CirRNAPL: A web server for the identification of circRNA based on extreme learning machine

Mengting Niu a,1, Jun Zhang b,1, Yanjuan Li c, Cankun Wang d, Zhaqian Liu d, Hui Ding e, Quan Zou a,e,*
Qin Ma d e

a Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu 610054, China
b Rehabilitation Department, Heilongjiang Province Land Reclamation Headquarters General Hospital, Harbin, China
c School of Information and Computer Engineering, Northeast Forestry University, Harbin 150040, China
d Department of Biomedical Informatics, Ohio State University, Columbus, OH, USA
e Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 610054, China

Abstract

Circular RNA (circRNA) plays an important role in the development of diseases, and it provides a novel idea for drug development. Accurate identification of circRNAs is important for a deeper understanding of their functions. In this study, we developed a new classifier, CirRNAPL, which extracts the features of nucleic acid composition and structure of the circRNA sequence and optimizes the extreme learning machine based on the particle swarm optimization algorithm. We compared CirRNAPL with existing methods, including blast, on three datasets and found CirRNAPL significantly improved the identification accuracy for the three datasets, with accuracies of 0.815, 0.802, and 0.782, respectively. Additionally, we performed sequence alignment on 564 sequences of the independent detection set of the third data set and analyzed the expression level of circRNAs. Results showed the expression level of the sequence is positively correlated with the abundance. A user-friendly CirRNAPL web server is freely available at http://server.malab.cn/CirRNAPL.

1. Introduction

Circular RNA (circRNA) is a newly identified RNA type that differs from conventional linear RNA in humans. It is a noncoding RNA molecule without having a 5-end cap or a 3-end tail, instead, forming a circular structure [1,2] (Fig. 1). CircRNA was first discovered in 1969 by Diener, while researching potato spindle tuber disease [3]. Electron microscopy revealed the formation of such closed-loop RNA, also known as a viroid. The subsequent emergence of high-throughput sequencing techniques (RNA-seq) enabled improved sequencing of circRNAs of various species, and many circRNAs have now been identified [4,5]. To date, more than 10,000 different circRNAs have been successfully identified from fruit flies and worms to mice and humans [6,7].

Against background researches and applications of circRNA, a series of databases have been built, such as circBase [8], circNet [9], circ2Traits [10], TSCD [11], circRNADb [12], and CirInterac-tome [13]. Studying the structure and function of circRNAs, researchers have revealed their importance in the pathogenesis of arteriosclerosis, nervous system disorders, diabetes, and tumors.
The unique endogenous characteristics of circRNAs enable their value as biomarkers, providing a new approach for drug development [6,17,18] and a new direction for the evolution of life. Xu used Qqiucomm sequencing to detect the expression profiles of three wheat samples (LH9, XN979, and YN29) and identified 33 differentially expressed circRNA [19]. Ye used published RNA-Seq data to perform genome-wide identification of circRNA in rice and Arabidopsis. Based on this, they compared the characteristics of plant and animal circRNA [20]. Moreover, Memczak analyzed circRNA from humans, mice, and nematodes, and detected thousands of well-represented, stable circRNAs, demonstrating their regulatory capabilities [21]. Given this landscape, accurate detection and quantification of circRNAs are essential for a deeper understanding of their function.

