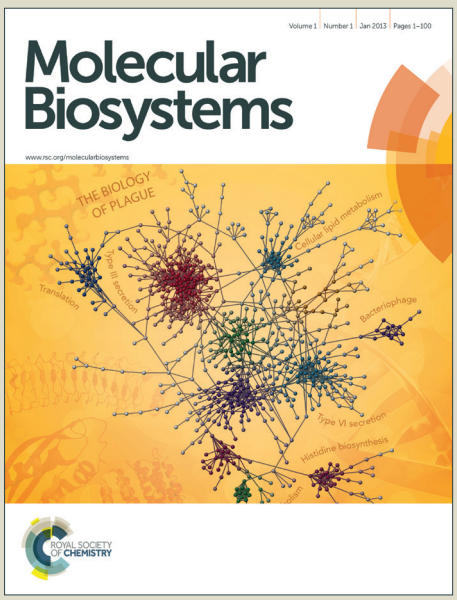
Molecular BioSystems

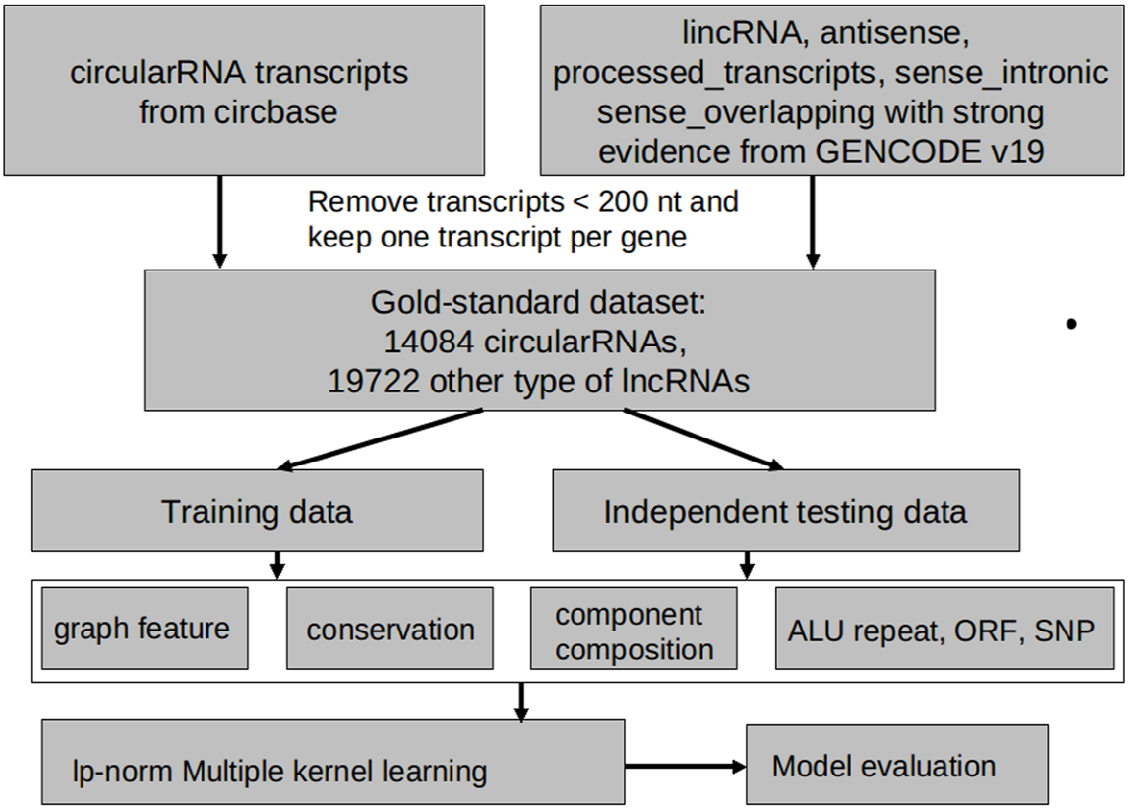
Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/molecularbiosystems

PredcircRNA present computational classification of circularRNA from other lncRNA using hybrid features based on multiple kernel learning.



