Computational Biology and Chemistry 61 (2016) 221–225

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Computational identification of circular RNAs based on conformational and thermodynamic properties in the flanking introns

Ze Liu[a], Jiuqiang Han[a,*], Hongqiang Lv[a], Jun Liu[a,b], Ruiling Liu[a]

[a] School of Electronics and Information Engineering, Xi'an jiaotong University, Xi'an 710049, PR China
[b] School of Electrical Engineering, Xi'an Jiaotong University, Xi'an 710049, PR China

ARTICLE INFO

Article history:
Received 4 August 2015
Received in revised form 3 February 2016
Accepted 3 February 2016
Available online 5 February 2016

Keywords:
Circular RNA
Competing endogenous RNA
Support vector machine
10-Fold cross-validation

ABSTRACT

Circular RNAs (circRNAs) were found more than 30 years ago, but have been treated as molecular flukes in a long time. Combining deep sequencing studies with bioinformatics technique, thousands of endogenous circRNAs have been found in mammalian cells, and some researchers have proved that several circRNAs act as competing endogenous RNAs (ceRNAs) to regulate gene expression. However, the mechanism by which the precursor mRNA to be transformed into a circular RNA or a linear mRNA is largely unknown. In this paper, we attempted to bioinformatically identify shared genomic features that might further elucidate the mechanism of formation and proposed a SVM-based model to distinguish circRNAs from non-circularized, expressed exons. Firstly, conformational and thermodynamic dinucleotide properties in the flanking introns were extracted as potential features. Secondly, two feature selection methods were applied to gain the optimal feature subset. Our 10-fold cross-validation results showed that the model can be used to distinguish circRNAs from non-circularized, expressed exons with an Sn of 0.884, Sp of 0.900, ACC of 0.892, MCC of 0.784, respectively. The identification results suggest that conformational and thermodynamic properties in the flanking introns are closely related to the formation of circRNAs. Datasets and the tool involved in this paper are all available at https://sourceforge.net/projects/predicircrnatool/files/.

1. Introduction

Circular RNAs (circRNAs) were first discovered in the cytoplasm of HeLa cells (Hsu and Coca-Prados, 1979). Since then, circRNA expressions at low levels have been found in a few expressed mammalian genes, such as ETS-1 in human cells and cytochrome P450 2C24 (CYPIIC24) in mouse cells (Cocquerelle et al., 1993 Zaphiropoulos, 1993). As only a small number of circRNAs were found, circRNAs were considered as molecular flukes with uncertain importance before the era of massively parallel sequencing. However, with the development of next-generation sequencing, thousands of circRNAs were found in different species and tissues (Glazar et al., 2014). Most of the circRNAs contain almost exclusively exonic sequences, located in the cytoplasm with high conservation and are more stable than linear RNAs (Salzman et al., 2012; Jeck et al., 2013). Besides, the expression levels of circRNAs are almost the same as linear RNAs and some can even exceed 10-fold of those associated linear transcripts (Salzman et al., 2012).

Although the functions of circRNAs in biological processes are largely unknown, researchers have found that the transcript antisence to the cerebellar degeneration-related protein 1 (CDR1as) is highly expressed in the cytoplasmic of nerve tissues, which contains about 70 miRNAs response elements (MRE) and can be interacted with miR-7 as miRNA sponges (Memczak et al., 2013; Hansen et al., 2013). What's more, circRNA expressions in human blood were also observed in the latest research, and the results suggest that circRNAs could be used as biomarker molecules in standard clinical blood samples (Memczak et al., 2015).

In order to elucidate the mechanism of formation, Jeck et al. (2013) investigated cis-sequence elements in the flanking introns and found that circRNAs are more likely to be generated by longer than average exons, flanked by longer than average introns that contain complementary ALU repeats, and bounded by a GT-AG pair of canonical splice sites. By using extensive mutagenesis of expression plasmids, Liang and Wilusz (2014) found that the intronic repeats and exonic sequences must collaborate with one another, and a functional 3′ end processing signal should be

* Corresponding author at: Room 163, 2nd East Building, 28 West Xianning Road, Xi'an, Shaanxi 710049, China. Fax: +86 29 82668665 181.
E-mail address: jqhan@mail.xjtu.edu.cn (J. Han).

http://dx.doi.org/10.1016/j.compbiolchem.2016.02.003
1476-9271/

required . However, the CisFinder program (Sharov and Ko, 2009) used by Jeck et al. (2013) to identify enriched motifs in the flanking introns needs a motif clustering step to remove redundant motifs, and cannot be restricted to find short, core motifs. And only three genes, ZKSCAN1, HIPK3, and EPHB4, were analyzed in the research of Liang and Wilusz (2014), which is hard to build a common model by using three genes only. Thus, the mechanism of formation is still largely unclear.

