Molecular BioSystems

PAPER

View Article Online
View Journal | View Issue

Cite this: *Mol. BioSyst.*, 2015, 11, 2219

PredcircRNA: computational classification of circular RNA from other long non-coding RNA using hybrid features†

Xiaoyong Pan* and Kai Xiong

Received 26th March 2015,
Accepted 18th May 2015

DOI: 10.1039/c5mb00214a

www.rsc.org/molecularbiosystems

Recently circular RNA (circularRNA) has been discovered as an increasingly important type of long non-coding RNA (lncRNA), playing an important role in gene regulation, such as functioning as miRNA sponges. So it is very promising to identify circularRNA transcripts from *de novo* assembled transcripts obtained by high-throughput sequencing, such as RNA-seq data. In this study, we presented a machine learning approach, named as PredcircRNA, focused on distinguishing circularRNA from other lncRNAs using multiple kernel learning. Firstly we extracted different sources of discriminative features, including graph features, conservation information and sequence compositions, ALU and tandem repeats, SNP densities and open reading frames (ORFs) from transcripts. Secondly, to better integrate features from different sources, we proposed a computational approach based on a multiple kernel learning framework to fuse those heterogeneous features. Our preliminary 5-fold cross-validation result showed that our proposed method can classify circularRNA from other types of lncRNAs with an accuracy of 0.778, sensitivity of 0.781, specificity of 0.770, precision of 0.784 and MCC of 0.554 in our constructed gold-standard dataset, respectively. Our feature importance analysis based on Random Forest illustrated some discriminative features, such as conservation features and a GTAG sequence motif. Our PredcircRNA tool is available for download at https://github.com/xypan1232/PredcircRNA.

1 Introduction

Non-coding RNA accounts for 98.8% of the transcribed genome, estimated as 70% of human genes.[1,2] Although it cannot encode for proteins, non-coding RNA plays a very crucial role in many cellular processes, such as gene regulation and RNA splicing. Long non-coding RNA is ncRNA with a size longer than 200 nt, which was previously considered to be experimental noise or artefacts. Now more and more evidence indicates that lncRNA has a range of biological functions,[3] whose dysfunction is closely related to epigenetic and post transcriptional control in diseases.[4,5]

With next-generation-sequencing developing, a huge volume of sequencing data is generated, and the discovery of lncRNAs is expanding. Experimental identification and annotation of these new sequences with enormous information is time-consuming and high-cost. So it is necessary to find alternative computational methods for analysing them, which can complement experimental techniques to identify new putative lncRNA candidates in the genome.

Currently there are many excellent computational approaches to distinguish lncRNA[6–10] from protein coding RNA with high accuracy for assembled transcripts from next-generation-sequencing. For example, iSeeRNA[7] used SVM to detect lncRNAs *via* integrating multiple features. lncRNA-MFDL[10] applied a deep learning[11] framework to enhance prediction accuracy. Previous methods were only focused on classifying lncRNAs from protein coding RNAs, but there exists multiple types of lncRNAs in the genome. In GENCODE,[1] lncRNA can be roughly catalogued into lincRNA, antisense, processed transcript, sense intronic and sense overlapping. Recently a new type of lncRNA (circularRNA) has received more and more attention, although it was discovered at least 20 years ago. Emerging evidence demonstrates that some circularRNAs may regulate miRNA function, such as through the miRNA sponge effect[12,13] and transcription regulation.[14,15] And thousands of circularRNAs are reported in recent works, which are collected in the circularRNA database circBase.[12] Different lncRNAs have very different characteristics and functions, so it is very promising to identify more exact lncRNA subgroups. There are some computational approaches to further classify small ncRNAs into subgroups, while still no method is available to further classify lncRNAs and thus effectively facilitate annotation. For example, with CoRAL,[16] a machine learning model was trained to identify classes of small non-coding RNAs, such as microRNAs, tRNAs, snRNAs and snoRNAs.

Department of Veterinary Clinical and Animal Sciences, University of Copenhagen, Denmark. E-mail: xypan172436@gmail.com; Tel: +45 52760908

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c5mb00214a

The identification of circularRNAs is very useful for further understanding regulatory mechanisms, and furthermore for their potential implications for therapeutic applications, such as functioning as miRNA sponges for oncogenic miRNAs. lncRNA is easily distinguished from other small ncRNAs, such as miRNA, siRNA and snoRNA, by using the simple property transcript size. However, for circularRNA identification from other lncRNAs, it is almost not possible to detect it only based on simple features. circularRNA has demonstrated some different sequence characteristics from other lncRNAs, such as a GT–AG pair of canonical splice sites, paired ALU repeats and a backsplice.[17] On the other hand, sequence features combined with machine learning are reported to be powerful to predict gene regulation, splicing sites and chromatin.[18] They promote that the sequence-based method may be utilized to identify circularRNA from other lncRNAs effectively.

In this study, we are focused on cataloguing circularRNA from other lncRNAs and proposed a computational method from a transcript sequence, which can be assembled from RNA-seq using Cufflinks,[19] to distinguish circularRNA from other lncRNAs. The proposed method trained a classifier based on experimentally verified circularRNAs and other lncRNAs using machine learning. We firstly extracted different sources of discriminative features from a transcript sequence, such as graph features, conservation, sequence compositions, ALU and tandem repeats, SNP densities and open reading frame (ORFs), which cope with the potential problem that a single feature cannot perfectly characterize circularRNA from other lncRNAs. Considering the heterogeneity of those extracted features, we applied l_p-norm multiple kernel learning[20] to integrate different sources of data representations, which can fuse them with greater flexibility, and weight the relative contribution of every type of feature for final predictions.

