in spatial dimensions can be discrete and continuous [37]. [38] combines empirical modal decomposition and multi-fractal detrended fluctuation analysis to study the fractal characteristics of harmonic signals. To address the multi-fractal characteristics of hydrographic data, Zhao established a wavelet method and a multi-fractal detrended fluctuation analysis model to study the multi-fractal characterization and simulation of river levels [39]. Additionally, the fractal dimension has been introduced in the studies of biological molecules. Yu exploited the hydrophobicity of amino acids and fractal analysis to classify the protein structure [40]. In [41], Yang

applied the fractal dimension to protein sequence similarity analysis and obtained a high degree of similarity, which has also been recently used to determine the distribution of purines and pyrimidines in the miRNAs of humans, gorillas, chimpanzees, mice and rats [42].

In this paper, we propose a novel combined algorithm. Based on the characteristics of DWT multi-resolution analysis and the universality of FD, DWT is used to analyze the feature vectors of RNA secondary structures, and FD is used to quantitatively characterize their geometric structure features. The purpose of this paper is to explore the application of DWT and fractal dimension in the comparison of RNA secondary structures.

The paper is organized as follows. We outline the basic concepts and theory of the fractal dimensions in Section 2. Our algorithm is presented in Section 3. Numerical experiments and the results are then given in Section 4. Section 5 is the conclusion.

Table 2 The Distance Matrix of Eight Virus Generated by DFA

	ALMV	CVV	CiLRV	TSV	APMV	LRMV	PDV	EMV
ALMV	0	0.4953	0.4728	0.5015	0.4860	0.5119	0.4815	0.4952
CVV		0	0.4945	0.4974	0.4857	0.5074	0.4789	0.4850
CiLRV			0	0.4816	0.4880	0.4921	0.4831	0.4837
TSV				0	0.4839	0.5073	0.4863	0.4804
APMV					0	0.5127	0.4725	0.4807
LRMV						0	0.5341	0.5175
PDV							0	0.4945
EMV								0

Results

In this section, eight RNA viruses with slight differences in structure are selected to judge the application of our method to sequences with small structural differences. Their full gene sequences are used here, and their secondary structures are predicted by the Vienna RNA package. The eight viruses include the alfalfa mosaic virus (ALMV), apple mosaic virus (APMV), citrus leaf rugose virus (CiLRV), citrus variegation virus (CVV), elm mottle virus (EMV), lilac ring mottle virus (LRMV), prune dwarf ilarvirus (PDV) and tobacco streak virus (TSV) [43]. Information on these eight viruses is presented in Table 1. Fig. 1 shows the DWT and DFA on the TV-Curve of the EMV virus sequence using our method. Constructing the experimental environment by setting the wavelet level to 3 and the window width to 17, we obtain the distance matrix of eight viruses generated by detrended fluctuation analysis in Table 2. Next, as shown in Fig. 2, our proposed method is compared with

Figure 2 The evolutionary tree of eight virus sequences generated by (a)DFA,(b)Li′s method,(c)RNAdistance, and (d)MEGA software.

RNAdistance, MEGA software [44] and Li′s method. The phylogenetic tree generated by the detrended fluctuation analysis method is based on the distance matrix

Table 3 11 Non-coding RNA sequences from the RFAM and NONCODEv5 databases

Name	Length	Family
NONHSAT000002.2	1653	NONCODEv5_human
NONHSAT000003.2	1483	
NONHSAT000004.2	632	
MF489813.1	849	tRNA
MF489812.1	535	
MF489811.1	549	
MF489810.1	566	
MF489809.1	549	
MF489808.1	552	
MF489806.1	554	
NT_033777.3	299	RNase P RNA

calculated in MATLAB. In MEGA software, the maximum likelihood method is used to generate the phylogenetic tree.

Table 4 The Distance Matrix of Eleven ncRNA Sequences produced by DFA

	NONHSAT000002.2	NONHSAT000003.2	NONHSAT000004.2	MF489813.1	MF489812.1	MF489811.1	MF489810.1	MF489809.1	MF489808.1	MF489806.1	NT_033777.3
NONHSAT000002.2	0	0.3882	0.4843	0.5123	0.5178	0.5153	0.5216	0.5212	0.5279	0.5394	0.5510
NONHSAT000003.2		0	0.4571	0.5337	0.5505	0.5504	0.5148	0.5173	0.5160	0.5252	0.5636
NONHSAT000004.2			0	0.5184	0.4199	0.4214	0.4698	0.5221	0.5199	0.4999	0.5348
MF489813.1				0	0.4989	0.4802	0.4905	0.4806	0.4824	0.5334	0.5929
MF489812.1					0	0.0204	0.3603	0.3634	0.4206	0.4557	0.4776
MF489811.1						0	0.3597	0.3519	0.4133	0.4616	0.4736
MF489810.1							0	0.3049	0.4157	0.4632	0.5165
MF489809.1								0	0.3731	0.4445	0.4902
MF489808.1									0	0.4139	0.4933
MF489806.1										0	0.4591
NT_033777.3											0

In addition, an experiment containing 11 non-coding RNAs is performed to test the applicability of our method in comparing the similarities between non-coding RNAs. Eight sequences are randomly selected from the ncRNAs in the RFAM database; additionally, three human ncRNA sequences from the NONCODE database are randomly sampled for the experiment. The information of these ncRNAs is provided in Table 3, and the distance matrix generated by the experiment is shown in Table 4. The results are shown in Fig. 3.

Discussion

A detailed explanation of the phylogenetic trees is given below; it is apparent from Fig. 2 that the phylogenetic tree shown in Fig. 2(a) is more similar to the standard phylogenetic tree, shown in Fig. 2(d), generated by MEGA software. First, PDV and APMV have a high similarity, which is also clearly reflected in Table 2. However, PDV and APMV in Fig. 2(b) and Fig. 2(c) are not the most similar of all sequences, and they deviate from the standard phylogenetic tree generated by MEGA software. In addition, CVV is also very similar to PDV and APMV, but CiLRV is not similar to them; hence, it is not reasonable to place CiLRV close to them in Fig. 2(b). Finally, Fig. 2(c) does not reveal a higher similarity between CiLRV and ALMV. In conclusion, compared with other methods, our proposed method can obtain the similarities of RNA secondary structures more accurately.

Furthermore, the phylogenetic trees were compared using the classical Robinson-Foulds (RF) metric [45] to calculate the Pearson correlation efficiency between the tip-to-tip distance matrices of the two trees. A and B are two $n \times n$ matrices generated from the two phylogenetic trees being compared. The Pearson correlation

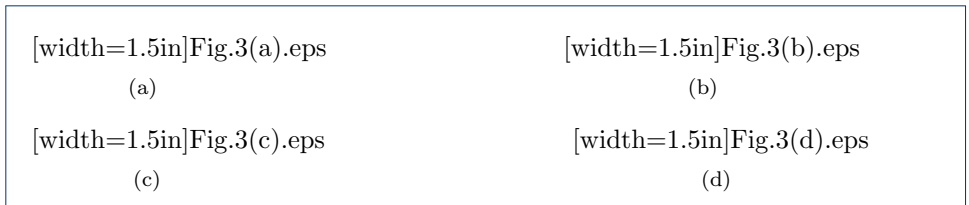

Figure 3 Phylogenetic tree of 11 non-coding RNA sequences generated by (a)DFA,(b)Li′s method,(c)RNAdistance,and (d)MEGA software.

coefficient between the two matrices is calculated as follows:

$$r = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(A\left(i,j\right)-\overline{A}\right)\left(B\left(i,j\right)-\overline{B}\right)}{\sqrt{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(A\left(i,j\right)-\overline{A}\right)^{2}}\sqrt{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(B\left(i,j\right)-\overline{B}\right)^{2}}} = \frac{1}{n-1}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(\frac{A\left(i,j\right)-\overline{A}}{\sigma_A}\right)\left(\frac{B\left(i,j\right)-\overline{B}}{\sigma_B}\right), \quad (1)$$

where $\overline{A}$ is the mean of matrix A and $\overline{B}$ is the mean of matrix B.

