OXFORD

Briefings in Bioinformatics, 2023, **24(2)**, 1–13
https://doi.org/10.1093/bib/bbad088
Advance access publication date 11 March 2023
Problem Solving Protocol

miWords: transformer-based composite deep learning for highly accurate discovery of pre-miRNA regions across plant genomes

Sagar Gupta and Ravi Shankar

Corresponding author. Ravi Shankar, Studio of Computational Biology & Bioinformatics, The Himalayan Centre for High-throughput Computational Biology (HiCHiCoB, A BIC supported by DBT, India), CSIR-Institute of Himalayan Bioresource Technology (CSIR-IHBT), Palampur, Himachal Pradesh 176061, India. E-mail: ravish@ihbt.res.in

Abstract

Discovering pre-microRNAs (miRNAs) is the core of miRNA discovery. Using traditional sequence/structural features, many tools have been published to discover miRNAs. However, in practical applications like genomic annotations, their actual performance has been very low. This becomes more grave in plants where unlike animals pre-miRNAs are much more complex and difficult to identify. A huge gap exists between animals and plants for the available software for miRNA discovery and species-specific miRNA information. Here, we present miWords, a composite deep learning system of transformers and convolutional neural networks which sees genome as a pool of sentences made of words with specific occurrence preferences and contexts, to accurately identify pre-miRNA regions across plant genomes. A comprehensive benchmarking was done involving >10 software representing different genre and many experimentally validated datasets. miWords emerged as the best one while breaching accuracy of 98% and performance lead of ∼10%. miWords was also evaluated across *Arabidopsis* genome where also it outperformed the compared tools. As a demonstration, miWords was run across the tea genome, reporting 803 pre-miRNA regions, all validated by small RNA-seq reads from multiple samples, and most of them were functionally supported by the degradome sequencing data. miWords is freely available as stand-alone source codes at https://scbb.ihbt.res.in/miWords/index.php.

Keywords: microRNA, transformers, CNN, deep learning, genomics

INTRODUCTION

microRNAs (miRNAs) are prime regulatory small RNAs (sRNAs) having ∼21 bases as length which are derived from longer precursor miRNA molecules (pre-miRNAs). Discovering these pre-miRNAs is the central to the problem of finding miRNAs. However, finding the pre-miRNAs remains a challenge, and more so in plants. Unlike animals, in plants mature miRNA formation from the precursors is a single-step process, with highly variable sequence and its secondary structural properties [1]. One can fathom the difficulties in the identification of plant miRNAs by the fact that in year 2018 miRBase had to scrap a large number of the reported plant miRNA data due to poor annotations [2]. The traditionally considered sequence and structural properties and features to identify miRNAs are also responsible for difficulty in identifying them. These traditional properties overlap a lot with other classes of RNAs also and differ significantly from those observed for animal pre-miRNAs. Figure 1 illustrates this while suggesting how much error prone the process of pre-microRNA discovery may become while relying on such traditional features.

In the identification of pre-miRNA, identification of secondary structure patterns, hairpin loops and their thermodynamic stability have been the most followed traditional approaches. Additionally, homology and conservation patterns were also used to locate a similar kind of precursors in other genomes. Compared to the conservation and rules-based methods, machine learning-based methods are mainly anchored on sequence and structure-based features of pre-miRNAs with more mature automated statistical learning processes. Most of the machine learning-based tools apply almost the similar set of features, as mentioned above, and differ mainly in the machine learning algorithms applied.

Off-late, sRNA-sequencing supported miRNA discovery has become popular where the sequencing reads mainly work as a support guide while at their core they use the same pre-miRNA discovery algorithms. These approaches have their own shortcomings and they are not immune to false identifications. Their dependence on sRNA-seq data makes them not easily approachable. This is well reflected by the skewed distribution of sRNA-sequencing studies reported from various nations

Sagar Gupta is a PhD research scholar in Bioinformatics at Studio of Computational Biology & Bioinformatics, HiCHiCoB, CSIR-Institute of Himalayan Bioresource Technology (CSIR-IHBT), India. His research interests are focused on computational genomics with machine and deep learning.
Ravi Shankar is a principal scientist and Coordinator of the Himalayan Centre for High-throughput Computational Biology (HiCHiCoB), a National Bioinformatics Center (BIC) of Department of Biotechnology, Ministry of Science & Technology, at Division of Biotechnology, CSIR-Institute of Himalayan Bioresource Technology (CSIR-IHBT), India. More about his research group can be found at https://scbb.ihbt.res.in.
Received: December 20, 2022. **Revised:** January 30, 2023. **Accepted:** February 15, 2023

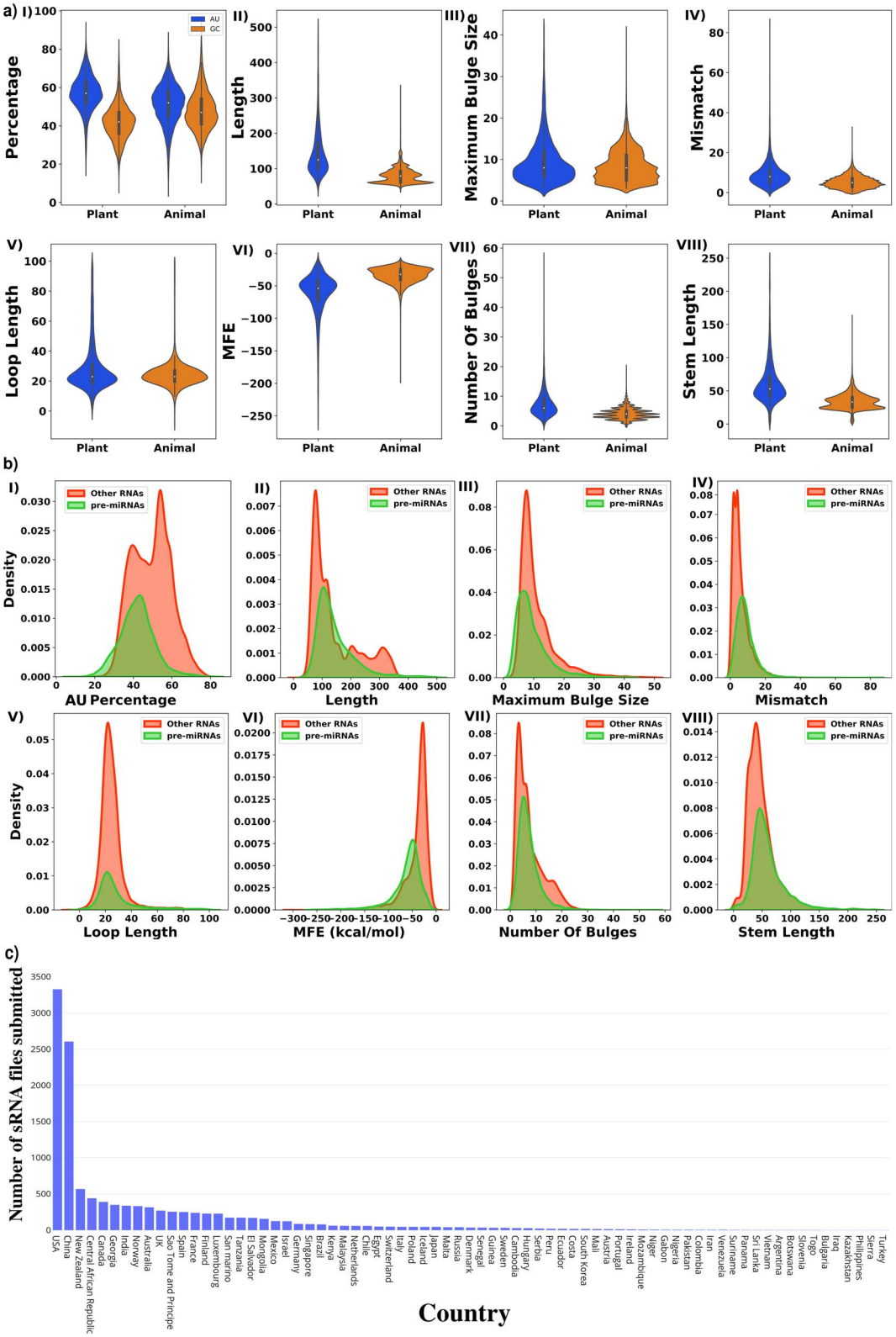

Figure 1. (**A**) Pattern of distribution comparison between animal and plant pre-miRNAs. Values differ significantly between animals and plants as unlike animal pre-miRNAs, plant pre-miRNAs display much more complexity and variability. (**B**) Pattern of distribution comparison between pre-miRNAs v/s other RNAs in plants. As can be seen clearly that most of these properties are actually not strong discriminators as lots of overlap in their values occur between pre-miRNAs and other RNAs. (**C**) Highly skewed distribution of sRNA-sequencing studies across the world. Barring a few nations, a majority of nations display lesser accessibility to sRNA-sequencing. This skew becomes much sharper if the values are normalized for the population of the country. Thus, tools expecting sRNA-seq data are indirectly not of much help for most of the world, as they limit miRNA biology to just few nations.

