MTSCCleav: a Multivariate Time Series Classification (MTSC)-based method for predicting human Dicer cleavage sites

First Author[1,2*], Second Author[2,3†] and Third Author[1,2†]

[1*]Department, Organization, Street, City, 100190, State, Country.
[2]Department, Organization, Street, City, 10587, State, Country.
[3]Department, Organization, Street, City, 610101, State, Country.

*Corresponding author(s). E-mail(s): iauthor@gmail.com;
Contributing authors: iiauthor@gmail.com; iiiauthor@gmail.com;
†These authors contributed equally to this work.

Abstract

Background: I am Coleman.
Results: I am Coleman.
Conclusion: I am Coleman. Code is available at: https://github.com/cyuab/time-series-classification-cleavage.

Keywords: miRNA, Dicer Cleavage Site, Genomic signal processing (GSP), (Multivariate) time Series Classification (MTSC, TSC)

1 Background

One of the most important theories in molecular biology is the central dogma. It depicts the flow of genetic information [1, 2]. Proteins are the functional units. And the information stored in DNA is used to make them up. This is not a one-step process. It involves transcription and translation. Genes (segments of DNA encoding proteins) in DNA are used as templates for messenger RNAs (mRNAs) synthesis. This synthesis process is called transcription. An mRNA acts as a set of instructions to assemble a chain of amino acids, which form a linear polypeptide. This construction process is called translation. This chain is not yet functional. To become biologically active, this chain is folded into a specific three-dimensional structure, a proper configuration. This

process is called protein folding. And this folded polypeptide is called a functional protein, or a protein in short.

This whole process is remarkably similar to the process of running a computer program on a computer. The source code does not function by itself. The source code turns into assembly code (less human-readable code) first, and finally into an executable that does the work the programmer intends. In this analogy, DNA acts as the whole source code, in which genes refer to the functions in the code (segments of the code). mRNA is the assembly code. And the final product proteins are the running executable.

These mRNAs are called "coding RNAs" because they code for proteins. There are other genes in which the final product is the RNA molecule itself. They are called non-coding RNAs (ncRNAs). They do functions by themselves, such as regulating gene expression. Gene expression refers to the process by which the genetic information is transcribed into RNA, which may be translated into protein or ncRNAs. Two types of small ncRNAs are particularly important. They are microRNAs (miRNAs) and small interfering RNAs (siRNAs). Their discovery was recognized with the 2006 Nobel Prize in Physiology or Medicine[1], awarded for work completed only eight years prior [1]. They play essential roles in gene expression regulation. In this manuscript, we focus on miRNAs. They are small RNAs with a length of about 22 nt. They regulate gene expression post-transcriptionally [3]. A miRNA can regulate the expression of several proteins. They are essential in diverse biological processes, such as development, differentiation, and disease [4, 5]. Hence, it is of great value to understand the formation or the biogenesis of miRNAs. It involves the processing of primary miRNAs (pri-miRNAs). A predicted secondary structure of a pri-miRNA's sequence is shown in Figure 1.

Recall that a pri-miRNA contains a hairpin loop, also called a stem loop. It is located on the bottom left part of Figure 1. A microprocessor complex comprising Drosa and DCGR8 cleaves the pri-miRNA to form a precursor miRNA (pre-miRNA) inside the nucleus. The stem-loop is still preserved, but the two arms become shorter. The two arms refer to the two ends of the sequence. The sequence or strand has the end-to-end chemical orientation. One end is called 5' end, and the other is called 3' end. After that, the pri-miRNA is transported by Exportin 5 from the nucleus to the cytoplasm (i.e., outside the nucleus). It is further cleaved by an enzyme called Dicer [6]. The Dicer cleaves the stem-loop from the two arms at the two cleavage sites, shown as the two scissors in figure 1. We call the bond between two nucleotides along the strand that is cleaved by Dicer the Dicer cleavage site. The stem-loop is removed. It results in a short double-stranded miRNA molecule. Furthermore, these molecules may be subjected to additional trimming. Some nucleotides are removed from the 5' (5p) and 3' (3'p) ends.

One strand of the double-stranded miRNA molecule is loaded into an RNA-induced silencing complex (RISC). This loaded strand guides the RISC to the target mRNA

[1] The Nobel Prize in Physiology or Medicine 2006 - NobelPrize.org: https://www.nobelprize.org/prizes/medicine/2006/summary/ (Accessed on: 2025-06-13).

[2] Its miRBase entry: https://mirbase.org/hairpin/MI0000060. (Accessed on: 2025-06-12).

[3] RNAfold web server: http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi. (Accessed on: 2025-06-12). The figure is viewed in forna. This view option can be chosen on the previous website.