The Pearson correlation efficiencies between the evolutionary trees generated by the different methods and those generated by MEGA software are shown in Table 5. The Pearson correlation coefficient between the evolutionary tree generated by Li′s method and the evolutionary tree generated by MEGA software is 0.4542, the Pearson correlation efficient between the evolutionary tree generated by the RNAdistance method and the evolutionary tree generated by MEGA software is 0.6334, and the Pearson correlation efficient between the evolutionary tree generated by the DFA method and the evolutionary tree generated by MEGA software is 0.6345. Obviously, our approach is closer to the evolutionary tree generated by MEGA software than the other approaches.

Table 5 The Pearson correlation efficiency between the phylogenetic trees generated by the different methods and those generated by MEGA software

Method	RNAdistance	Li′s method	DFA method
Pearson correlation efficient	0.6334	0.4542	0.6345

In the second experiment, as indicated in Fig. 3, the experimental sequences are accurately divided into three families in Fig. 3(d), tRNA, RNase P RNA, and a group of human ncRNA in the NONCODE database. Fig. 3(d) shows that the $NT\_033777.3$ sequence has a high similarity to the tRNA family, which is also reflected in Fig. 6(a) and Table 4. In Fig. 3(b), $NT\_033777.3$ and $MF489813.1$ are classified into a group, and $NT\_033777.3$ is the sequence that is the least similar to the others in the tRNA family, which is unreasonable. $NT\_033777.3$ was mistakenly placed in a group close to the NONCODE human non-coding RNA in Fig. 3(c). Based on the above analysis, the method based on DFA and DWT is an effective algorithm for RNA secondary structure similarity analysis. The Pearson correlation efficiencies between the evolutionary trees generated by the DFA method, RNAdistance, Li′s method and the evolutionary tree generated by MEGA software are

Table 6 The Pearson correlation efficiency between the phylogenetic trees generated by the different methods and those generated by MEGA software

Method	RNAdistance	Li's method	DFA method
Pearson correlation efficient	0.5013	0.5904	0.6505

shown in Table 6. Evidently, the results obtained by our method in this experiment are more accurate.

Conclusions

This paper proposes a hybrid method for the similarity comparison of RNA secondary structures. The algorithm is based on the existing RNA triple vector representation, uses DWT to process the feature sequences, and captures the fractal characteristics using the DFA method. Compared with several commonly used RNA comparison methods, the approximate relationships between the RNA secondary structures obtained by the DFA and wavelet transform method are close to the actual relationships. However, the secondary structures predicted by the minimum free energy in the Vienna RNA package is not optimal, and finding the optimal secondary structure for RNA rapidly and efficiently remains a challenging problem. In addition, the method is not yet excellent for analyzing shorter RNAs, and systematic studies will be carried out on RNAs of different lengths and characteristics.

[width=9cm]Fig.4.png

Figure 4 The flowchart of our method.

Methods

In this paper, a combination of the discrete wavelet transform, detrended fluctuation analysis and sliding window is proposed to evaluate the similarity of RNA secondary structures based on the TV-Curve of RNA. The TV-Curve of RNA is used to construct the feature vector, which is analyzed by the discrete wavelet transform, and then the sliding window is introduced. For a signal of fixed length within the window, the fractal dimension is calculated using the detrended fluctuation analysis (DFA) method to obtain the distance matrix. Finally, the corresponding phylogenetic tree is obtained. The algorithmic process is shown in Fig. 4.

Basic concepts of fractal dimension

Theoretical Fractal Dimension

The German mathematician F. Hausdorff studied the properties and quantities of singular aggregates and proposed the concept of fractal dimension called the Hausdorff dimension in 1918 [46]. It is one of the most important fractal dimensions. The Hausdorff dimension can be used for any set. The theory of the Hausdorff dimension is as follows: Suppose a non-empty subset A of an n-dimensional Euclidean space $\mathbb{R}^n$, where the diameter of A is the maximum distance of any two points in A, then:

$$|A| = \sup\{|x - y| : x, y \in A\}, \tag{2}$$

Algorithm 1. Analysis of RNA secondary structures by the discrete wavelet transform and DFA method

Input: N RNA primary sequences and their secondary structures indicated by a dot-bracket representation.
Output: Distance matrix T.
1) The TV-Curve of an RNA secondary structure is represented as feature vector S_i .
2) Discrete wavelet transform deals with feature sequences using the Haar wavelet function.
3) At level M, set the window width k and slide a fixed length window along the signal to generate a k_i sequence for each S_i:
$W_i^{(m)} = [S_i(m), S_i(m+1), ..., S_i(m-k+1)], m = 1...K_i, i = 1...N$
and
$K_i = length(S_i) - k + 1$
4) Perform detrended fluctuation analysis for each window:
$F_2(m) = DFA(S(m : m+k-1), \min(\left\lfloor \frac{k}{2} \right\rfloor, \max\_level), 1)$
and
$F_{2i} = [F_2(1), F_2(2), ..., F_2(K_i)], i = 1...N$
5) Compute the distance of $\{F_{2i}, F_{2j}\}$:
$d(F_{2i}, F_{2j}) = 1 - \frac{\langle F_{2i}, F_{2j} \rangle}{|F_{2i}| \cdot |F_{2j}|}, i = 1...N, j = 1...N$

where $sup\{\cdot\}$ indicates the supremum of $\{\cdot\}$. If $C \subset \bigcup_{i=1}^{\infty} A_i$ and $0 \le |A_i| \le \delta$, for any i, $\{A_i\}$ is called a δ-cover of C.

Assume $C \subset \mathbb{R}^n$ and $0 \le s \le \infty$, for any $\delta > 0$,

$$H_\delta^s(C) = \inf\left\{ \sum_{i=1}^{\infty} |U_i|^s : \{A_i\} \ is\ a\ \delta\ -\ cover\ of\ C \right\}, \tag{3}$$

The above equation refers to the coverage of C whose diameter does not exceed δ, and the sum of these diameters is minimized to the power of s. The symbol $inf\{\cdot\}$ takes the minimum value of $\{\cdot\}$.

As δ decreases, the clusters covering C in equation (3) also decrease, and the infimum $H_\delta^s(S)$ increases accordingly. Moreover, when $\delta \to 0$, it tends to a limit, denoted as

$$H^s(C) = \lim_{\delta \to 0} H_\delta^s(C), \tag{4}$$

For any subset C in $\mathbb{R}^n$, the limit exists, and the limiting value is commonly 0 or ∞, and $H^s(C)$ is called the $s - dimensional\ Hausdorff\ measure$ of C.

In equation (3), for a given set $C \subset \mathbb{R}^n$ and $\delta < 1$, $H_\delta^s(C)$ does not increase with respect to s, so it can be shown from equation (4) that $H^s(C)$ does not increase either. Further, if $t > s$, and $\{A_i\}$ is the δ-cover of C, then

$$\sum_i |A_i|^t \le \sum_i |A_i|^{t-s} |A_i|^s \le \delta^{t-s} \sum_i |A_i|^s, \tag{5}$$

Take the infimum,

$$H_\delta^t(C) \le \delta^{t-s} H_\delta^s(C), \tag{6}$$

In the case of $t > s$, let $\delta \to 0$; if $H^s(C) < \infty$, then $H^t(C) = 0$; thus, the existence of a critical point of s makes $H^s(C)$ "jump" from ∞ to 0. This critical value is known as the $Hausdorff\ dimension$ of F, recorded as $\dim_H C$, and occasionally called the $Hausdorff - Basicovitch\ dimension$.