Table 1. List of some published tools for pre-miRNAs identification

S. No.	Software	Algorithm	Year [Ref.]	sRNA-Seq	Webserver (W)/Standalone (S)
1	MiRFinder	SVM	2004 [3]	N	S
2	MIRcheck	Target Identification	2004 [4]	N	S
3	FindMiRNA	K-mer-based sequence similarity	2005 [5]	N	S
4	MiMatcher	SVM	2005 [6]	N	S
5	PalGrade	Scoring hairpins by thermodynamic stability and structural features	2005 [7]	N	-
6	*Triplet-SVM*	SVM	2005 [8]	N	S
7	MicroHARVESTOR	Sequence similarity	2006 [9]	N	S
8	RNAmicro	SVM	2006 [10]	N	S
9	miPred	Random forest	2007 [11]	N	W/S
10	microPred	SMOTE	2009 [12]	N	W/S
11	miRanalyzer	sRNA-Seq-based filtering	2009 [13]	Y	W/S
12	MIReNA	sRNA-Seq-based filtering	2010 [14]	Y	S
13	mirDeep-P	sRNA-Seq-based filtering	2011 [15]	Y	S
14	PlantMiRNAPred	SVM	2011 [1]	N	S
15	miR-BAG	Bagging ensemble (SVM, BF-Tree and Naive Bayes)	2012 [16]	Y	W/S
16	mirDeep2	sRNA-Seq-based filtering	2012 [17]	Y	S
17	mirDeep*	sRNA-Seq-based filtering	2013 [18]	Y	W/S
18	HuntMi	Random forest	2013 [19]	N	S
19	ShortStack	sRNA-Seq-based filtering	2013 [20]	Y	S
20	MiPlantPreMat	SVM	2014 [21]	N	S
21	miR-PREFeR	sRNA-Seq-based filtering	2014 [22]	Y	S
22	plantMirP	Random forest	2016 [23]	N	S
23	DP-miRNA	Boltzmann machine-based DL	2017 [24]	N	S
24	deepSOM	DL-based self-organizing maps	2017 [25]	N	W/S
25	deepMiRGene	LSTM	2017 [26]	N	S
26	miRNAs	Semi-supervised and transductive learning	2018 [27]	N	S
27	DeepMiR	CNN	2019 [28]	N	S
28	mirDNN	Convolutional deep residual networks	2021 [29]	N	S

Note: These tools may be categorized broadly into similarity based, rules based, probability based, shallow learning based and deep learning based.

(Figure 1C). And in terms of per capita, this skew becomes much sharper. Also, sRNA-seq can capture only those miRNAs which express in any given condition and not all. Many of these algorithms capture non-miRNAs as well as discard genuine miRNAs.

Table 1 highlights the various categories of the core algorithms employed to discover pre-miRNA regions across genome. Compared to homology, rules and probability-based methods, the machine learning methods performed much superior. However, not much development has happened thereafter. Very recently, deep learning (DL) techniques have emerged highly successful for various applications as they have been very effective in digging out better but hidden features for model building which are otherwise difficult to detect manually [30]. DL approaches like convolutional neural networks (CNNs) and recurrent neural networks have shown huge success in image recognition and natural language processing (NLP) [31–33].

Limited forays have been made into DL-based tools to detect pre-miRNAs while there remains a lot of voids to be filled. First is the inconsistent performance where huge gaps were found when benchmarked across different datasets. Secondly, most of them are still based on direct reading of the input sequences with four-state nucleic acid sequence inputs and three-states secondary structure inputs. Third, barring CNN, DL approaches like Long Short-Term Memory (LSTM) and Restricted Boltzmann Machine (RBM) DL require lots of compute power, time, and resources. They can't be parallelized and are compute resources exhaustive. Further to this, they fail to detect long ranged associations effectively. Plant pre-miRNAs have much larger sequences than animals which complicates learning. Also, the existing software pool still perform poorly in their practical application for genome annotation for miRNAs, as noted by some recent studies [34]. Recently, a new revolutionary DL architecture, Transformers, has been introduced, which has emerged as a highly efficient architecture for language processing tasks [35]. It uses a self-attention mechanism on the input which can be processed in parallel while more effectively capturing the long-distanced associations and contexts within any sequential data. Very recently, transformers have been used for genome-wide pre-miRNAs discovery with miRe2e [36]. It used CNN-transformers to learn upon sequence, structure and MFE. Though miRe2e is specific to animals, it demonstrated a significant leap transformer-based approaches are capable of. However, miRe2e too uses single-character words and Minimal Free Energy (MFE) which may not be effective for much complex plant genomes. Consideration of appropriately bigger word sets with additional learning upon the global transformer scoring patterns may improve such approach further and enhance its scope.

The present work proposes a novel composite DL system where the multiheaded transformers define the first phase to generate the classification decision score (*T*-score). Unlike most of the existing approaches, miWords sees a genome sequence as a set of sentences composed of words made from monomers, dimers, pentamers and structural triplets capturing sequence, structure and shape-based information and associations among themselves. The classification score made by the above-mentioned system can be used to classify an independent sequence as well as to convert a genomic sequence into a *T*-score sequence which in turn captures long-ranged genomic contextual information, on which further deep learning through CNN was done to enhance the

performance and applicability for genomic annotation purposes. To this, one can also optionally use sRNA-seq data without any binding while maintaining a very high level of accuracy even without sRNA-seq data availability. Several levels of benchmarking studies have been done and miWords consistently emerged as the best performing tool for plant pre-miRNA discovery.

MATERIALS AND METHODS

Primary dataset creation

A total of 5685 pre-miRNAs from miRBase, covering 27 plant species, formed the initial positive dataset [37]. Considering the central base of the terminal loop of these pre-miRNAs, sequence encompassing the 200 base long flanking regions formed the positive instances representing the pre-miRNAs. The negative instances were derived from mRNAs, rRNAs, snRNAs, tRNAs and other noncoding RNAs using the same approach while considering the central base in their pseudo-hairpins. The dataset formed this way was called Dataset 'A'.

Filtering of Dataset 'A' for the most confident pre-miRNA instances was done using sRNAnno, pmiREN and PNRD [38–40]. A total of 3923 pre-miRNAs qualified this filtering and formed the positive instances. An equal number of randomly picked negative instances from Dataset 'A' formed the negative instances. Together they formed Dataset 'B'.

Another Dataset 'C' was formed by unifying the datasets used by various published tools and removing the redundancy. A total of 9214 pre-miRNAs and equal number of non-miRNAs formed this dataset.

In addition to these datasets, the benchmarking dataset used by Bugnon et al. [41] was also used. Full description on these datasets can be found in the associated Supplementary Materials and Methods details available online at http://bib.oxfordjournals.org/.

Sequence representations and tokenization

Each sequence may be seen as a sequence of various words of size *K (k-mers)*. For example, the sequence ATTGGCAG may be represented as a sequence of trimeric words of ATT, TTG, TGG, GGC, GCA and CAG. A total of 64 3-mer words, 1024 unique 5-mer words, 16 2-mer words are uniquely possible from the alphabet of A = {'A', 'T', 'G', 'C'}. Similarly 27 unique structural triplet words are possible from the alphabet S = {'(', '.', ')'}. All these words can be assigned unique integers as an ID, which is called token. The same token is used to convert the sequences into their integral tokenized representation when they are converted into some k-mer representation as shown above. The tokenized form of the sequence becomes the entry point to the transformers where their integral representation makes it easy to convert them into numeric vectors and matrices, called embedding.

Implementation of the transformer models of miWords

Each tokenized sequence was converted into a two-dimensional matrix whose rows were determined by the encoding vector size for each token (28 in the present case), and the total number of tokenized words determined the number of columns. This is called word embedding. Besides word embedding, the positional embedding of the words was done in parallel, and the combined word and positional embedding of the tokenized sequence became the input to the transformers which had five key steps.

In these five steps, at first the embedded word matrix becomes a source to which three different weight matrices namely W(Query), W(Key), W(Value) were separately multiplied to obtain Query, Key and Value matrices. The weights were iteratively updated during the training process. At the second step, inner product between Query and Key matrices provided an association map of proximity between various words in the sentence. At the third step, scaling of this product reduced the extreme value gradients, and its normalization through softmax at the fourth step fixed the values in the range of 0–1 besides neutralizing meaningless associations. The final attention scores between the words were obtained by multiplying the value matrix to the softmax normalized matrix at the fifth step.

A multi-headed transformer with 14 attention heads was implemented in the present study and the above step was repeated by all of them to obtain their respective attention score vectors. These attention score vectors from each transformer head were finally concatenated and passed on to a multi-layered system having dropout, normalization, feed forward, global average pooling, dense layers and XGBoost classification layers.

Hyperparameter optimization of the transformer system was done using Bayesian optimization while random search optimization was used for its XGBoost part. Full implementation details of the transformer system of miWords are given in the Supplementary Materials and Methods section while Supplementary Table S1 Sheet 1–2, available online at http://bib.oxfordjournals.org/, details about the hyperparameters and their optimization. Figure 2 illustrates the full Transformer system implementation of miWords.