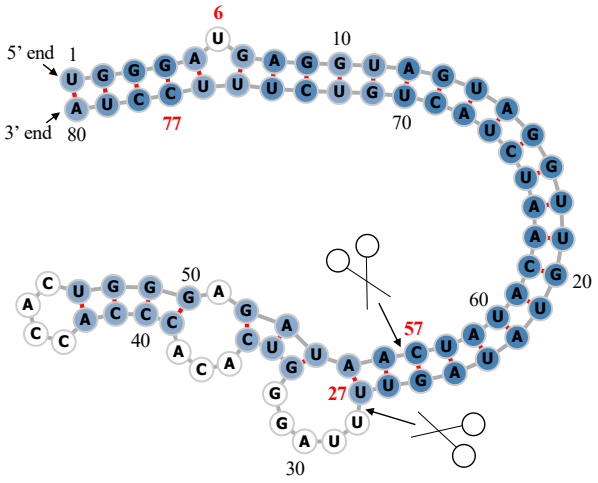

Fig. 1 Secondary structure of the pri-miRNA "hsa-let-7a-1"[2]. We denote the sequence as S. Experimental evidence suggests that the two deviated mature miRNAs are $UGA\cdots GUU$ and $CUA\cdots UUC$. They are $S[6:27]$ and $S[57:77]$. Since $S[6:27]$ ($S[57:77]$) is near the 5' (3') end, we call it "5p (3p) mature miRNA". The starting and ending indices of these two subsequences are indicated in **bold and red** in the figure. It suggests that the two cleavage sites are the two bonds immediately after the 27^{th} nucleotide and before the 57^{th} nucleotide. The two scissors indicate the two cleavage sites. The color intensity of the nodes reflects their base pair probability in this predicted secondary structure configuration. The unpaired nodes are uncolored. The raw figure is generated by RNAfold web server[3].

to silence it, and it results in gene silencing. This loaded strand is called the guide strand or mature miRNA. The other stand, which is called the passenger strand, is usually degraded. Note that both miRNA single strands, resulting from the unwinding of the double-stranded miRNA molecule, can become the guide strand. For example, "hsa-let-7a-1" has two guide strands or mature miRNA products.

Dicer plays an important role in the biogenesis of miRNAs. It would be of great benefit to understand how Dicer selects cleavage sites from the neighborhood information near the cleavage sites. The neighborhood information refers to both the sequence and the secondary structure information. Studies [7–9] revealed that the secondary structures of the sequence are essential for cleavage site determination. Hence, to predict or classify whether a subsequence, extracted from pri-miRNAs, contains a cleavage site, we need to make use of both the sequence and secondary structure information.

1.1 Related work

Cleavage site prediction is not only defined on human dicer. Other examples include the calpain [10] and caspasesl [11], which are proteases that cleave proteins. Most of the studies about cleavage site prediction focus on protein cleavage sites. A wide range of computational methods has been applied on this topic, including support vector machine [11], deep neural networks [12]. In this manuscript, we focus on human dicer cleavage sites. We review the works on human dicer cleavage sites here. PHDcleav employed support vector machines (SVM) leveraging sequence and structure-based

features [13]. LBSizeCleav improved upon it by considering the loop and buldge lengths [14]. [15] proposed a ensemble learning approach, used gradient boosting machine for better accuracy. [16] developed a deep learning model namely DiCleave. This model used an autoencoder to learn the secondary structure embeddings of pre-miRNA.

In summary, our contributions are shown as follow.

1. To the best of our knowledge, we are the first to frame the prediction of the cleavage sites, not only Dicer cleavage sites, as a multivariate time series classification problem. By using time series modeling, it bridges bioinformatics and time series which allows us to use state-of-the-art time series classification algorithms. Besides, the visualization power of time series allows us understand the data more intuitively.
2. We propose of making use the base-pair probabilities in the predicted secondary structure in the prediction. To our surprise, this information has been ignored in the existing works. We leverage this base-pair probabilities to our novel transformation scheme for RNA sequence and its complementary sequence.
3. We conduct extensive experiments on different transformation methods and convolution based classification methods.

2 Methods

This manuscript aims to transform the problem of predicting human dicer cleavage sites into a multivariate time series classification problem. This allows algorithms from the time series community to be run, and it gives us more ways to analyze and visualize such data. Time series data is good for visualization. In this session, we first discuss how to prepare the dataset. Since we want to transform the data into time series representations, we briefly review time series and then propose transformation algorithms. After that, time series classifiers are discussed. Finally, we discuss how to evaluate the performance of the time series classifiers.

Our main goal is to build a classifier that checks whether an input string contains a cleavage site of Dicer.

2.1 Data preparation

We use miRBase database [17][4]. The database consists of miRNA data from different organisms, such as humans, mice, and C. elegans [18]. Each data entry refers to a miRNA sequence, along with other properties such as name, accession (the identifier used in miRBase), organism (from which it comes), and information on its derivative miRNA products. In this manuscript, we are interested in pri-miRNA in humans. The derivative miRNA products are the mature miRNAs. The database also annotates the location of the mature miRNA in the original pri-miRNA and states whether or not its existence has experimental evidence.

The database contains 38589 miRNA records. Table 1 shows four representative records from miRBase database. We use Table 1 to elucidate our selection criteria. The

[4]The current website is www.mirbase.org, and the newest version of the database is Release 22.1 (accessed on Feb 2025).