The definition is expressed as

$$\begin{aligned} \dim_H C &= \inf\{s \geq 0 : H^s(C) = 0\} \\ &= \sup\{s : H^s(C) = \infty\}, \end{aligned} \tag{7}$$

Therefore,

$$H^s(C) = \begin{cases} \infty, & if \quad 0 \leq s < \dim_H C \\ 0, & if \quad s > \dim_H C \end{cases} \tag{8}$$

In addition, if $s = \dim_H C$, $H^s(C)$ satisfies,

$$0 \leq H^s(C) \leq \infty$$

Algorithms of Fractal Dimension Calculation

Although the fractal dimension is widely defined, it is subject to some limitations in practical applications. Due to its high computational complexity and the discrete and finite nature of the scale factor, we usually use a number of algorithms to approximate the fractal dimension. To date, many methods have been developed to compute fractal dimensions, such as Katz′s algorithm [47], Petrosian′s algorithm [48], Higuchi′s algorithm [49], multifractal detrended fluctuation analysis (MFDFA) [50], detrended fluctuation analysis (DFA) [51] and fluctuation analysis (FA) [52]. In terms of processing time, Katz′s algorithm and the DFA method are relatively slow. However, in terms of processing accuracy, Higuchi′s algorithm, detrended fluctuation analysis (DFA) and fluctuation analysis (FA) are relatively accurate. In the following, we will introduce Higuchi′s algorithm, fluctuation analysis (FA) and detrended fluctuation analysis (DFA).

Higuchi′s Algorithm

Treating an RNA sequence as a one-dimensional signal $s_1, s_2, ..., s_N$, construct a subsequence as

$$s_m^k = \left\{ s(m), s(m+k), s(m+2k), ..., s\left(m + \left\lfloor \frac{N-m}{k} \right\rfloor k\right) \right\}, m = 1...k \tag{9}$$

where m indicates the starting position of the signal and k is the measurement scale. $\left\lfloor \frac{N-m}{k} \right\rfloor$ means the integer part of $\frac{N-m}{k}$, which is the number of terms in s_m^k.

Calculate the average length $L_m(k)$ of the curve with starting position m according to the measurement scale k.

For $m = 1, ..., k$,

$$L_m(k) = \frac{\sum_{i=1}^{\left\lfloor \frac{N-m}{k} \right\rfloor} |s(m+ik) - s(m+(i-1)k)|(N-1)}{\left\lfloor \frac{N-m}{k} \right\rfloor k^2} \tag{10}$$

where N is the length of the signal, and $\frac{(N-1)}{\left(\left\lfloor \frac{N-m}{k} \right\rfloor k\right)}$ is the normalization term. With the given measurement scale, the approximate length of the signal with starting position m is computed, $m = 1...k$.

Thus, the approximated signal length can be obtained by:

$$L(k) = \frac{1}{k}\sum_{m=1}^{k} L_m(k), k = 1...k_{\max} \tag{11}$$

Finally, the fractal dimension h^* of the signal is calculated by the least squares method [53]:

$$h^* = \arg\min_{h} \sum_{k=1}^{K} \left(h\log\left(\frac{1}{k}\right) - \log\left(L(k)\right) + c \right)^2 \tag{12}$$

where c is the bias.

Fluctuation Analysis (FA) method

The fluctuation analysis (FA) method is commonly used to calculate the *Hurst parameter* for time series. It works as follows.

Suppose $X = \{X_i, i = 1, 2, ..., N\}$ is a random process with a mean value of μ and a variance of σ^2. First, we remove the mean from the signal and represent the new signal as $x = \{x_i, i = 1, 2, ..., N\}$, where $x_i = X_i - \mu$.

Then, construct a new one-dimensional signal $y = \{y_n, n = 1, 2, ..., N\}$ with the sum of the first n terms of x,

$$y(n) = \sum_{i=1}^{n} x_i, \tag{13}$$

Test whether the following formula satisfies the power law formula,

$$F^{(2)}(m) = \left\langle \left| y(n+m) - y(n) \right|^2 \right\rangle^{1/2} \sim m^H, \tag{14}$$

That is, whether $\log_2\left(F^{(2)}(m)\right)$ and $\log_2(m)$ are satisfied:

$$\log_2\left(F^{(2)}(m)\right) = H * \log_2(m) + c, \tag{15}$$

where c is a constant.

From the above equation, we can obtain the *Hurst parameter* H.

In the end, we can obtain FD by the following formula:

$$FD = 2 - H \tag{16}$$

Detrended fluctuation analysis (DFA) method

Owing to the complicated characterization of long-range correlations and power-law statistics that RNA TV-Curves have, it follows that comprehensively describing and studying the internal features of non-stationary signals such as RNA TV-Curves by traditional methods is difficult. For complex, highly non-stationary signals, time series analysis methods derived from statistical physics are currently used. The DFA method can effectively eliminate various unknown trends in the signal, thus

avoiding the interference caused by noise and signal instability. The DFA method has also been widely used in image processing and other fields [54]. The DFA method has been proven to be an effective method to analyze non-stationary signals. In summary, as the FA, the signal $x_r, r = 1, ..., N$ is preliminarily processed, and the mean value $\langle x \rangle$ is removed. The sum of the first term i of x_r as is the i of a new sequence:

$$Y(i) = \sum_{r=1}^{i} \left[x_r - \langle x \rangle \right], i = 1, 2, ..., N, \tag{17}$$

Next, the signal $Y(i)$ is equally split into data boxes of length k. k, which is the measurement scale, is the number of points per data box.

In each box of length k, a least-squares line is fit to the data. As a result, fitting data box $y_k(j)$ is obtained. Then, we detrend the signal Y,

$$\widetilde{Y_k}(j) = Y(j) - y_k(j) \tag{18}$$

In the end, the different data obtained for each box are integrated and averaged,

$$F_2(k) = \left[\frac{1}{N} \sum_{j=1}^{N} F_{DFA}^2(k) \right]^{1/2}, \tag{19}$$

where, $F_{DFA}^2(k) = \frac{1}{k} \sum_{j=1}^{k} \widetilde{Y_k}^2(j)$.

The scaling exponent α is the slope of the line of fit of $F_2(k)$ and k.

The fractal dimension is calculated by:

$$FD = 2 - \alpha \tag{20}$$

Graphical characterization of RNA secondary structure

Dot-bracket representation

The sequence of RNA consists of the bases C, G, U, and A. The primary sequence does not contain structural information. During the exploration of the RNA function, two bases combine with each other to form base pairs. Typically, the base pairs are A-U, G-C. Thus, the secondary structure of RNA is formed. Currently, 'the dot-bracket′ representation is commonly used to signify the secondary structure of

[width=8cm]Fig.5.png

Figure 5 Graphical representation of the eight nucleotides.

RNA, the dissociative bases and base pairs are indicated by dots and parentheses, respectively. Among them, open brackets express the base pairs near the 50-terminal of the RNA chain, while closed brackets indicate the bases close to the 30-terminal. This representation can be obtained via the Vienna RNA package. However, this linear representation is degenerate in some cases in which different RNA secondary

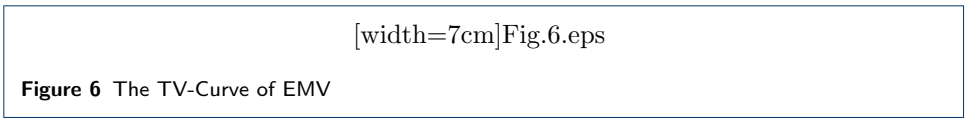

Figure 6 The TV-Curve of EMV

structures have the same characteristic sequence. Nevertheless, this circumstance usually occurs in short RNA sequences, which can be ignored in the analysis of long and complex RNA secondary structures.