CNN-based deep learning on Transformer's scoring profiles for genomic scanning capabilities

The above-mentioned Transformer system generates the score (*T*-score) which captures the potential of the region to host pre-miRNA. If run across the genome, for every base such *T*-score is generated. This sequence of *T*-score is passed to two CNNs to learn from the global and flanking region contexts to identify pre-miRNA regions across genomic sequences with enhanced accuracy.

Two different types of CNNs are involved. The first one converts the *T*-score sequence into one-hot encoding matrix of dimension 280 × 10, which becomes input to a CNN made up of two Convolution, one max pooling, four batch normalization and four hidden layers.

The second CNN is a bimodal CNN which works if the sRNA-seq profile for the genomic region is available. Its first part's input is reads per million (RPM) representation of each base position which passes through a 1-D convolution, one max pooling, two batch normalization and two dense layers. The second part is the *T*-score CNN as mentioned above. Full implementation details of the CNNs and their hyperparameter optimization are given in the Supplementary Materials and Methods section available online at http://bib.oxfordjournals.org/.

Validation of the identified pre-miRNAs using miRBase, RNA-seq and degradome-seq

The identified pre-miRNAs across the tea genome were first screened through the miRBase database for similarity-based validation from the experimentally reported miRNAs. Further to this, sRNA-seq read data from different studies covering 104 samples of *Camellia sinensis* were collected from Gene Expression Omnibus (GEO)/Sequence Read Archive (SRA). Reads were checked for quality against which scanning for the identified pre-miRNAs was done while applying certain criteria to be considered to call a pre-miRNA supported by sRNA reads.

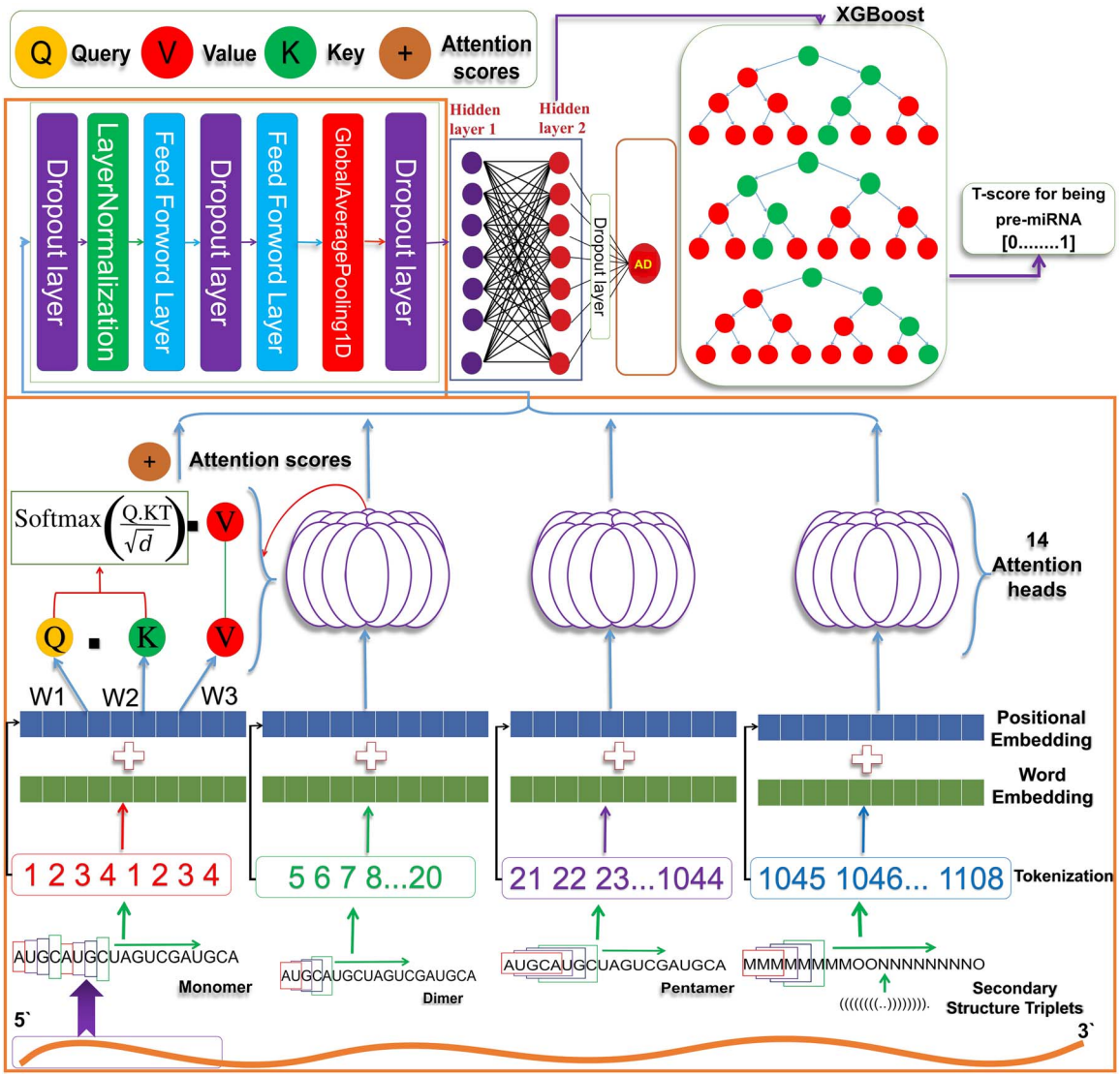

Figure 2. Implementation of the transformer-based module to identify pre-miRNAs. The image provides the brief outline of the entire computation protocol implemented to develop the Transformer-XGBoost based model to identify pre-miRNAs. This illustrates how a genomic sequence can be seen as a sentence composed of words and their related arrangements which can be efficiently learned through multi-headed transformers. The various nucleotides k-mers and RNA secondary structural triplets define the words for any given regions (the sentence). The words and their attention scores are evaluated through query, key and value matrices which are then passed to different layers of a deep learning protocol to present its learning for classification job through XGBoost.

Functional validation of the pre-miRNAs was done using tea degradome-seq data for 15 different samples and CleaveLand [42]. Further details are given in the Supplementary Materials and Methods section and Supplementary Table S1 Sheet 3 available online at http://bib.oxfordjournals.org/.

miWords implementation

miWords source codes have been made publicly available at GitHub (https://github.com/SCBB-LAB/miWords). miWords has been developed using Python, Keras, and works on Linux/Mac OS. Figure 3 provides the illustration of the running workflow of miWords system. A user has three different choices:

(i) Scan sequences ≤400 bases with transformers only, run the commands for module 1 (M1):
sh M1.sh path_of_execution_folder input_sequence_file_in_fasta_format.

(ii) Scan large genomic sequences without sRNA-seq reads: Here the sequence is converted into transformer score sequence on which CNN finally identifies the pre-miRNA regions. Run the following command:
sh M2.sh execution_folder_path input_sequence_file (t2 in the present example) module-subtype ('A').

(iii) Scan large genomic sequences with sRNA-seq reads: Run the same command as given above with module subtype 'B', calling Transformers with bimodal CNN.

A detailed step-by-step interactive guide with examples and files for all these three conditions is given at the tutorial page of miWords server at https://scbb.ihbt.res.in/miWords/tutorial.php.

Full methods details are given in the Supplementary Materials and Methods section available online at http://bib.oxfordjournals.org/ and readers are highly encouraged to go through it for comprehensive details of the methods.

