A Representation Learning Approach for Predicting circRNA Back-Splicing Event via Sequence-Interaction-Aware Dual Encoder

Chengxin He*, Lei Duan*, Huiru Zheng†, Xinye Wang*, Lili Guan*, and Jiaxuan Xu*
*School of Computer Science, Sichuan University, Chengdu, China
†School of Computing, Ulster University, Northern Ireland, United Kingdom

I. APPENDIX

A. Zero-shot Back-splicing Discovery Analysis

To further prove the effectiveness of SIDE, we also followed *JEDI* [1] to construct an experiment on zero-shot back-splicing discovery. Details of the specific dataset and experimental setup used can be found in [1]. First, we performed a five-fold cross-validation on the "Human isoform level" dataset, and then uses this trained model to directly test it on the "Mouse isoform level" dataset. The training experimental results on the "Human isoform level" dataset, and experimental validation results on the "Mouse isoform level" dataset are shown in Table I, which show that SIDE has the best or second best results in most of the reported metrics, which also demonstrates the effectiveness of our model. It can be found that training all models on the "Human isoform level" dataset yields results where the performance of SIDE is second best in most of the metrics, whereas our model performs best in most of the metrics when we perform a zero-shot validation on the "Mouse isoform level" dataset. From this point, we can see that our model has an advantage in the generalization of prediction. At the same time, the non-optimal performance of our model trained on the "Human isoform level" dataset may be due to the fact that our model is more likely to be overfitted when the training samples are too large (The "Human isoform level" dataset has the largest sample size of the datasets used in our experiments), resulting in a decrease in the performance of our model. In contrast, *JEDI* model only focuses on localized sequence fragments, so there is no excessive redundant information when the training samples are too large, and overfitting is not easy to occur.

B. Sequence Modeling Approach Analysis

In this work, the key to the identification of circRNA back-splicing events lies in the feature extraction of RNA transcript sequences, and for this reason, the SIDE model processes and extracts important discriminative features of circRNA loop formation from the perspectives of both global feature of sequences and local interaction features. Different approaches of modeling the sequence, i.e., different encoding ways, can produce different results, especially when the sequence is considered from different perspectives. Therefore, to illustrate the importance of the choice of modeling sequences under global and local perspectives in our work, this subsection verifies the effect of different ways of modeling sequences under different perspectives on the performance of the model by varying the ways of encoding to input sequences in the global sequence feature extraction and local interaction feature extraction parts.

In this experiment, we analyze and discuss three common coding approaches for RNA sequences [2], *i.e.*, "one-hot", "*k*-mer", and "word2vec". In particular, for the *k*-mer encoding approach, the default optimal parameter of the SIDE model, *i.e.*, $k = 5$, is chosen for the setting of k. For the word2vec encoding approach, the six-base character used in the *CircCNN* [3] model to represent a word in a bag of words is referenced.

Table II shows the experimental results of the SIDE model for the prediction of circRNA back-splicing events on the Human dataset under different combinations of the three sequence modeling approaches as part of the global sequence feature extraction and local context interaction feature extraction. Similarly, Table II shows the results of the evaluation metrics of AUC, ACC, and MCC for the Human dataset only. The other metrics and the experimental results under the other two datasets are similar to the results analyzed in Table II. The experimental results presented in Table II can be observed:

The performance of extracting features to identify circRNA back-splicing events is optimal when the global sequence features are extracted using "one-hot" and the local interaction features are extracted using "*k*-mer" encoding to model the input sequences. This also indicates that when extracting global features of the RNA sequence, it is beneficial to retain as much recognition of each base as possible in order to capture the perceptual signals of circRNA back-splicing events in the sequence as a whole; this is also evident in the case of the global sequence feature extraction using coding methods other than "one-hot". This can also be seen in the results of the global sequence feature extraction part where the model performance decreases significantly when using encoding approaches other than "one-hot", and it is worth noting that the first and penultimate rows of the results are different from this conclusion, which demonstrates the importance of using a suitable sequence modeling approach in the local interaction feature extraction. In the local interaction feature extraction part, it is important to focus on the local segments