The TV-Curve representation of RNA secondary structures

Based on the 'dot-bracket′ representation introduced previously, in this part, we will show a method for the feature extraction of RNA secondary structures, called 'RNA triple vector curve′ representation. This method was proposed by Li in 2012, as a 2-D graphical representation of RNA secondary structures, in which RNA sequence information and structural information are considered. The bases in the RNA secondary structure are divided into two types: nucleotide bases paired by hydrogen bonds and unpaired nucleotide bases. The four unpaired nucleotide bases are denoted by C (cytosine), G (guanine), U (uracil), and A (adenine); in addition, C′, G′, U′, and A′ denote paired nucleotide bases. As shown in Fig. 5, each of the eight symbols is described by three vectors,

$$\begin{aligned}
&(1,-1),(1,1),(1,-1)\Rightarrow C,(1,1),(1,-1),(1,-1)\Rightarrow C'\\
&(1,-1),(1,-1),(1,-1)\Rightarrow G,(1,1),(1,-1),(1,-1)\Rightarrow G'\\
&(1,1),(1,-1),(1,1)\Rightarrow U,(1,-1),(1,-1),(1,1)\Rightarrow U'\\
&(1,1),(1,1),(1,1)\Rightarrow A,(1,-1),(1,-1),(1,1)\Rightarrow A'
\end{aligned} \tag{21}$$

The feature sequence is read from 50-terminal to 30-terminal, and the vectors are sequentially connected to obtain the TV-Curves. For example, Fig. 6 demonstrates the secondary structure of EMV and the corresponding TV-Curve. An RNA sequence of length N generates a TV-Curve with an X-axis length of 3N. An RNA sequence and its secondary structure may produce only one TV-Curve, and similarly, a TV-Curve may represent only one RNA secondary structure. Moreover, the TV-Curve has no information missing from the process of feature extraction of the RNA.

Comparison of RNA secondary structures based on the RNA TV-Curve by detrended fluctuation analysis

Discrete Wavelet Transform

The wavelet transform decomposes data, functions or operators into components of different frequencies, and then studies each component on the corresponding scale [55,56]. We now study the wavelet transform within the range of signal analysis. Similar to the discrete Fourier transform, the discrete wavelet transform is also a time-frequency description method. The signal is decomposed into two characters: the approximation coefficients (AC) and the detailed coefficients (DC).

For the signal $x(t)$, we use two functions at the same time, wavelet function $\Psi(x)$ and scaling function $\Phi(x)$. $\Phi(x)$ for the s(t) overview approximation and $\Psi(x)$

details the approximate function of $x(t)$. $\Phi(x)$ can be formulated as:

$$\Phi(x) = \sqrt{2} \sum_{n} h_n \Phi(2x - n), \tag{22}$$

where h_n is the low-pass filter. The wavelet function $\Psi(x)$ can be calculated by:

$$\Psi(x) = \sqrt{2} \sum_{n} g_n \Phi(2x - n), \tag{23}$$

Two sets of orthogonal functions can be extracted from the wavelet transform: shifted wavelet functions $\Psi(x - k)$ and scaling functions $\Phi(x - k)$.

For any signal $x = \left\{a_k^j\right\}$, the approximation coefficients (AC) and detailed coefficients (DC) in the next level can be quickly calculated as follows:

$$a_k^{j+1} = \sum_{i \in Z} a_i^j \overline{h}_{i-2k}, j = 0, 1, 2, ... \tag{24}$$

and

$$d_k^{j+1} = \sum_{i \in Z} a_i^j \overline{g}_{i-2k}, j = 0, 1, 2, ... \tag{25}$$

The decomposition is taken level by level. The lengths of AC and DC obtained after the progressive decomposition are always half the length of the level sequence antecedently. The processing of RNA secondary structures using DFA allows for the description of the overall trends and general characteristics of the signal, as well as the local details.

Windowed detrended fluctuation analysis (DFA) method

If the fractal dimension is calculated by the DFA of the whole signal, only one scalar can be obtained. The sliding window only calculates the signal values in a fixed length window [57]. We introduce a sliding window that moves along the signal and calculates the fractal dimension within the window. Calculating the fractal dimensions along a set sliding window yields a feature vector, not just a number. The length of this feature vector is $N - k - 1$, where N is the signal length and k is the window width. m indicates the starting position of the signal. A least-squares line is fit to the data in each window of length k. As a result, a fitting data box $y_k^m(j)$ is obtained. Then, we detrend signal Y:

$$\widetilde{Y}_k(j) = Y(j) - y_k^m(j) \tag{26}$$

For each window, calculate the following formula,

$$F_{DFAm}^2(k) = \frac{1}{k} \sum_{j=1}^{k} \widetilde{Y_k}^2(j). \tag{27}$$

Next, the average is calculated,

$$F_2(k) = \left[\frac{1}{N} \sum_{j=1}^{N} F_{DFAm}^2(k) \right]^{1/2}, \tag{28}$$

From the above, the $lb-lb$ graph of $F_2(k)-k$ is drawn for different values of k, it is fit to a straight line, and the scaling exponent α is obtained. The fractal dimension is calculated by:

$$FD = 2 - \alpha$$

Abbreviations
DFA: Detrended fluctuation analysis; DWT: Discrete wavelet transform; TV-Curve: Triple vector curve; FD: Fractal dimension.

Declarations
The manuscript has not been published before and is not being considered for publication elsewhere. All authors have contributed important intellectual content for the creation of this manuscript and have read and approved the final manuscript. We declare that there are no conflicts of interest.

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Availability of data and materials
The datasets analyzed during the current study are available from the RFAM and NONCODE databases, http://rfam.xfam.org/, http://www.noncode.org/.

Competing interests
The authors declare that they have no competing interests.

Funding
This work was financially supported by the National Natural Science Foundation of China under Grant, the Guangxi Natural Science Foundation, and Scientific Research Foundation of Guangxi University.

Authors' contributions
YLN proposed the method and derived the formula. LY drafted the manuscript and designed the implementation of the method. PW revised the manuscript. LXC helped to improve the manuscript. All authors read and approved the final manuscript.

Acknowledgments
Not applicable

Footnotes
Not applicable

Author details
[1]Computer and Electronic Information, Guangxi University, 530004 Nanning, China. [2]Computer and Information Science, Northeastern University, 02115 Boston, USA. [3]Guangxi Normal University for Nationalities, Chongzuo, China.

References
1. Borges dos Anjos LR, Do Prado Assunção L, Lima Freitas B, Pereira Campos N, Sousa Silva N, Araújo Monteiro JM, et al. Popularização da Ciência: Desmistificando o Dogma Central da Biologia Molecular. Revista de Ensino de Bioquímica. 2019;16(2):71-86.
2. Ke Y, Jiahua R, Zhao H, Lu Y, Xiao N, Yang Y. Accurate Prediction of Genome-wide RNA Secondary Structure Profile Based On Extreme Gradient Boosting2019.
3. Lu W, Tang Y, Wu H, Huang H, Fu Q, Qiu J, et al. Predicting RNA secondary structure via adaptive deep recurrent neural networks with energy-based filter. BMC Bioinformatics. 2019;20(25):684.
4. Zhou T, Routh A. Mapping RNA-capsid interactions and RNA secondary structure within virus particles using next-generation sequencing. Nucleic Acids Research. 2019;48.
5. Wang F, Akutsu T, Mori T. Comparison of Pseudoknotted RNA Secondary Structures by Topological Centroid Identification and Tree Edit Distance. Journal of Computational Biology. 2020.