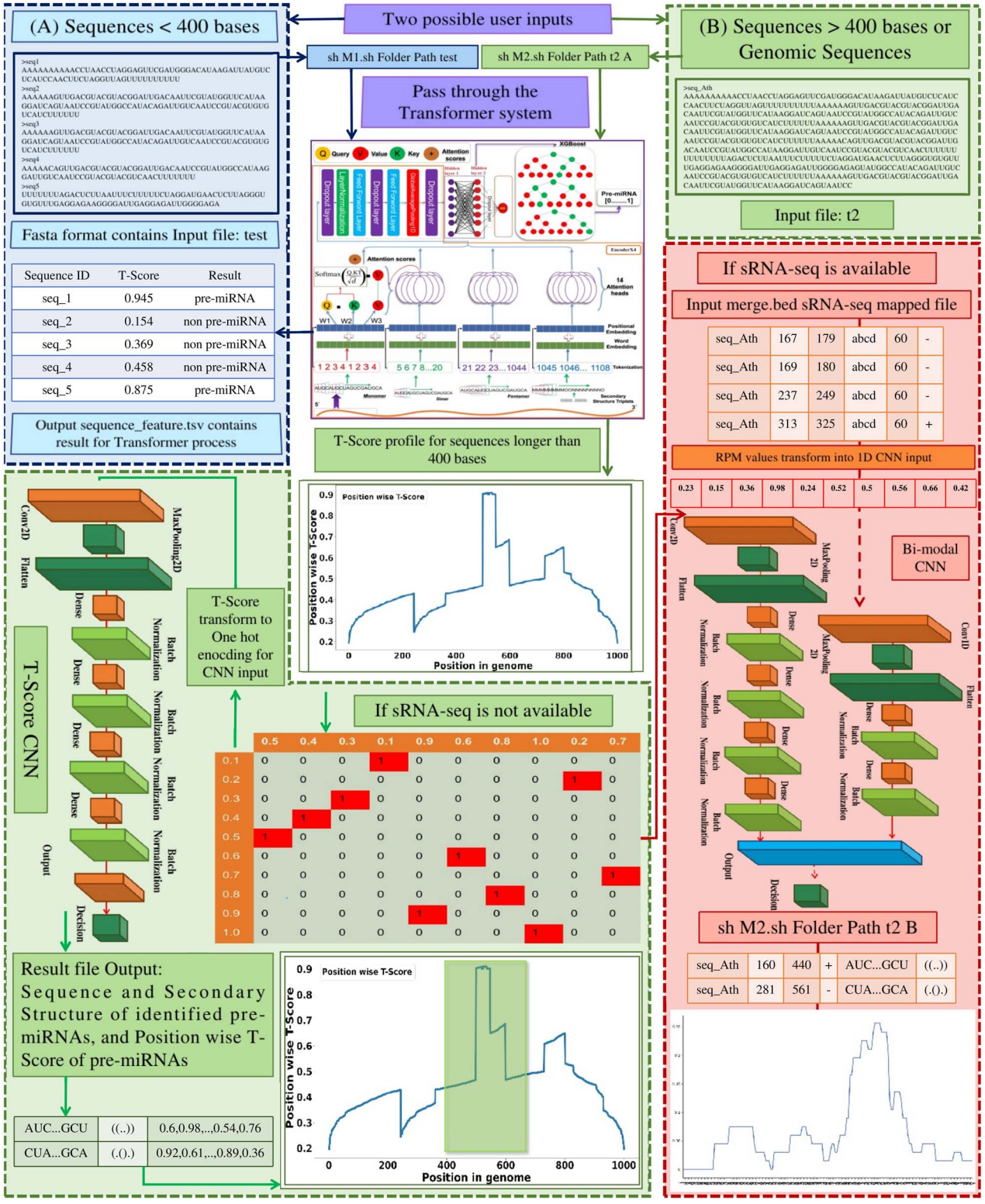

Figure 3. Illustration of the running workflow of miWords system. The user can provide either shorter sequences on which only the transformer-based module runs and gives its classification with *T*-score. The second one is for longer and genomic sequences, which passes the entire sequence's *T*-score representation to a CNN module and finally provides the candidate pre-miRNAs across the genome. If the sRNA-seq data is available, then the user can run the bimodal CNN which takes the *T*-score profile as well as sRNA-seq mapped genome to identify the pre-microRNAs.

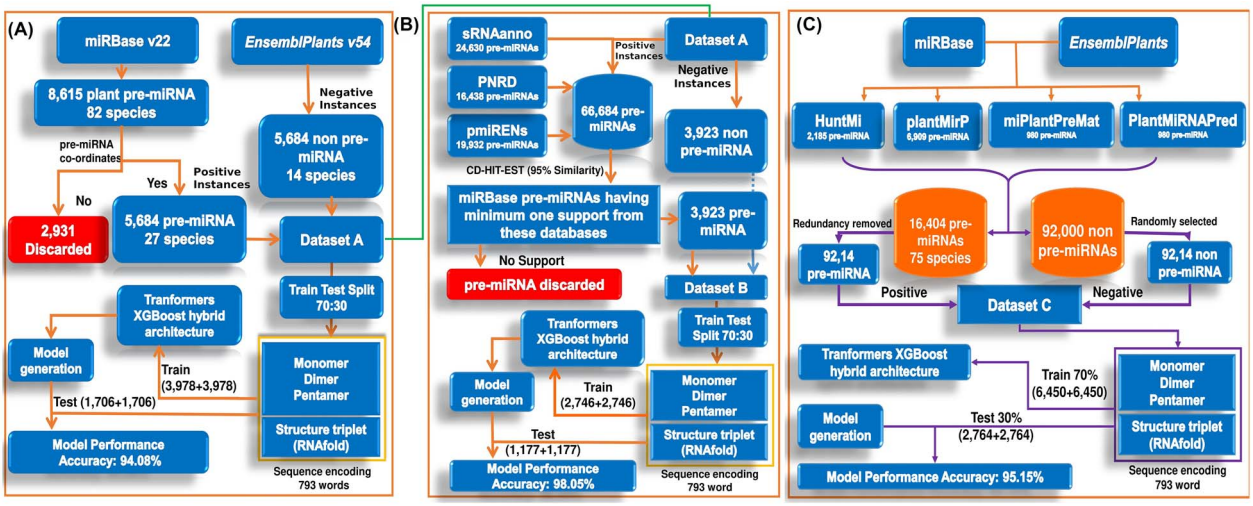

Figure 4. Flowchart representation of dataset processing and formation. (**A**) The protocol followed for Dataset 'A' creation, (**B**) protocol followed for Dataset 'B' creation and (**C**) protocol followed for Dataset 'C' creation.

RESULTS AND DISCUSSION

Sentences, Words and Attention: Seeing genome as a pool of sentences through transformers delivers high accuracy

While building the models, three different datasets were considered initially: Datasets 'A', 'B' and 'C' as described in the methods section. Figure 4 illustrates how these datasets were built and used. For the building of a universal model for plant pre-miRNA regions for its characterization against the other types of RNAs, we used 13 different combinations for various sequence representations. They were evaluated for performance through the raised transformer encoder-based model. An assessment was made for each representation considered where Dataset 'A' was split into 70:30 ratio to form the train and test datasets. At first, the transformers were trained and tested without the XGBoost gradient boosting to evaluate its performance. The observed accuracy for monomeric representation was just 72.36%. This was followed by introduction of dimeric, trimeric and pentameric sequence representations which returned the accuracy of 73.21%, 75.36% and 79.01%, respectively, while covering a total of 199, 198 and 196 words per sequence window, respectively. Besides the above-mentioned sequence-based properties, the secondary structure stem-loop-based structural triplets were also used for the representation (198 words), as pre-miRNAs exist in the stem-loop hairpin form. This fetched an accuracy of 77.09%. As can be seen here, individually all these properties did not score much and needed information sharing with each other.

The next step was observing the influence of combining these sequence and structure derived word representations for the sequences. Combination of the various representations of sequences was done in a gradual manner in order to see their additive effect on the classification performance. These combinations yielded a better result than using any single representation, as can be seen from Figure 5 performance plots for the various combinations of the sequence representations. Monomers + dinucleotides + trinucleotides + structural triplets (795 words), Monomers + dinucleotides + trinucleotides + pentanucleotides + structural triplets (991 words), and monomers + dinucleotides + pentamers + structural triplets (793 words) combinations yielded the accuracy of 93.67%, 93.84% and 93.96%, respectively, with the latter one having the better balance between the sensitivity and specificity values. Thus, the combinations of different representations for the genomic sequences markedly improved the performance through the NLP approach of transformers.

The Transformers built from the combination of monomer + dimer + pentamer + structure triplet encodings delivered a good accuracy of 93.96%. There was a gap of 0.9% between sensitivity and specificity, though not a big gap, yet we tried to reduce it further. In doing so, the output layer of the transformer having the LeakyReLu activation function was replaced by XGBoost for the classification purpose. This hybrid deep-shallow model reduced the performance gap between the sensitivity and specificity to just 0.46% while also increased the accuracy slightly to 94.08% (Supplementary Table S2 Sheet 1 available online at http://bib.oxfordjournals.org/). Likewise, another model was derived from Dataset 'B' which was based on filtered high confidence positive instances. This model attained an accuracy of 98.04% on its test set along with a specificity of 98.56% and a sensitivity of 97.54%. This all became the first part of the transformer-based pre-miRNA identification system, which can even work independently and can be used directly for pre-miRNA region identification. The full architectural details including hyperparameter optimizations are elaborated in the Supplementary Materials and Methods section available online at http://bib.oxfordjournals.org/.

Consistent performance across different validated datasets reinforces miWords as a universal classifier for plant pre-miRNAs

As mentioned in the methods details and Figure 4, initially, for performance testing three different datasets 'A', 'B' and 'C' were created. Dataset 'A' had 5684 positive and 5684 negative instances, totalling 11 368 instances. Dataset 'B' had 3923 positive and 3923 negative instances, totalling 7846 instances. Seventy percent of datasets 'A' and 'B' were used for training purpose and 30% were kept aside as totally unseen test set instances in a mutually exclusive manner to ensure unbiased performance testing with no scope for memory from data instances.