Accession	Name	Organism	Sequence	Mature miRNA 1	Mature miRNA 2
MI0000001	cel-let-7	Caenorhabditis elegans	UACAC...UUCGA	cel-let-7-5p 17:38 experimental	cel-let-7-3p 60:81 experimental
MI0000060	hsa-let-7a-1	Homo sapiens	UGGGA...UCCUA	hsa-let-7a-5p 6:27 experimental	hsa-let-7a-3p 57:77 experimental
MI0000114	hsa-mir-107	Homo sapiens	CUCUC...ACAGA	hsa-miR-107 50:72 experimental	NA
MI0000238	hsa-mir-196a-1	Homo sapiens	GUGAA...UUCAC	hsa-miR-196a-5p 7:28 experimental	hsa-miR-196a-1-3p 45:65 not experimental

Table 1 Selected representative records from miRBase. For the last two columns, there are three lines in each cell. The first line shows the name of this Mature miRNA product. The second line shows its location in the original sequence. We use the $x : y$ notation to denote that this product is located from the x position to the y position in the original sequence, inclusively.

records are rows in the table. We selected the records from humans (Homo sapiens). It resulted in 1917 records. In order to have the actual locations of the two cleavage sites in the pri-miRNA sequence that have the experimental evidence, we selected the records with the two mature miRNAs resulting from cleavage at the 5p arm and the 3p arm that have experimental support. According to the above selection criteria, only "MI0000060" would be selected for our further analysis in the four records in Table 1.

Sequence	Secondary Structure (In Dot-bracket notation)
1 UGGGA**UGAGGUAGUAGGUUGUAUAGUU** 27 28 UUAGGGUCACACCCACCACUGGGAGAU 54 55 AA**CUAUACAAUCUACUGUCUUUC**CUA 80	1 (((((.(((((((((((((((((((27 28 UUAGGGUCACACCCACCACUGGGAGAU 54 55))))))))))))))))))))))))) 80
Base-pair probabilities sequence (the first 10 bases)	
1 (0.549, 0.946, 0.987, 0.987, 0.904) 5 6 (**-1.000**, 0.841, 0.974, 0.981, 0.890) 10	

Table 2 The whole sequence of "hsa-let-7a-1" with the locations of its two mature miRNA and its predicted secondary structure. We have numbered each line with the staring and ending positions. The corresponding positions of the two mature miRNAs and the probability of the unpaired 'U' are **bolded and in red**.

We use "hsa-let-7a-1" as our running example. The whole sequence of it and its necessary information for our downstream analysis is listed in Table 2.

After the selection process, we selected 827 experimental validated pre-miRNA sequences together with its two mature miRNA products, and this formed our dataset.

2.1.1 Argument the dataset with Secondary Structure information

We want to use the domain knowledge about pre-miRNA sequences to improve our classifier's accuracy. We leverage the secondary structure of these sequences to achieve it. Recall that a specific three-dimensional (3D) structure is required for DNA, RNA, and protein to perform functions [19]. However, finding these 3D structures using experimental methods such as X-ray crystallography or nuclear magnetic resonance

(NMR) is costly and time-consuming. Hence, the prediction methods on such 3D structures are necessary and helpful for the downstream analysis. However, prediction of such 3D structures is difficult. One of the reasons is that there are some "non-conventional" base-pair interactions (e.g., A-G) that allow an RNA structure to fold into a 3D structure. It makes the search space for prediction much larger than, in the 2D case, the secondary structure. The local structure of the 3D structures, the secondary structures, only focus on the conventional (G-C and A-U) base-pair interactions [2]. It makes the prediction of the secondary structure easier and more applicable than predicting the 3D structure. Secondary structure can still shed light on some of these functions in the structure-function relationships. We employ RNAfold from the ViennaRNA Package[5]. RNAfold returns the secondary structure in the dot-bracket notation and a matrix for the base-pair probabilities. Equipped with the matrix, we can construct the base-pair probability sequence of the original sequence. The first ten entries of the probability sequence of our running example are shown in 2.

Definition 1 (Dot-bracket notation) Dot-bracket notation is a way of representing the secondary structure of the given string s. One of the following symbols is assigned to each base in s.

- Open parentheses (indicates that the base is paired with a complementary base further along in s.
- Close parentheses) indicates that the base is paired with a complementary base earlier in s.
- Dot . indicates that the base is unpaired.

The secondary structure of "hsa-let-7a-1" in dot-bracket notation has been shown in Table 2. The visualization of it has been shown in Figure 1[6]. The nodes' colors show the base-pair probabilities. The deeper the color, the higher the probability. The unpaired nodes are in white.

2.1.2 Extract cleavage patterns

The locations of the two mature miRNAs on the main sequence indicate the probable locations of the two cleavage sites. The 5p cleavage site (i.e., the cleavage site near the 5p end) must be beyond and near the ending location of the 5p mature miRNA. For example, the ending position of the 5p mature miRNA for "hsa-let-7a-1" (i.e., "hsa-let-7a-5p") is 27, as shown in Figure 1. So, the 5p cleavage site would be one of the bonds beyond the 27^{th} nucleotide, as indicated by a scissor in the Figure. We deemed the immediate bond next to the ending position of the 5p mature miRNA the 5p cleavage site with the knowledge that the actual cleavage site may not be this immediate bond but the nearby bonds after it. The same applies to the 3p cleavage site. It is at the immediate bond before the starting position of the 3p mature miRNA, which is 57.

[5]The latest stable release is Version 2.7.0, accessed on Feb 2025) to predict the secondary structure for a given pri-miRNA [20].