6. Michela Q, Luca T, Bioinformatics MEJ. ASPRAlign: a tool for the alignment of RNA secondary structures with arbitrary pseudoknots. 2020.
7. Havgaard JH, Torarinsson E, Gorodkin J. Fast pairwise structural RNA alignments by pruning of the dynamical programming matrix. PLoS Comput Biol. 2007;3(10):1896-908.
8. Shapiro BA, Zhang KZJCAB. Comparing Multiple RNA Secondary Structures Using Tree Comparisons. 1990;6(4):309-18.
9. Allali J, Sagot MF. A new distance for high level RNA secondary structure comparison: IEEE Computer Society Press; 2005.
10. Sankoff D. Simultaneous Solution of the RNA Folding, Alignment and Protosequence Problems. Siam Journal on Applied Mathematics - SIAMAM. 1985;45.
11. Mathews DJB. Predicting a set of minimal free energy RNA secondary structures common to two sequences. 2005;21(10):p. 2246-53.
12. Hofacker I, Bernhart S, Stadler PJB. Alignment of RNA base pairing probability matrices. 2004;20(14):p. 2222-7.
13. Holmes IJBB. Accelerated probabilistic inference of RNA structure evolution. 2005;6.
14. Hull HJ, B LR, Stormo GD, Jan GJB. Pairwise local structural alignment of RNA sequences with sequence similarity less than 40
15. Will S, Reiche K, Hofacker IL, Stadler PF, Backofen R. Inferring Noncoding RNA Families and Classes by Means of Genome-Scale Structure-Based Clustering. PLOS Computational Biology. 2007;3(4):e65.
16. Will S, Schmiedl C, Miladi M, Möhl M, Backofen R, editors. SPARSE: Quadratic Time Simultaneous Alignment and Folding of RNAs without Sequence-Based Heuristics. Proceedings of the 17th international conference on Research in Computational Molecular Biology; 2013.
17. Siebert S, Backofen RJB. MARNA: multiple alignment and consensus structure prediction of RNAs based on sequence structure comparisons. 2005;21(16):p.3352-9.
18. Torarinsson E, Havgaard JH, Gorodkin, Bioinformatics JJ. Multiple structural alignment and clustering of RNA sequences. 2007;23(8):926-32.
19. Kiryu H, Tabei Y, Kin T, Asai KJB. Murlet: a practical multiple alignment tool for structural RNA sequences. 2007;23(13):1588-98.
20. Sorescu D, Möhl M, Mann M, Backofen R, Will S. CARNA—alignment of RNA structure ensembles. Nucleic acids research. 2012;40:W49-53.
21. Do CB, Chuan-Sheng F, Serafim BJB. A max-margin model for efficient simultaneous alignment and folding of RNA sequences. 2008(13):13.
22. Hochsmann M, Toller T, Giegerich R, Kurtz S, editors. Local similarity in RNA secondary structures. Computational Systems Bioinformatics Csb IEEE Bioinformatics Conference Csb; 2003.
23. Guignon V, Chauve C, Hamel S. An Edit Distance Between RNA Stem-Loops: Springer Berlin Heidelberg; 2005.
24. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P. Fast folding and comparison of RNA secondary structures. Monatshefte für Chemie / Chemical Monthly. 1994;125(2):167-88.
25. Jörg F, Pavankumar V, Andrea B, Bérénice B, A DM, Tomas K, et al. The RNA workbench 2.0: next generation RNA data analysis. 2019(W1):W1.
26. Miladi M, Raden M, Will S, Backofen R, editors. Fast and Accurate Structure Probability Estimation for Simultaneous Alignment and Folding of RNAs. 19th International Workshop on Algorithms in Bioinformatics (WABI 2019); 2019.
27. Bo, Liao, Weiyang, Chen, Xingming, Sun, et al. A binary coding method of RNA secondary structure and its application. 2009.
28. Letters MRJCP. On characterization of DNA primary sequences by a condensed matrix. 2000;317(1-2):29-34.
29. Yao YH, Liao B, Wang TMJJoMST. A 2D graphical representation of RNA secondary structures and the analysis of similarity/dissimilarity based on it. 2005;755(1-3):131-6.
30. Randi M, Letters DPJCP. Novel spectral representation of RNA secondary structure without loss of information. 2009;476(4-6):277-80.
31. Li Y, Duan M, Liang YJBB. Multi-scale RNA comparison based on RNA triple vector curve representation. 2012;13(1):280.
32. Siyu S, Stephen M, Ruqiang Y, Informatics BPJITo. Highly-Accurate Machine Fault Diagnosis Using Deep Transfer Learning. 2018:1-.
33. Wang S, Wang X. Prediction of protein structural classes by different feature expressions based on 2-D wavelet denoising and fusion. BMC Bioinformatics. 2019;20(Suppl 25):701.
34. Oliveira RB, Pereira AS, Tavares JOMRSJNC, Applications. Computational diagnosis of skin lesions from dermoscopic images using combined features. 2018.
35. Pando J, Sands L, Shaheen SEJPRESN, Physics SM. Detection of protein secondary structures via the discrete wavelet transform. 2009;80(5):051909.
36. Esteller R, Vachtsevanos G, Echauz J, Lilt B, editors. A comparison of fractal dimension algorithms using synthetic and experimental data. IEEE International Symposium on Circuits & Systems; 1999.
37. Mandelbrot BB, Van Ness JW. Fractional Brownian Motions, Fractional Noises and Applications. SIAM Review. 1968;10(4):422-37.
38. Li J, Ma X, Zhao M, Cheng XJE. A Novel MFDFA Algorithm and Its Application to Analysis of Harmonic Multifractal Features. 2019;8(2).
39. Zhao T, Wu L, Li D, Ding YJDDiN, Society. Multifractal Analysis of Hydrologic Data Using Wavelet Methods and Fluctuation Analysis. 2017;2017:1-18.
40. Yu Z, Lau K, Zhou L. Clustering of protein structures using hydrophobic free energy and solvent accessibility of proteins. Physical review E, Statistical, nonlinear, and soft matter physics. 2006;73:031920.
41. Yang L, Tang YY, Lu Y, Luo H. A Fractal Dimension and Wavelet Transform Based Method for Protein Sequence Similarity Analysis. IEEE/ACM Trans Comput Biol Bioinform. 2015;12(2):348-59.
42. Kumar DJ, Pal CP, Adwitiya C, Sarif HS, Pallab BJSR. Analysis of Purines and Pyrimidines distribution over