2 Method and materials

Data source

We used human circularRNA data from the circBase database.[21] This dataset collects more than 90 000 experimentally verified circularRNA transcripts along with their genomic coordinates. After removing transcripts shorter than 200 nt and overlapped transcripts from the same gene, we got 14 084 circularRNAs as positive data. circBase collects genome-wide circularRNAs, and GENCODE[1] also provides genome-wide experimentally verified and high-quality gene annotation including protein coding RNA and non-coding RNA, and it's widely used for gene annotation in public data sources, such as Ensembl.[22] So GENCODE is used to construct a corresponding genome-wide gold-standard negative dataset. For constructing a high-quality gold-standard negative dataset, we extracted other types of lncRNA defined in GENCODE, such as lincRNA, antisense, processed transcript, sense intronic and sense overlapping, as a negative dataset with strong experimental evidence. The annotated lncRNAs in GENCODE have three confidence levels for RNA annotation (level 1: validated; level 2: manual annotation; level 3: automated annotation). We only selected annotated transcripts of level 1 and level 2, which are experimentally verified by RT-PCR and sequencing or HAVANA manual annotation. After removing overlapped transcripts existing in circBase and other preprocessing steps for circularRNAs, we obtained 19 722 lncRNAs as the negative dataset. We generated the training and independent testing datasets from the above constructed gold-standard dataset; 10 000 circularRNAs and the same number of other lncRNAs are randomly selected for model training, and the remaining 4084 circularRNAs and 9722 lncRNAs were constructed to be the independent testing dataset.

Feature extraction

Extracting discriminative features is a very crucial step in building machine learning classifiers. As shown in Fig. 1, simple features, such as GC content and transcript size, cannot obviously distinguish circularRNA from other lncRNAs. In order to achieve more obvious discrimination, we extracted different sources of features from transcript sequences to build a machine learning model, including graph features from sequence, conservation, component composition, ALU and tandem repeats, and ORF features. Besides, as reported in ref. 23, circularRNA has a significant decrease in SNPs at its miRNA binding sites, so SNP density is also included in our extracted features. Taken together, 188 features are extracted for our model training and testing.

Graph features from RNA structure and sequence. RNA structure plays crucial roles in gene regulation, polyadenylation and splicing,[24,25] especially as different lncRNAs are spliced differently, such as during exon scrambling for circularRNA. A graph can represent the sequence and structure of an RNA molecule and express two levels of relations: one is between nucleotides, the other is abstract structure annotations predicted from RNA shapes,[26] such as multi-loops, hairpins, bulges and stems. An RNA graph uses nodes to represent the nucleotides and edges to represent the backbone or bond relationships between the nucleotides. More details can be seen in ref. 27.

Graph features are very high-dimensional, with more than 30 000 dimensions from GraphProt.[27] To reduce the computational cost and possible dimension curse, here we also applied Random Forest[28] to rank feature importance for graph features based on a small randomly selected subset, and only the top 101 features are kept for the following experiments.

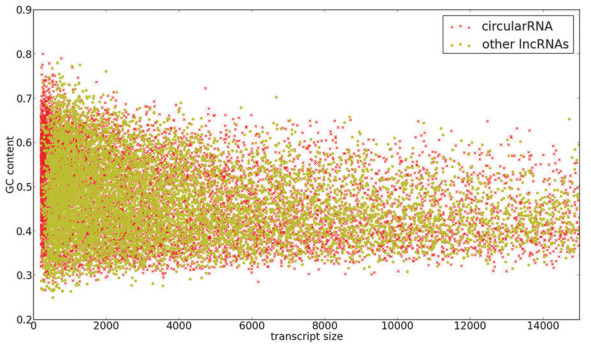

Fig. 1 GC content and transcript size comparison between circularRNAs and other lncRNAs.

Conservation score features. Firstly per-base phyloP (phylogenetic *p*-values)[29] conservation score tracks are downloaded from UCSC. The conserved features are extracted as follows: (1) calculate the mean, maximum and variance of the conservation score within the genomic region of each transcript; (2) count the frequencies of bases whose conservation score is greater than 0.3, 0.6 and 0.9 respectively, and the frequencies of bases smaller than 0.9; (3) most of the circularRNAs have very similar conserved motif sequences, which correspond to a large number of conserved docking sites for miRNA,[30] such as for ciRS-7 which has about 63 conserved binding sites for miR-7.[12] So we count the frequencies of consecutive bases (such as 4, 5, 6, 7, 8) whose scores are greater than 0.3. A total of 12 conservation score features are included in this study.

Component composition features. As shown in a paper,[7] tri-nucleotide composition has a very strong discriminating ability for detecting lincRNAs from protein coding RNAs, which is one type of lncRNA collected as a golden negative dataset. Beside the tri-nucleotide feature, other sequence component composition features are also extracted, such as GC content, sequence length, frequencies of GT, AG, GTAG and AGGT (GT/AG sequence motifs were closely related to backsplice[14]).

