Hindawi
Complexity
Volume 2021, Article ID 8841822, 15 pages
https://doi.org/10.1155/2021/8841822

Research Article

Graph-Based Analysis of RNA Secondary Structure Similarity Comparison

Lina Yang,[1] Yang Liu,[1] Xiaochun Hu,[2] Patrick Wang,[3] Xichun Li,[4] and Jun Wu[5]

[1]*Computer and Electronic Information, Guangxi University, Nanning 530004, China*
[2]*School of Information and Statistics, Guangxi University of Finance and Economics, Nanning 530007, China*
[3]*Computer and Information Science, Northeastern University, Boston 02115, USA*
[4]*Guangxi Normal University for Nationalities, Chongzuo, China*
[5]*School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China*

Correspondence should be addressed to Yang Liu; 1813302002@st.gxu.edu.cn

Received 14 September 2020; Revised 16 February 2021; Accepted 10 March 2021; Published 23 March 2021

Academic Editor: Jia Wu

In organisms, ribonucleic acid (RNA) plays an essential role. Its function is being discovered more and more. Due to the conserved nature of RNA sequences, its function mainly depends on the RNA secondary structure. The discovery of an approximate relationship between two RNA secondary structures helps to understand their functional relationship better. It is an important and urgent task to explore structural similarities from the graphical representation of RNA secondary structures. In this paper, a novel graphical analysis method based on the triple vector curve representation of RNA secondary structures is proposed. A combinational method involving a discrete wavelet transform (DWT) and fractal dimension with sliding window is introduced to analyze and compare the graphs derived from feature extraction; after that, the distance matrix is generated. Then, the distance matrix is analyzed by clustering and visualized as a clustering tree. RNA virus and noncoding RNA datasets are applied to perform experiments and analyze the clustering tree. The results show that the proposed method yields more accurate results in the comparison of RNA secondary structures.

1. Introduction

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 [5–9].

The Sankoff algorithm used a free-energy minimization method to synchronize the folding and alignment of two or more RNA sequences [9]. Combining the information from folding and alignment optimizes the objective function when processing the problem separately, while generalizing the alignment method to the original sequence reconstruction problem and selecting appropriate, existing constraints for possible solutions. However, this suffers from a problem. The time cost is high and about O (n^3k^N) (N is the number of sequences, n is the maximum sequence length, and the value of k is small enough). In practice, it is

impossible to solve and satisfy both the time complexity problem and the larger memory requirements, which, together with the impractical constraints, are all difficult to achieve.

Therefore, to be more effective in RNA secondary structure alignment, a number of improved algorithms were subsequently developed, such as Consan [10] and, Dynalign [11], PMcomp [12], Stemloc [13], Foldalign [14], locARNA [15], SPARSE [16], MARNA [17], FoldAlignM [18], Murlet [19], CARNA [20], and RAF [21].

Consan is based on Sankoff algorithm and was proposed by Dowell to address two problems not solved by Sankoff: the fact that it cannot be generalized to RNA structures of different lengths and the need to deal with the complexity of the algorithm. First, the method of modelling secondary structures and alignments using random context-independent grammars (SCFG) [22] is extended to construct pair-SCFG descriptions and build unconstrained algorithms. Second, the complete pair-SCFG algorithm is almost infeasible due to the $O(N^4)$ memory requirement and $O(N^6)$ time consumption. The structure algorithm is constrained by adding pins. Long sequence comparison is achieved using anchors fixed partial comparison. Also, the time and space complexity of the algorithm is effectively reduced. However, it is clear that Consan syntax lacks the handling of more complex statistical functions [10].

PMcomp used McCaskill's algorithm to obtain the base-pairing probability matrix and, next, the alignment probability matrix. This algorithm computes the input pairing matrix independently, unlike other combinatorial alignment and folding procedures (e.g., Dynalign and FoldAlign). Promising candidate sequences are preselected using a pattern search procedure, and an algebraic dynamic programming approach is used to scan the candidate sequences. It is folded by McCakill's algorithm, and finally, each individual is compared with each other separately using familiar techniques. The method reached a lightweight energy computation and can be described as simplified Sankoff's algorithm [12].

LocARNA is a novel RNA local comparison tool that has achieved high sensitivity in multiple experiments based on different datasets, but its complexity is still high [15]. For this feature of LocARNA, a sparsified prediction and alignment of RNAs based on their structure ensembles (SRARSE) is introduced. It supports the operation of loop structures for the first time and retains the full flexibility of Sankoff's algorithm. In the benchmark test, high accuracy and speed were both maintained [16].

Non-Sankoff algorithms divide the process into two parts: folding and alignment. Among them, some tree-based methods, such as RNAforester [23], RNAdistance [8], RNAStrAt [24], and RNApdist [25], are widely used.

The RNAforester tree comparison algorithm computes the local similarity in RNA secondary structures. The method provides a new representation of RNA secondary structure in forest representation and provides elaborate dynamic programming implementations using dense two-dimensional tables, which can greatly reduce space requirements in practice. 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 [26, 27]. Considering the structure with pseudoknots, Michela established ASPRAlign and obtained a method for comparing ASPRA distances [28]. A comparison of algebraic RNA trees and structural RNA trees [29] is made with ignoring the primary sequence and considering only the structure of the molecule. The execution time of the ASPRAlign workbench tool to process molecular pairs was compared experimentally. The experiments show that the time complexity of the algorithm is $O(n^2)$. 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 [30]. 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 [31], Yao used four horizontal lines and eight symbols to represent the RNA secondary structure [32]. 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. Randic proposed a method to visualize the secondary structure of RNA without loss of information [33]. The method is based on [30] labeling bases before and after hydrogen bonds, using 12 symbols to represent bases at different positions and, in turn, using numbers and graphs to represent RNA sequences with structural information. To illustrate the validity of the method, two RNAs that are similar were selected, and visualization of RNA sequences using the method enables visual observation of the sequence differences. In [34], Li proposed a novel graphical representation of RNA secondary structures (TV-Curve) and introduced a wavelet decomposition to compare RNA secondary structures. Based on the feature representation of RNA secondary structure with eight symbols proposed in [30], each character is represented by three vectors, and the vectors are connected sequentially to generate a unique graphical representation of RNA secondary structure. After feature extraction, RNA similarity is estimated using weighted correlations of wavelet domains at different scales. The method integrates the global and local structures. The effect of the program was verified using 100 RNA sequences from the RFAM database and 9 viruses to obtain a similarity metric closer to the actual data. Based on the approach of [34], Li further created a web server for multiscale similarity RNA secondary structure comparison based on the curve representation of triple vectors (RNA-TV-curve), including

RNA visualization, mutation analysis, and multiple RNA structure comparison [35].

Neutral components (NCs) in the genotype-phenotype (GP) maps were found in [36] to be able to influence the biological sequence function. Thus, based on a method that reveals the hierarchical community structure, which reveals the hierarchical community structure solely from the sequence constraints and composition of the genotypes that form a given NC, community detection has been used to analyze RNA secondary structure, and recent works on community detection have been studied in [37]. Some novel classifiers and learning models have also been proposed for multigraph analysis. Multigraph feature-based representation and learning methods for multigraph classification were used to achieve an effective learning model for multigraph classification [38]. Subsequently, to address the complexity of graph data generalization in multigraph classification, Wu proposed a boosting-based multigraph classification framework (bMGC). The framework uses a two-level weighting strategy to balance label ambiguity and introduces dynamic subgraph selection criteria to solve the structured data representation problem [39].