[6]The main body of the figure is created by RNAfold web server (http://rna.tbi.univie.ac.at//cgi-bin/RNAWebSuite/RNAfold.cgi, accessed on Feb 2025), included in [20]

We extract a 14-string (a.k.a, string with length = 14) with the cleavage site located at the center. The first 7 nt (nucleotide) before the center are **bolded and in red**. In our running example, it would be "**UAUAGUU**UUAGGU" for the 5p cleavage site and "**GAGAUAA**CUAUACA" for the 3p cleavage site. We call these 14-strings cleavage patterns as they contains the cleavage sites. We can also generate non-cleavage patterns by selecting a 14-string with the center 6 nt away from the corresponding cleavage sites towards the corresponding mature miRNA. For example, the bond that is 6nt away from the 5p cleavage site towards 5p mature miRNA is the bond between 21^{st} and 22^{nd} nucleuotides. It is based on the assumption that the dicer is less likely to cut the middle of the mature miRNA than the opposite side. So, in our example, the 5p non-cleavage pattern would be "**AGGUUGU**AUAGUUU". The center of the 3p non-cleavage pattern is the bond between 62^{nd} and 63^{rd} nucleuotides. The 3p non-cleavage pattern would be "**ACUAUAC**AAUCUAC".

In conclusion, for a given pri-miRNA, we can generate two cleavage patterns (positive samples) and two non-cleavage patterns (negative samples). We also call these four patterns simply the "four strings" of a given pri-miRNA. The four strings of "hsa-let-7a-1" are listed in Table 3. We could construct the complementary strand of each of the

	5p-cleav	non-5p-cleav	3p-cleav	3p-non-cleav
Input strand	UAUAGUUUUAGGGU	AGGUUGUAUAGUUU	GAGAUAACUAUACA	ACUAUACAAUCUAC
Complementary strand	AUAUCAA_____UA	C_CUGUUGAUAUGU	UCUAACAUAUCAA_	UGAUAUGUUGGAUG

Table 3 The first row shows the four strings of "hsa-let-7a-1". If they are regarded as the input strands, the complementary strands are shown in the second row.

string/ strand in the "four strings" by finding the corresponding paired base for each of the bases in the input strand by considering the secondary structure information. Bases refer to the nucleotides. We will use these two terms interchangeably. We use '_' to denote the unpaired base in the complementary strand. For example, in Figure 1, the sub-string "UUAGG" in the 5p-cleavage pattern "UAUAGUUUUAGGGU" are unpaired while other bases do pair with some bases in the complementary strand, the resulting complementary strand is "AUAUCAA_____UA". Note that the five underscores indicate that "UUAGG" is unpaired. There is a loop/ budge there.

The strand and its complementary strand together can then encode the loop/ budge information. We call the four original input strands and the constructed four complementary strands together as the "eight strings" of the input pre-miRNA. We are now ready to transform the "eight strings" into time series.

2.2 Time Series

Definition 2 (Time Series) A time series $T = t_1, t_2, ..., t_n$ is a sequence of real-valued numbers with length = n.

A short contiguous region of T is called a subsequence.

Definition 3 (Subsequence) A subsequence $T(i:j)$ of a time series T is a shorter time series that starts from position i and ends at position j. Formally, $T(i:j) = t_i, t_{i+1}, ..., t_j$, where $1 \leq i \leq j \leq n$.

The above two notations are also used to represent string and its subsequence.

2.2.1 Transform strings into time series

Both strings and time series are temporal sequences. The order in a sequence usually represents time ordering. They are the values on the x-axis if we plot them on an x-y plane.

The only difference between strings and time series is the behavioral attributes [21]. They are the values on the y-axis. For strings, also known as words, a y-value is a symbol from a predefined set called the alphabet. Thus, we also refer to the symbols as letters. For example, the alphabet is $\{A, C, G, T\}$ in the DNA string, while $\{A, C, G, U\}$ in the RNA string. For time series, a y-value is a scalar number. The number can be an integer or a real number. The important point is that they are naturally ordered. The "greater than" and "smaller than" are well-defined. There is no ordering in the alphabet unless some external domain knowledge is introduced to explain why a letter is smaller (greater) than another letter.

In the bioinformatics community, the study of applying signal processing techniques to genomic data, which includes DNA and RNA strings, is called "Genomic Signal Processing" (GSP) [22]. In the field of GSP, the time series representations of DNA strings are called DNA numeric representations (DNR). Many DNRs have been proposed in the field of GSP, with applications including identifying protein-coding regions in DNA sequences [23], biological sequence querying [24], and finding similarities between DNA sequences [25]. We noted that DNA strings and RNA strings are the same from the computational point of view. They are simply strings with different alphabets. Recall that the alphabet of DNA is $\{A, C, G, T\}$ and that of RNA is $\{A, C, G, U\}$. Many transformation methods designed for DNA are applicable to RNA by simply substituting T for U. In this manuscript, our default alphabet is $\{A, C, G, U\}$. One of the simple, if not the simplest, transformations is to map the letters in the alphabet into integer numbers without considering any domain knowledge about the nucleotides. Method 1 in Table 4 shows this approach. We call this method "Toy'. This method belongs to a group of methods called the "single-value" approach [22, 26–29]. One single value is assigned to each of the letters. A more reasonable approach in this category is to employ the domain knowledge during assignments. For example, [30] employs the atomic number of each nucleotide as the transformed values where $\{78, 70, 58, 66\}$ is assigned to $\{G, A, C, T\}$ respectively. [31] uses "electron-ion" interaction potential representation (EIIP) as such values. These values are $\{0:0806, 0:1260, 0:1340, 0:1335\}$. Our goal is to transform the input strand and its complementary strand into two-time series and aim to capture the information of these sequences and the secondary structure. So, we need to employ