ALU and tandem repeat, ORF, and SNP. Base pairing ALU repeats may enable the splice sites to recognize each other, thus promoting circularization.[17] Annotated ALU repeat sites were downloaded from the UCSC Genome Browser's RepeatMasker track using the table viewer December 2011, which gives the coordinates of the ALU repeat on genomes. We count the number of ALU repeats for each transcript. Besides, circularRNAs are formed by head-to-tail splicing of exons, and tandem duplications[14] generating duplicated exons within a gene can promote apparent backsplice. In this study, Tandem Repeats Finder[31] was employed to detect tandem repeats, and the frequency of tandem repeats was extracted. txCdsPredict from the UCSC genome browser was used to obtain the ORF for each transcript; ORF length and proportion are extracted, which are reportedly useful for lincRNA classification.[7] Splice variants may produce circularRNAs[32] and a significant decrease in SNPs at miRNA targets,[23] therefore SNP density was also considered in this study. SNP data with coordinates in the genome is downloaded from the 1000 Genomes Project, and SNP density was calculated on the genomic region of each transcript.

Random Forest

Random Forest (RF)[28] is an aggregation of multiple unpruned decision trees grown from separate bootstrap samples of the training data and a feature subset sampled independently from the original feature space, and it is applied widely in bioinformatics.[33–35] It has very few parameters to tune and has better expandability when compared to other algorithms, such as the support vector machine (SVM).[36]

In this study, RF is applied to analyse the importance of extracted features. During the RF training process, bootstrap sampling will take out about 1/3 of the training data as the out-of-bag data points, whose averaged error is calculated over the constructed forest by the other 2/3 of the data points. Then the out-of-bag error is calculated again based on the new trained forest after the values of each feature are exchanged among the 2/3 of the training data points. The importance score for each feature is the mean of the difference of the out-of-bag error before and after the permutation over forest.

l_p-norm multiple kernel learning

Kernel learning is firstly applied in the SVM, which used a kernel matrix to encode similarity between samples in their respective space instead of the original feature space, and it can transfer a non-linear model in the original feature space to a linear model in the kernel space. Considering multiple feature representations of the same data, combining them together to get better feature representations is very useful for machine learning algorithms. One traditional way to combine heterogeneous features from different sources is to directly concatenate them into a single high-dimensional feature, which easily leads to not only the curse of dimensionality problem, but also to feature heterogeneity disappearance. On the contrary, multiple kernel learning can decouple the original data by combining the kernel similarity matrix in a respective space, so it is an appealing strategy in this study. A simple way for combining different kernels is linearly weighting kernels, which is often sensitive to noisy kernels, so l_p-norm multiple kernel learning[20] is applied to robustly integrate individual kernels.

In l_p-norm multiple kernel learning, the kernel mixing coefficients are optimized through regularized loss minimization with additional norm constraints when integrating multiple kernels. Given M different reproducing kernel k_m constructed from M different sources of features, the problem is formulated into a weighted linear combination of base kernels under some regularization constraints:

$$k_\theta = \sum\{\theta_m k_m,\ \theta_m \geq 0\} \tag{1}$$

To get k_θ, it can be formulated as an optimization problem of p-norm MKL as follows:[20]

$$\min_\theta \max_\alpha \left\{ \mathbf{1}^{\mathbf{T}}\alpha - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j \sum_m \theta_m k_m \right\}$$
$$\text{subject to } \|\theta\|_p \leq 1,\ \theta \geq 0,\ \mathbf{Y}^{\mathbf{T}}\alpha = 0,\ \alpha \geq 0,\ \alpha \leq C \tag{2}$$

where i and j are training data indexes, and p is the norm of the vector controlling kernel weight regularization ($p = 1$ promotes sparse combination of kernels). The above optimization can be solved iteratively by optimizing α and θ alternately. For more details and its implementation we refer to ref. 20 and the SHOGUN package.[37]

In this study, the extracted 4 views of features are incorporated into a Gaussian base kernel, then l_p-norm multiple kernel learning is used to calculate optimized weights to fuse them together.

Experimental setting

In this study, we compared the 5-fold cross-validation performance of 3 different models MKL, SVM and RF on our constructed golden dataset. Here SVM and RF implementation from Scikit-learn[38] are used. For SVM, we used the grid search best regularization

parameter C and a Gaussian kernel width g, and using 5-fold cross-validation we obtained the best value of C = 3 and g = 0.75. For RF, we set the parameter number of trees as 100 and the other parameters as their default values. For MKL, we used implementation from the SHOGUN package,[37] and kernel width 0.5 for the Gaussian kernel and a p norm of 3.5 are used. Our method accepted a BED file as the input format, which should give the coordinates of transcripts on the genome.

To provide an intuitive picture, a flowchart diagram about gold-standard dataset generation and an applied pipeline is given in Fig. 2.

Evaluation criteria

In order to compare with previous proposed methods, a 5-fold cross-validation test was used to evaluate the predicted performance. We follow their evaluation measure by means of classification accuracy, sensitivity, specificity, precision and the Matthews correlation coefficient (MCC) as defined respectively by:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (3)$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (4)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (5)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (6)$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \quad (7)$$

where TP, TN, FP, and FN represent the true positive, true negative, false positive, and false negative, respectively.