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

Xiaoyong Pan[a,*] **and Kai Xiong**[a]

Received Xth XXXXXXXXXX 20XX, Accepted Xth XXXXXXXXX 20XX
First published on the web Xth XXXXXXXXXX 200X
DOI: 10.1039/b000000x

Recently circular RNA (circularRNA) has been discovered as a growing 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 feature, conservation information and sequence compositions, ALU and tandem repeat, SNP density and open reading frame (ORF) from transcripts. Secondly, to better integrate features from different sources, we proposed a computational approach based on 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 on our constructed gold-standard dataset, respectively. Our feature importance analysis based on random forest illustrated some discriminative features, such as conservation features and GTAG sequence motif. Our PredcircRNA tool is available for download at `https://github.com/xypan1232/PredcircRNA`.
Keywords: circularRNA, lncRNA, multiple kernel learning, graph feature, conservation information, sequence composition

1 INTRODUCTION

Non-coding RNA accounts for 98.8% of transcribed genome estimated as 70% of human genes[1,2]. Although it cannot encode for proteins, non-conding RNA plays a very crucial role in many cellular processes, such as gene regulation and RNA splicing. Long non-coding RNA is ncRNA with size longer than 200 nt, which is previously considered to be experimental noises and artefact. Now more and more evidence indicates that lncRNA plays 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 lncRNAs are expanding to be discovered. While 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 with experimental techniques to identify new putative lncRNA candidates in 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 deep learning[11] framwork to enhance prediction accuracy. Previous methods were only focused on classifying lncRNAs from protein coding RNAs, but there exists multiple types of lncRNAs in genome. In GENCODE[1], lncRNA can be roughly catalogued into lincRNA, antisense, processed transcript, sense intronic and sense overlapping. Recently a new type of lncRNAs (circularRNA) get more and more attention, although it has been discovered at least 20 years ago. Emerging evidence demonstrates that some circularRNAs may regulate miRNA function, such as miRNA sponge effect[12,13] and transcription regulation[14,15]. And thousand of circularRNAs are reported in recent works, which are collected in circularRNA database circbase[12]. For different lncRNAs, they 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 to subgroups, while still no method is available to further classify lncRNAs, and thus facilitates annotation effectively. For example, CoRAL[16], it trained a machine learning model to identify class of small non-coding RNAs, such as microRNAs, tRNAs, snRNAs and snoRNAs.
The identification of circularRNAs is very useful for further understanding regulatory mechanisms, furthermore for potential implications for therapeutic applications, such as function-

[a]*Department of Veterinary Clinical and Animal Sciences, University of Copenhagen, Denmark. Tel: +4552760908; Xiaoyong Pan(xypan172436@gmail.com)*

ing as miRNA sponges for oncogenic miRNAs. lncRNA is easily distinguished from other small ncRNA, such as miRNA, siRNA and snoRNA, by using simple property transcript size. However, for circularRNA identification from other lncRNAs, it is almost not possible to detect them only based on simple features. While circularRNA has demonstrated some different sequence characteristics from other lncRNAs, such as GT-AG pair of canonical splice sites, paired Alu repeat and backsplice[17]. On the other hand, sequence features combining with machine learning is reported to be powerful to predict gene regulation, splicing sites and chromatin[18]. They promote that sequence-based method maybe 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 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 transcript sequence, such as graph feature, conservation, sequence composition, ALU and tandem repeat, SNP density and open reading frame (ORF), which cope with the potential problem that 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 relative contribution for every view of features to final predictions.

2 METHOD AND MATERIALS

Data source

We used human circularRNA data from circbase database[21]. This dataset collects more than 90000 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 14084 circularRNAs as positive data. circbase collected 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 source, such as Ensembl[22]. So GENCODE is used to construct corresponding genome-wide gold-standard negative dataset. For constructing high-quality gold-standard negative dataset, we extracted another types of lncRNA defined in GENCODE, such as lincRNA, antisense, processed transcript, sense intronic and sense overlapping, as 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 with 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 19722 lncRNAs as negative dataset. we generated the training and independent testing datasets from above constructed gold-standard dataset, 10000 circularRNAs and the same number of other lncRNAs are randomly selected for model training, the remaining 4084 circularRNAs and 9722 lncRNAs were constructed to be independent testing dataset.