the complementarity property during the transformation [32]. Recall that in the base-pairing rules, 'A' pairs with 'U'[7] to form two hydrogen bonds while 'G' pairs with 'C' to form three hydrogen bonds. Hence, 'A' ('C') can be regarded as the "inverse" of 'U' ('G'). Recall that we call '-1' as the inverse of '1' and vice versa under addition in Algebra. We can preserve these base-pairing rules in the time series representation by assigning A (G) and U (C) opposite values. The time series of the complementary strand would then be a flipped version along the y-axis of the original strand.

We can group the four nucleotides by their chemical structures. 'A' and 'G' have a two-ring structure. They are purines. 'U' and 'C' have a one-ring structure. They are pyrimidines. We put 'A' and 'G' ('U' and 'C') in the same group. These two groups would be on the two sides of the number line with zero in the middle.

Now the only remaining question before assigning $\{-2, -1, +1, +2\}$ to $\{A, C, G, U\}$ is which nucleotide in the same group we should assign a larger absolute value. In other words, for 'A' and 'G', which one should be assigned a larger absolute value? In this manuscript, we adopted the "A = 2 and G = 1" assignment. The reasoning is as follows. The main goal of this manuscript is to find the two cleavage sites on the pre-miRNA. The cleavage sites are the bonds along the strands, the phosphodiester linkages, or bonds. After this pre-miRNA is cleaved, the resulting double-stranded miRNA molecule is unwound to form the guide strand and the passenger strand. The stability of the double strand would affect the unwinding process. There are three hydrogen bonds in C-G pairs and two in A-U pairs. C-G pairs are more stable than A-U pairs. This means that in regions with more C-G pairs, the double strands hold more tightly, and the unwinding process is less likely to occur than in regions with more A-U pairs. So, we want to emphasize the existence of such less stable A-U pairs in the double strands in our time series representations. We assign A and U with larger absolute values than those of G and C. With this reasoning, we propose the second transformation method, Method 2, in Table 4. It is called "Single-value". It is our baseline transformation method in the single-value category.

Instead of looking at the values in the time series one by one, we can accumulate the values assigned to each nucleotide. The "cumulative version" allows us to focus on analyzing the "trend" such as increasing and decreasing, by accumulating what has happened in the past. The "original version" (Single-value) allows us to focus on the absolute values of the alphabet and ignore what has happened in the past.

Until now, we have used only one time series to encode the dynamic of the four letters in the original string; we can also represent the dynamic using two time series. Each time series only represents the dynamic of the occurrences of one nucleotide and its inverse. One time series encodes the dynamic of "A and U", and the other encodes that of "G and C". There are two variations of the multivariate methods. One is the "Multivariate with different length" and the other is "Multivariate with same length". This method is shown in row 4 in Table 4 In table 4, transformations 1, 2, 3, 5 are lossless while transformation 4 is lossy. Lossy transformation refers to a transformation that does not allow us to restore the original string by the new time series representation.

[7] In DNA, 'A' pairs with 'T'

	Name	Numeric representation	Example for $s = G, A, G, A, U, A, A, C, U, A$
1	Toy	for $i = 1$ to $\|s\|$: $t_i = \begin{cases} 0 & \text{if } s_i = A \\ 1 & \text{if } s_i = C \\ 2 & \text{if } s_i = G \\ 3 & \text{if } s_i = U \end{cases}$	$t = 2, 0, 2, 0, 3, 0, 0, 1, 3, 0$
2	Single-value	for $i = 1$ to $\|s\|$: $t_i = \begin{cases} 2 & \text{if } s_i = A \\ 1 & \text{if } s_i = G \\ -1 & \text{if } s_i = C \\ -2 & \text{if } s_i = U \end{cases}$	$t = 1, 2, 1, 2, -2, 2, 2, -1, -2, 2$
3	Cumulative	$t_1 = 0$ for $i = 1$ to $\|s\|$: $t_{i+1} = \begin{cases} t_i + 2 & \text{if } s_i = A \\ t_i + 1 & \text{if } s_i = G \\ t_i - 1 & \text{if } s_i = C \\ t_i - 2 & \text{if } s_i = U \end{cases}$	$t = 0, 1, 3, 4, 6, 4, 6, 8, 7, 5, 7$
4	Multivariate with different length	$t_{1,0} = 0, t_{2,0} = 0$ $j = 0, k = 0$ for $i = 1$ to $\|s\|$: $t_{1,j+1} = \begin{cases} t_j + 1; & j = j + 1 \quad \text{if } s_i = A \\ t_j - 1; & j = j + 1 \quad \text{if } s_i = U \end{cases}$ $t_{1,k+1} = \begin{cases} t_k + 1; & k = k + 1 \quad \text{if } s_i = G \\ t_k - 1; & k = k + 1 \quad \text{if } s_i = C \end{cases}$	$t_1 = 0, -1, 0, -1, 0, 1, 2, 3$ $t_2 = 0, -1, 0, -1$
5	Multivariate with same length	$t_{1,0} = 0, t_{2,0} = 0$ for $i = 1$ to $\|s\|$: $t_{1,i+1} = \begin{cases} t_i + 1 & \text{if } s_i = A \\ t_i & \text{if } s_i = G \\ t_i & \text{if } s_i = C \\ t_i - 1 & \text{if } s_i = U \end{cases}$ $t_{2,i+1} = \begin{cases} t_i & \text{if } s_i = A \\ t_i + 1 & \text{if } s_i = G \\ t_i - 1 & \text{if } s_i = C \\ t_i & \text{if } s_i = U \end{cases}$	$t_1 = 0, -1, 0, -1, 0, 0, 1, 2, 2, 2, 3$ $t_2 = 0, 0, 0, 0, 0, -1, -1, -1, 0, -1, -1$