3 Results

Analyzing feature importance

For verifying the importance of extracted features in distinguishing circularRNAs, we applied Random Forest feature selection to rank the importance of them, the top 50 features are shown in Fig. 3. The top 5 features are all conservation features (conservation score variance of each transcript, mean conservation score, frequencies of 8 consecutive bases greater than 0.3, max. conservation score and frequencies of conservation scores greater than 0.9 respectively). This analysis indicated that conservation features have very powerful discriminative ability. Besides, the GTAG sequence motif has the highest importance score among the sequence composition features, which has been demonstrated to play a key role in backsplicing.[14] However only 6 of 64 tri-nucleotide frequency features are in the top 50 features, which shows that there is no obvious difference between circularRNA and other lncRNAs. RNAcon[39] also indicates that tri-nucleotide frequencies are unable to classify different classes of ncRNA. In addition, extracted ORF length and proportion, ALU repeats, and SNP density are all ranked in the top 50 features among the extracted 188 features (all features' importance scores are given in the ESI†).

Comparison between MKL, SVM and RF

Here we classified circularRNA from other types of lncRNAs using all sources of biological features, and three classifiers (MKL, SVM and RF) were implemented and compared. We concatenated all different sources of features into a single

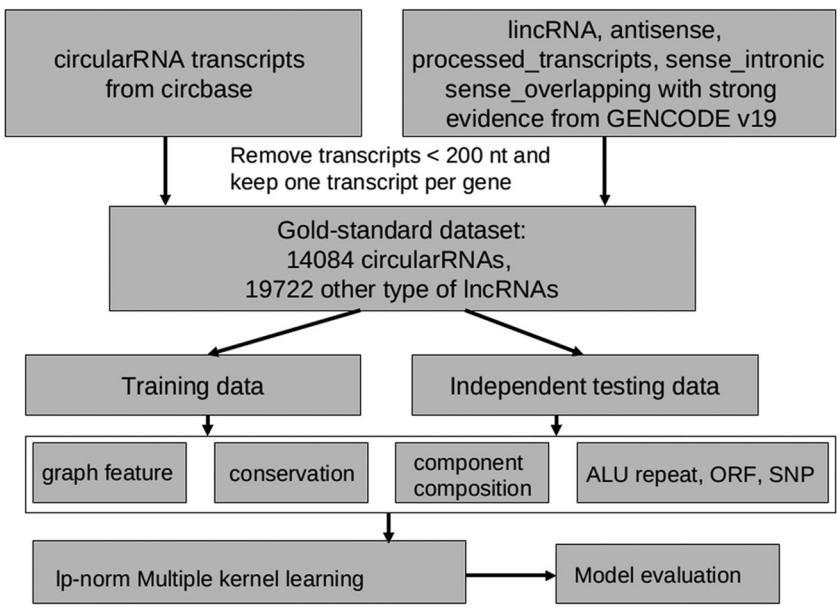

Fig. 2 Flowchart of the proposed method. Gold-standard datasets were split into training and independent testing datasets; the training data consists of 10 000 circularRNAs, 3500 lincRNAs, 3500 processed transcripts, 2700 antisense, 200 sense intronic and 100 sense overlapping. The remaining are the independent testing dataset, which were then used for independent data evaluation.

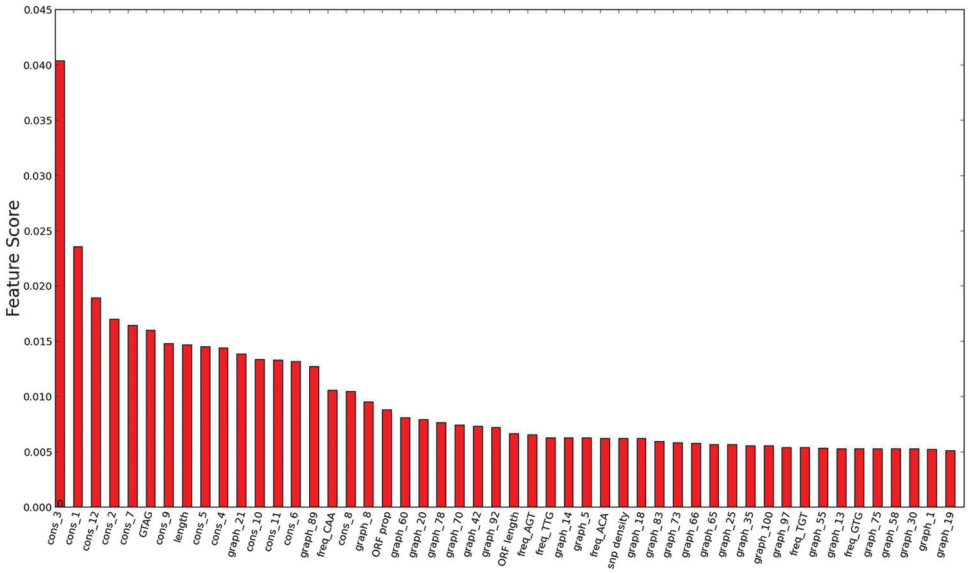

Fig. 3 Top 50 features from Random Forest importance ranking. For *X*-axis label, cons: conservation score feature; graph: graph feature; freq: tri-nucleotide frequencies feature, followed by the index for each group feature.