Feature extraction

Extracting discriminative features is a very crucial step in building machine learning classifiers. As shown in Figure 1, simple featues, 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 machine learning model, including graph feature from sequence, conservation, component composition, ALU and tandem repeat, and ORF features. Besides, as reported in[23], circularRNA has 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.

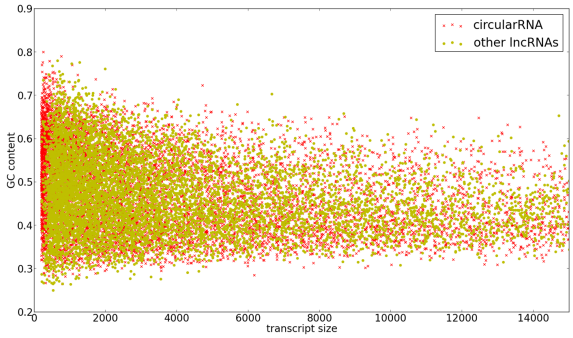


Fig. 1 GC content and transcript size comparison between circularRNAs and other lncRNAs.

Graph features from RNA structure and sequence. RNA structure plays crucial roles in gene regulation, polyadenylation and splicing[24,25], especially different lncRNA are spliced differently, such as 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 RNAshapes[26], such as multi-loops, hairpins, bulges and stems. RNA graph uses node to represent the nucleotides and edge to represent the backbone or bond relationships between the nucleotides. More details can be seen in[27].

Graph feature is very high-dimensional, more than 30000 dimension from GraphProt[27]. To reduce computational cost and possible dimension curse, Here we also applied Random Forest[28] to rank feature importance for graph features based on small random selected subset, and only the top 101 features are kept for following experiments.

Conservation score features. Firstly per-base phyloP (phylogenetic p-values)[29] conservation score track are downloaded from UCSC. The conserved features are extracted as follows: 1) calculate the mean, max and variance of 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, the frequencies of bases smaller than 0.9; 3) Most of circularRNAs have very similar conserved motif sequences, which correspond to large number of conserved docking sites for miRNA[30], such as ciRS-7 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 score are greater than 0.3. Total 12 conservation score features are included in this study.

Component composition features. As shown in paper[7], tri-nucleotides composition has very strong discriminating ability for detecting lincRNAs from protein coding RNAs, which is one type of lncRNAs collected as 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, 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 repeat 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 repeat was extracted. txCdsPredict from UCSC genome browser was used to obtain the ORF for each transcript, ORF length and proportion are extracted, which is reported useful for lincRNA classification[7]. Splice variants may produce circularRNAs[32] and significant decrease in SNPs at miRNA targets[23], therefore SNP density was also considered in this study. SNP data with coordinates in genome is download 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 have better expandability when compared to other algorithms, such as 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 training data as the out-of-bag data points, whose averaged error is calculated over the constructed forest by other 2/3 data points. Then the out-of-bag error is calculated based on new trained forest again after the values of the each feature are exchanged among the 2/3 training data points. The importance score for each feature is the mean of difference of out-of-bag error before and after the permutation over forest.

l_p-norm Multiple kernel Learning

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

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

$$k_\theta = \sum\{\theta_m k_m,\ \theta_m \geq 0\} \quad (1)$$

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

$$\min_{\theta}\max_{\alpha}\{\mathbf{1}^{\mathbf{T}}\alpha-\frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j\sum_m\theta_m k_m\}$$
$$\text{subject to } \|\theta\|_p\le 1,\theta\ge 0,\mathbf{Y}^{\mathbf{T}}\alpha=0,\alpha\ge 0,\alpha\le C \quad (2)$$

where i, j is training data index, p is Norm of vector controlling kernel weights regularization, p = 1 promote sparse combination of kernels. The above optimization can be solved iteratively using optimizing α and θ alternately. More details and its implementation we refer to[20] and SHOGUN package[37]

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

Experimental setting

In this study, we compared 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 grid search best regularization parameter C and Gaussian kernel width g using 5-fold cross-validation, we obtained best C=3 and g=0.75. For RF, we set parameter number of tree as 100 and other parameters as default value. For MKL, we used implementation from SHOGUN package[37], and kernel width 0.5 for Gaussian kernel and p norm 3.5 are used. Our method accept BED file as input format, which should give the coordinate of transcripts on genome.