Besides, the classification and comparison of noncoding RNAs is also gaining attention. The deep-learning-based model ncRDeep is a recently proposed tool to efficiently classify ncRNA families [40]. It uses a simple convolutional neural network and RNA sequence information only to predict the class of ncRNAs. The experiments in the article evaluate the performance of ncRDeep using a benchmark dataset. Ultimately, the method was effective in improving the accuracy of prediction.

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 nonstationary signals should have different resolutions at different positions in the time-frequency plain; in other words, it should be a multiresolution analysis method. Wavelet transform is a multiresolution analysis method. Furthermore, wavelet is a fairly simple and effective mathematical tool that has been widely used. A combination of wavelet transform and neural network has been proposed to achieve highly accurate machine fault diagnosis [41]. In [42], 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. 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 [43].

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 [44]. 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 [45]. Reference [46] combines empirical modal decomposition and multifractal detrended fluctuation analysis to study the fractal characteristics of harmonic signals. 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 [47]. In [48], Yang applied the fractal dimension to protein sequence similarity analysis and obtained a high degree of similarity. Yang also integrated the fractal dimension with empirical mode decomposition to compare protein sequences and optimise the feature extraction process. Experiments show that the combination of the two methods extracts higher quality features [49]. The fractal dimension has also been recently used to determine the distribution of purines and pyrimidines in the miRNAs of humans, gorillas, chimpanzees, mice, and rats [50].

In this paper, we propose a novel combined algorithm. Based on the characteristics of discrete wavelet transform (DWT) multiresolution analysis and the universality of FD, DWT is used to analyze the graphical representation 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 theory of the fractal dimensions and the graphical characterization of RNA secondary structure in Section 2. Numerical experiments and the results are then given and discussed in Section 3. Section 4 is the conclusion.

2. Materials and 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 graphical representation 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 clustering analysis is performed, and the corresponding cluster analysis graph is obtained. The algorithmic process is shown in Figure 1.

2.1. Basic Concepts of Fractal Dimension

2.1.1. 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* [51]. 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 nonempty 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,

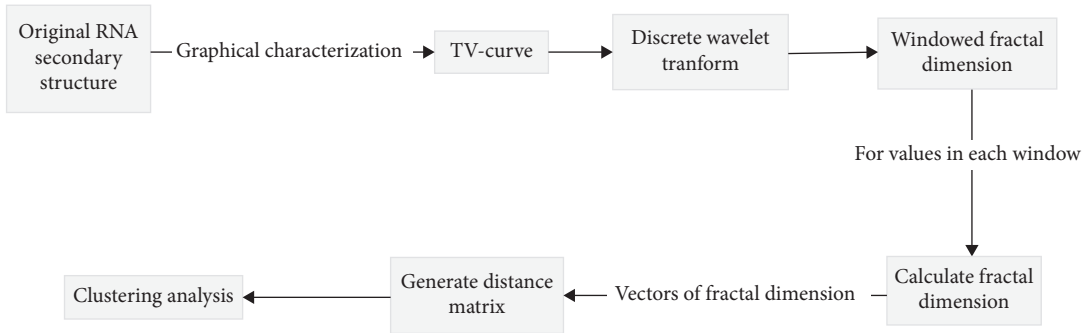

Figure 1: The flowchart of our method.

$$|A| = \sup\{|x - y|: \ x, y \in A\}, \tag{1}$$

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

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

$$H_\delta^s(C) = \inf\left\{\sum_{i=1}^{\infty} |U_i|^s: \ \{A_i\} \text{ is a } \delta - \text{cover of } C\right\}. \tag{2}$$

The abovementioned 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 (2) also decreases, and the infimum $H_\delta^s(C)$ increases accordingly. Moreover, when $\delta \longrightarrow 0$, it tends to a limit, denoted as

$$H^s(C) = \lim_{\delta \longrightarrow 0} H_\delta^s(C). \tag{3}$$

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 (2), 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 (3) that $H^s(C)$ does not increase either. Furthermore, if $t > s$ and $\{A_i\}$ is the δ-cover of C, then

$$\sum_i |A_i|^t \le \sum_i |A_i|^{t-s} |A_i|^s \le \delta^{t-s} \sum_i |A_i|^s. \tag{4}$$

Taking the infimum,

$$H_\delta^t(C) \le \delta^{t-s} H_\delta^s(C). \tag{5}$$

In the case of $t > s$, let $\delta \longrightarrow 0$; 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 *Haus do r f f − Basicovitch* dimension.

The definition is expressed as

$$\begin{aligned} \dim_H C &= \inf\{s \ge 0: \ H^s(C) = 0\}, \\ &= \sup\{s: \ H^s(C) = \infty\}. \end{aligned} \tag{6}$$

Therefore,

$$H^s(C) = \begin{cases} \infty, & \text{if } 0 \le s < \dim_H C, \\ 0, & \text{if } s > \dim_H C. \end{cases} \tag{7}$$

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

$$0 \le H^s(C) \le \infty. \tag{8}$$

2.2. 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 [52], Petrosian's algorithm [53], Higuchi's algorithm [54], multifractal detrended fluctuation analysis (MFDFA) [55], detrended fluctuation analysis (DFA) [56], and fluctuation analysis (FA) [57]. 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).

2.2.1. Higuchi's Algorithm. Treating an RNA sequence as a one-dimensional signal $s_1, s_2, \ldots, s_N$, we construct a subsequence as

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

where m indicates the starting position of the signal and k is the measurement scale. $\lfloor (N-m)/k \rfloor$ means the integer part of $\lfloor (N-m)/k \rfloor$, which is the number of terms in s_m^k.

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

For $m = 1, \ldots, k$,

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

where N is the length of the signal and $(N-1)/(((N-m)/k)\,k)$ is the normalization term. With the given measurement scale, the approximate length of the signal with starting position m is computed, $m = 1, \ldots, k$.

Thus, the approximated signal length can be obtained by

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

Finally, the fractal dimension h^* of the signal is calculated by the least-squares method [58]:

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

where c is the bias.

2.2.2. 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 $X = \{X_i, i = 1, 2, \ldots, N\}$ 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 $x = \{x_i, i = 1, 2, \ldots, N\}$, where $x_i = X_i - \mu$.

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

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

We test whether the following formula satisfies the power-law formula:

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

That is, we test whether $\log_2(F^{(2)}(m))$ 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 abovementioned equation, we can obtain the Hurst parameter H.

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

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

2.2.3. 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 nonstationary signals such as RNA TV-curves by traditional methods is difficult. For complex, highly nonstationary 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 and instability.

The DFA method has also been widely used in image processing and other fields [59]. The DFA method has been proven to be an effective method to analyze nonstationary signals. In summary, as the FA, the signal x_r, $r = 1, \ldots, N$ is preliminarily processed, and the mean value x is removed. The summation of i terms before the series x_r as the term i in the new sequence is

$$Y(i) = \sum_{r=1}^{i} [x_r - < x >], \quad i = 1, 2, \ldots, N. \tag{17}$$

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.

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

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

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

The fractal dimension is calculated by

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

2.3. Graphical Characterization of RNA Secondary Structure

2.3.1. 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 RNA, and 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 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.

2.3.2. The TV-Curve Representation of RNA Secondary Structure. 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 [34]. 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 Figure 2, 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, Figure 3 demonstrates the secondary structure of tRNA_Y00055.1 : 4327–4494 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.