miRNAs of Human, Gorilla, Chimpanzee, Mouse and Rat. 2018;8(1):9974-.
43. Reusken CBEM, Bol JF. Structural Elements of the 3'-Terminal Coat Protein binding Site in Alfalfa Mosaic Virus RNAs. Nucleic Acids Research. 1996;24(14):2660-5.
44. Sudhir K, Glen S, Li M, Christina K, Koichiro TJMB, Evolution. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. 2018(6):6.
45. Robinson DF, Foulds LR. Comparison of phylogenetic trees. bellman prize in mathematical biosciences. 1981;53:131-47.
46. Hausdorff FJMA. Dimension und äßeres Maß. 1918;79(1):157-79.
47. Michael, J., Biology KJCi, Medicine. Fractals and the analysis of waveforms. 1988.
48. Petrosian A, editor Kolmogorov complexity of finite sequences and recognition of different preictal EEG patterns. IEEE Symposium on Computer-based Medical Systems; 1995.
49. Higuchi T. Approach to an irregular time series on the basis of the fractal theory. 1988;31(2):277-83.
50. Paweł, Oświęcimka, Jarosław, Kwapień, Stanisław, E DJPR. Wavelet versus detrended fluctuation analysis of multifractal structures. 2006.
51. Peng, C.-K., Havlin, Shlomo, Stanley, H., et al. Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. 1995.
52. Jing H, Wen-Wen T, Jianbo G. Detection of low observable targets within sea clutter by structure function based multifractal analysis. IEEE Transactions on Antennas and Propagation. 2006;54(1):136-43.
53. Savitzky A, Golay MJE. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. Analytical Chemistry. 1964;36(8):1627-39.
54. Alvarez-Ramirez J, Rodriguez E, Cervantes I, Echeverria JCJPASM, Applications I. Scaling properties of image textures: A detrending fluctuation analysis approach. 2006;361(2):677-98.
55. Esser C, Jaffard SJAiM. Divergence of wavelet series: A multifractal analysis. 2018;328:928-58.
56. Zhang Z. Multivariate Wavelets. Multivariate Time Series Analysis in Climate and Environmental Research. Cham: Springer International Publishing; 2018. p. 97-148.
57. Sun Z, Zhang Z, Chen Y, Liu S, Song Y. Frost Filtering Algorithm of SAR Images With Adaptive Windowing and Adaptive Tuning Factor. IEEE Geoscience and Remote Sensing Letters. 2020;17(6):1097-101.

Additional Files

Additional file 1 — Fig.1(a)
Feature extraction of the RNA secondary structure.

Additional file 2 — Fig.1(b)
The third-level discrete wavelet transform.

Additional file 3 — Fig.1(c)
The DFA fractal dimension of the wavelet obtained with a sliding window width of 17.

Additional file 4 — Fig.2(a)
The phylogenetic tree of eight virus sequences generated by the DFA method.

Additional file 5 — Fig.2(b)
The phylogenetic tree of eight virus sequences generated by Li′s method.

Additional file 6 — Fig.2(c)
The phylogenetic tree of eight virus sequences generated by RNAdistance.

Additional file 7 — Fig.2(d)
The phylogenetic tree of eight virus sequences generated by MEGA software.

Additional file 8 — Fig.3(a)
The phylogenetic tree of 11 non-coding RNA sequences generated by the DFA method.

Additional file 9 — Fig.3(b)
The phylogenetic tree of 11 non-coding RNA sequences generated by Li′s method.

Additional file 10 — Fig.3(c)
The phylogenetic tree of 11 non-coding RNA sequences generated by RNAdistance.

Additional file 11 — Fig.3(d)
The phylogenetic tree of 11 non-coding RNA sequences generated by MEGA software.

Additional file 12 — Fig.4
The flowchart of our method.

Additional file 13 — Fig.5
Graphical representation of the eight nucleotides.

Additional file 13 — Fig.6
The TV-Curve of EMV.

Figures

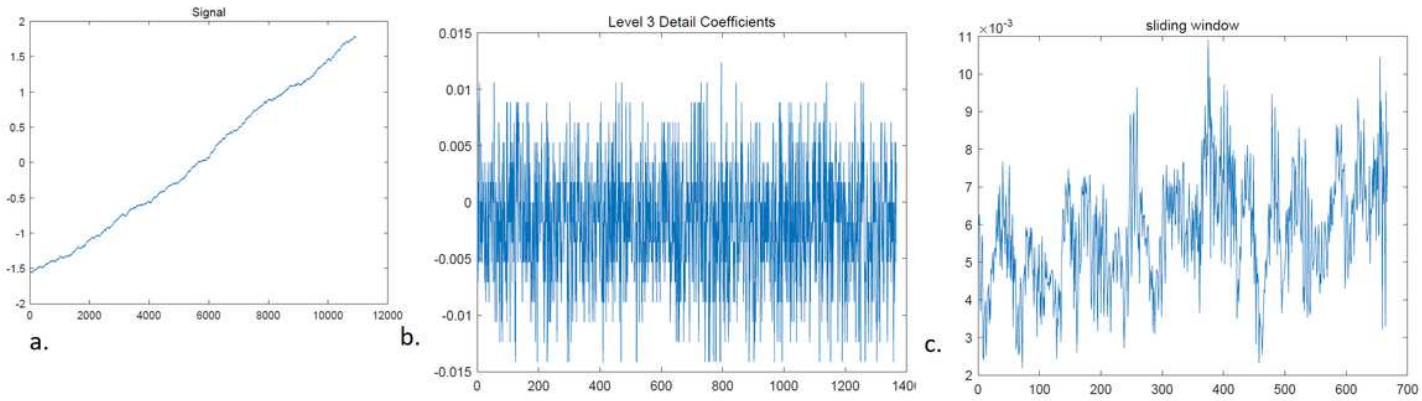

Figure 1

(a) Feature extraction of RNA secondary structure. (b) The third-level discrete wavelet transform. (c) The DFA fractal dimension of (b) obtained with a sliding window width of 17.