In addition to the abundance of RNA-seq data, a variety of algorithms for calculating and visualizing circRNA have recently been developed [2,17,22]. Li Chen developed the CIRCexplorer software based on RNA-seq data of Arabidopsis and rice, providing more comprehensive and sensitive circRNA prediction method for plants [23]. Song et al. developed a computational pipeline, called UROBORUS, to detect circRNA in RNA-seq data. They successfully verified 24 circRNA from 27 randomly selected circRNA. Additionally, based on RNA-seq, Danan developed a circRNA-seq method for reading circRNA in an unbiased, genome-wide manner. Moreover, they mapped the transcriptome of solfataricus [24]. In another study, Jeck et al. used high-throughput sequencing of ribosomal-deficient RNA and identified more than 25,000 loops in human fibroblasts containing non-collinear exons (“reverse splicing”) [25]. Gao et al. proposed a new algorithm, CIRI, based on cross-shear signal, which uses a variety of filtering strategies to accurately detect circRNA from transcriptome data [26]. Zhang et al. used non-polyadenylation and raser-treated RNA-seq data from H9 human embryonic stem cells to predict reverse splicing junctions, and they systematically characterized circRNA using the newly developed pipeline CIRCexplorer [27]. Furthermore, Vo et al. used the exome capture RNA sequencing protocol to detect and characterize circRNA across >2000 cancer samples, establishing the most comprehensive catalog of circRNA species to date (MiOncoCirc) [28]. You et al. proposed the method, Acfs, allowing for the redefinition and accurate and rapid identification of circRNA from single-ended and double-ended RNA-seq data, as well as their abundance [29]. Furthermore, Zhang et al. sequenced Branchiostoma belcheri circRNA and identified 1859 circRNAs using the find_circ and CIRI algorithms [30]. At present, many available recognition methods are based on RNA-Seq data, and it is very important to use bioinformatics with direct circRNA sequence training to achieve more accurate recognition. In contrast to the tools that use RNA-seq data as input, Pan et al. extracted the characteristics of the sequence and used the random forest (RF) algorithm to identify circRNAs, subsequently building the online identification server WebCircRNA [31]. PredcircRNA classified circRNA and IncRNA based on a computational method of a multi-core learning framework with multiple functional training [32]. H-ELM used a hierarchical extreme learning machine algorithm with feature selection to extract the same features from other Incrna and classify circular RNAs [33]. Mohamed Chaabane proposed an end-to-end deep learning framework circDeep, which integrates shared representations across different modes and improves circularRNA for classification [34].

Many studies for protein recognition [35] and site detection [36] have been performed based on machine learning, such as RF [37] and Artificial Neural Network [38]. By contrast, few studies focus on the identification of circRNAs. Therefore, there is a need for studies on how to use sequence information to achieve more accurate identification by using the characteristics of RNA sequences. We propose a new method that mainly uses the structural features of RNA and nucleotide composition to optimize the extreme learning machine (ELM) based on the particle swarm optimization algorithm (PSO). Based on these, a classification system was built to optimize the effect of circRNA identification.

2. Material and methods

In this study, we used the ELM method to identify circRNA. The flowchart of the identification framework is shown in Fig. 2.

2.1. Datasets

circRNAs are evolutionarily conserved among different species, and the detection of circular RNAs is important for further understanding the biological origin and purpose of circular RNAs [39,40]. The datasets that we used were presented in the literature of Pan et al. [31] (Table 1). The literature downloaded 92,375 circRNA transcripts from circBase. After deleting transcripts of less than 200nt in length and the overlapping circular RNA transcripts, 14,084 circRNA data were obtained. The database source of the 9533 PCGs and 19,723 IncRNAs dataset is GENCODEv19, in which overlapping circRNA sequences were removed from the PCG data. The third dataset includes 2082 circRNAs from stem cells expressed in H1hsec and the same number of identically derived circRNAs from other cells that were not expressed in H1hsec. And the circRNA vs IncRNA dataset also applied to the literature of Mohamed et al. [34], Pan et al. [32] and Chen et al. [33].

The construction of the classifier used in this study involved ten-fold cross-validation. To verify the classifier, independent training and test set verification was also performed. For the dataset of 14,084 circRNAs, the number used for the training set was 10,000, and that for the test set was 4084. For the IncRNA dataset of 19,722, the number for the training set was 10,000, and that for the test set was 9722. For the PCG training set of 9533, the number for the training set was 8000, and that for the test set was 1533. Finally, for the two datasets of 2082 circRNAs of each of stem cells and other cells, the training set consisted of 1800 circRNAs, and the test set consisted of the remaining 282 circRNAs.

2.2. Feature representation methods

To identify circRNAs, four features of the sequence data were extracted, including Ribonucleic acid composition, Autocorrelation, Pseudo-ribonucleic acid composition, and Predicted structure composition [41]. These four features comprised a total of 15 modes, wherein Ribonucleic acid composition included k-mer (the param-
eter k is 2 and 3), Mismatch, and Subsequence; Autocorrelation included DAC, DCC, DACC, MAC, GAC, and NMBAC; Pseudo-nucleotide composition included General parallel correlation pseudo-dinucleotide composition (PC-PseDNC-General) and General series correlation pseudo-dinucleotide composition (SC-PseDNC-General); and Predicted structure composition included Local structure-sequence triplet element (Triplet), PseSSC Pseudo-structure status, and PseDPC. A total of 520-dimensional features were obtained through these 15 modes (Fig. 3).