2.4. Comparison of RNA Secondary Structures Based on the RNA TV-Curve by Detrended Fluctuation Analysis

2.4.1. 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 [60, 61]. 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 $s(t)$ overviews 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

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

where gn is the high-pass filter

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:

$$\begin{aligned}
a_k^{j+1} &= \sum_{i\in Z} a_i^j \overline{h}_{i-2k}, \quad j = 0,1,2,\ldots,\\
d_k^{j+1} &= \sum_{i\in Z} a_i^j \overline{g}_{i-2k}, j = 0,1,2,\ldots.
\end{aligned} \tag{24}$$

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.

2.4.2. 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 [62]. 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:

$$\widetilde{Y}_k(j) = Y(j) - y_k^m(j). \tag{25}$$

For each window, we calculate the following formula:

$$F^2_{\mathrm{DFA}_m}(k) = \frac{1}{k}\sum_{j=1}^{k} \widetilde{Y}_k^{\,2}(j). \tag{26}$$

Next, the average is calculated:

$$F_2(k) = \left[\frac{1}{N}\sum_{j=1}^{N} F^2_{\mathrm{DFA}_m}(k)\right]^{1/2}. \tag{27}$$

Transform the value of k, plot the *lb-lb* graph of $F_2(k) - k$, fit the straight line, calculate the slope, and obtain the scaling exponent α. The fractal dimension is calculated by

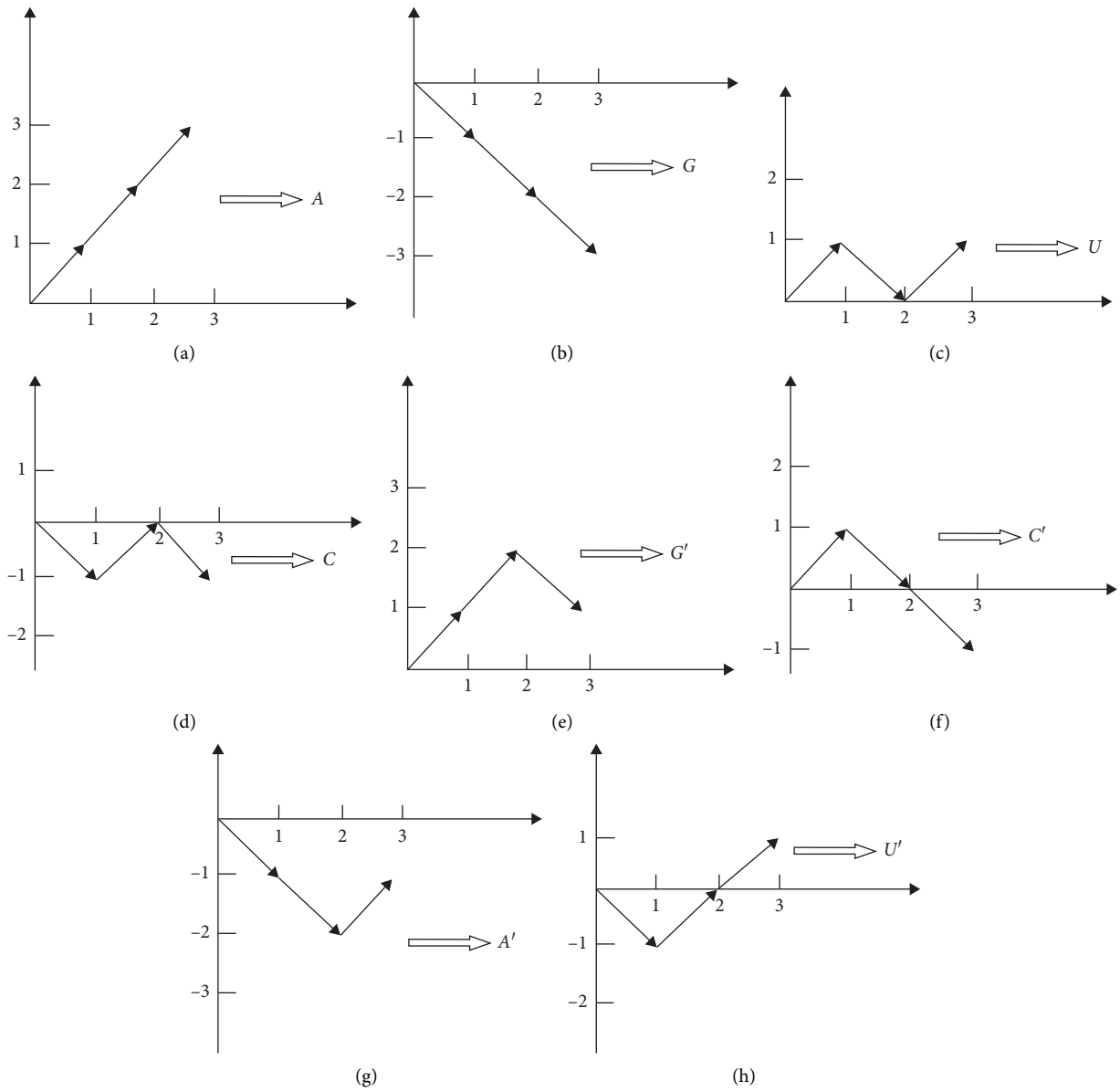

Figure 2: Graphical representation of the eight nucleotides.

$$FD = 2 - \alpha. \tag{28}$$

3. Results and Discussion

3.1. Results. In this section, the optimal experimental setting was obtained by the trial-and-error method: the level of wavelet decomposition was set to 3 and the window width was set to 17. Figure 4 shows the DWT and DFA on the TV-curve graphical representation of a RNA secondary structure using our method. Nine commonly used viruses were selected in this experiment. These viruses have conserved secondary structures, and the differences between them are subtle enough to verify the effectiveness of the experimental method in the classification of secondary structures with slight differences. Information on these nine viruses is presented in Figure 5. The nine 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), asparagus virus II (AVII), and tobacco streak virus (TSV) [63].

The distance matrix of the nine viruses obtained by the method of this paper is shown in Table 1. Next, as shown in Figure 6, the proposed method is compared with the results obtained by MEGA software [64], while the classification effect was compared with classical methods such as RNApdist and the innovative methods ASPRA distance and Li's method. The cluster analysis graph generated by the detrended fluctuation analysis method is based on the distance matrix calculated in MATLAB. In MEGA software, the maximum likelihood method is used to generate the cluster analysis graph.