Table 4 Time series transformation for RNA string s

2.2.2 Incorporating base-pair probabilities

We propose a novel time series transformation for RNA sequences incorporating the predicted secondary structures and the base-pair probabilities. RNA secondary structure prediction algorithm is a kind of RNA folding algorithm. The RNA folding algorithms attempt to solve a harder problem: predicting the 3D structure of a given RNA sequence. Meanwhile, the RNA secondary structure prediction algorithm returns the predicted secondary structure of a given RNA sequence, a planar graph such as the minimum free energy structure. It also returns a probability matrix for the base-pair probabilities, which denote the probabilities of pairing one nucleotide with the other nucleotide. In our case, the bases in an RNA sequence will be either paired up with another base in the same sequence or remain unpaired. For example, the 1^{st} base 'U' is paired up with the 80^{th} base 'A'. The 6^{th} base 'U' is unpaired. The width and height

	Name	Numeric representation	Example for $s = C, \_, C, U, G, U, U, G, A, U$ with $s^p = 0.843, -1, 0.807, 0.807, 0.793, 0.914, 0.982, 1.000, 0.993, 0.999$
1	Single-value	for $i = 1$ to $\|s\|$: $t_i = \begin{cases} 2 \cdot s_i^p & \text{if } s_i = A \\ 1 \cdot s_i^p & \text{if } s_i = G \\ -1 \cdot s_i^p & \text{if } s_i = C \\ -2 \cdot s_i^p & \text{if } s_i = U \\ 0 & \text{if } s_i = \_ \end{cases}$	Without base-pair probability: $t = -1, 0, -1, -2, 1, -2, -2, 1, 2, -2$ With base-pair probability: $t = -0.843, 0.000, -0.807, -1.614, 0.793, -1.829, -1.963, 1.000, 1.999, -1.998$
2	Cumulative	$t_1 = 0$ for $i = 1$ to $\|s\|$: $t_{i+1} = \begin{cases} t_i + 2 \cdot s_i^p & \text{if } s_i = A \\ t_i + 1 \cdot s_i^p & \text{if } s_i = G \\ t_i - 1 \cdot s_i^p & \text{if } s_i = C \\ t_i - 2 \cdot s_i^p & \text{if } s_i = U \\ t_i & \text{if } s_i = \_ \end{cases}$	Without base-pair probability: $t = 0, -1, -1, -2, -4, -3, -5, -7, -6, -4, -6$ With base-pair probability: $t = 0.000, -0.843, -0.843, -1.650, -3.265, -2.471, -4.300, -6.263, -5.264, -3.265, -5.263$
3	Multivariate with same length	$t_{1,0} = 0, t_{2,0} = 0$ for $i = 1$ to $\|s\|$: $t_{1,i+1} = \begin{cases} t_i + 1 \cdot s_i^p & \text{if } s_i = A \\ t_i \cdot s_i^p & \text{if } s_i = G \\ t_i \cdot s_i^p & \text{if } s_i = C \\ t_i - 1 \cdot s_i^p & \text{if } s_i = U \\ t_i \cdot s_i^p & \text{if } s_i = \_ \end{cases}$ $t_{2,i+1} = \begin{cases} t_i \cdot s_i^p & \text{if } s_i = A \\ t_i + 1 \cdot s_i^p & \text{if } s_i = G \\ t_i - 1 \cdot s_i^p & \text{if } s_i = C \\ t_i \cdot s_i^p & \text{if } s_i = U \\ t_i \cdot s_i^p & \text{if } s_i = \_ \end{cases}$	Without base-pair probability: $t_1 = 0, 0, 0, 0, 0, -1, -1, -1, -2, -1, -1$ $t_2 = 0, 1, 1, 2, 1, 1, 0, -1, -1, -1, -2$ With base-pair probability: $t_1 = 0.000, 0.000, 0.000, 0.000, 0.000, -0.793, -0.793, -0.793, -1.793, -0.794, -0.794$ $t_2 = 0.000, 0.843, 0.843, 1.650, 0.843, 0.843, -0.071, -1.053, -1.053, -1.053, -2.052$
4	Multivariate with different length	$t_{1,0} = 0, t_{2,0} = 0$ $j = 0, k = 0$ for $i = 1$ to $\|s\|$: $t_{1,j+1} = \begin{cases} t_j + 1 \cdot s_i^p; & j = j + 1 \quad \text{if } s_i = A \\ t_j - 1 \cdot s_i^p; & j = j + 1 \quad \text{if } s_i = U \end{cases}$ $t_{1,k+1} = \begin{cases} t_k + 1 \cdot s_i^p; & k = k + 1 \quad \text{if } s_i = G \\ t_k - 1 \cdot s_i^p; & k = k + 1 \quad \text{if } s_i = C \end{cases}$	Without base-pair probability: $t_1 = 0, -1, -2, -1$ $t_2 = 0, -1, -2, -1$ With base-pair probability: $t_1 = 0.000, -0.793, -1.793, -0.794$ $t_2 = 0.000, 0.843, 1.650, 0.843, -0.071, -1.053, -2.052$