Besides raising the trained models and testing it, as mentioned in the methods details and the section above, 10-fold random train-test trials have also been done to evaluate the consistency of the transformer system on Datasets 'A' and

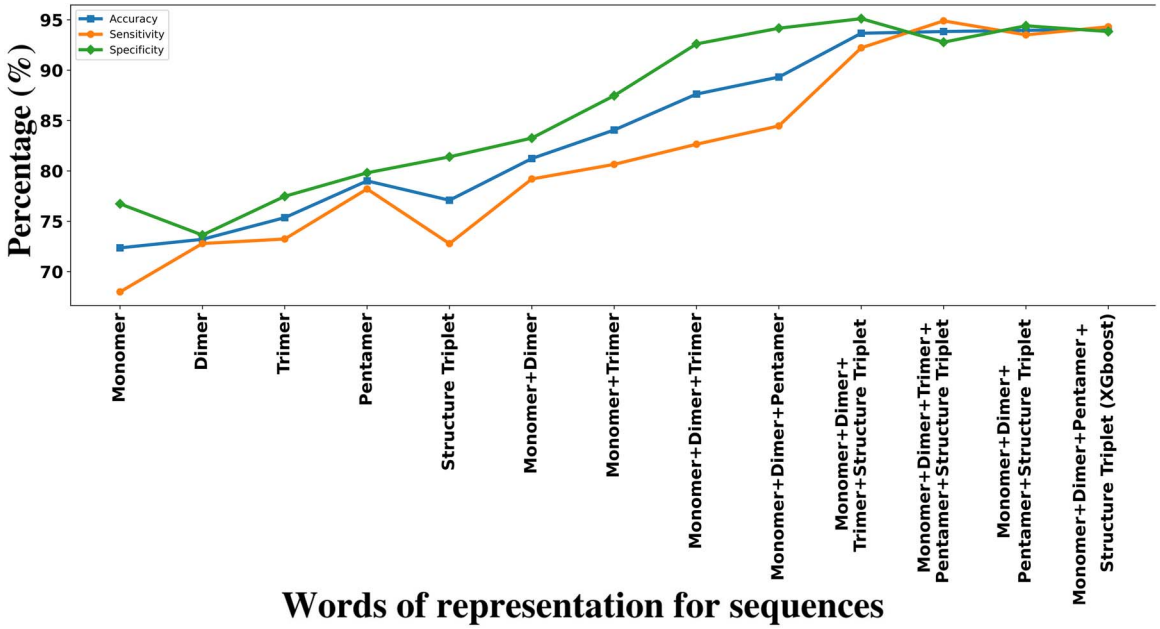

Figure 5. Ablation analysis for five main properties in discriminating between the negative and positive instances. Impact of combination of the monomer, dimer, trimer, pentamers and structure triplet properties-based sequence word representations. These word representations appeared highly additive and complementary to each other as the performance in accurately identifying pre-miRNAs increased substantially as they combined together.

'B'. This 10-fold random trials concurred with the above-observed performance level and scored in the same range consistently. All of them achieved good quality receiver operating characteristic curves with high Area Under Curve (AUC) values in the range of 0.9294 to 0.9436 (Dataset 'A') and 0.9734 to 0.9779 (Dataset 'B') while maintaining reasonable balance between specificity and sensitivity (Supplementary Table S2 Sheet 2–3; Supplementary Figure S1A and B available online at http://bib.oxfordjournals.org/).

miWords consistently outperforms all the compared tools for pre-miRNA discovery

This study has performed a series of different comparative benchmarkings. The first four are covered in this section. In this comparative benchmarking, the performances of eight compared software were studied across Datasets 'A', 'B' and 'C'. The compared tools covered some best performing machine learning and recently developed DL approaches for pre-miRNA discovery. Besides measuring the performance of miWords of this neutral and totally unseen testing part of Datasets 'A' and 'B', performance of the other eight tools was also benchmarked. The performance measure on the test set of Datasets 'A' and 'B' gave an idea how the compared algorithms in their existing form perform. The third dataset 'C' was used to carry out objective comparative benchmarking, where each of the compared software was trained as well as tested across a common dataset in order to fathom exactly how their learning algorithms differed in their comparative performance.

All these eight tools were tested across both the datasets ('A' and 'B') where miWords outperformed all of them across both the datasets, for all the performance metrics considered (Figure 6A and B). As already reported above for the Dataset 'A' and 'B' test set, miWords scored the accuracy of 94.08% and 98.04% with MCC values of 0.8816 and 0.9610, respectively, while displaying a very good balance between sensitivity and specificity with a difference of just 0.4% on Dataset 'A' and 1.0% on Dataset 'B'. On the same Datasets 'A' and 'B', the second best performing tool was plantmiRP-Rice which scored an accuracy of 87.89% and 92.56%, respectively, and Matthews Correlation Coefficient (MCC) values of 0.75 and 0.8560, respectively. The values were significantly behind those observed for miWords. A Chi-square test confirmed that miWords significantly outperformed the second best performing tool on Dataset 'A' comparative benchmarking (P-value $\ll 0.01$).

On Dataset 'C', all these tools were trained on the same common training dataset and tested across the common testing dataset in order to achieve the objective comparative benchmarking of the algorithms. However, two tools, microPred and miPlantPreMat, could not be included in this part of benchmarking as both these tools do not give provision to train on another dataset and rebuild models. In this benchmarking also, miWords outperformed all the compared tools with a significant margin with the similar level of performance (Figure 6C). The second best performing tool was HuntMi, which attained an accuracy of 90.5% and an MCC value of 0.81 but displayed a much higher gap of ~7% between sensitivity and specificity scores. A Chi-square test done here also confirmed that miWords significantly outperformed the second best performing tool, HuntMi (P-value $\ll 0.01$).

Besides this all, one more interesting objective comparative benchmarking analysis was done on an imbalanced dataset recently provided by Bugnon et al. [41]. In their benchmarking study, they strongly attracted the attention on the fact that how most of the existing pre-miRNA discovery software performed very poorly in an actual scenario where class imbalance exists naturally. The dataset was split into 70:30 ratio to train and test the model for miWords, maintaining 1:1616 ratio of positive and negative instances. Here also, miWords scored the highest for all the performance metrics with a big lead margin from the rest of the compared six software (Figure 6D). The full details and data for this benchmarking study are given in Supplementary Table S3 Sheet 1–4 available online at http://bib.oxfordjournals.org/.

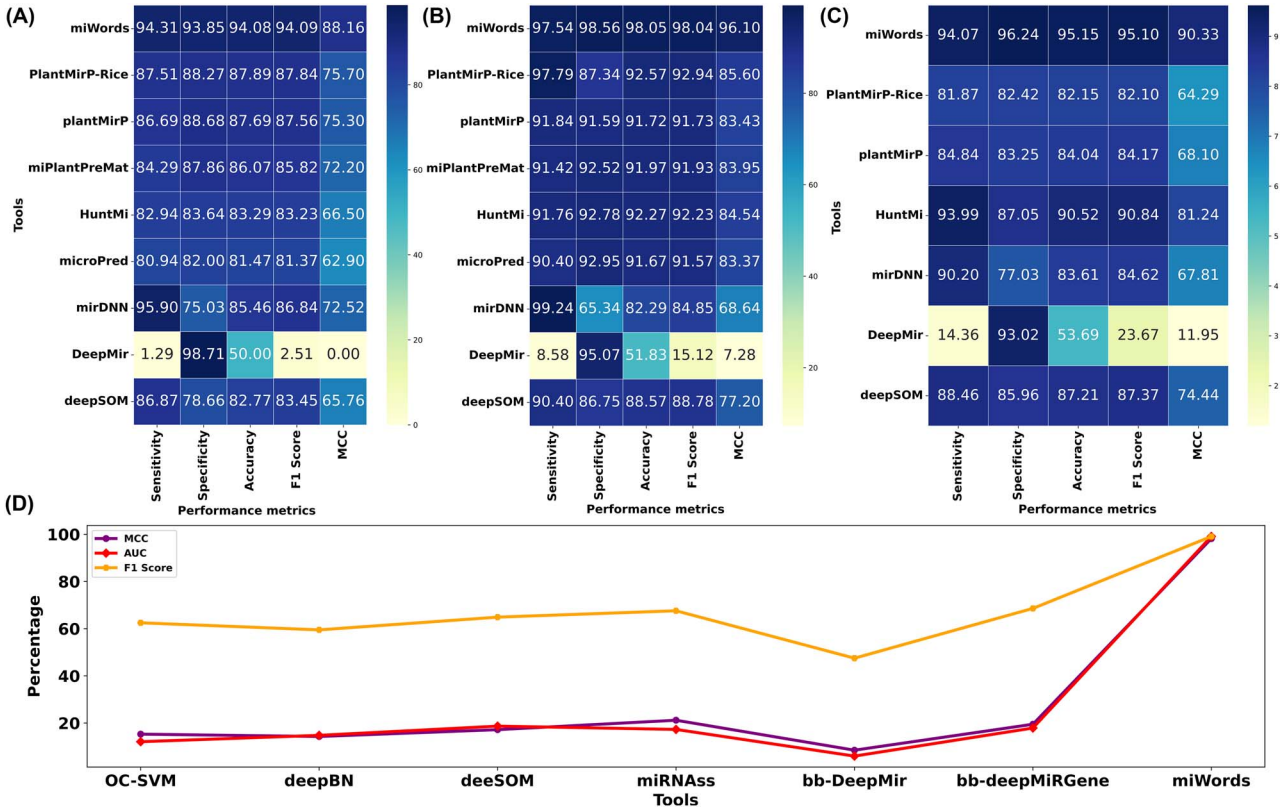

Figure 6. Comparative benchmarking results for miWords for different datasets. (**A**) Benchmarking result on Dataset 'A'. Here all the compared tools were tested on the testing dataset part of Dataset 'A' which was totally unseen and untouched for all the compared tools including miWords. This gives a view of how the compared software would behave in their existing form and models. (**B**) Benchmarking result on Dataset 'B'. These datasets contained the high confidence refined and filtered entries from miRBase while taking support and evidence from three other databases (sRNAanno, PmiREN and PNRD). (**C**) Objective comparative benchmarking on Dataset 'C'. Here, all the compared tools were first trained on a common dataset for training and then tested on a common mutually exclusive dataset for their performance. This gave a clear view on the performance of each of the compared algorithms. (**D**) Comparative benchmarking done on the imbalanced dataset introduced by Bugnon et al. [41]. All the compared tools were trained and tested on this common dataset for objective comparative benchmarking for imbalanced dataset performance. The logic for such dataset is that in usual genomic annotation conditions, the negative instances are manifold higher than the pre-miRNA regions. A capable software should perform good on such imbalanced dataset. Here also, miWords outperformed all the compared software. From the plots it is clearly visible that for all these datasets and associated benchmarkings, miWords consistently and significantly outperformed the compared tools for all the compared metrics.