In this paper, we attempted to bioinformatically identify shared genomic features that might further elucidate the mechanism of formation and proposed a machine learning method to distinguish circRNAs from non-circularized, expressed exons. Firstly, we extracted different length of sequences in the flanking introns and calculated the thermodynamic and conformational dinucleotide properties as the original features. And then two feature selection methods, Minimum Redundancy and Maximum Relevance (mRMR) and Random Forest (RF), were used to construct the optimized feature subset. For predictions, the performance of three algorithms, Support Vector Machine (SVM), Artificial Neural Network (ANN) and RF were compared on the training dataset and the independent dataset, respectively. Our results suggest that conformational and thermodynamic properties in the flanking introns are closely related to the formation of circRNAs.

2. Materials and methods

2.1. Datasets

The circRNA dataset proposed in PredcircRNA (Pan and Xiong, 2015) was used as the positive dataset. This dataset contains 14,084 circRNAs, all of which are longer than 200 nt and non-overlapped from the same gene. The HEXEvent (Busch and Hertel, 2013) provides all the splice events based on EST information from the UCSC Genome Browser. In this research, only constitutive exons, showing non-circularized, were selected from HEXEvent as negative samples. In this way, 139,180 constitutive exons, showing non-circularized, were collected as negative samples. We randomly selected 19,022 exons from the negative samples and removed the overlapped samples with the circRNA dataset. Thus, the whole datasets contain 14,084 circRNAs and 18,970 non-circularized exons. To avoid overfitting, we split the whole datasets into three parts: 3000 circRNAs and 3000 non-circularized exons were randomly selected for feature selection; 7000 circRNAs and 7000 non-circularized exons were randomly selected for model training; the remaining 4084 circRNAs and 8970 non-circularized exons were used to construct the independent testing dataset.

2.2. Feature extraction

Researchers have found that specific elements, such as ALU elements in the flanking introns, may be required for circularization (Jeck et al., 2013). And in order to allow sufficient time for a backsplice to occur, the intron immediately preceding the circularizing exons likely must be spliced slower than the average intron (Liang and Wilusz, 2014). These results suggest that nucleic acid properties in the flanking introns may be very important for the formation of circRNAs. In this research, the DiProDB database, collects conformational and thermodynamic dinucleotide properties, was used to calculate the nucleic acid properties in the flanking introns (Friedel et al., 2009). As sequences in the flanking introns might collaborate with each other for the formation of circRNAs, we extracted different length of sequences in the flanking introns and combined each of them to comprehensive analysis the corresponding nucleic acid properties. Thus, a total of 125 features were extracted in the original feature space.

2.3. Feature selection

Feature selection can remove irrelevant or redundant characteristics, so as to reduce the running time of training process and to improve the accuracy of model. On the other hand, the selection of true relevant features will simplify the model, which makes it is easier for researchers to clearly understand the research objects. Recently, there exists a large number of feature selection methods, such as RF (Strobl et al., 2007; Ganz et al., 2015; Janitza et al., 2016), Principal Component Analysis (PCA) and minimum redundancy maximum relevance (mRMR) algorithm (Peng et al., 2005). As each method have its own advantages and disadvantages, a comparative study is required. In this research, two feature selection methods, mRMR and RF, were applied to the feature selection dataset to obtain the optimal feature subset. We selected different number of features into the optimal feature subset and tested the performance of each methods on the training dataset.

2.4. Support vector machine (SVM)

Support vector machine (SVM), proposed by Vapnik, is developed based on the VC dimension theory and the structural risk minimization principle. It shows many special advantages in solving small sample, nonlinear and high dimensional pattern recognition, such as the identification of lincRNAs (Wang et al., 2014), and the identification of transmembrane protein topology (Nugent and Jones, 2009). In this paper, the LIBSVM program (Chang and Lin, 2011) was adopted, which can be downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/.

2.5. Performance assessment

In order to test the performance of our model, we compared the performance of three different models, SVM, RF and artificial neural network (ANN), on the training dataset and the independent testing dataset, respectively. Here RF implementation from http://cran.r-project.org/web/packages/randomForest/ and ANN implementation in MATLAB were used. Standard indices, such as sensitivity (S_n), specificity (S_p), accuracy (ACC) and Matthews correlation coefficient (MCC), were used to evaluate the performance of our model.

$$S_n = \frac{TP}{TP + FN} \tag{1}$$

$$S_p = \frac{TN}{TN + FP} \tag{2}$$

where, Ture Positive (TP) and False Negative (FN) denote the numbers of positive samples that are predicted to be positive and negative, respectively. Analogously, True Negative (TN) and False Positive (FP) denote the number of negative samples that are predicted as negative and positive, respectively. S_n represents the ability to identify the positive samples, while S_p represents the ability to identify the negative samples. The ACC is defined as:

$$ACC = \frac{TP + TN}{TP + FN + TN + FP} \tag{3}$$

and the MCC is defined as:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{4}$$