2.2.1. Ribonucleic acid composition

For the nucleic acid composition feature of RNA sequences, three features, including Basic k-mer, Mismatch, and Subsequence, were used, of which k-mer [42] is the simplest method for expressing RNA. Basic k-mer can then be used to calculate the frequency of occurrence of k adjacent nucleic acids. Mismatch [43, 44] can calculate the frequency of k-length adjacent nucleic acids, which differ by up to m mismatches (m < k); Subsequence is a method that allows discontinuous matching [45].

Suppose an RNA sequence R is as follows:

$$R = R_1R_2R_3 \cdots R_n$$

(1)

wherein R1 represents the first nucleic acid in R, R2 represents the second nucleic acid in R, and so on.

Table 1

| Datasets | Positive Data | Negative Data |
|----------|---------------|---------------|
| circRNA vs PCG | 14,084 circRNAs | 9,533 PCGs |
| circRNA vs IncRNA | 14,084 circRNAs | 19,722 IncRNAs |
| Stem cell vs not | 2082 circRNAs | 2082 circRNAs |

Fig. 2. Flowchart of CirRNAPL. CirRNAPL identifies circRNA in four main steps: data, feature, classifier, and circRNA identification. Data: This involves dataset construction. According to the bed data file and the hg38 human genome, we write Python script files to extract the corresponding sequence data. Features: This involves the extraction of features. In this work, information such as the structural characteristics of the RNA sequence is used as the feature to be extracted, including four parts, with a total of 14 calculation models. Classification: This involves the construction of the classifier. Here, the extreme learning machine based on particle swarm optimization is used as the classification algorithm. The classifier CirRNAPL is constructed by a tenfold cross-validation method, and the final classification model is output. CircRNA identification: The RNA sequence to be labeled is identified using the classifier CirRNAPL.

Fig. 3. Feature expression method and feature dimension histogram.
L = \left( \sum_{a_1} \phi^{(1)}(a_1), \sum_{a_2} \phi^{(2)}(a_2), \ldots, \sum_{a_l} \phi^{(l)}(a_l) \right)

(3)

When \( a_i \) is exact matching, \( s(a_i) = 0 \); when \( a_i \) is non-contiguous matching, \( s(a_i) = |a_i| \). Here, \( \phi \) represents the attenuation coefficient, \( \phi \in [0, 1], j = 1, 2, \ldots, 4^k \).

2.2.2. Autocorrelation

For the autocorrelation feature of RNA sequences, Dinucleotide-based auto-covariance (DAC) [46], Dinucleotide-based cross-covariance (DCC), Dinucleotide-based auto-cross-covariance (DACC) [47], Moran autocorrelation (MAC), Geary autocorrelation (GAC), and Normalized Moreau–Broto autocorrelation (NMBAC) are used [48]. DAC measures the correlation of the same chemical index between two dinucleotides along the sequence interval \( l \). DCC measures the correlation of two different physico-chemical indicators between two dinucleotides separated by 1 nucleic acids. MAC measures the correlation of the same property between two residues along with sequence interval \( l \). GAC measures the correlation of the same properties between two residues of distance \( l \). NMBAC measures the correlation of the same properties between two residues at a distance \( l \).

2.2.3. Pseudo-ribonucleic acid composition

For the Pseudo-ribo nucleic acid composition feature of RNA sequences, PC-PseDNC-General and SC-PseDNC-General are used. In the PC-PNC-General method, the user can select not only 22 built-in physical and chemical indicators, but also upload their own indicators to generate PC-PNC-General feature vectors. SC-PseDNC-General is a variant of PC-PseDNC-General [49].

2.2.4. Predicted structure composition

For the Predicted structure composition feature of RNA sequences, Triplet, Pseudo-structure status composition (PsessSC), and Pseudo-distance structure status pair composition (PseDPC) are used. Triplet (24) is an early method based on RNA sequence-structure information, and it shows better circRNA identification performance than other sequence-based methods [50].