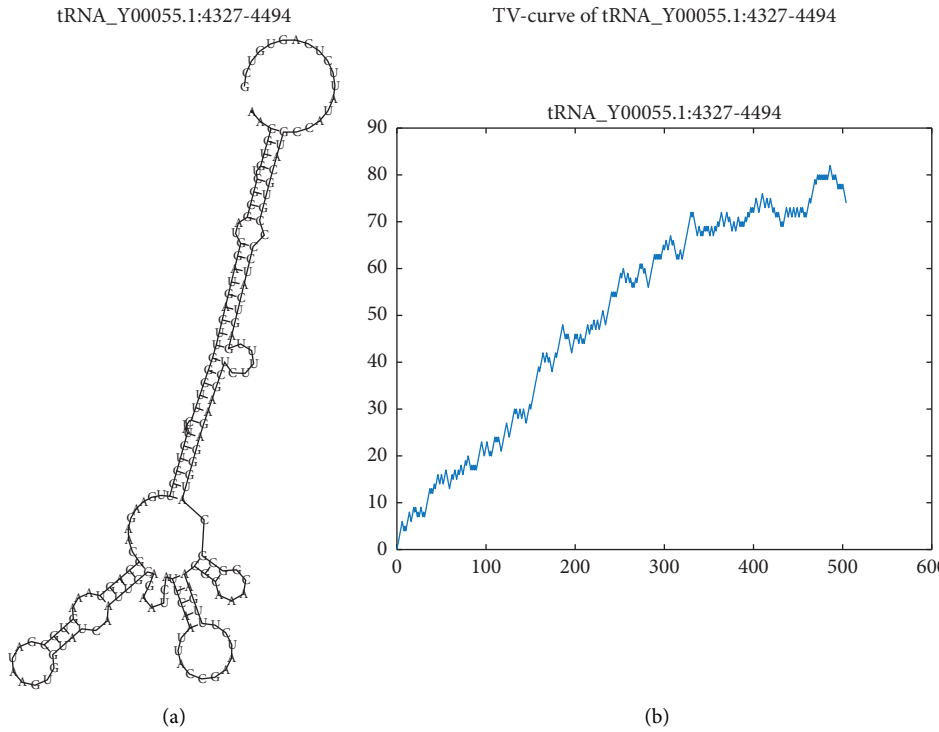

Figure 3: The TV-curve of tRNA_Y00055.1 : 4327–4494.

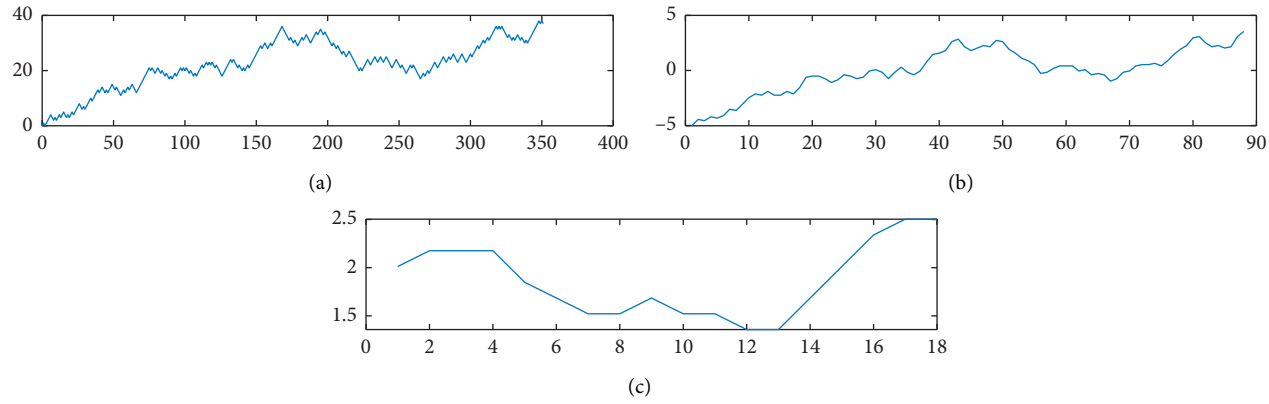

Figure 4: (a) Graphical representation of RNA secondary structure. (b) The third-level discrete wavelet transform. (c) The DFA fractal dimension of (b) obtained with a sliding window width of 17.

In addition, an experiment containing 11 noncoding RNAs is performed to test the applicability of our method in comparing the similarities between noncoding 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 2, and the distance matrix generated by the experiment is shown in Table 3. Based on this dataset, the results of the classification using ncRDeep are shown in Table 4.

Unfortunately, ncRDeep does not provide the corresponding distance matrix as well as the cluster analysis graph like the other methods, which would not provide a complete comparison in this paper. The remaining experimental results, including the proposed method, RNApdist, ASPRAlign, MEGA software, and Li's method, are shown in Figure 7.

To get rid of the specificity, we randomly extracted 100 ncRNA sequences from four families in the RFAM database, including tRNA, RNase P arch, miRNA, and 5s rRNA. We constructed three cluster analysis graphs using our method, Li's method, RNApdist, and ASPRAlign (see Figure 8: Figure S1, Figure S2, and Figure S3). As shown in Figure 8, the new dataset was applied to build a cluster analysis graph using the proposed method. Four branches can be clearly shown in Figure 8, where the 5 s rRNA family is thoroughly distinguished, while four ncRNAs in tRNA failed to successfully assign with most of the same family. As well, in the

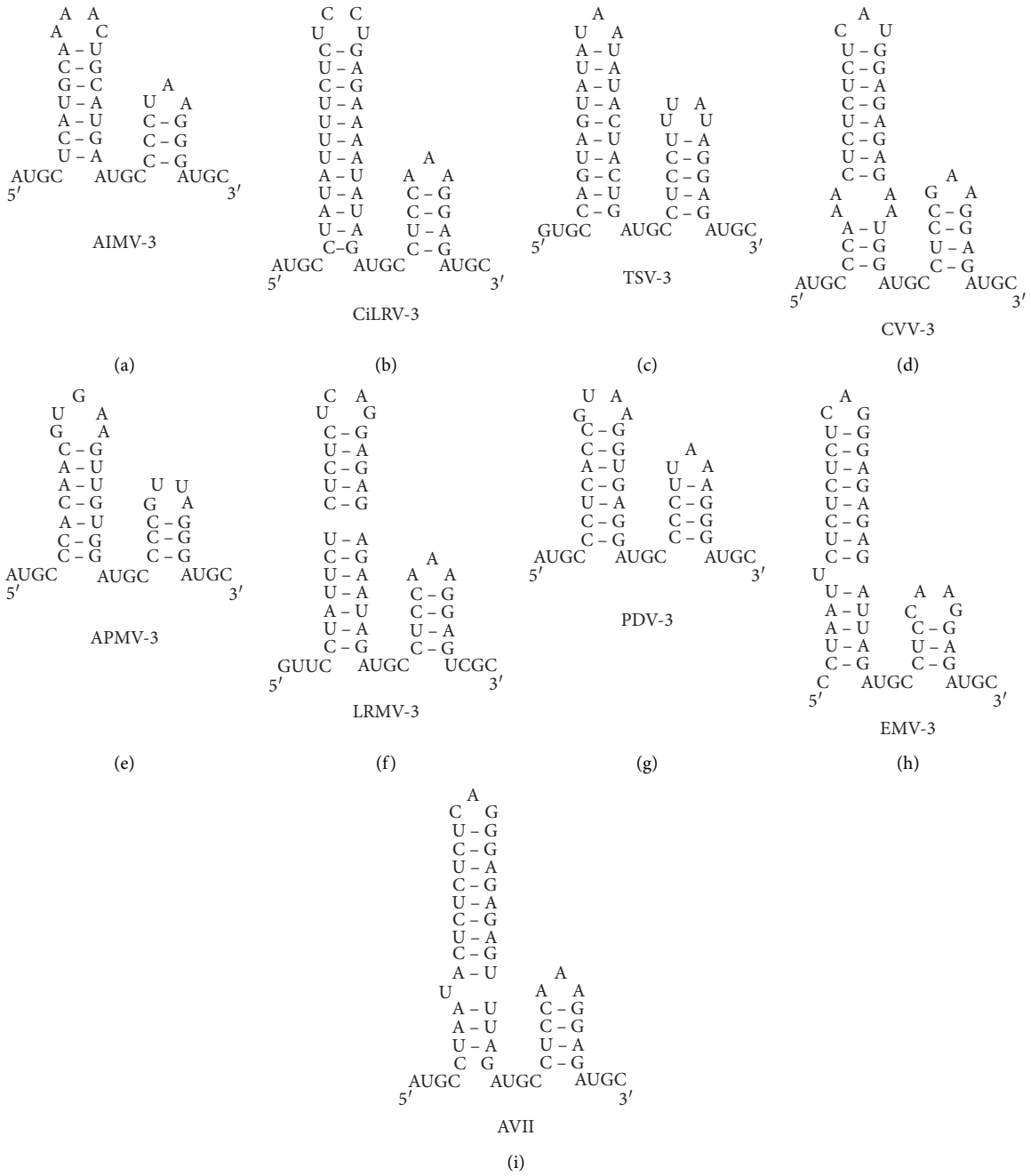

Figure 5: Secondary structure information of nine viruses.

miRNA and RNase P arch families, there are small numbers of ncRNAs that are distant from their major branches. However, among the method proposed by us, RNApdist, Li's method, and ASPRAlign, the classification effect of the proposed method is clearly superior to that of the other three methods.