Table 1 5-fold cross-validation performance comparison between MKL, SVM and RF on the training dataset

Classifier	Accuracy	Sensitivity	Specificity	Precision	MCC
SVM	0.773	0.780	0.767	0.784	0.551
RF	0.767	0.769	0.773	0.768	0.541
MKL	0.778	0.781	0.770	0.784	0.554

high-dimensional feature for RF and SVM when model training and testing. As indicated in Table 1, MKL achieved the best accuracy of 0.778, sensitivity of 0.781, specificity of 0.770, precision of 0.784 and MCC of 0.554, which indicated that MKL can better integrate different sources of features, likely due to the feature heterogeneity. And SVM and RF classifiers yield a comparable performance, which also indicated that our extracted features and training dataset are very robust. Although the model performance is acceptable, it still needs to be improved from the following factors: (1) the features currently extracted are insufficient for perfectly distinguishing circularRNAs from other lncRNAs; (2) only one isoform is used for every gene, other isoforms need to be integrated into training data with more data without leading to model over-fitting.

To further demonstrate the robustness of our proposed methods, we applied our trained model to the independent testing dataset, the result is shown in Table 2. Similarly, MKL also achieved the best performance accuracy of 0.866, which is a better performance than the 5-fold cross-validation, demonstrating the robustness of our approaches. It is because the training dataset size is larger than the dataset used in doing 5-fold cross-validation, whose model can be trained with better generalization and performance. To achieve better performance, another promising direction is to fuse different models together using ensemble learning, but it will be much more time-consuming.[40,41]

Table 2 Performance evaluation on the independent testing dataset

Classifier	Accuracy	Sensitivity	Specificity	Precision	MCC
SVM	0.862	0.864	0.859	0.865	0.724
RF	0.844	0.849	0.837	0.852	0.689
MKL	0.866	0.870	0.861	0.872	0.734

We also trained the MKL model on all collected datasets consisting of 14 084 circularRNA transcripts with another 19 722 lncRNA transcripts. The 5-fold cross-validation result achieved an accuracy of 0.806, sensitivity of 0.811, specificity of 0.798, precision of 0.814 and MCC of 0.613. The results are better than those obtained on our constructed gold-standard dataset. One reason is also that more training data is used during the 5-fold cross-validation, and the trained models have better generalization power. The same situation also happens with RF (accuracy of 0.793, sensitivity of 0.795, specificity of 0.790, precision of 0.797 and MCC of 0.587) and SVM (accuracy of 0.801, sensitivity of 0.807, specificity of 0.792, precision of 0.813 and MCC of 0.607). RF performed a little worse than the other classifiers, but it runs faster than SVM and MKL.

Meanwhile we randomly selected 10 000 negative transcripts from GENCODE, including protein coding RNA, lincRNA, antisense, processed transcript, sense intronic and sense overlapping. The new random negative dataset not only contains other lncRNAs, it also includes protein coding transcripts. Our method achieved an accuracy of 0.759, sensitivity of 0.780, specificity of 0.720, precision of 0.797 and MCC of 0.519. The model performance is a little worse than negative data only from other lncRNAs. It is because our goal is to classify circularRNA from other lncRNAs using specifically curated features. On the other hand, our method can be easily integrated with other genome-wide lncRNA prediction tools, such as iSeeRNA, which aims to discriminate lncRNAs from protein coding RNAs with high accuracy.

Discriminative power between different lncRNAs

Here we performed multiclass classification for various types of lncRNAs, such as lincRNA, circularRNA, antisense, and processed transcripts, using the one-*vs.*-another strategy for multiclass classification. A balanced subset randomly selected from the original data is constructed, which consists of 2700 antisense RNAs, 2700 lincRNAs, 2700 circularRNAs and 2700 processed transcripts respectively. They were used to train a MKL multiclass model to evaluate how well different types of lncRNAs were separately classified using our extracted features. We did not include sense overlapping and sense intronic because of their quantity limit compared to other lncRNAs. Our method achieved an overall accuracy of 0.604. As seen in the following confusion matrix (Fig. 4), circularRNA is almost equally misclassified as other lncRNAs, and lincRNA is misclassified as circularRNA with the largest number. On the other hand, the result indicated that circularRNA is to the same extent different from other lncRNAs. In order to cover more negative samples, golden negative samples are constructed based on the combination of various lncRNAs in our final model.

Performance on fusing different views of features

We also compared performance between combining different sources of features using MKL. Firstly, we compared the performance for each of the extracted 4 types of features, which simply used SVM with a Gaussian kernel. And we also evaluate the classification performance of combining different types of group features using MKL. As indicated in Table 3, a combination of all features can achieve the best performance. Individual groups of features are associated with circularRNA to different extents. For individual views of features, component composition achieved the best performance. When four views of features are concatenated into one single high-dimensional feature, conservation features have higher feature importance indicated in Fig. 3, showing that conservation features may override some sequence composition features when they are fused. The above result demonstrated that different views of features have some interrelationship. It is also observed that different types of features have a different preference to circularRNA and other lncRNAs. ATOS can achieve the best precision (TP/(TP + FP)), which means it does not misclassify other lncRNAs as circularRNAs, CF achieves the best specificity, and CC yields the best sensitivity. Therefore, fusing them together can take complementary information of individual group features into consideration when training the model.