Genomic context learning on transformer scores delivers significantly good results on genome-wide annotations

Performance over standard testing datasets may be claimed good, as has been done by most of the published software in the past. However, during the real-world application of genome annotation, a huge performance gap exists, far below the acceptable limits. Some recent reports have highlighted the high degree of poor performance by a majority of the existing categories of miRNA discovery during the process of genomic annotations where most of them end up reporting very high proportion of false positives [2, 41].

One major drawback of these existing tools is that they hardly acknowledge the role of relative information from the flanking regions for the identification of the miRNA regions during genome scanning. While in actual the relative scoring patterns between miRNA regions and neighbourhood regions may become highly informative for more accurate discrimination. A high scoring pre-miRNA region is expected to display higher scoring distribution across its bases along with a gradual decline when compared to its non-miRNA flanking regions where scoring is also expected to exhibit a random and sharper trend. A *t*-test between the *T*-score distribution for the flanking regions and pre-miRNA regions supported this view (P-value < 0.05). Thus, it became another important aspect of miRNA regions to refine their discovery with genomic context.

Therefore, we conducted the next part of this study and trained a CNN-based deep learning module on the obtained *T*-scoring profiles across the genomic sequences from the Transformer-XGBoost system run across the genomes. Two different CNN models were raised (Models 'A' and 'B') whose details are given in methods details above and in Supplementary Methods available online at http://bib.oxfordjournals.org/. For Model 'A', an accuracy of 78.6% with sensitivity of 79.21% and specificity of 77.99% was observed. When the same test was carried out for Model 'B', the accuracy of *T*-score CNN attained 90.4% with 91% sensitivity and 89.7% specificity.

From here it transpires that if the Transformer-XGBoost classification system's scoring scheme is learned with genomic context using the above-mentioned *T*-score CNN, the actual application in genomic annotations would benefit a lot as defining the boundaries and pre-miRNA regions became much accurate. Also, since Model 'B' based on the refined Dataset 'B' performed much better, in all the following stages of the present study Dataset 'B' derived models were implemented, including the present *T*-score CNN module.

The next important aspect to evaluate was the raised model's ability to correctly identify the miRNAs and non-miRNA regions

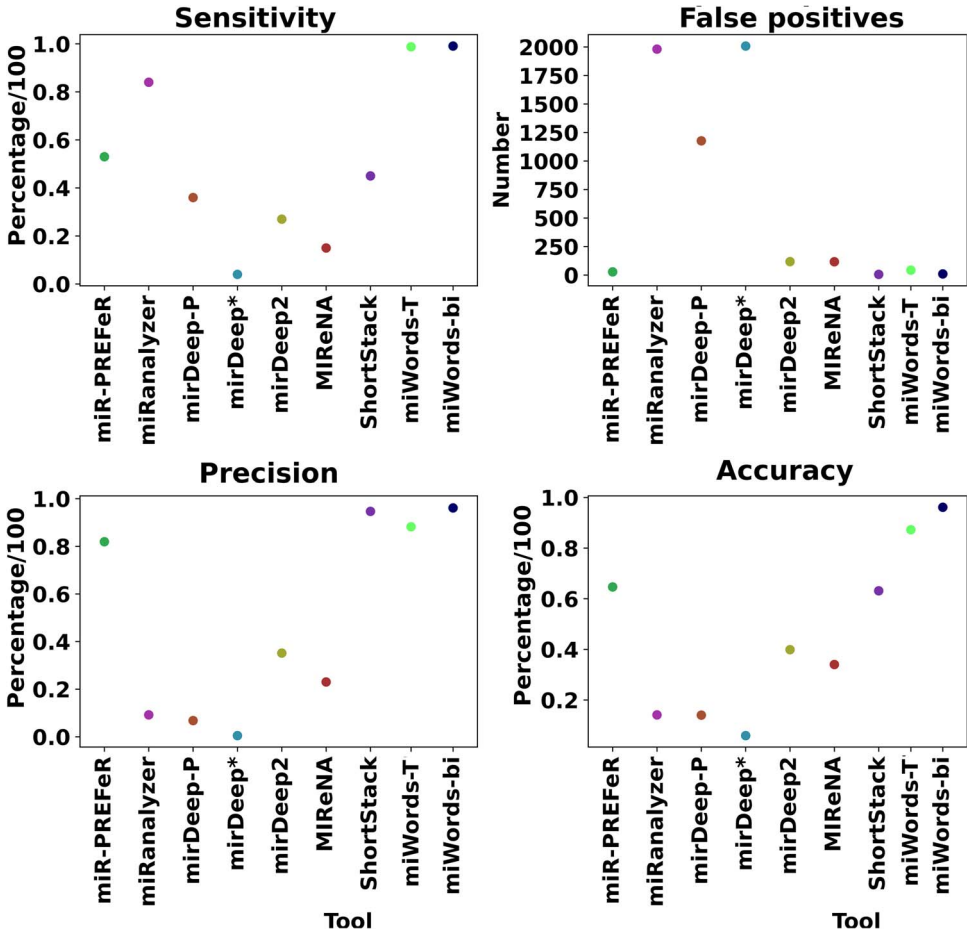

Figure 7. Comparative benchmarking for genomic annotation capability and performance. Most of the existing software perform poorly in the actual application of genome annotation and end up reporting a large number of false-positive cases. It has been recommended to assess performance of any such tool across well-annotated genomes like Arabidopsis. Any reporting of novel miRNAs on such genomes should be considered as a false-positive case and accordingly the performance of a software may be rated. In this performance benchmarking, miWords was compared to the tools that are most preferred ones for genomic annotation at present, as they use sRNA-sequencing read data as a help guide to reduce their false-positive predictions. As can be seen from this benchmarking plot, all the forms of miWords (miWords-T: working without any sRNA-seq reads' help and directly with genomic *T*-score CNN; miWords-Bi: the bimodal CNN form where *T*-score and sRNA-seq derived per base RPM representations of sequences) outperformed all the compared tools for all these performance metrics. 'A' and 'B' are for the datasets used in the present study to derive the models. In *Arabidopsis*, miWords identified all of its pre-miRNAs correctly except three of them, and reported only 10 false positives, the lowest of all.

in some very well-annotated and studied genomes. This would provide the clear picture about the usability of such software in their practical application of discovery of miRNA regions across genomes. For this, the *Arabidopsis* genome was taken with its full annotations reported in miRBase V.22.0. In the first phase, for the entire genome for each base position the transformer score (*T*-score) was generated. This became the input to the *T*-scoring-based CNN. A total of 322 out of the annotated 326 pre-miRNAs of *Arabidopsis* were detected successfully from the raised model. Hence, it was clear that discovering miRNA regions by the above-mentioned system was highly accurate even for genome annotation purposes.

The next important question was that how much novel miRNAs were identified across this genome, which could be most probably the false-positive cases? A total of 759 pre-miRNAs regions were suggested by the transformers. Though this number is very much lesser than what the currently existing pre-miRNA discovery tools and approaches report (including some Next Generation Sequencing (NGS) sRNA-seq data dependent tools) [24], yet even this number may be considered substantially high. However, we got an exceedingly good result when the transformers scores were passed through the above-mentioned *T*-scoring-based CNN. Just 43 novel candidates, the potentially false-positive cases, were obtained for the entire *Arabidopsis* genome.

The existing tools which were run across the *Arabidopsis* genome with sRNA-seq read data supports identified at least 11 false-positive pre-miRNA candidates and went predicting up to 12 306 miRNAs despite having sRNA-seq reads support [22]. This clearly underlines that despite having sRNA-seq read data as their guide, due to inefficiency in their existing core algorithms to identify pre-miRNA regions, in actual these tools do not benefit significantly from sRNA-seq read data guidance and end up identifying a large number of false-positive instances. Also, in general, they grossly missed to identify a big number of actual miRNAs with their sensitivity values falling as low as 3% and attaining utmost 86%.