3.2. Discussion. A detailed explanation of the cluster analysis graphs is given below; it is apparent from Figure 6 that the cluster analysis graph shown in Figure 6(a) is more similar to the standard cluster analysis graph (Figure 6(e)) generated by MEGA software. First, AVII and APMV and ALMV and PDV have a high similarity, which is also clearly reflected in Figure 6(a). However, it is not reflected in the other cluster analysis graphs, such as Figures 6(b)–6(d), and they deviate from the standard cluster analysis graph generated by MEGA software. Obviously, there are three branches in Figure 6(e). CiLRV is closer to LRMV, EMV, and TSV than the remaining four, and Figures 6(a) and 6(b) have similar representations which is not reflected in Figures 6(c) and 6(d). In addition, EMV is closer to LRMV and CVV in all viruses, which is visible at a glance in Figure 6(e), while Figure 6(a) has the same branch. They are, however, separated in Figures 6(b) and 6(c), as well as in Figure 6(d). In

Table 1: The distance matrix of nine viruses generated by our method.

	ALMV	APMV	CiLRV	TSV	CVV	PDV	LRMV	EMV	AVII
ALMV	0	0.718	0.641	0.769	0.667	0.538	0.692	0.846	0.692
APMV		0	0.718	0.718	0.667	0.846	0.718	0.821	0.641
CiLRV			0	0.692	0.462	0.718	0.487	0.718	0.590
TSV				0	0.744	0.795	0.744	0.769	0.769
CVV					0	0.769	0.282	0.615	0.718
PDV						0	0.718	0.744	0.718
LRMV							0	0.590	0.692
EMV								0	0.846
AVII									0

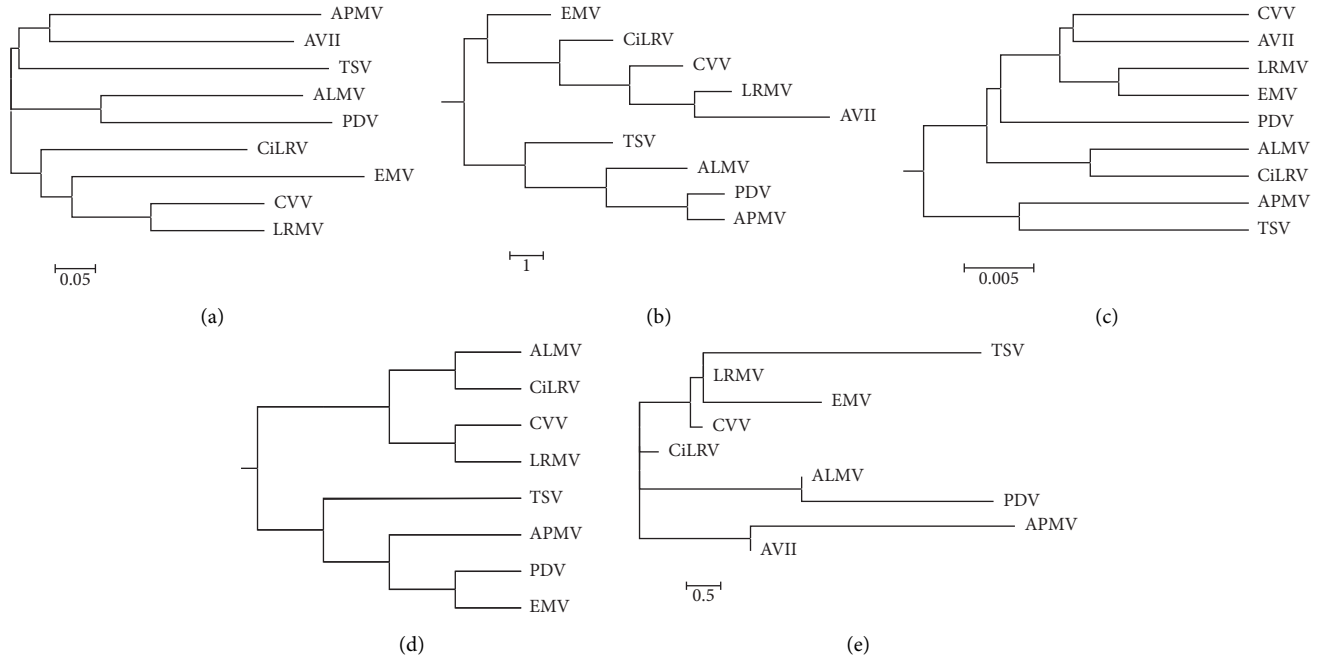

Figure 6: The cluster analysis graph of nine virus sequences generated by (a) our method, (b) Li's method, (c) RNApdist, (d) ASPRAlign, and (e) MEGA software.

Table 2: 11 noncoding 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

conclusion, compared with other methods, the proposed method can obtain the similarities of RNA secondary structures more accurately.

The experimental results of the same dataset, the cluster analysis graphs, and the distance between them are measured using the classical Robinson–Foulds (RF) metric [65]. In [65], the smallest number of steps to transform the topology of one of the trees into the other tree by the necessary operation (α or α^{-1}) is noted as the distance between the two trees.

α is the contraction of disjoint, possibly empty set, or zero distance nodes with other normal nodes into a single node while the label changes.

$$T_2 = \alpha(T_1, p_r p_s), \tag{29}$$

where trees T_1 and T_2 have the same set of species, $p_r p_s$ is the edge, and T_2 is obtained from T_1 by shrinking the edge $p_r p_s$.

In turn, α^{-1} is the inverse operation of α, which means dispersing a node into two nodes.

In this paper, the metric was used to calculate the distance between evolutionary trees. Thus, the classification efficiency of the method is verified.

The distance between evolutionary trees generated by the different methods and those generated by MEGA software is

Table 3: The distance matrix of eleven ncRNA sequences produced by our method.

	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

Table 4: The classification results of ncRDeep.