2.3. Extreme machine learning

To identify circRNAs, the ELM is used here as the basic classification algorithm. It has also been applied to biomeric identification [40, 51–55].

ELM is a generalized single hidden layer feedforward network. This algorithm randomly assigns input weights and hidden layer thresholds and directly calculates output layer weights by least squares. The entire learning process is completed once, and no iteration is required. Therefore, the algorithm learns rapidly.

2.4. Particle swarm optimization algorithm

Particle swarm optimization (PSO) is often used to neural network optimization due to its simple rules, rapid convergence, less parameter adjustability, and strong ability for search [56, 57]. PSO is based on the behavior of group foraging, where the particles are the solution to be optimized [58]. PSO is used to optimize the input weight and hidden layer deviation of ELMs, which can improve the generalizability of the methods. The particle swarm ELM algorithm relies on fewer hidden layer nodes to achieve higher precision.

2.5. ELM of PSO optimization

The kernel function of the extreme learning machine has a significant influence on the performance of the algorithm. The kernel parameter \( \sigma \) and penalty coefficient \( C \) in the kernel function have an important impact on the performance of the ELM. \( \sigma \) affects the scope of the kernel function, and \( C \) affects the stability of the model. The paper used PSO to optimize the parameters \( \sigma \) and \( C \). The search space of the PSO corresponds to the parameters of the ELM. The position of the particles represents the parameter value, and the accuracy is used as the fitness function of PSO. The steps of the PSO optimization ELM are as follows.

1. Initialization. The number of iterations and the overall size were set to 50 and 50, respectively. A particle population was randomly generated. Each particle in the population consists of a set of \( \sigma \) and \( C \).
2. Calculation of fitness function value. The fitness function value of the PSO algorithm is the accuracy of the ELM.
3. Updating the best position and particle position of the population, as well as the speed and position of the particles according to the formula (4) and (5) [58].

\[ v_i(t + 1) = \omega \cdot v_i(t) + c_1 R_1 [p_{best,i} - p_i(t)] + c_2 R_2 [G_{best} - p_i(t)] \]  
\[ p_i(t + 1) = p_i(t) + v_i(t + 1) \]  

where \( p_i(t) \) and \( v_i(t) \) are the position and velocity of the \( t \)-th iteration of particle, respectively; \( p_{best,i} \) is the optimal solution for particle \( i \); \( G_{best} \) is the optimal solution for the population; \( \omega \) is the weight; \( c1 \) and \( c2 \) are the acceleration factors; \( R1 \) and \( R2 \) are random numbers between 0 and 1; \( t \) is the number of iterations.

4. Checking termination conditions. If the maximum fitness value or the maximum number of iterations is reached, go to step (5). Otherwise, return to step (2).
5. Obtaining the optimal ELM parameter with the largest fitness value. In this way, the parameters of the ELM model can be obtained, and the circRNA is identified using the optimized ELM.

2.6. Performance measurement

To evaluate the identification effect of the constructed classifier, four commonly used indicators were selected here: SE (Sensitivity), SP (Specificity), ACC (Accuracy), and MCC (Matthews Correlation Coefficient). SE indicates the rate of correct prediction of positive sequences. SP indicates the rate of correct prediction of the counterexample. ACC indicates the correct rate of classification. MCC indicates the reliability of the classifier, which can reflect the prediction ability more fairly. A larger MCC reflects better reliability.

\[ SE = \frac{TP}{TP + FN} \]  
\[ SP = \frac{TN}{TN + FP} \]  
\[ ACC = \frac{TN + TP}{TN + FP + TP + FN} \]  
\[ MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \] 

Here, TP indicates the number of circRNAs predicted correctly; FP indicates the number of non-circRNAs predicted correctly; TN indi-
cates the number of incorrectly predicted circRNAs; FN indicates the number of incorrectly predicted non-circRNAs.