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

Evaluation Criteria

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

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

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

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

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

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

Where TP, TN, FP, and FN represents 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 is shown in Figure 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 score greater than 0.9 respectively.). This analysis indicated that conservation features have very powerful discriminative ability. Besides, GTAG sequence motif has the highest importance score among sequence composition features, which is demonstrated to play a key role in backsplice[14]. However only 6 of 64 tri-nucleotide frequencies features are in the top 50 features, which shows no obvious difference between circularRNA and other lncRNAs. RNAcon[39] also indicates tri-nucleotide frequencies is unable to classify different classes of ncRNA. In addition, extracted ORF length and proportion, ALU repeat, SNP density are all ranked in top 50 features among the extracted 188 features (all features' importance score is given in supplement file 1).

Comparison between MKL, SVM and RF.

Here we classified circularRNA from other type of lncRNAs using all sources of biological features, three classifiers(MKL, SVM and RF) were implemented and compared. We concatenated all different sources of features into a single high dimensional features 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 MKL can better integrate different sources of features, likely due to the feature heterogeneity. And SVM and RF classifiers yield comparable performance, which also indicated that our extracted features and training dataset is very robust. Although the model performance is acceptable, it still need to be improved from 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 be integrated into training data with more data without leading model over-fitting.

To further demonstrate the robustness of our proposed

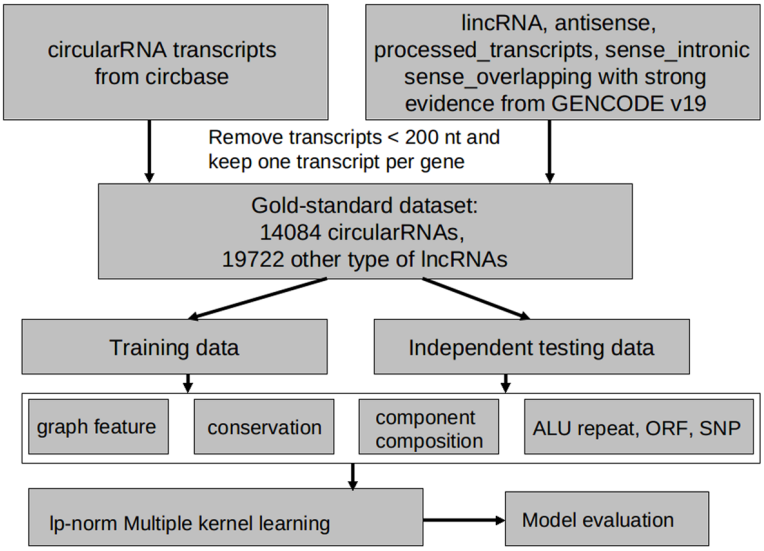


Fig. 2 Flowchart of proposed method. Gold-standard datasets were split into training and independent testing datasets, training data consists of 10000 circularRNAs, 3500 lincRNAs, 3500 processed transcripts, 2700 antisense, 200 sense intronic and 100 sense overlapping. The remaining are independent testing dataset, which were then used for independent data evaluation.