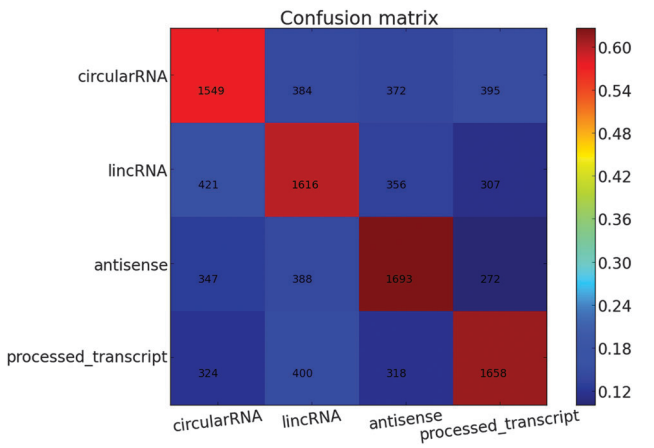

Fig. 4 Confusion matrix for 4 different lncRNAs, circularRNA, lincRNA, antisense and processed transcript using MKL multiclass classification.

Table 3 Performance comparison between combining different types of features using MKL classification. Abbreviations; CF: conservation feature; GF: graph feature; CC: component composition; ATOS: ALU and tandem repeat, ORF, and SNP

Feature	Accuracy	Sensitivity	Specificity	Precision	MCC
CF	0.703	0.696	0.721	0.685	0.406
GF	0.688	0.687	0.692	0.684	0.377
CC	0.720	0.726	0.719	0.728	0.447
ATOS	0.554	0.668	0.215	0.893	0.147
CF + GF	0.760	0.761	0.762	0.761	0.524
CF + GF + CC	0.769	0.775	0.765	0.778	0.549
CF + GF + ATOS	0.763	0.764	0.760	0.766	0.526
CF + GF + ATOS + CC	0.778	0.781	0.770	0.784	0.554

4 Discussion

In this study, we presented a novel machine learning method PredcircRNA to distinguish circularRNA from other lncRNA using different sources of features, which is the first method to further classify circularRNAs from other lncRNAs. PredcircRNA can achieve an accuracy of 0.778, sensitivity of 0.781, specificity of 0.770, precision of 0.784 and MCC of 0.554 on our gold-standard dataset. And it also shows similar performance on independent testing data. We also investigated the contribution of different sources of features to model performance. As the results showed, the conservation feature, GATG motif and component composition feature have strong discriminating power for circularRNA classification. In addition, classifiers can achieve better performance using all the available features rather than only one type of feature, which indicated their complementary property between different sources of features. PredcircRNA has the following advantages over existing lncRNA prediction tools: (1) to the best of our knowledge, it is the first study to further distinguish circularRNA from other lncRNAs using machine learning; (2) it extracted new sources of discriminative features for model training, such as conservation features and graph features; (3) it applied multiple kernel learning to better fuse different sources of extracted features.

PredcircRNA demonstrated good performance at identifying circularRNAs from other lncRNAs. Nevertheless, compared to other machine learning-based models to identify lncRNAs from protein coding RNA, which achieve a high accuracy of more than 90%, such as iSeeRNA,[7] it is still to some extent difficult to discriminate different lncRNAs. That's because circularRNA and other lncRNA have a much smaller difference than that between lncRNA and protein coding RNA. On the other hand, circularRNA is expressed in specific tissues and in a developmental manner,[12] which is ignored in our model training. Hence in future work, the proposed model will be expected to

further improve circularRNA predictions by introducing other sources of features instead of only sequences, such as expression data in different tissues or cell lines. Meanwhile, instead of training models for whole circularRNA from different tissues or cell lines, a tissue-wise classifier can be trained on a circularRNA subset from an individual tissue or cell line, which is better for aligning with tissue-specific characteristics of circularRNAs.

Recently there have also been a growing number of circularRNAs discovered in other species, but currently our classifier is only trained on human transcripts. In future work, it should be extended to other species. PredcircRNA accepts the BED format input, so it can also be smoothly integrated with a genome-wide tool for identification of lncRNAs and protein coding gene transcripts, such as PhyloCSF,[42] CPC[43] and iSeeRNA. For instance, iSeeRNA can be firstly applied to check if candidate transcripts are lncRNAs or not, then it can be fed into our tool PredcircRNA to further predict if it is circularRNA or not. And it also can be considered as a filtering tool for other circularRNA prediction tools, such as circBase,[21] which can be firstly used to screen a whole genome, then applied to our PredcircRNA to filter out false positives. This integrated pipeline can be used to find genome-wide circularRNA candidates, which can be further experimentally verified.

5 Conclusions

In this study, we presented a computational method for classifying circularRNAs from other lncRNAs based on a multiple kernel learning framework integrating hybrid features. Our experimental results indicated its efficiency both on a constructed gold-standard dataset and an independent dataset. We also compared the performance of the model with only one feature source and the different combinations of features, demonstrating that different sources of features can complement each other to improve the model's performance. And we also analysed the importance of extracted features, which indicated that the conservation feature and GTAG sequence motif have strong discriminative power for circularRNA from other lncRNAs. Python implementation of PredcircRNA is available at https://github.com/xypan1232/PredcircRNA.