3. Results

3.1. Feature importance analysis

Taking into account the efficiency of the classifier operation, MRMD (Maximum-Relevance-Maximum-Distance) was used for feature selection and the output of feature score sorting results [59]. We analyzed the top 20 features based on the output feature scores (Fig. 4(A)), and get the 20-dimensional feature set obtained by feature extraction. On the circRNA vs PCG dataset, the first dimension is F496, and the 17-dimensional distribution is around F200. On the CircRNA vs IncRNA dataset, the first dimension feature is F404, and most are located between F200 and F300-F450. For the last dataset, the first dimension is characterized by F98 and F404 is the second dimension. Most of the feature distributions are the same as the second one. Moreover, we can get the common points of the 20-dimensional feature distribution on the three data sets. That is, most of the features are around F200 and F356. Finally, we analyzed which features are important for the identification of circRNA based on feature extraction.

To analyse the distribution of local features, we divided the 520 features into four intervals ([1,148), [148,318), [318,364), and [364,520]) according to the order and results of using the feature expression method. The distribution of the top 20 features in each interval is given in Fig. 4(B). According to the top 20 features of the score and the interval map of the distribution, four falls into the interval [148,318), seven into [318,364), and nine into [364,520], in which the scores in the intervals [318,364) and [364,520] were also ranked higher. This shows that the structural features of RNA and Pseudo-ribonucleic acid composition are most important, and the proportions of Autocorrelation and Ribonucleic acid composition are relatively less important. Therefore, we concluded that the structural features and Pseudo-ribonucleic acid composition contain more feature information than the other two feature regions.

3.2. Optimization of kernel function of ELM

For ELM with excellent classification performance, it is important to choose an appropriate and stable hidden layer activation function. This section presents a discussion of the different classification performances of different activation functions and how to find the optimal activation function. For optimization of the kernel function, five common activation functions are selected here: sigmoid, sine, hardlim, tribas, and radial basis function (RBF) [60].

Five activation functions were used to the three datasets, respectively. The classification results are evaluated based on four indicators: ACC, SE, SP, and MCC. The recognition results were obtained under tenfold cross-validation (Fig. 5(A)). We found that the values of MCC, SP, and SE of RBF are not the best for the dataset of circRNAs expressed and not expressed in stem cells. However, the effect is better than the other four functions for the other two datasets. The results for the three datasets are as follows: accuracy of 0.784, 0.794, and 0.749; SE of 0.762, 0.679, and 0.749; and SP of 0.75, 0.839, and 0.739, respectively. In general, the RBF function achieves the best recognition of the overall effect and proves the validity of RBF for recognizing circRNA. Therefore, RBF was selected as the activation function of ELM.

To further prove the stability of the classification effect, the activation function was verified on an independent test set. The results are shown in Fig. 5(B). According to the trend of the fold line, the RBF function is at a higher position for the three datasets in terms of accuracy, and the accuracy of the independent test set is higher than the ten-fold cross-validation effect. This also proves that the recognition effect of RBF has certain stability and effectiveness. Therefore, in the following experiments, RBF is used as an activation function of the hidden layer.

3.3. Optimization of ELM using PSO algorithm

The traditional ELM lacks effective training methods, and it has the disadvantage of poor prediction accuracy [61]. In contrast, PSO, as an intelligent optimization algorithm, can improve the performance of ELM. Here, the PSO algorithm was used to optimize ELM and it was compared with basic ELM results and the optimization effect of genetic algorithm [62]. The results of experiments of tenfold cross-validation show that, compared with the recognition results of the optimized ELM and the basic ELM algorithm, GA-ELM and CirRNAPL improved to some extent (Fig. 5(C)). In terms of the classification effect, ELM and CirRNAPL achieved better results than GA-ELM. On the three datasets, CirRNAPL achieved accurate ACC values of 0.815, 0.822, and 0.782. Experiments showed that

Fig. 4. Feature importance analysis. A) Top 20-dimensional feature distribution on three data sets. B) Feature Importance Analysis: On the three data sets, 520-dimensional features were obtained by feature selection. The 520-dimensional feature distribution was organized.
the PSO effectively improved the prediction accuracy and generalizability of the ELM network. Therefore, the improved ELM was used as a classification algorithm to identify circRNA.

3.4. Comparison with other classifiers

To verify the validity of the CirRNAPL, the classifier was compared to other classification algorithms. Based on Section 3.2 and Section 3.3, RBF is the selected activation function and the ELM was improved using the PSO algorithm. Next, the classification effect of the improved ELM algorithm was compared with those commonly used algorithms, such as CNN (Convolutional Neural Networks) [63], RF, support vector machine (SVM), J48 [64], and ZeroR [65]. The comparison results are shown in Fig. 5(D).