Sequence ID	Result
NONHSAT000002.2	Ribozymes
NONHSAT000003.2	Ribozymes
NONHSAT000004.2	Ribozymes
MF489813.1	Intron_gpl
MF489812.1	Intron_gpl
MF489811.1	Intron_gpl
MF489810.1	Intron_gpl
MF489809.1	Intron_gpl
MF489808.1	Intron_gpl
MF489806.1	Intron_gpl
NT_033777.3	Ribozymes

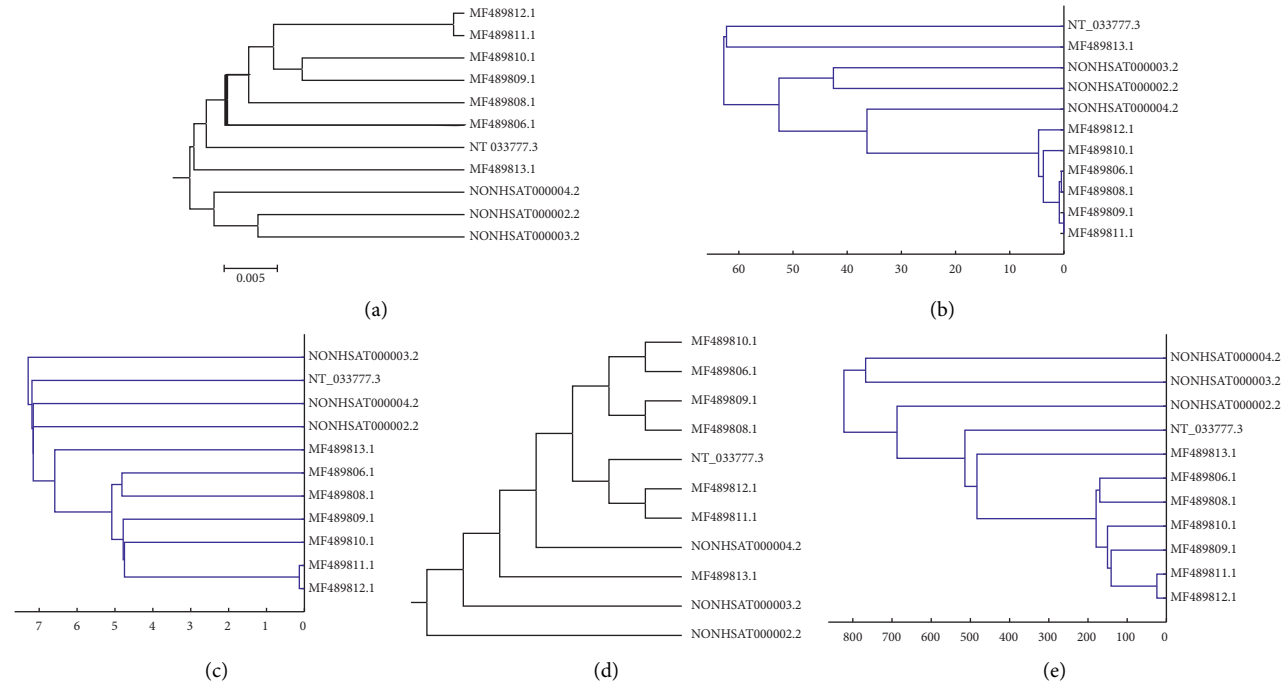

Figure 7: The cluster analysis graphs of 11 noncoding RNA sequences generated by (a) our method, (b) Li's method, (c) RNApdist, (d) ASPRAlign, and (e) MEGA software.

shown in Table 5. Clearly, the cluster analysis graph generated by our approach is closer to the cluster analysis graph generated by MEGA software than the other approaches.

In the second experiment, as indicated in Figure 7, the experimental sequences are accurately divided into three families in Figure 7(e), tRNA, RNase P RNA, and a group of human ncRNA in the NONCODE database. Figure 7(e) shows that the NT_033777.3 sequence has a high similarity to the tRNA family, which is also reflected in Figure 7(a) and Table 3. In Figure 7(b), NT_033777.3 and MF489813.1 are classified into a group, with NT_033777.3 being the most distant from the other sequences in the tRNA family, which is unreasonable. NT_033777.3 is mistakenly placed in a group close to the NONCODE human noncoding RNA in Figure 7(c). More unreliably, the grouping in Figure 7(d) does not completely separate the three groups of RNAs. On analysis of Table 4, it is apparent that ncRDeep did not succeed in accurately distinguishing ncRNAs. Based on the abovementioned analysis, the method based on DFA and DWT is an effective algorithm for RNA secondary structure similarity analysis. The distance between evolutionary trees generated by the method proposed by us, RNApdist, and Li's method, ASPRAlign, and the cluster analysis graph generated by MEGA software are shown in Table 6. Evidently, the results obtained by our method in this experiment are more accurate.

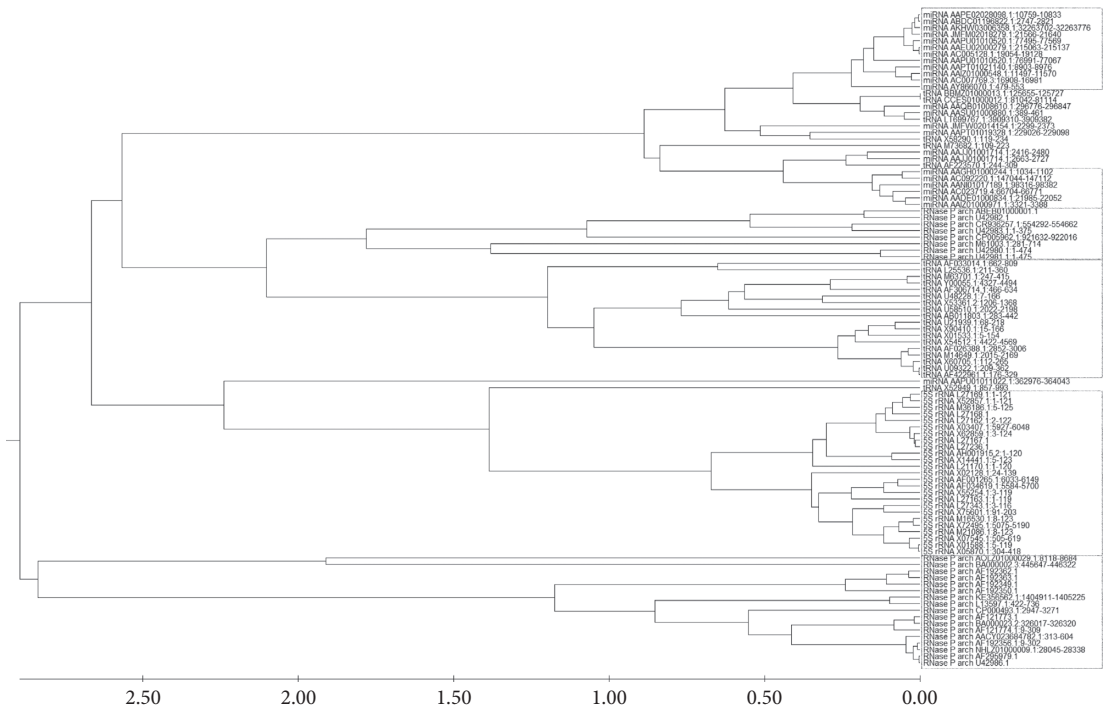

Figure 8: The cluster analysis graph of 100 noncoding RNA sequences generated by our method.

Table 5: The Pearson correlation efficiency between the cluster analysis graphs generated by the different methods and those generated by MEGA software.

Method	RNApdist	Our method	Li's method	ASPRAlign
Distance	0.1549	**0.1288**	0.2973	0.4709

Table 6: The Pearson correlation efficiency between the cluster analysis graphs generated by the different methods and those generated by MEGA software.

Method	RNApdist	Our method	Li's method	ASPRAlign
Distance	0.2494	**0.1748**	0.2048	0.2722

4. 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 graphical 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 are 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 hoped to be carried out on RNAs of different lengths and characteristics.