Acknowledgements

This research is supported by the Innovation Fund Denmark, Fellowship from the Faculty of Health and Medical Sciences, University of Copenhagen and the China Scholarship Council.

References

1 J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, I. Barnes, A. Bignell, V. Boychenko, T. Hunt, M. Kay, G. Mukherjee, J. Rajan, G. Despacio-Reyes, G. Saunders, C. Steward, R. Harte, M. Lin, C. Howald, A. Tanzer, T. Derrien, J. Chrast, N. Walters, S. Balasubramanian, B. Pei, M. Tress, J. M. Rodriguez, I. Ezkurdia, J. van Baren, M. Brent, D. Haussler, M. Kellis, A. Valencia, A. Reymond, M. Gerstein, R. Guigo and T. J. Hubbard, *Genome Res.*, 2012, **22**, 1760–1774.
2 G. Storz, *Science*, 2002, **296**, 1260–1263.
3 F. F. Costa, *BioEssays*, 2010, **32**, 599–608.
4 G. Chen, Z. Wang, D. Wang, C. Qiu, M. Liu, X. Chen, Q. Zhang, G. Yan and Q. Cui, *Nucleic Acids Res.*, 2013, **41**, D983–D986.
5 P. J. Batista and H. Y. Chang, *Cell*, 2013, **152**, 1298–1307.
6 Y. Okazaki, M. Furuno, T. Kasukawa, J. Adachi, H. Bono, S. Kondo, I. Nikaido, N. Osato, R. Saito, H. Suzuki, I. Yamanaka, H. Kiyosawa, K. Yagi, Y. Tomaru, Y. Hasegawa, A. Nogami, C. Schönbach, T. Gojobori, R. Baldarelli, D. P. Hill, C. Bult, D. A. Hume, J. Quackenbush, L. M. Schriml, A. Kanapin, H. Matsuda, S. Batalov, K. W. Beisel, J. A. Blake, D. Bradt, V. Brusic, C. Chothia, L. E. Corbani, S. Cousins, E. Dalla, T. A. Dragani, C. F. Fletcher, A. Forrest, K. S. Frazer, T. Gaasterland, M. Gariboldi, C. Gissi, A. Godzik, J. Gough, S. Grimmond, S. Gustincich, N. Hirokawa, I. J. Jackson, E. D. Jarvis, A. Kanai, H. Kawaji, Y. Kawasawa, R. M. Kedzierski, B. L. King, A. Konagaya, I. V. Kurochkin, Y. Lee, B. Lenhard, P. A. Lyons, D. R. Maglott, L. Maltais, L. Marchionni, L. McKenzie, H. Miki, T. Nagashima, K. Numata, T. Okido, W. J. Pavan, G. Pertea, G. Pesole, N. Petrovsky, R. Pillai, J. U. Pontius, D. Qi, S. Ramachandran, T. Ravasi, J. C. Reed, D. J. Reed, J. Reid, B. Z. Ring, M. Ringwald, A. Sandelin, C. Schneider, C. A. M. Semple, M. Setou, K. Shimada, R. Sultana, Y. Takenaka, M. S. Taylor, R. D. Teasdale, M. Tomita, R. Verardo, L. Wagner, C. Wahlestedt, Y. Wang, Y. Watanabe, C. Wells, L. G. Wilming, A. Wynshaw-Boris, M. Yanagisawa, I. Yang, L. Yang, Z. Yuan, M. Zavolan, Y. Zhu, A. Zimmer, P. Carninci, N. Hayatsu, T. Hirozane-Kishikawa, H. Konno, M. Nakamura, N. Sakazume, K. Sato, T. Shiraki, K. Waki, J. Kawai, K. Aizawa, T. Arakawa, S. Fukuda, A. Hara, W. Hashizume, K. Imotani, Y. Ishii, M. Itoh, I. Kagawa, A. Miyazaki, K. Sakai, D. Sasaki, K. Shibata, A. Shinagawa, A. Yasunishi, M. Yoshino, R. Waterston, E. S. Lander, J. Rogers, E. Birney and Y. Hayashizaki, *Nature*, 2002, **420**, 563–573.
7 K. Sun, X. Chen, P. Jiang, X. Song, H. Wang and H. Sun, *BMC Genomics*, 2013, **14**(suppl 2), S7.
8 M. N. Cabili, C. Trapnell, L. Goff, M. Koziol, B. Tazon-Vega, A. Regev and J. L. Rinn, *Genes Dev.*, 2011, **25**, 1915–1927.
9 J. Lv, H. Liu, Z. Huang, J. Su, H. He, Y. Xiu, Y. Zhang and Q. Wu, *Nucleic Acids Res.*, 2013, **41**, 10044–10061.
10 X.-N. Fan and S.-W. Zhang, *Mol. BioSyst.*, 2015, **11**, 892–897.
11 G. E. Hinton, G. E. Hinton, S. Osindero, S. Osindero, Y. W. Teh and Y. W. Teh, *Neural computation*, 2006, **18**, 1527–1554.
12 S. Memczak, M. Jens, A. Elefsinioti, F. Torti, J. Krueger, A. Rybak, L. Maier, S. D. Mackowiak, L. H. Gregersen, M. Munschauer, A. Loewer, U. Ziebold, M. Landthaler, C. Kocks, F. le Noble and N. Rajewsky, *Nature*, 2013, **495**, 333–338.
13 T. B. Hansen, T. I. Jensen, B. H. Clausen, J. B. Bramsen, B. Finsen, C. K. Damgaard and J. R. Kjems, *Nature*, 2013, **495**, 384–388.
14 W. R. Jeck and N. E. Sharpless, *Nat. Biotechnol.*, 2014, **32**, 453–461.