As shown in Fig. 5(D), from the comparison of the results of ACC, SE, SP, and MCC, compared with the CNN, RF, SVM, J48, and ZeroR algorithms, the classifier CirRNAPL constructed here achieved good results. For the three datasets, CirRNAPL achieves recognition accuracy of 0.815, 0.822, and 0.782, and also demonstrates the effectiveness of PSO-ELM for identifying circRNA.

After verifying the classifier CirRNAPL under ten fold cross-validation, to further explain the effect, CirRNAPL was further validated here on an independent test set. The results of the independent test set verification are shown in Fig. 5(E). According to the accuracy values of the five algorithms on the three datasets, the classifier CirRNAPL constructed in this study achieved a high accuracy rate. The accuracy rates for the three datasets are 0.887, 0.906, and 0.841, and the accuracy of the independent test set is higher than the accuracy of the tenfold cross-validation. The findings also prove that the classification accuracy of the tenfold cross-validated classifier is reliable and stable.

3.5. Comparison of the results with the traditional blast method

This section mainly describes a comparison of the recognition effects of the constructed classifier CirRNAPL and the traditional blast sequence alignment. In biological research, if a previously unknown gene is found, then a standard method, such as the base alignment search tool blast, is applied to obtain useful information through comparative analysis. With the development of machine learning algorithms and information technology, an increasing abundance of methods is becoming available for applying machine learning algorithms to identify unknown genes. To combine the effect of the traditional blast and the identification effect of the machine learning method, we compared the identification effect between CirRNAPL and blast. When performing blast, all sequence files are database files queried by blast, and each sequence is used as a query file in turn. Each sequence can get a blast result file, and the results are arranged in descending order of identity/length. Then, the prediction category of the query sequence is the category corresponding to the sequence with the largest identity/length value. The final result is shown in Fig. 5(F). We found that the classification accuracy of blast is clearly lower than that of CirRNAPL. The recognition accuracy of blast is 0.439, 0.605, and 0.611, while the classification accuracy of CirRNAPL is 0.815, 0.802, and 0.782, respectively. Given that BLAST only compares certain keywords that are more or less important in the sequence, it is not surprising that the accuracy is slightly lower. Thus, there is no doubt that the sequence data-based CirRNAPL classification methods will have increasingly broad validity and usability in research.

3.6. Comparison with state-of-the-art methods

To test the effectiveness of the identification of the proposed classifier CirRNAPL, we compared its identification performance...
with that of existing classifiers: WebCircRNA, PredcircRNA, H-ELM, circDeep (Fig. 6(A)). Among these models, the literature [31] only used SE and SP indicators among the four commonly used evaluation indicators. PredcircRNA, H-ELM, and circDeep only used the circRNA vs IncRNA dataset.

Firstly, we compared the results of CirRNAPL with WebCircRNA. From Fig. 6(A), we found that cirRNAPL achieved better performance on the “stem cell vs not” and “circRNA vs PCG” datasets than WebCircRNA, and in contrast, the performance on “circRNA vs lncRNA” was slightly worse. We analyzed the causes of the performance on the “circRNA vs IncRNA” data set from the perspective of features. According to the score generated after feature selection, the paper selected the two-dimensional features that effect is better in the results of each feature expression method and used violin plots to show the data distribution on the positive and negative example data sets (Fig. 6(B)). From the Fig. 6(B), we can see...
that the difference of feature distribution is relatively small between positive and negative examples, showing that the feature representation method we used cannot represent the characteristics of circRNA and lncRNA well. This result indicates that the feature representation method, which can make a significant difference in the distribution of positive and negative samples, would greatly contribute to the final classification result. It brings new challenges on how to use more efficient features and improve the performance of the classifier.

Secondly, we compared the results of CirRNAPL with PredcircRNA, H-ELM, and circDeep, respectively. From Fig. 6(A), we can know that the performance of cirRNAPL is better than PredcircRNA and H-ELM in the three indicators of ACC, SE, and MCC, while lower than the latest circDeep. The ACC value of the circDeep method is higher than cirRNAPL. By comparison, we can know that although cirRNAPL has achieved some progress, more advanced techniques are required for better identification results in future research.