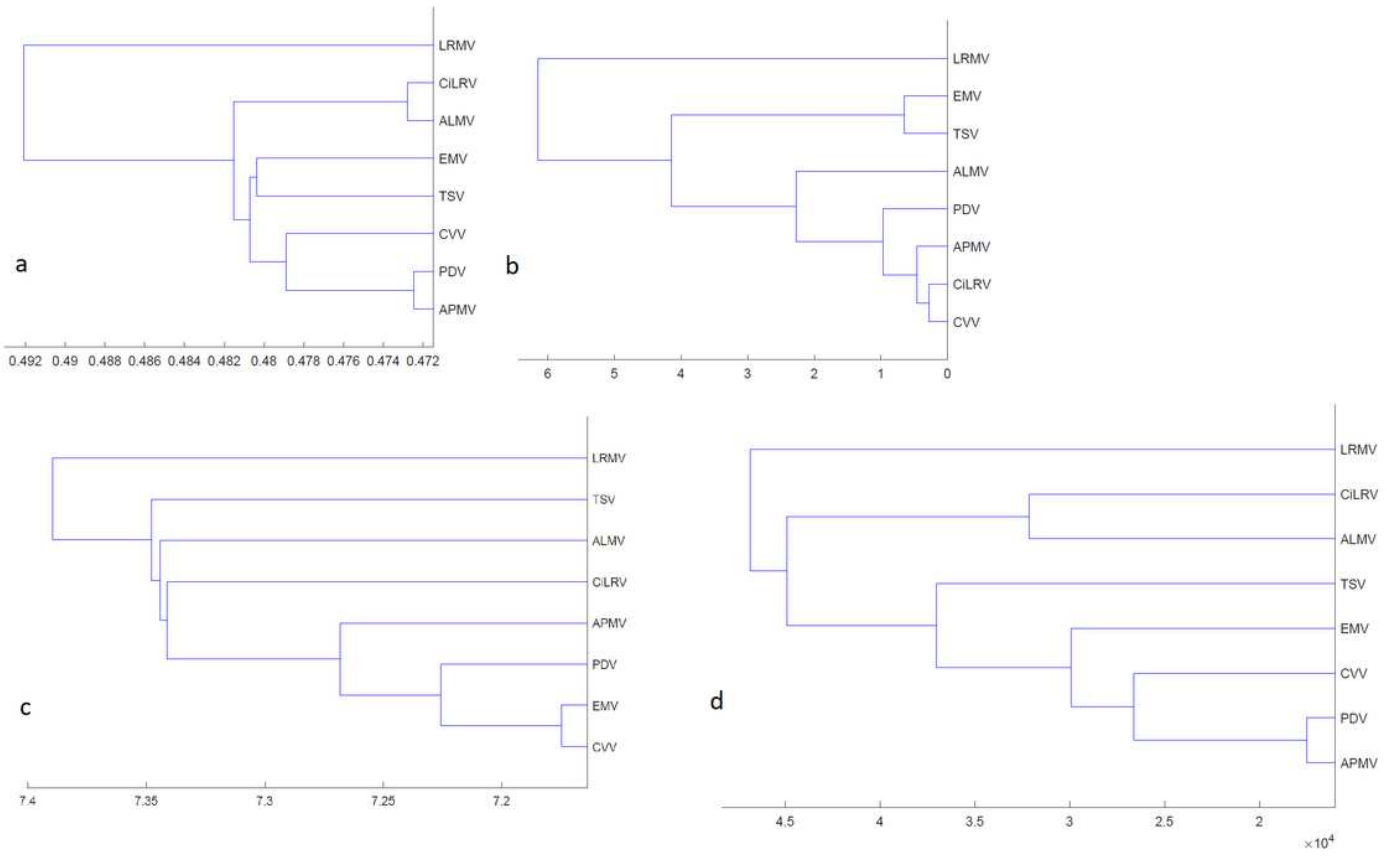

Figure 2

The evolutionary tree of eight virus sequences generated by (a)DFA,(b)Li0s method,(c)RNAdistance, and (d)MEGA software.

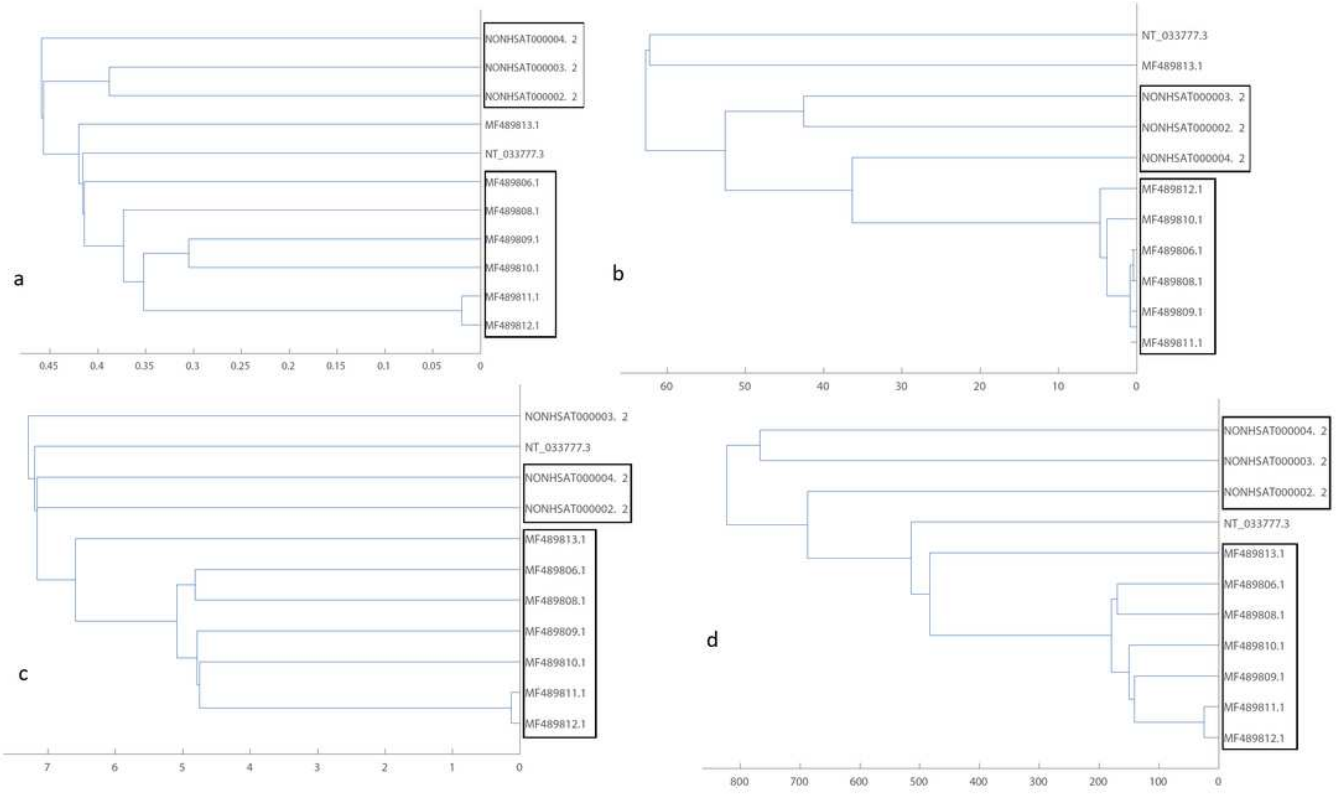

Figure 3

Phylogenetic tree of 11 non-coding RNA sequences generated by (a)DFA,(b)Li0s method, (c)RNAdistance,and (d)MEGA software.

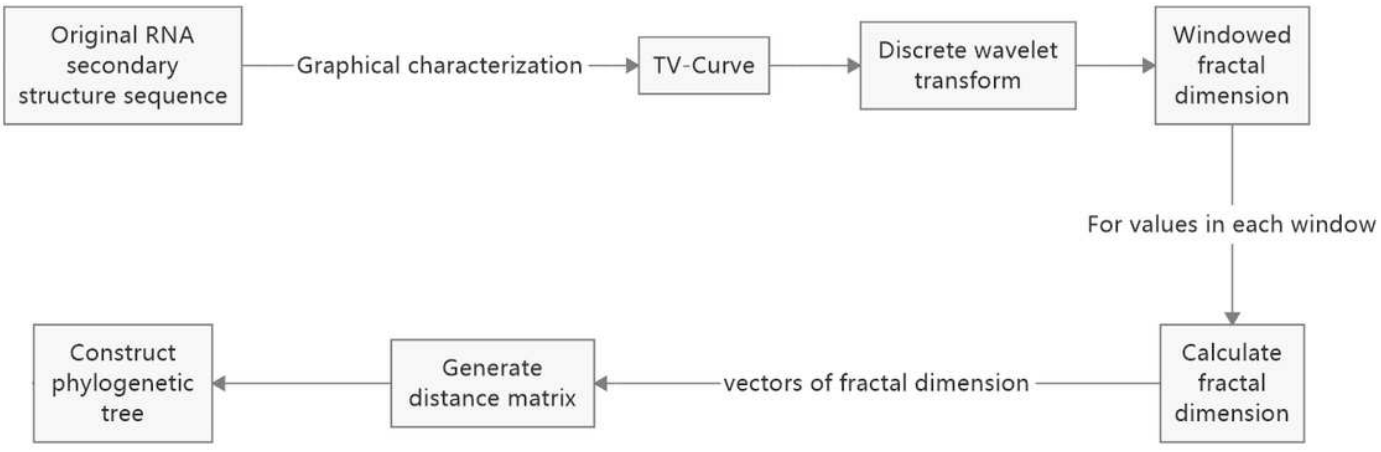

Figure 4

The flowchart of our method.