Table 5 Time series transformation for RNA complementary string s with its probability time series s^p

of the probability matrix are the length of the RNA sequence because it is a pair-wise matrix. From this matrix, we can construct the base-pair probability time series of the whole sequence. The unpaired base will have very small probabilities in any base-pair. In our formulation, we denote these small probabilities as -1 to emphasize these bases are deemed to be unpaired in the resulting predicted secondary structure. Table 2 shows the base-pair probabilities for the first ten bases of the sequence of ""hsa-let-7a". Since the 6^{th} base 'U' is unpaired, its probability is assigned as "-1".

Table 5 incorporates the base-pair probability time series into the transformation methods listed in Table 4.

2.2.3 Accumulating from the beginning of the pre-miRNA sequence

"Cumulative" would return different result if we choose different starts for the accumulation.

Considering the time series representation of $s(6:10) = UUGAU$ of the running example s in Table 5, the "Cumulative" transformation without base-pair probability of it starting at the beginning of s (i.e., s_1) would be $t(7:11) = -3, -5, -7, -6, -4, -6$. If we start the "Cumulative" at the beginning of $s(6:10)$ (i.e., $s(6)$), the resulting time series would be $0, -2, -4, -3, -1, -3$. Note that these two time series have the same trend, but they start at different values. The first one starts at -3 while the latter one starts at 0. Figure 2 shows how to use the discussed notion in transformations.

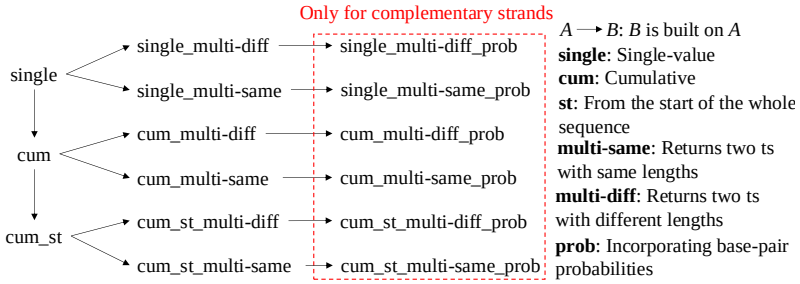

Fig. 2 Relationship of the proposed transformation methods. ts stands for time series.

2.2.4 Time series classification

2.3 Evaluation

References

[1] Urry, L.A., Cain, M.L., Wasserman, S.A., Minorsky, P.V., Orr, R.B., Campbell, N.A.: Campbell Biology, Twelfth edition edn., New York, NY (2020)

[2] Alberts, B.: Molecular Biology of the Cell, Seventh edition edn., New York (2022)

[3] Bartel, D.P.: Micrornas: Genomics, biogenesis, mechanism, and function. Cell **116**(2), 281–297 (2004)

[4] Iorio, M.V., Ferracin, M., Liu, C.-G., Veronese, A., Spizzo, R., Sabbioni, S., Magri, E., Pedriali, M., Fabbri, M., Campiglio, M., Ménard, S., Palazzo, J.P., Rosenberg, A., Musiani, P., Volinia, S., Nenci, I., Calin, G.A., Querzoli, P., Negrini, M., Croce, C.M.: Microrna gene expression deregulation in human breast cancer. Cancer Research **65**(16), 7065–7070 (2005)

[5] He, H., Jazdzewski, K., Li, W., Liyanarachchi, S., Nagy, R., Volinia, S., Calin, G.A., Liu, C.-g., Franssila, K., Suster, S., Kloos, R.T., Croce, C.M., de la Chapelle, A.: The role of microrna genes in papillary thyroid carcinoma. Proceedings of the National Academy of Sciences **102**(52), 19075–19080 (2005)

[6] Lee, Y., Jeon, K., Lee, J.-T., Kim, S., Kim, V.N.: Microrna maturation: Stepwise processing and subcellular localization. The EMBO Journal **21**(17), 4663–4670 (2002)

[7] Gu, S., Jin, L., Zhang, Y., Huang, Y., Zhang, F., Valdmanis, P.N., Kay, M.A.: The loop position of shrnas and pre-mirnas is critical for the accuracy of dicer processing in vivo. Cell **151**(4), 900–911 (2012)

[8] Feng, Y., Zhang, X., Graves, P., Zeng, Y.: A comprehensive analysis of precursor microrna cleavage by human dicer. RNA **18**(11), 2083–2092 (2012)