3.7. Blast sequence alignment expression analysis based on RNA-seq data

After the prediction of circular RNA, we further consider whether it is possible to analyze the expression of circular RNA based on the available dataset. We explore the relationship between the level of sequence expression and the abundance of the sequence (the number of reads on the alignment). So, we downloaded the RNA-seq data from the GEO database (https://www.ncbi.nlm.nih.gov/geo), which is used for similarity detection. The independent test set of the third dataset in the paper was used as a query sequence and included 282 positive samples that were expressed as H1hesc in the stem cell and 282 negative samples that were not in the stem cell and were not expressed as H1hesc.

Similarity sequence alignment was then performed by local blast, with different E values, including 1.0, 0.1, 0.01, 1e–3, and 1e–4, respectively. For the two groups of 282 query sequences, we recorded the values of different sequence alignments under different E-values (Fig. 7(A)). When the E-value was at least 1e–3, the numbers of sequence alignments on the two datasets were 64 and 69. In the case of the stem and no stem, the changing trend of e value and expression level is the same. It can be seen that the influence. Then we calculated the abundance of each circRNA and then obtained a scatter plot of the relationship between expression and abundance (Fig. 7(B)). The scatter plot implicates that the abundance of the sequence is high when the expression level of the sequence is high. Then, the polynomial fitting of the obtained data points was carried out, and two fitting curves were obtained. According to the trend analysis of the fitted curve of Fig. 7(B), the relationship between the expression level of circRNA and the sequence abundance is the same in both cases: the expression level increases with the increase of abundance. The correlation coefficients in the two cases were 0.840 and 0.899, respectively, and the correlation of not stem was higher. We can also see that there are more sequence abundances on the not stem data. When the expression reaches a certain level, it does not change even though the sequence abundance increases. It shows that the expression level and abundance of the sequence are positively correlated.

After analyzing the expression level, we randomly selected 8 sequences for multiple sequence alignment. The multi-sequence alignment of the circRNA was obtained by the Clustal Omega online alignment tool, and some of the conserved regions were displayed (Fig. 7(C)). The results showed there are different colorations of nucleic acids in conserved regions while the consensus logo in the conservative region [66]. The nucleic acid distribution characteristics of the conserved regions of the circRNA were revealed by the displayed logo map.

4. Conclusion

We constructed a new classifier for the accuracy of identification of circRNA, and the effectiveness of the identification of circRNA has been demonstrated based on three publicly available datasets. Through the analysis of feature importance, we found that the structural features and Pseudo-ribonucleic acid composition feature showed better performance. However, the performance to distinguish circRNA and lncRNA is not very good, and it is still a widespread problem to distinguish them. More effective features and data are urgently required. Additionally, sequence alignments of circRNAs were preferably analyzed based on RNA-seq data. The comparison with the circDeep method also prompted us to use more efficient and updated techniques to improve the performance of our model.

Planned future work includes not only finding more useful features but also using advanced parallel technology to identify the growing number of circRNA sequences. Both of them will improve the efficiency of identification. Meanwhile, given that the functions of most circRNAs are still unknown, and little work has been performed on the large-scale discovery of disease-related circRNAs, future studies should focus on the biological significance of circRNAs.

5. Data accession numbers

The RNA-seq data is from the GEO database (https://www.ncbi.nlm.nih.gov/geo). The GEO number is GSE63823.

CRediT authorship contribution statement

Mengting Niu: Writing - original draft, Methodology. Jun Zhang: Conceptualization, Software. Yanjuan Li: Project administration. Cankun Wang: Formal analysis, Validation. Zhaqiaoqian Liu: Writing - review & editing. Hui Ding: Data curation. Quan Zou: Funding acquisition. Qin Ma: Supervision.

Acknowledgements

We thank all members in the Zou lab for helping in paper writing and server support. We also thank Marlena Merling for language help.

Funding

The work was supported by the National Key R&D Program of China, China (2018YFC0910405), and the Natural Science Foundation of China, China (No. 61771331, 61300098). The work was supported by the Fundamental Research Funds for the Central Universities, China (No. 2572017CB33).