Thereafter, the natural question arises that how good it would perform if sRNA-sequencing data are also provided? To answer this, at the first step a 1D-CNN module was implemented using RPM normalized expression values while using short read mapping data from 201 sRNA-seq studies across every considered genome in Dataset B. This way every sequence at every base was converted into its corresponding RPM value, capturing its

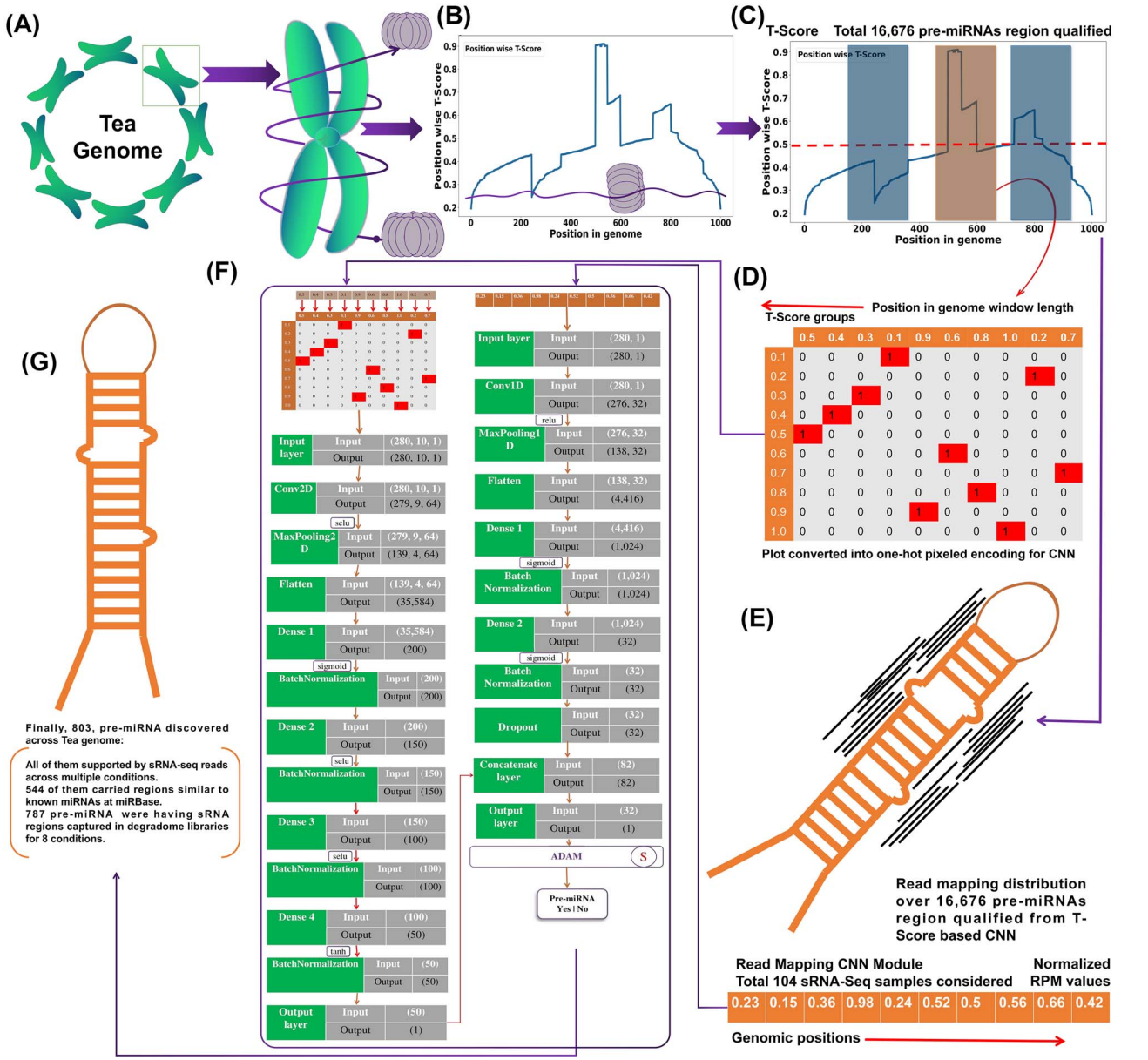

Figure 8. miWords complete architectural implementation for its bimodal option and its steps to annotate the tea genome for pre-miRNAs. Parts A-C display the T-score conversion of the genome. Parts D-F illustrate the subsequent implementation of T-Score and RPM CNNs, using which a total of 803 pre-miRNAs were identified in *C. sinensis*. All of the identified miRNA regions had sRNA-seq reads supports from multiple experimental conditions. >500 matched to already validated known miRNAs, while 787 of them exhibited the presence of their regions in degradome-seq data, approving their functional role as miRNAs involved in targeting.

contribution towards any sRNA level for the given region. On the corresponding test set, this RPM-based CNN scored an accuracy of 85% with sensitivity of 85% and specificity of 85%. Further details of this module are provided in the Supplementary Materials and Methods section and Supplementary Figure S2B available online at http://bib.oxfordjournals.org/. Though, just based on sRNA-sequencing data, the RPM CNN module's stand-alone performance was found lesser than the one obtained from the *T*-score-based CNN by ~5%, it still performed reasonably good. Thus, it also became suggestive of the possible benefit of combined application of both the modules, *T*-score CNN and RPM CNN.

The combinations of *T*-score CNN and RPM-CNN could be made in two ways: (i) connecting in a serial manner where *T*-score CNN output could be passed for filtering by RPM CNN for the final decision, and (ii) a bimodal CNN parallel architecture, where the input goes in parallel to *T*-score CNN and RPM CNN modules and the decision is made in a combined manner after passing through a common dense layer system.

It was found that the performance of the serial architecture dipped down to just 80.4% accuracy. However, the performance of the bimodal CNN model was found increased and better than the stand-alone of both *T*-score and RPM CNNs, with an accuracy of ~91% and far balanced sensitivity and specificity values of 90.4% and 91%, respectively.

Across the Arabidopsis genome, the bimodal CNN reported 323 out of 326 pre-miRNAs and just 10 novel pre-miRNAs, a performance far superior than the current lot of published tools. Figure 7 provides the comparative benchmarking of miWords and its various versions with seven best performing tools that use sRNA-seq data across the Arabidopsis genome, clearly suggesting the top-notch performance by miWords. This also may be noted that in a previous benchmarking study on these

compared software, it was found that they are sensitive towards the size of sRNA-seq data and number of studies included. As this data volume and number of studies increase, the number of potential false positives by these software was reported to increase also [22]. In the present study, we had considered comparatively much bigger sRNA-seq data for *Arabidopsis*, a total of 88 samples, and yet did not see such an overshooting effect for miWords which reported only 10 novel pre-miRNAs.

Application and revalidation: using miWords for plant genome annotation

To exhibit the applicability of miWords in a practical scenario of genome scanning for pre-miRNA discovery, miWords was run across the *C. sinensis* genome whose size is 3.06 GB. Though its genome has been revealed, to this date, there are no entries for tea miRNAs in miRBase.

The first run of miWords, which was the transformer part (Model 'B'), identified 16 676 pre-miRNA regions in the tea genome. This was followed by the bimodal CNN scanning on the generated *T*-score profiles for the above-mentioned pre-miRNA candidates, which reported finally 803 pre-miRNAs in tea. Substantial validation of the identified pre-miRNAs was done through three ways: sRNA-seq read mapping data, similarity to experimentally validated known miRNAs in miRBase, and functional validation through degradome-seq data. All of these potential pre-miRNA regions exhibited sRNA-seq read data mapping to them across multiple samples (total 104 samples, 34 conditions) where at least five reads mapped in each condition. Of total, 544 of these identified pre-miRNAs contained regions similar to known miRNAs in miRBase. Remaining 259 pre-miRNAs were novel candidates. A total of 787 pre-miRNAs exhibited regions reflected in the degradome-seq data suggesting them participating in targeting and functionally valid. All this gave strong support evidence to the identified pre-miRNAs by miWords and its algorithm. Figure 8 provides schematic details of pre-miRNA discovery in tea using miWords. It also illustrates the entire architectural details of the miWords pre-miRNA discovery system.

miWords application was also demonstrated across the recently published genome of *Picrorhiza kurroa* ($2n = 3.4$ GB), an endangered Himalayan herb of very high medicinal value due to its picrosides content [43]. This was done to demonstrate the applicability of miWords in characterizing medicinal plants through miRNAs regions specific to it. A total of 905 pre-miRNA regions were found across the genome, of which 573 regions carried regions similar to already reported miRNAs in miRBase. Of total, 332 pre-microRNAs were found novel and specific to Picrorhiza which may be used for its characterization and analysis for pathways specific to its medicinal values. A target analysis against the picrosides pathway genes identified four such novel miRNAs targeting three genes of picrosides pathways. Associated details for both the studies are given in the Supplementary Table S1 Sheet 3–4 available online at http://bib.oxfordjournals.org/.

CONCLUSION

The miWords algorithm presented here brings a totally new approach to see the miRNAs across the genome using transformer-based composite deep learning. The performance benchmarking across the several types of datasets and with several software underlined a never-seen-before performance leap. With this all the miWords approach ensures far superior annotation of plant genomes for miRNAs which has been almost stalled and limited in the lack of the reliable software system.