[9] MacRae, I.J., Zhou, K., Doudna, J.A.: Structural determinants of rna recognition and cleavage by dicer. Nature Structural & Molecular Biology **14**(10), 934–940 (2007)

[10] duVerle, D.A., Ono, Y., Sorimachi, H., Mamitsuka, H.: Calpain cleavage prediction using multiple kernel learning. PLOS ONE **6**(5), 19035 (2011)

[11] Wee, L.J., Tan, T.W., Ranganathan, S.: Svm-based prediction of caspase substrate cleavage sites. BMC Bioinformatics **7**(5), 14 (2006)

[12] Liu, Z.-X., Yu, K., Dong, J., Zhao, L., Liu, Z., Zhang, Q., Li, S., Du, Y., Cheng, H.: Precise prediction of calpain cleavage sites and their aberrance caused by mutations in cancer. Frontiers in Genetics **10** (2019)

[13] Ahmed, F., Kaundal, R., Raghava, G.P.: Phdcleav: A svm based method for predicting human dicer cleavage sites using sequence and secondary structure of mirna precursors. BMC Bioinformatics **14**(14), 9 (2013)

[14] Bao, Y., Hayashida, M., Akutsu, T.: Lbsizecleav: Improved support vector machine (svm)-based prediction of dicer cleavage sites using loop/bulge length. BMC Bioinformatics **17**(1), 487 (2016)

[15] Liu, P., Song, J., Lin, C.-Y., Akutsu, T.: Recgbm: A gradient boosting-based method for predicting human dicer cleavage sites. BMC Bioinformatics **22**(1), 63 (2021)

[16] Mu, L., Song, J., Akutsu, T., Mori, T.: Dicleave: A deep learning model for predicting human dicer cleavage sites. BMC Bioinformatics **25**(1), 13 (2024)

[17] Griffiths-Jones, S., Saini, H.K., van Dongen, S.: mirbase: Tools for microrna genomics. Nucleic Acids Research **36**(suppl_1), 154–158 (2008)

[18] Xu, T., Su, N., Liu, L., Zhang, J., Wang, H., Zhang, W., Gui, J., Yu, K., Li, J., Le, T.D.: mirbaseconverter: An r/bioconductor package for converting and retrieving mirna name, accession, sequence and family information in different versions of mirbase. BMC Bioinformatics **19**(19), 514 (2018)

[19] Zvelebil, M.J., Baum, J.O., Zvelebil, M.: Understanding Bioinformatics, New York (2008)

[20] Lorenz, R., Bernhart, S.H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F., Hofacker, I.L.: Viennarna package 2.0. Algorithms for Molecular Biology **6**(1), 26 (2011)

[21] Aggarwal, C.C.: Data Mining: The Textbook, Cham (2015)

[22] Mendizabal-Ruiz, G., Román-Godínez, I., Torres-Ramos, S., Salido-Ruiz, R.A., Morales, J.A.: On dna numerical representations for genomic similarity computation. PLOS ONE **12**(3), 0173288 (2017)

[23] Sahu, S.S., Panda, G.: Identification of protein-coding regions in dna sequences using a time-frequency filtering approach. Genomics, Proteomics & Bioinformatics **9**(1), 45–55 (2011)

[24] Ravichandran, L., Papandreou-Suppappola, A., Spanias, A., Lacroix, Z., Legendre, C.: Time-frequency based biological sequence querying. In: 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4174–4177 (2010)

[25] Yin, C., Yin, X.E., Wang, J.: A novel method for comparative analysis of dna sequences by ramanujan-fourier transform. Journal of Computational Biology **21**(12), 867–879 (2014)

[26] Rakthanmanon, T., Campana, B., Mueen, A., Batista, G., Westover, B., Zhu, Q., Zakaria, J., Keogh, E.: Searching and mining trillions of time series subsequences under dynamic time warping. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '12, pp. 262–270, New York, NY, USA (2012)

[27] Cristea, P.D.: Conversion of nucleotides sequences into genomic signals. Journal of Cellular and Molecular Medicine **6**(2), 279–303 (2002)

[28] Chakravarthy, N., Spanias, A., Iasemidis, L.D., Tsakalis, K.: Autoregressive modeling and feature analysis of dna sequences. EURASIP Journal on Advances in Signal Processing **2004**(1), 1–16 (2004)

[29] Zhao, J., Yang, X.W., Li, J.P., Tang, Y.Y.: Dna sequences classification based on wavelet packet analysis. In: Wavelet Analysis and Its Applications, pp. 424–429 (2001)

[30] Holden, T., Subramaniam, R., Sullivan, R., Cheung, E., Schneider, C., Jr, G.T., Flamholz, A., Lieberman, D.H., Cheung, T.D.: Atcg nucleotide fluctuation of deinococcus radiodurans radiation genes. In: Instruments, Methods, and Missions for Astrobiology X, vol. 6694, pp. 402–411 (2007)

[31] Nair, A.S., Sreenadhan, S.P.: A coding measure scheme employing electron-ion interaction pseudopotential (eiip). Bioinformation **1**(6), 197–202 (2006)

[32] Akhtar, M., Epps, J., Ambikairajah, E.: On dna numerical representations for period-3 based exon prediction. In: 2007 IEEE International Workshop on Genomic Signal Processing and Statistics, pp. 1–4 (2007)