References

[1] Hao S, Lv J, Yang Q, Wang A, Li Z, Guo Y, et al. Identification of key genes and circular RNAs in human gastric cancer. J Med Sci Monitor. 2019;25:2488–504.
[2] Bogard B, Francastel C, Hubé F. A new method for the identification of thousands of circular RNAs. J Non-coding RNA Investigation 2018:2.
[3] Diener T. Spindle tuber ‘virus’: IV. A replicating, low molecular weight RNA. J Virol. 1971:45:411–28.
[4] Cheng J, Metge F, Dieterich C. Specific identification and quantification of circular RNAs from sequencing data. J Bioinformatics 2015:32:1094–6.
[5] Dori M, Ailooi HHA, Cavalli D, Massalin S, Lesche M, Dahl A, et al. Sequence and expression levels of circular RNAs in progenitor cell types during mouse corticogenesis. J Life Sci Alliance 2019:2.e201900354.
[6] Song X, Zhang N, Han F, Moon B-S, Lai RK, Wang K, et al. Circular RNA profile in gliomas revealed by identification tool UROBORUS. J Nucleic Acids Res 2016:44:e87.
[7] Li S, Han L. Circular RNAs as promising biomarkers in cancer: detection, function, and beyond. J Genome Med. 2019;11:1:15.
Liu Y-C, Li J-R, Sun C-H, Andrews E, Chao R-F, Lin F-M, et al. CircNet: a database for circular RNAs. J RNA 2014;20:1666–70.

Liu Y-C, Li J-R, Sun C-H, Andrews E, Chao R-F, Lin F-M, et al. CircNet: a database of circular RNAs derived from transcriptomic sequencing data. J Nucleic Acids Res 2015;44:D209–15.

Ghosal S, Das S, Sen R, Basak P, Chakrabarti J. CircTraits: A comprehensive database for circular RNA potentially associated with disease and traits. J Front Gen 2013;4:263.

Xia S, Feng J, Lei L, Hu J, Xia L, Wang J, et al. Comprehensive characterization of tissue-specific circular RNAs in the human and mouse genomes. J Bioinforms Bioinformatics 2016;18:984–92.

Chen X, Han P, Zhou T, Guo X, Song X, Li Y. circRNAdb: a comprehensive database for human circular RNAs with protein-coding annotations. J Sci Rep 2016;6:34085.

Dudekula DB, Panda AC, Grammatikakis I, De S, Abdelmohsen K, Gorospe M. Pse-in-One: a web server for predicting 3’ body parts of circular RNA. J Proteomics Bioinformatics 2017;10:33–42.

Luo L, Li D, Zhang W, Tu S, Zhu X, Tian G. Accurate prediction of transposon-derived piRNAs by integrating various sequential and physicochemical features. J PloS One 2016;11:e0153268.

Zhang W, Niu Y. Predicting flexible length linear b-cell epitopes using pairwise sequence similarity. In: 2010 3rd International Conference on Bioomedical Engineering and Informatics, Volume 6. IEEE; 2010. p. 2338–42.

Lodhi H, Saunders C, Shawe-Taylor J, Cristianini N, Watkins C. Text classification using string kernels. J Machine Learning Res 2002;2:419–44.

Dong Q, Zhou S, Guan J. A new taxonomy-based protein fold recognition approach based on autoencoder-covariance transformation. J Bioinformatics Bioinformatics 2019;10:25–65.

Friedel M, Nikolajewska S, Sühnel J, Wilhelm T. DvirDB: a database for dinucleotide properties. J Nucleic Acids Res 2008;37:D37–40.

Chen W, Zhang X, Brooker J, Lin H, Zhang L, Chou K-C. PseKNC-General: a cross-platform package for generating various modes of pseudo nucleotide compositions. J Bioinformatics Bioinformatics 2014;31:119–20.

Chen W, Feng P-M, Lin H, Chou K-C. iSPort-PseDNC: identiﬁcation recombination spots with pseudo dinucleotide composition. J Nucleic Acids Res 2013;41:e68.

Luo L, Li D, Zhang W, Tu S, Zhu X, Tian G. Accurate prediction of transposon-derived piRNAs by integrating various sequential and physicochemical features. J PloS One 2016;11:e0153268.