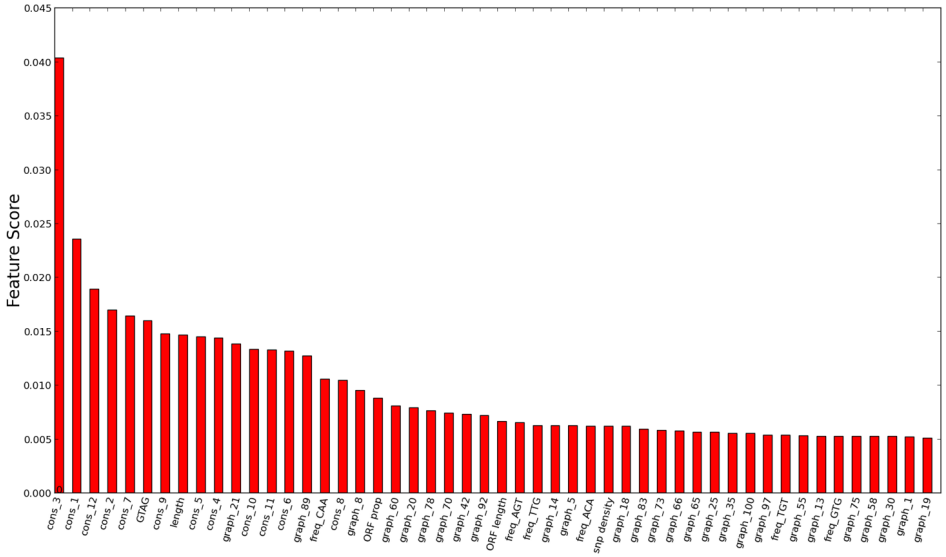


Fig. 3 Top 50 features from Random Forest importance ranking. For X-axis label, cons: conservation score feature; graph:graph feature; freq: tri-nucleotides frequencies feature, followed by index for each group feature.

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

Classifier	Accuracy	Sensitivity	Specificity	Precision	MCC
SVC	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

methods, we applied our trained model on independent testing dataset, the result is shown in Table 2. Similarly, MKL also achieved the best performance of accuracy 0.866, which performed better than 5-fold cross-validation, demonstrating the robust of our approaches. It is because that training dataset size is larger than dataset 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 independent testing dataset.

Classifier	Accuracy	Sensitivity	Specificity	Precision	MCC
SVC	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 MKL model on all collected dataset consisting of 14084 circularRNA transcripts with another 19722 lncRNA transcripts, 5-fold cross-validation result achieved accuracy of 0.806, sensitivity of 0.811, specificity of 0.798, precision of 0.814 and MCC of 0.613. It is better than 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 to 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 other classifiers, but it runs faster than SVM and MKL.

Meanwhile we randomly selected 10000 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 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 that 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 multi-class classification for various types of lncRNAs, such as lincRNA, circularRNA and antisense, processed transcripts, using one-vs-other strategy for multi-class classification. A balanced subset randomly selected from original data is constructed, which consists of 2700 antisense, 2700 lincRNAs, 2700 circularRNAs and 2700 processed transcripts respectively. They were used to train a MKL multiclass model to evaluate how well separately classify between different types of lncRNAs 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 overall accuracy of 0.604. As seen in following confusion matrix (Figure 4), circularRNA is almost equally misclassified as other lncRNAs, and lincRNAs 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 combination of various lncRNAs in our final model.

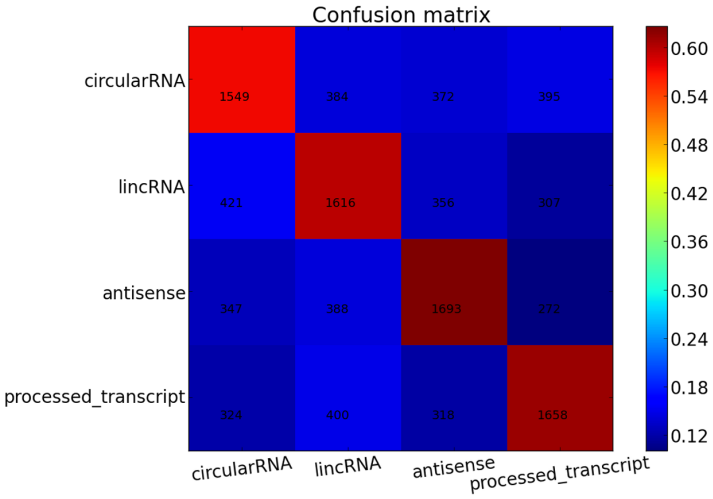


Fig. 4 Confusion matrix for 4 different lncRNAs, circularRNA, lincRNA, antisense and processed transcript using MKL multiclass classification.