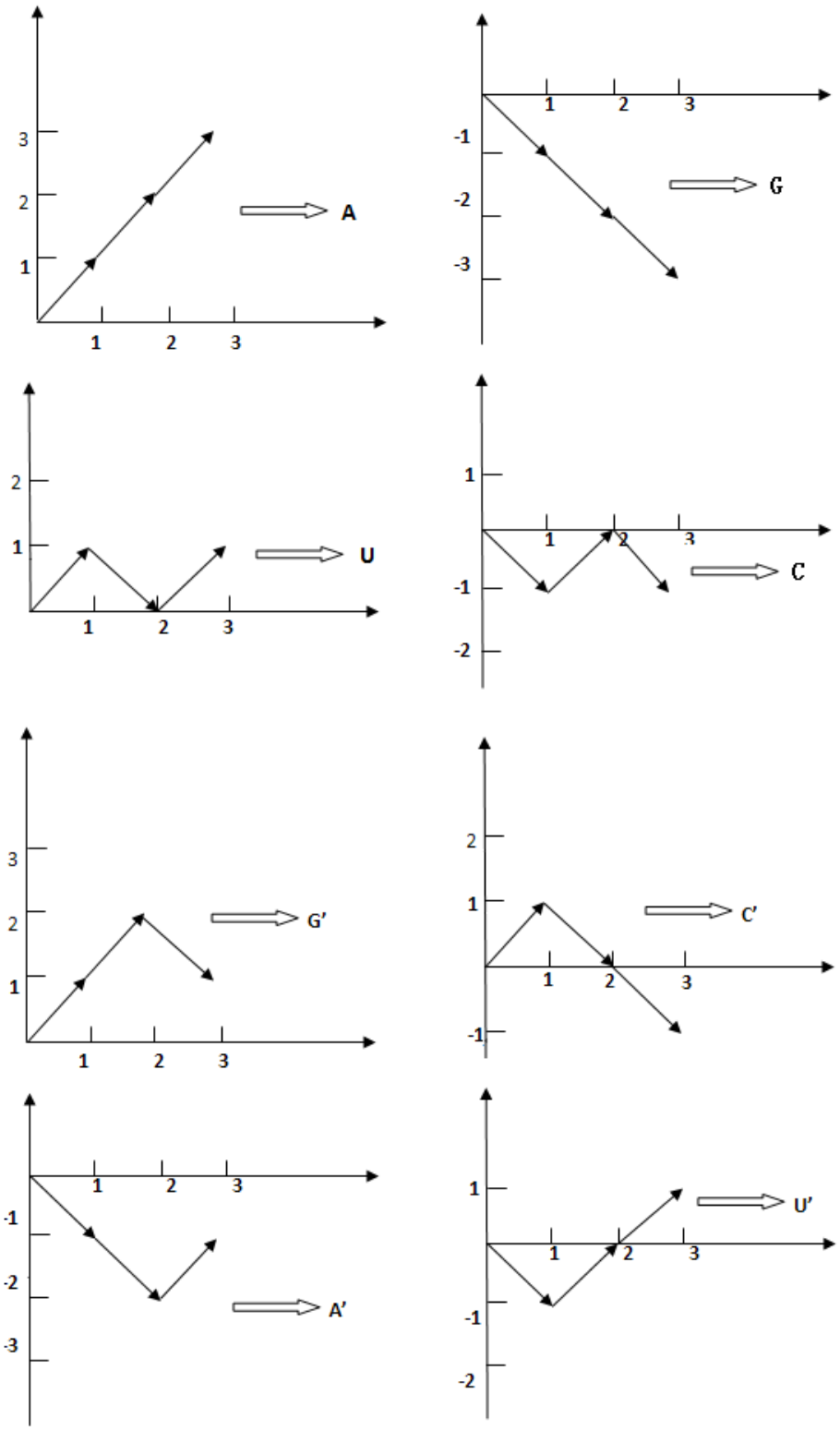

Figure 5

Graphical representation of the eight nucleotides.

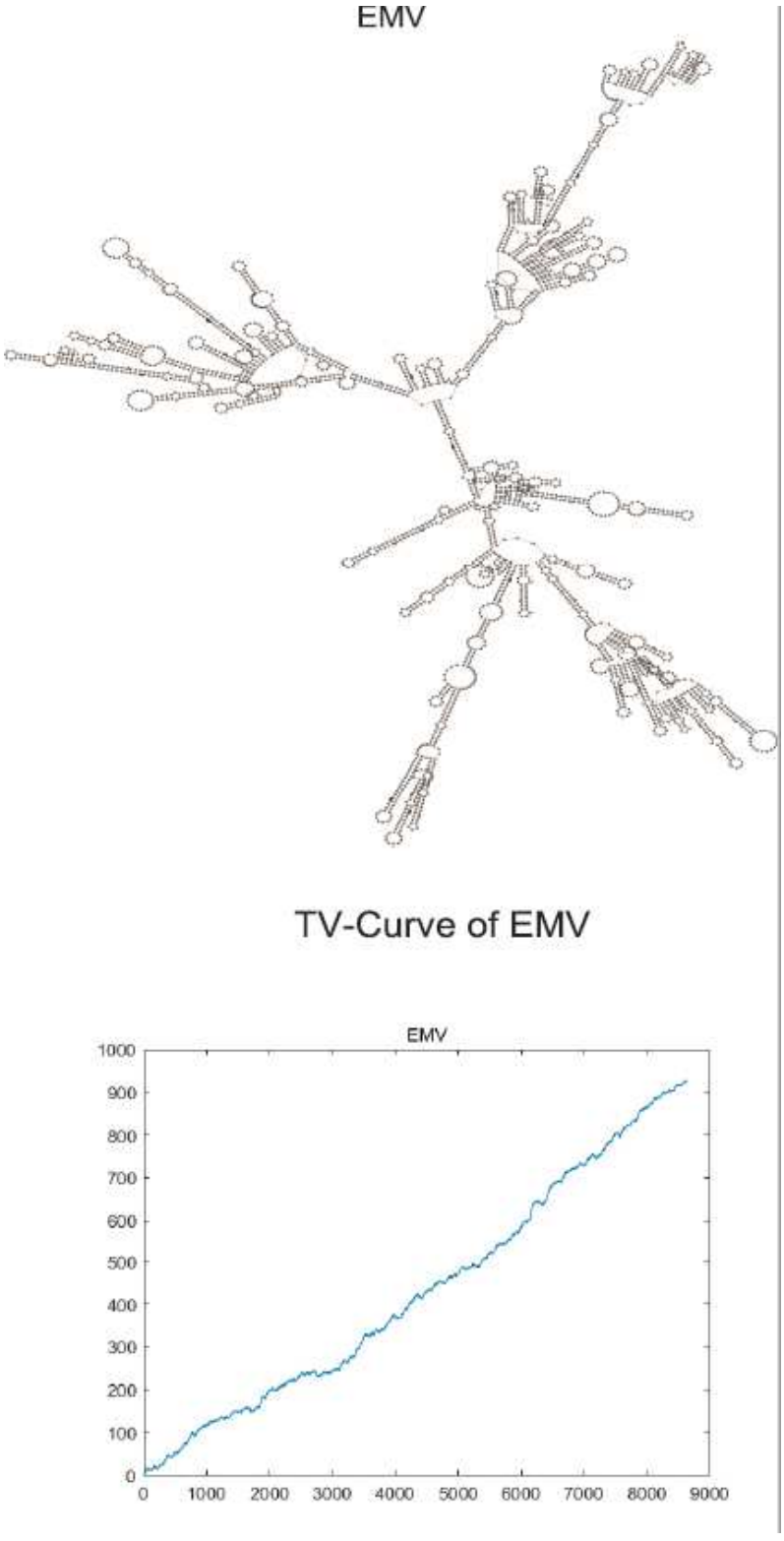

Figure 6

The TV-Curve of EMV