15 Z. Li, C. Huang, C. Bao, L. Chen, M. Lin, X. Wang, G. Zhong, B. Yu, W. Hu, L. Dai, P. Zhu, Z. Chang, Q. Wu, Y. Zhao, P. X. Ya Jia, H. Liu and G. Shan, *Nat. Struct. Mol. Biol.*, 2015, **22**, 256–264.
16 P. Ryvkin, Y. Y. Leung, L. H. Ungar, B. D. Gregory and L. S. Wang, *Methods*, 2013, 28–35.
17 W. R. Jeck, J. A. Sorrentino, K. Wang, M. K. Slevin, C. E. Burd, J. Liu, W. F. Marzluff and N. E. Sharpless, *RNA*, 2013, **19**, 141–157.
18 H. Y. Xiong, B. Alipanahi, L. J. Lee, H. Bretschneider, D. Merico, R. K. Yuen, Y. Hua, S. Gueroussov, H. S. Najafabadi and T. R. Hughes, *et al.*, *Science*, 2015, **347**, 6218.
19 A. Roberts, H. Pimentel, C. Trapnell and L. Pachter, *Bioinformatics*, 2011, **27**, 2325–2329.
20 M. Kloft, *Journal of Machine Learning Research*, 2011, **12**, 953–997.
21 P. Glažar, P. Papavasileiou and N. Rajewsky, *RNA*, 2014, **20**, 1666–1670.
22 T. Hubbard, D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen and T. Down, *et al.*, *Nucleic Acids Res.*, 2002, **30**, 38–41.
23 L. F. Thomas and P. l. S trom, *Bioinformatics*, 2014, 1–4.
24 J. A. Cruz and E. Westhof, *Cell*, 2009, **136**, 604–609.
25 Y. Ding, Y. Tang, C. K. Kwok, Y. Zhang, P. C. Bevilacqua and S. M. Assmann, *Nature*, 2014, **505**, 696–700.
26 P. Steffen, B. Voß, M. Rehmsmeier, J. Reeder and R. Giegerich, *Bioinformatics*, 2006, **22**, 500–503.
27 D. Maticzka, S. J. Lange, F. Costa and R. Backofen, *Genome Biol.*, 2014, **15**, R17.
28 L. U. O. C. Breiman, *Random Forest*, 1999, vol. 45, pp. 1–35.
29 K. S. Pollard, M. J. Hubisz, K. R. Rosenbloom and A. Siepel, *Genome Res.*, 2010, **20**, 110–121.
30 J. O. Westholm, P. Miura, S. Olson, S. Shenker, B. Joseph, P. Sanfilippo, S. E. Celniker, B. R. Graveley and E. C. Lai, *Cell Rep.*, 2014, **9**, 1966–1980.
31 G. Benson, *Nucleic Acids Res.*, 1999, **27**, 573–580.
32 A. Gschwendtner, S. Bevan, J. W. Cole, A. Plourde, M. Matarin, H. Ross-Adams, T. Meitinger, E. Wichmann, B. D. Mitchell, K. Furie, A. Slowik, S. S. Rich, P. D. Syme, M. J. MacLeod, J. F. Meschia, J. Rosand, S. J. Kittner, H. S. Markus, B. D. Mitchell and M. Dichgans, *Ann. Neurol.*, 2009, **65**, 531–539.
33 Y. Li, M. Wang, H. Wang, H. Tan, Z. Zhang, G. I. Webb and J. Song, *Sci. Rep.*, 2014, **4**, 5765.
34 X. Y. Pan, Y. N. Zhang and H. B. Shen, *J. Proteome Res.*, 2010, **9**, 4992–5001.
35 X. Pan, L. Zhu, Y.-X. Fan and J. Yan, *Comput. Biol. Chem.*, 2014, **53**, 324–330.
36 V. N. Vapnik, *The Nature of Statistical Learning Theory*, 1995, vol. 8, p. 188.
37 R. Gunnar, *Journal of Machine Learning Research*, 2010, **22**, 2006.
38 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, *Journal of Machine Learning Research*, 2012, **12**, 2825–2830.
39 B. Panwar, A. Arora and G. P. Raghava, *BMC Genomics*, 2014, **15**, 127.
40 X. Y. Pan, Y. Tian, Y. Huang and H.-B. Shen, *Genomics*, 2011, **97**, 257–264.
41 L. Nanni, S. Branham, N. Lazzarini and C. Fantozzi, 2013 Annual Meeting of the Northeast Decision Sciences Institute, 2013, 523--535.
42 M. F. Lin, I. Jungreis and M. Kellis, *Bioinformatics*, 2011, **27**, i275–i282.
43 L. Kong, Y. Zhang, Z. Q. Ye, X. Q. Liu, S. Q. Zhao, L. Wei and G. Gao, *Nucleic Acids Res.*, 2007, **35**, W345–W349.