Key Points

- miWords is a revolutionary algorithm based on Transformers and CNNs for pre-miRNA discovery in plants. miWords sees genome as a syntax and set of sentences made up of words of sequences and structure.
- miWords achieved never-seen-before performance and accuracy where it outperformed major pre-miRNA discovery tools by a huge margin under one of the most comprehensive benchmarking analyses done.
- For actual application like genomic annotation of pre-miRNAs, it attains very high accuracy where it can work with and without sRNA-sequencing data unlike most of the existing software, and yet outperformed them by leaps for both conditions.
- Its application across tea genome for pre-miRNA annotation strongly underlines the performance accuracy of miWords where three different experimental evidences (sRNA-seq, known miRNAs at miRBase and degradome-seq) supported its reported pre-miRNAs.
- The proposed approach is expected to remarkably change the scenario of the area of miRNA discovery and plant miRNA biology, which will also be democratized due to miWords ability to work with and without NGS experiment read data support.

SUPPLEMENTARY DATA

Supplementary data are available online at http://bib.oxfordjournals.org/.

DATA AVAILABILITY

All the secondary data used in the present study were publicly available and their due references and sources have been provided in Supplementary Tables S1–S4 available online at http://bib.oxfordjournals.org/. The software has also been made available at GitHub at https://github.com/SCBB-LAB/miWords as well as at the companion web page at https://scbb.ihbt.res.in/miWords/ (all related datasets in the study are hosted here).

AUTHORS' CONTRIBUTIONS

S.G. carried out the computational part and benchmarking of the study. R.S. conceptualized, designed, analyzed and supervised the entire study. S.G. and R.S. wrote the MS.

ACKNOWLEDGEMENTS

We are thankful to CSIR ANB for iPRESS. S.G. is thankful to CSIR and DBT, India for financial support as project associateship. This MS has CSIR-IHBT MSID 5137.

FUNDING

The work was carried out under the aegis of The Himalayan Centre for High-throughput Computational Biology (HiCHiCoB), a BIC supported by DBT, Govt. of India [BT/PR40122/BTIS/137/30/2021].

References

1. Xuan P, Guo M, Liu X, *et al.* PlantMiRNAPred: efficient classification of real and pseudo plant pre-miRNAs. *Bioinformatics* 2011;**27**:1368–76.
2. Taylor RS, Tarver JE, Foroozani A, *et al.* MicroRNA annotation of plant genomes—do it right or not at all. *Bioessays* 2017;**39**:1–6.
3. Bonnet E, Wuyts J, Rouzé P, *et al.* Detection of 91 potential conserved plant microRNAs in Arabidopsis thaliana and Oryza sativa identifies important target genes. *Proc Natl Acad Sci U S A* 2004;**101**:11511–6.
4. Jones-Rhoades MW, Bartel DP. Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. *Mol Cell* 2004;**14**:787–99.
5. Adai A, Johnson C, Mlotshwa S, *et al.* Computational prediction of miRNAs in Arabidopsis thaliana. *Genome Res* 2005;**15**:78–91.
6. Lindow M, Krogh A. Computational evidence for hundreds of non-conserved plant microRNAs. *BMC Genomics* 2005;**6**:119.
7. Bentwich I, Avniel A, Karov Y, *et al.* Identification of hundreds of conserved and nonconserved human microRNAs. *Nat Genet* 2005;**37**:766–70.
8. Xue C, Li F, He T, *et al.* Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinformatics* 2005;**6**:310.
9. Dezulian T, Remmert M, Palatnik JF, *et al.* Identification of plant microRNA homologs. *Bioinformatics* 2006;**22**:359–60.
10. Hertel J, Stadler PF. Hairpins in a haystack: recognizing microRNA precursors in comparative genomics data. *Bioinformatics* 2006;**22**:e197–202.
11. Ng KLS, Mishra SK. De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures. *Bioinformatics* 2007;**23**:1321–30.
12. Batuwita R, Palade V. microPred: effective classification of pre-miRNAs for human miRNA gene prediction. *Bioinformatics* 2009;**25**:989–95.
13. Hackenberg M, Sturm M, Langenberger D, *et al.* miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments. *Nucleic Acids Res* 2009;**37**:W68–76.
14. Mathelier A, Carbone A. MIReNA: finding microRNAs with high accuracy and no learning at genome scale and from deep sequencing data. *Bioinformatics* 2010;**26**:2226–34.
15. Yang X, Li L. miRDeep-P: a computational tool for analyzing the microRNA transcriptome in plants. *Bioinformatics* 2011;**27**:2614–5.
16. Jha A, Chauhan R, Mehra M, *et al.* miR-BAG: bagging based identification of MicroRNA precursors. *PloS One* 2012;**7**:e45782.
17. Friedländer MR, Mackowiak SD, Li N, *et al.* miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res* 2012;**40**:37–52.
18. An J, Lai J, Lehman ML, *et al.* miRDeep*: an integrated application tool for miRNA identification from RNA sequencing data. *Nucleic Acids Res* 2013;**41**:727–37.
19. Gudyś A, Szcześniak MW, Sikora M, *et al.* HuntMi: an efficient and taxon-specific approach in pre-miRNA identification. *BMC Bioinformatics* 2013;**14**:83.
20. Axtell MJ. ShortStack: comprehensive annotation and quantification of small RNA genes. *RNA* 2013;**19**:740–51.
21. Meng J, Liu D, Sun C, *et al.* Prediction of plant pre-microRNAs and their microRNAs in genome-scale sequences using structure-sequence features and support vector machine. *BMC Bioinformatics* 2014;**15**:423.
22. Lei J, Sun Y. miR-PREFeR: an accurate, fast and easy-to-use plant miRNA prediction tool using small RNA-Seq data. *Bioinformatics* 2014;**30**:2837–9.
23. Yao Y, Ma C, Deng H, *et al.* plantMirP: an efficient computational program for the prediction of plant pre-miRNA by incorporating knowledge-based energy features. *Mol Biosyst* 2016;**12**:3124–31.
24. Thomas J, Thomas S, Lee Sael. DP-miRNA: an improved prediction of precursor microRNA using deep learning model. *2017 IEEE International Conference on Big Data and Smart Computing (BigComp)* 2017; 96–99.
25. Stegmayer G, Yones C, Kamenetzky L, *et al.* High class-imbalance in pre-miRNA prediction: a novel approach based on deepSOM. *IEEE/ACM Trans Comput Biol Bioinform* 2017;**14**:1316–26.
26. Park S, Min S, Choi H-S, *et al.* Deep recurrent neural network-based identification of precursor microRNAs. *Adv Neural Inf Process Syst* 2017;**30**:2891–2900.
27. Yones C, Stegmayer G, Milone DH. Genome-wide pre-miRNA discovery from few labeled examples. *Bioinformatics* 2018;**34**:541–9.
28. Tang X, Sun Y. Fast and accurate microRNA search using CNN. *BMC Bioinformatics* 2019;**20**:646.
29. Yones C, Raad J, Bugnon LA, *et al.* High precision in microRNA prediction: a novel genome-wide approach with convolutional deep residual networks. *Comput Biol Med* 2021;**134**:104448.
30. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;**521**:436–44.
31. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Commun ACM* 2017;**60**:84–90.
32. Vieira JPA, Moura RS. *An analysis of convolutional neural networks for sentence classification. 2017 XLIII Latin American Computer Conference (CLEI)* **2017**; 1–5
33. Mandic DP, Chambers JA. *Recurrent Neural Networks for Prediction: learning algorithms, architectures and stability*. John Wiley & Sons, 2001. https://doi.org/10.1002/047084535.
34. Axtell MJ, Meyers BC. Revisiting criteria for plant MicroRNA annotation in the era of big data. *Plant Cell* 2018;**30**:272–84.
35. Vaswani A, Shazeer N, Parmar N, *et al.* Attention is all you need. *Adv Neural Inf Process Syst* 2017;**30**:5998–6008.
36. Raad J, Bugnon LA, Milone DH, *et al.* miRe2e: a full end-to-end deep model based on transformers for prediction of pre-miRNAs. *Bioinformatics* 2022;**38**:1191–7.
37. Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: from microRNA sequences to function. *Nucleic Acids Res* 2019;**47**:D155–62.
38. Chen C, Li J, Feng J, *et al.* sRNAanno—a database repository of uniformly annotated small RNAs in plants. *Hortic Res* 2021;**8**:45.
39. Guo Z, Kuang Z, Zhao Y, *et al.* PmiREN2.0: from data annotation to functional exploration of plant microRNAs. *Nucleic Acids Res* 2022;**50**:D1475–82.
40. Yi X, Zhang Z, Ling Y, *et al.* PNRD: a plant non-coding RNA database. *Nucleic Acids Res* 2015;**43**:D982–9.
41. Bugnon LA, Yones C, Milone DH, *et al.* Genome-wide discovery of pre-miRNAs: comparison of recent approaches based on machine learning. *Brief Bioinform* 2021;**22**:bbaa184.
42. Addo-Quaye C, Miller W, Axtell MJ. CleaveLand: a pipeline for using degradome data to find cleaved small RNA targets. *Bioinformatics* 2009;**25**:130–1.
43. Gahlan P, Singh HR, Shankar R, *et al.* De novo sequencing and characterization of Picrorhiza kurrooa transcriptome at two temperatures showed major transcriptome adjustments. *BMC Genomics* 2012;**13**:126.