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 extracted 4 types of features, which

simply used SVM with Gaussian kernel. And also evaluate the classification performance of combining different types of group features using MKL. As indicated in Table 3, all features combination can achieve best performance. Individual group of features are associated with circularRNA to different extents. For individual view of features, component composition achieve the best performance. While when four views of features are concatenated into one single high-dimensional feature, conservation features have higher feature importance indicated in Figure 3, showing conservation feature 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 has different preference to circularRNA and other lncRNAs. ATOS can achieve best precision (TP/(TP + FP)), which means it does not misclassify other lncRNAs as circularRNAs, CF achieve best specificity, CC yield best sensitivity. Therefore, fusing them together can take complementary information of individual group features into consideration when training model.

4 Discussions

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 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, respectively. And it also show similar performance on independent testing data. We also investigated contribution of different sources of features to model performance. As result showed, conservation feature, GATG motif and component composition feature has strong discriminating power for circularRNA classification. In addition, classifiers can achieve better performance using all the available features than only one type of features, which indicated their complementary property between different sources of features. PredcircRNA has the following advantages over existing lncRNA prediction tools: 1) To best of our knowledge, it is for 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 on identifying circularRNAs from other lncRNAs. Nevertheless, compared to other machine learning based models to identify lncRNAs from protein coding RNA, which achieve high accuracy of more than 90%, such as iSeeRNA[7], it is still to some extent diffcult to discriminate different lncRNAs. That's because circularRNA and other lncRNA have much smaller difference than between lncRNA and protein coding RNA. On the other hand, circularRNA is expressed in specific tissue and 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, tissue-wise classifier can be trained on circularRNA subset from individual tissue or cell line, which is better to align with tissue-specific characteristics of circularRNAs.

Recently there are also a growing number of circularRNAs discovered in other specifies, but currently our classifier is only trained on human transcripts. In future work, it should be extended to other specifies. predcircRNA accept BED format input, so it can also be smoothly integrated with 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 feed into our tool predcircRNA to further predict it is circularRNA or not. And it also can be considered as filtering tool for other circularRNA prediction tools, such as circbase[21], which can be firstly used to screen whole genome, then applied 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 CONCLUSION

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

6 ACKNOWLEDGMENT

This research is supported by Innovation Fund Denmark, Fellowship from Faculty of Health and Medical Sciences, Uni-

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

versity of Copenhagen and 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 research*, 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 Research*, 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 & development*, 2011, **25**, 1915–1927.

9 J. Lv, H. Liu, Z. Huang, J. Su, H. He, Y. Xiu, Y. Zhang and Q. Wu, *Nucleic Acids Research*, 2013, **41**, 10044–10061.

10 F. XN and Z. SW, *Molecular BioSystems*, 2015, **11**, 892–7.

11 G. E. Hinton, G. E. Hinton, S. Osindero, S. Osindero, Y. W. Teh and Y. W. Teh, *Neural computation*, 2006, **18**, 1527–54.

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**, 333–8.

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**, 384–8.

14 W. R. Jeck and N. E. Sharpless, *Nature biotechnology*, 2014, **32**, 453–61.

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, *Nature Structural & Molecular Biology*, 2015, **22**, 256–264.

16 P. Ryvkin, Y. Y. Leung, L. H. Ungar, B. D. Gregory and L. S. Wang, 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, 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, *PhD thesis*, 2011.

21 G. P, P. P and R. N, *RNA*, 2014, **20**, 1666–70.

22 T. Hubbard, D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down *et al.*, *Nucleic acids research*, 2002, **30**, 38–41.

23 L. F. Thomas and P. l. Sæ 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. Vo??, 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 research*, 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 Reports*, 2014, **9**, 1966–1980.

31 G. Benson, *Nucleic Acids Research*, 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, *Annals of Neurology*, 2009, **65**, 531–539.

33 Y. Li, M. Wang, H. Wang, H. Tan, Z. Zhang, G. I. Webb and J. Song, *Scientific reports*, 2014, **4**, 5765.

34 X. Y. Pan, Y. N. Zhang and H. B. Shen, *Journal of Proteome Research*, 2010, **9**, 4992–5001.

35 X. Pan, L. Zhu, Y.-X. Fan and J. Yan, *Computational biology and chemistry*, 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**, 2006–2006.

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 Research*, 2007, **35**, W345–W349.