Data Availability

The RNA sequence data used to support the findings of this study are deposited in the RFAM database.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Authors' Contributions

Lina Yang proposed the ideas and supervised the process of the research and the validity of the results. Yang Liu performed the practical experiments, recorded the experimental results, and wrote the manuscript. Xiaochun Hu guided the knowledge and application of the detrended fluctuation analysis (DFA) method. Patrick Wang provided guidance on the discrete wavelet transform. Xichun Li improved the language of the manuscript, and Jun Wu focused on the standardization of the manuscript format. All authors read and approved the final manuscript.

Acknowledgments

This work was financially supported by the National Natural Science Foundation of China under Grant No. 61862005.

Supplementary Materials

Supplementary Figure S1: The cluster analysis graph of 100 noncoding RNA sequences generated by Li's method. Supplementary Figure S2: The cluster analysis graph of 100 noncoding RNA sequences generated by *RNApdist*.

Supplementary Figure S3: The cluster analysis graph of 100 noncoding RNA sequences generated by *ASPRAlign*. (*Supplementary Materials*)

References

[1] L. R. Borges Dos Anjos, L. Do Prado Assunção, B. Lima Freitas et al., "Popularização da Ciência: desmistificando o Dogma Central da Biologia Molecular," *Revista de Ensino de Bioquímica*, vol. 16, no. 2, pp. 71–86, 2019.

[2] Y. Ke, R. Jiahua, H. Zhao, Y. Lu, N. Xiao, and Y. Yang, "Accurate prediction of genome-wide rna secondary structure profile based on extreme gradient boosting." 2019.

[3] W. Lu, Y. Tang, H. Wu et al., "Predicting RNA secondary structure via adaptive deep recurrent neural networks with energy-based filter," *BMC Bioinformatics*, vol. 20, no. 25, p. 684, 2019.

[4] T. Zhou and A. Routh, "Mapping RNA-capsid interactions and RNA secondary structure within virus particles using next-generation sequencing," *Nucleic Acids Research*, vol. 48, 2019.

[5] F. Wang, T. Akutsu, and T. Mori, "Comparison of pseudoknotted RNA secondary structures by topological centroid identification and tree edit distance," *Journal of Computational Biology*, vol. 55, 2020.

[6] J. H. Havgaard, E. Torarinsson, and J. Gorodkin, "Fast pairwise structural RNA alignments by pruning of the dynamical programming matrix," *PLOS Computational Biology*, vol. 3, no. 10, pp. 1896–1908, 2007.

[7] B. A. Shapiro and K. Zhang, "Comparing multiple RNA secondary structures using tree comparisons," *Bioinformatics*, vol. 6, no. 4, pp. 309–318, 1990.

[8] J. Allali and M. F. Sagot, *A New Distance for High Level RNA Secondary Structure Comparison*, IEEE Computer Society Press, Washington, DC,USA, 2005.

[9] D. Sankoff, "Simultaneous solution of the RNA folding, alignment and protosequence problems," *Siam Journal on Applied Mathematics-SIAMAM*, vol. 45, 1985.

[10] R. D. Dowell and S. R. J. B. B. Eddy, "Efficient pairwise RNA structure prediction and alignment using sequence alignment constraints," *BMC Bioinformatics*, vol. 7, no. 1, pp. 1–18, 2006.

[11] D. H. Mathews, "Predicting a set of minimal free energy RNA secondary structures common to two sequences," *Bioinformatics*, vol. 21, no. 10, pp. 2246–2253, 2005.

[12] I. L. Hofacker, S. H. F. Bernhart, and P. F. Stadler, "Alignment of RNA base pairing probability matrices," *Bioinformatics*, vol. 20, no. 14, pp. 2222–2227, 2004.

[13] I. J. B. B. Holmes, "Accelerated probabilistic inference of RNA structure evolution," *BMC Bioinformatics*, vol. 6, 2005.

[14] H. J. Hull, G. D. Stormo, and G. J. B. Jan, "Pairwise local structural alignment of RNA sequences with sequence similarity less than," *Bioinformatics*, vol. 40, no. 9, p. 9, 2005.

[15] S. Will, K. Reiche, I. L. Hofacker, P. F. Stadler, and R. Backofen, "Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering," *PLOS Computational Biology*, vol. 3, no. 4, p. e65, 2007.

[16] S. Will, C. Schmiedl, M. Miladi, M. Möhl, and R. Backofen, "SPARSE: quadratic time simultaneous alignment and folding of RNAs without sequence-based heuristics," in *Proceedings of the 17th International Conference on Research in Computational Molecular Biology*, Santa Monica, USA, April 2013.

[17] S. Siebert and R. Backofen, "MARNA: multiple alignment and consensus structure prediction of RNAs based on sequence structure comparisons," *Bioinformatics*, vol. 21, no. 16, pp. 3352–3359, 2005.

[18] E. Torarinsson, J. H. Havgaard, and J. Gorodkin, "Multiple structural alignment and clustering of RNA sequences," *Bioinformatics*, vol. 23, no. 8, pp. 926–932, 2007.

[19] H. Kiryu, Y. Tabei, T. Kin, and K. Asai, "Murlet: a practical multiple alignment tool for structural RNA sequences," *Bioinformatics*, vol. 23, no. 13, pp. 1588–1598, 2007.

[20] D. A. Sorescu, M. Möhl, M. Mann, R. Backofen, and S. Will, "CARNA--alignment of RNA structure ensembles," *Nucleic Acids Research*, vol. 40, no. W1, pp. W49–W53, 2012.

[21] C. B. Do, F. Chuan-Sheng, and B. J. B. Serafim, "A max-margin model for efficient simultaneous alignment and folding of RNA sequences," *Bioinformatics*, vol. 13, p. 13, 2008.

[22] R. D. Dowell, *RNA Structural Alignment Using Stochastic Context-Free Grammars*, Washington University, St. Louis, MO, USA, 2004.

[23] M. Hochsmann, T. Toller, R. Giegerich, and S. Kurtz, "Local similarity in RNA secondary structures," *Computational Systems Bioinformatics Csb IEEE Bioinformatics Conference Csb*, vol. 62, 2003.

[24] V. Guignon, C. Chauve, and S. Hamel, *An Edit Distance Between RNA Stem-Loops*, Springer, Berlin, Germany, 2005.

[25] I. L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster, "Fast folding and comparison of RNA secondary structures," *Monatshefte F and R Chemie Chemical Monthly*, vol. 125, no. 2, pp. 167–188, 1994.

[26] F. Jörg, V. Pavankumar, B. Andrea et al., "The RNA workbench 2.0: next generation rna data analysis," *Nucleic Acids Research*, vol. 47, 2019.

[27] M. Miladi, M. Raden, S. Will, and R. Backofen, "Fast and accurate structure probability estimation for simultaneous alignment and folding of RNAs," in *Proceedings of the 19th International Workshop on Algorithms in Bioinformatics*, Niagara Falls, NY, USA, September 2019.

[28] Q. Michela, T. Luca, and M. E. J. Bioinformatics, "ASPRAlign: a tool for the alignment of rna secondary structures with arbitrary pseudoknots," 2020.

[29] M. Quadrini, L. Tesei, and E. Merelli, "An algebraic language for RNA pseudoknots comparison," *Bmc Bioinformatics*, vol. 20, no. S4, 2019.

[30] L. Bo, C. Weiyang, S. Xingming et al., "A binary coding method of rna secondary structure and its application," *Journal of Computational Chemistry*, vol. 15, 2009.

[31] M. R. J. C. P. Letters, "On characterization of DNA primary sequences by a condensed matrix," *Journal of Chemistry*, vol. 317, no. 1, pp. 29–34, 2000.

