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: MicroRNAs (miRNAs) are small non-coding RNAs (ncRNAs) that regulate gene expression at the post-transcriptional level and hence play essential roles in diverse biological processes such as development and differentiation. The biogenesis of miRNAs requires Dicer, which is an enzyme, to cleave at specific sites on the precursor miRNAs (pre-miRNAs). These sites are called cleavage sites. Several machine learning approaches, such as ReCGBM and DiCleave, have been proposed to predict human dicer cleavage sites. Given an input sequence, these classifiers predict whether it contains a cleavage site. Despite the advances made, existing studies have several limitations. They have ignored the probabilities of the base pairs in the secondary structure predicted by RNA secondary structure prediction tools in the classification. Besides, they rely heavily on complicated feature engineering or opaque deep neural models. It results in a lack of generalizability and interoperability. Also, they have a long running time. There is a need for alternative modeling paradigms that are simple, fast, and provide comparable accuracy while offering better model transparency.
Results: We propose a novel approach to predict human Dicer cleavage sites by reframing the task as a multivariate time series classification problem. To reframe, we have proposed different transformation schemes to convert the RNA sequence and the information about its secondary structure into time series. Hence, the data can be represented in the form of a multivariate time series. We also proposed a novel transformation scheme that involves the probabilities of the base pairs in the predicted secondary structure, which have been long ignored in

the literature. Computational experiments show that our proposed scheme can achieve comparable results using a simpler, intuitive model and less computation time. Besides, we test our models with perturbation-based experiments. We found that the regions or subsequences close to the cleavage sites and hence to the center of pre-miRNAs are essential to the predictions of the human dicer cleavage sites. Of note, the transformation of RNA data to time series allows the use of many state-of-the-art algorithms in the time series community and can relate some novel problem definitions in time series, such as motifs, discords, and chains, to the computational study of RNA data that paves novel ways using the well-established tools in the time series community.
Conclusion: Our proposed scheme allows us to study this problem in a new way. By transforming the RNA sequence and its secondary structure information into time series and using simple state-of-the-art time series classifiers, we obtain comparable or even superior performance in a simpler, faster way. We introduce a novel RNA transformation method that leverages the base pair probabilities. We also analyze the importance of the subsequences of the multivariate time series to the classification task, which hints that the regions that are close to the center are essential for this problem. 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 process 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 to perform desired functions. 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. In this analogy, DNA acts as the whole source code, in which genes refer to the functions in the code (segments of the code) [3]. mRNA is the assembly code. And the final product proteins are the running executables.

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 function by themselves, such as by 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 study, we focus on miRNAs. They are small RNAs with a length of about 22 nt. They regulate gene expression post-transcriptionally [4]. A miRNA can regulate the expression of several proteins. They are essential in diverse biological processes, such as development, differentiation, and disease [5, 6]. Hence, it is of great value to understand the formation or the biogenesis of miRNAs. It involves the processing of primary miRNAs (pri-miRNAs). RNAs are 3D molecules, but it is hard to measure the 3D structure (tertiary structure). However, we can understand their behavior by analyzing their 1D sequence. It is easy to obtain through sequencing. A predicted secondary structure of a pri-miRNA's sequence is shown in Figure 1.

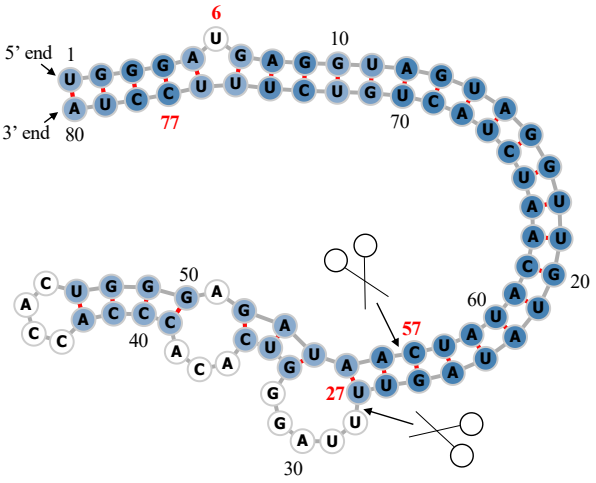

Fig. 1 Predicted secondary structure of the pri-miRNA "hsa-let-7a-1"[2]. We denote the sequence as S. Experimental evidence suggests that the two deviated mature miRNAs are $UGA\cdots GUU$ and $CUA\cdots UUC$. They are $S[6:27]$ and $S[57:77]$. Since $S[6:27]$ ($S[57:77]$) is near the 5' (3') end, we call it "5p (3p) mature miRNA". The starting and ending indices of these two subsequences are indicated in **bold and red**. 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].