[32] Y. H. Yao, B. Liao, T. M. J. Wang, and S. T. JoM, "A 2D graphical representation of RNA secondary structures and the analysis of similarity/dissimilarity based on it," *Journal of Molecular Structure*, vol. 755, no. 1, pp. 131–136, 2005.

[33] M. Randi and D. P. J. C. P. Letters, "Novel spectral representation of RNA secondary structure without loss of information," *Genomic Analysis and Structural Prediction of DNA*, vol. 476, no. 4–6, pp. 277–280, 2009.

[34] Y. Li, M. Duan, and Y. Liang, "Multi-scale RNA comparison based on RNA triple vector curve representation," *BMC Bioinformatics*, vol. 13, no. 1, p. 280, 2012.

[35] Y. Li, X. Shi, Y. Liang, J. Xie, Y. Zhang, and Q. Ma, "RNA-TVcurve: a Web server for RNA secondary structure comparison based on a multi-scale similarity of its triple vector curve representation," *BMC Bioinformatics*, vol. 18, no. 1, p. 51, 2017.

[36] M. Weiss and S. Ahnert, "Neutral components show a hierarchical community structure in the genotype–phenotype map of RNA secondary structure," *Journal of The Royal Society Interface*, vol. 17, 2020.

[37] F. Liu, S. Xue, J. Wu et al., "Deep learning for community detection: progress, challenges and opportunities," 2020.

[38] J. Wu, X. Zhu, C. Zhang, and P. S. Yu, "Bag constrained structure pattern mining for multi-graph classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 10, pp. 2382–2396, 2014.

[39] J. Wu, S. Pan, X. Zhu, and Z. J. I. Cai, "Boosting for multi-graph classification," *IEEE Trans Cybern*, vol. 45, no. 3, pp. 416–429, 2014.

[40] T. Chantsalnyam, D. Y. Lim, H. Tayara, and K. T. Chong, "ncRDeep: non-coding RNA classification with convolutional neural network," *Computational Biology and Chemistry*, vol. 88, Article ID 107364, 2020.

[41] S. Siyu, M. A. Stephen, Y. Ruqiang et al., "Highly accurate machine fault diagnosis using deep transfer learning," *IEEE Transactions on Industrial Informatics*, vol. 63, no. 5, p. 1, 2019.

[42] S. Wang and X. Wang, "Prediction of protein structural classes by different feature expressions based on 2-D wavelet denoising and fusion," *BMC Bioinformatics*, vol. 20, no. Suppl 25, p. 701, 2019.

[43] J. Pando, L. Sands, S. E. J. P. R. E. S. N. Shaheen, and S. M. Physics, "Detection of protein secondary structures via the discrete wavelet transform," *Statistical, Nonlinear, and Soft Matter Physics*, vol. 80, no. 5, Article ID 051909, 2009.

[44] R. Esteller, G. Vachtsevanos, J. Echauz, and B. Lilt, "A comparison of fractal dimension algorithms using synthetic and experimental data," *IEEE International Symposium on Circuits & Systems*, vol. 5, 1999.

[45] B. B. Mandelbrot and J. W. Van Ness, "Fractional brownian motions, fractional noises and applications," *SIAM Review*, vol. 10, no. 4, pp. 422–437, 1968.

[46] J. Li, X. Ma, M. Zhao, and X. J. E. Cheng, "A novel MFDFA algorithm and its application to analysis of harmonic," *Multifractal Features*, vol. 8, no. 2, 2019.

[47] Z. Yu, K. Lau, and L. Zhou, "Clustering of protein structures using hydrophobic free energy and solvent accessibility of proteins," *Physical Review E, Statistical, Nonlinear, and Soft Matter Physics*, vol. 73, Article ID 031920, 2006.

[48] L. Yang, Y. Y. Tang, Y. Lu, and H. Luo, "A fractal dimension and wavelet transform based method for protein sequence similarity analysis," *IEEE/ACM Trans Comput Biol Bioinform*, vol. 12, no. 2, pp. 348–359, 2015.

[49] L. Yang, P. Wei, C. Zhong et al., "A fractal dimension and empirical mode decomposition-based method for protein sequence analysis," 2019.

[50] D. J. Kumar, C. P. Pal, C. Adwitiya, H. S. Sarif, and B. J. S. R. Pallab, "Analysis of purines and pyrimidines distribution over miRNAs of human, Gorilla, chimpanzee, mouse and rat," *Scientific Report*, vol. 8, no. 1, p. 9974, 2018.

[51] F. Hausdorff, "Dimension und *äβ*eres Ma*β*," *Mathematische Annalen*, vol. 79, no. 1, pp. 157–179, 1918.

[52] J. Michael, "Biology KJCi, Medicine," *Fractals and the Analysis of Waveforms*, vol. 18, no. 3, 1988.

[53] A. Petrosian, "Kolmogorov complexity of finite sequences and recognition of different preictal EEG patterns," in *Proceedings of the IEEE Symposium on Computer-Based Medical Systems*, Lubbock, TX, USA, June 1995.

[54] T. Higuchi, "Approach to an irregular time series on the basis of the fractal theory," *Physica D: Nonlinear Phenomena*, vol. 31, no. 2, pp. 277–283, 1988.

[55] O. . Paweł, K. Jarosław, and E. D. J. P. R. Stanisław, "Wavelet versus detrended fluctuation analysis of multifractal structures," *Statistical, Nonlinear, and Soft Matter*, vol. 16, 2006.

[56] C.-K. Peng, S. Havlin, H. Stanley et al., "Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series," 1995.

[57] H. Jing, T. Wen-Wen, and G. Jianbo, "Detection of low observable targets within sea clutter by structure function based multifractal analysis," *IEEE Transactions on Antennas and Propagation*, vol. 54, no. 1, pp. 136–143, 2006.

[58] A. Savitzky and M. J. E. Golay, "Smoothing and differentiation of data by simplified least squares procedures," *Analytical Chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964.

[59] J. Alvarez-Ramirez, E. Rodriguez, I. Cervantes, J. Carlos Echeverria, and I. Applications, "Scaling properties of image textures: a detrending fluctuation analysis approach," *Physica A: Statistical Mechanics and Its Applications*, vol. 361, no. 2, pp. 677–698, 2006.

[60] C. Esser and S. Jaffard, "Divergence of wavelet series: a multifractal analysis," *Advances in Mathematics*, vol. 328, pp. 928–958, 2018.

[61] Z. Zhang, *Multivariate Wavelets. Multivariate Time Series Analysis in Climate and Environmental Research*, Springer International Publishing, Berlin, Germany, 2018.

[62] Z. Sun, Z. Zhang, Y. Chen, S. Liu, and Y. Song, "Frost filtering algorithm of SAR images with adaptive windowing and adaptive tuning factor," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 6, pp. 1097–1101, 2020.

[63] C. B. E. M. Reusken and J. F. Bol, "Structural elements of the 3′-terminal coat protein binding site in alfalfa mosaic virus RNAs," *Nucleic Acids Research*, vol. 24, no. 14, pp. 2660–2665, 1996.

[64] K. Sudhir, S. Glen, M. Li, K. Christina, and T. J. M. B. Koichiro, "Evolution. MEGA X," *Molecular Evolutionary Genetics Analysis Across Computing Platforms*, vol. 6, 2018.

[65] D. F. Robinson and L. R. Foulds, "Comparison of phylogenetic trees," *Mathematical Biosciences*, vol. 53, no. 1-2, pp. 131–147, 1981.