[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.

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 [7]. 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 two ends.

One strand of the resulting molecule is loaded into an RNA-induced silencing complex (RISC). This loaded strand guides the RISC to the target mRNA to silence it, and it results in gene silencing. This mechanism was discovered in the aforementioned Nobel Prize. 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. Hence, accurate cleavage of pre-miRNAs by Dicer is crucial for gene silencing. It is reasonable to argue that the structure of the pre-miRNAs informs Dicer about the cleavage process (i.e., where the cleavage sites are). A recent study shows that a particular secondary RNA structure, namely 22-bulge, enhances the accuracy of miRNA biogenesis experimentally [8].

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 [9–11] 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 [12] and caspasesl [13], 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 to this topic, including support vector machine [13], deep neural networks [14]. In this study, we focus on human dicer cleavage sites. We review the studies on human dicer cleavage sites here. PHDcleav employed support vector machines (SVM), leveraging sequence and structure-based features [15] for the prediction. LBSizeCleav improved upon it by considering the loop and bulge lengths [16]. [17] proposed an ensemble learning approach, using a gradient boosting

machine for better accuracy. [18] developed a deep learning model, namely DiCleave. This model used an autoencoder to learn the secondary structure embeddings of pre-miRNAs from all the species (not only human) presented in the miRBase database and leveraged this information in the prediction of human dicer cleavage sites. These methods begin with curated pre-miRNA sequences from the miRBase database. Their secondary structures are predicted using tools such as Quickfold or RNAfold. These secondary structures are stored in the dot bracket notation. They are also sequences. Patterns are extracted from these 2-tuple sequences. These patterns are called cleavage patterns. If the patterns contain a cleavage site, they are called positive patterns. If not, they are called negative patterns. They create the positive patterns by setting the cleavage sites at the middle of the patterns. Details of creating cleavage patterns would be discussed in Section **??**. One exception is the follow-up work of [18], which creates the cleavage pattern by allowing cleavage sites to appear at any position within the pattern, instead of the middle [19]. Then, these works use different encoding schemes to represent these patterns and use different machine learning models to perform classification on them.

Despite the advances made, which mainly focus on the prediction accuracy, these models suffer several limitations. They rely heavily on complicated feature engineering and opaque deep learning models [18, 19]. It results in a lack of generalizability and interoperability. Also, they have a long running time. There is a need to design a simpler model so that it can be easily extended to other prediction tasks on RNA data. The RNA data in miRBase are sequence data. And the predicted secondary structure can also be presented as sequences. These sequences are strings because the entries are discrete. One of the ways to analyze string data is to transform it into time series data where the entries are continuous. By doing that, it builds a strong bridge between RNA analysis and time-series data mining. In response to that, we introduce the transformation method of RNA data to time series. In addition, we leverage the base pair probabilities in the predicted secondary structure in the transformation. To the best of our knowledge, this information has not yet been leveraged in the transformation. For the resulting time series data, we employ state-of-the-art convolution based classifiers, whose superiority has been experimentally tested on data from diverse domains [20]. Besides, we also perform perturbation-based experiments of classification on the resulting time series to investigate the importance of the subsequences to the prediction.

In summary, our contributions are shown as follows.

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. Time series modeling 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 to understand the data more intuitively.
2. We propose making use of 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 these base-pair probabilities in 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.
4. We test our models with perturbation-based experiments. It shows that the regions that are close to the cleavage sites, the 3' arm and 5' arm that are close to the cleavage sites, as shown in figure 1, are important for the human dicer cleavage sites prediction. It agrees with the existing study [17].

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] Cohen, W.W.: A Computer Scientist's Guide to Cell Biology: A Travelogue from a Stranger in a Strange Land, New York, NY (2007)

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

[5] 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)

[6] 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)

[7] 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)

[8] Nguyen, T.D., Trinh, T.A., Bao, S., Nguyen, T.A.: Secondary structure rna elements control the cleavage activity of dicer. Nature Communications **13**(1), 2138 (2022)

[9] 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)

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

[11] 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)

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

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

[14] 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)

[15] 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)

[16] 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)

[17] 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)

[18] 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)

[19] Mu, L., Akutsu, T.: DiCleavePlus: A Transformer-Based Model to Detect Dicer Cleavage Sites within Cleavage Patterns. Submitted, under review (2025). https://github.com/MGuard0303/DiCleavePlus

[20] Middlehurst, M., Schäfer, P., Bagnall, A.: Bake off redux: A review and experimental evaluation of recent time series classification algorithms. Data Mining and Knowledge Discovery **38**(4), 1958–2031 (2024)