TABLE I
THE PERFORMANCES OF COMPARISON WITH BASELINES ON ZERO-SHOT BACK-SPLICING DISCOVERY ANALYSIS

Method	Human isoform level					Mouse isoform level				
	AUC	ACC	Sens	Spec	MCC	AUC	ACC	Sens	Spec	MCC
PredcircRNA	0.5882	0.6550	0.5949	0.7202	0.3169	0.6067	0.5696	0.5056	0.6437	0.1501
SVM	0.6729	0.7279	0.8932	0.4526	0.4031	0.7059	0.7328	0.8108	0.6011	0.4196
CircCNN	0.7283	0.7325	0.8259	0.5988	0.4815	0.7241	0.7324	0.7857	0.6327	0.4235
RF	0.7296	0.7607	0.8610	0.5982	0.4804	0.6726	0.7186	0.8523	0.4929	0.3733
circDeep	0.7395	0.8748	0.8161	0.9407	0.7584	0.7669	0.6140	0.6982	0.7509	0.4491
nRC	0.8280	0.7557	0.8410	0.6193	0.4781	0.8097	0.7410	0.8455	0.5647	0.4298
DeepCirCode	0.8994	0.8997	0.9021	0.8967	0.7914	0.8304	0.8129	0.7620	**0.8989**	0.6392
JEDI	0.9872	**0.9878**	**0.9906**	**0.9836**	**0.9742**	0.8621	0.8654	0.8749	0.8493	**0.7162**
SIDE	**0.9937**	0.9646	0.9847	0.9060	0.9059	**0.9196**	**0.8665**	**0.9140**	0.7903	0.7153

The best results are highlighted in bold, and the second-best result is marked with an underline.

TABLE II
RESULTS OF DIFFERENT MODELING WAYS FOR RNA SEQUENCES ON THE HUMAN DATASET

Sequence Encoding Approach on		AUC	ACC	MCC
Section.III.B.1	Section.III.B.2			
one-hot	one-hot	0.6085	0.6188	0.2433
one-hot	***k*-mer**	**0.9216**	**0.8806**	**0.7538**
one-hot	word2vec	0.8617	0.7996	0.5991
k-mer	one-hot	0.5993	0.6010	0.2260
k-mer	*k*-mer	0.8095	0.7361	0.4821
k-mer	word2vec	0.7838	0.6990	0.4051
word2vec	one-hot	0.6286	0.6102	0.2413
word2vec	*k*-mer	0.8948	0.8392	0.6989
word2vec	word2vec	0.8439	0.7266	0.4686

of the sequence. If the sequence segments are not organized in the modeling approach (e.g., the "one-hot" approach only considers the organization of individual bases), then the loop characteristics of circRNAs within and between segments will be easily lost, and thus the local interaction features cannot be extracted. Therefore, in this part of the modeling, the performance is improved by using the coding method that can divide and organize the sequence fragments, and the performance of the model is basically poor in the case of using the "one-hot" coding approach in this part, as can be seen from the results in Table II.

C. Parameter Sensitivity

SIDE contains several hyper-parameters, which have been tested to evaluate their impacts on SIDE measured by AUC, ACC, and MCC, including the dimension of the final representation $\mathbf{Z}^{\mathbf{f}}$, the k in k-mer, the temperature coefficient τ, and the epoch. In this part, all experiments are performed on three datasets. When comparing a parameter, we keep the other parameters unchanged. Their performances are presented in Fig. 1.

1) Effect of dimension: The size of the dimension will be related to the adequacy of the expression of the key discriminative features extracted by the model, and too small a dimension setting will not be able to reflect the amount of information of the important features. Therefore, to verify the effect of dimension on SIDE, we change the dimension in {16, 32, 64, 128, 256}. From Fig. 1, we can be observed that as the size of dimension increases, the performance of the model is also gradually improved, and when the dimension is 64, the performance of the model reaches the optimal, with the size of dimension continue to increase the performance also tends to stabilize, and even in the Fruit Fly dataset also appears to have a downward trend. From these experimental results, it can be shown that the SIDE model is sufficient to express the information of important features when the dimension is set to 64, and increasing the dimension size may bring redundancy or noise to affect the performance improvement.

2) Effect of k*:* The size of k represents the division of the input sequence on local segments in the local interaction feature extraction part, *i.e.*, the subsequence segments are represented by k number of bases, and in this part, the size of k is chosen from {3, 5, 7, 9, 11}. From the experimental results presented in Fig. 1, it can be seen that the model performs optimally on the three datasets when k is taken to be 5, while it performs poorly on all the datasets when it is taken to be 3. This also shows that in the part of the local interaction feature extraction, the embodiment of the local segments of the sequence cannot be set too small, or else it may not be able to capture the information about the interactions between the local segments. At the same time, it should not be set too large,

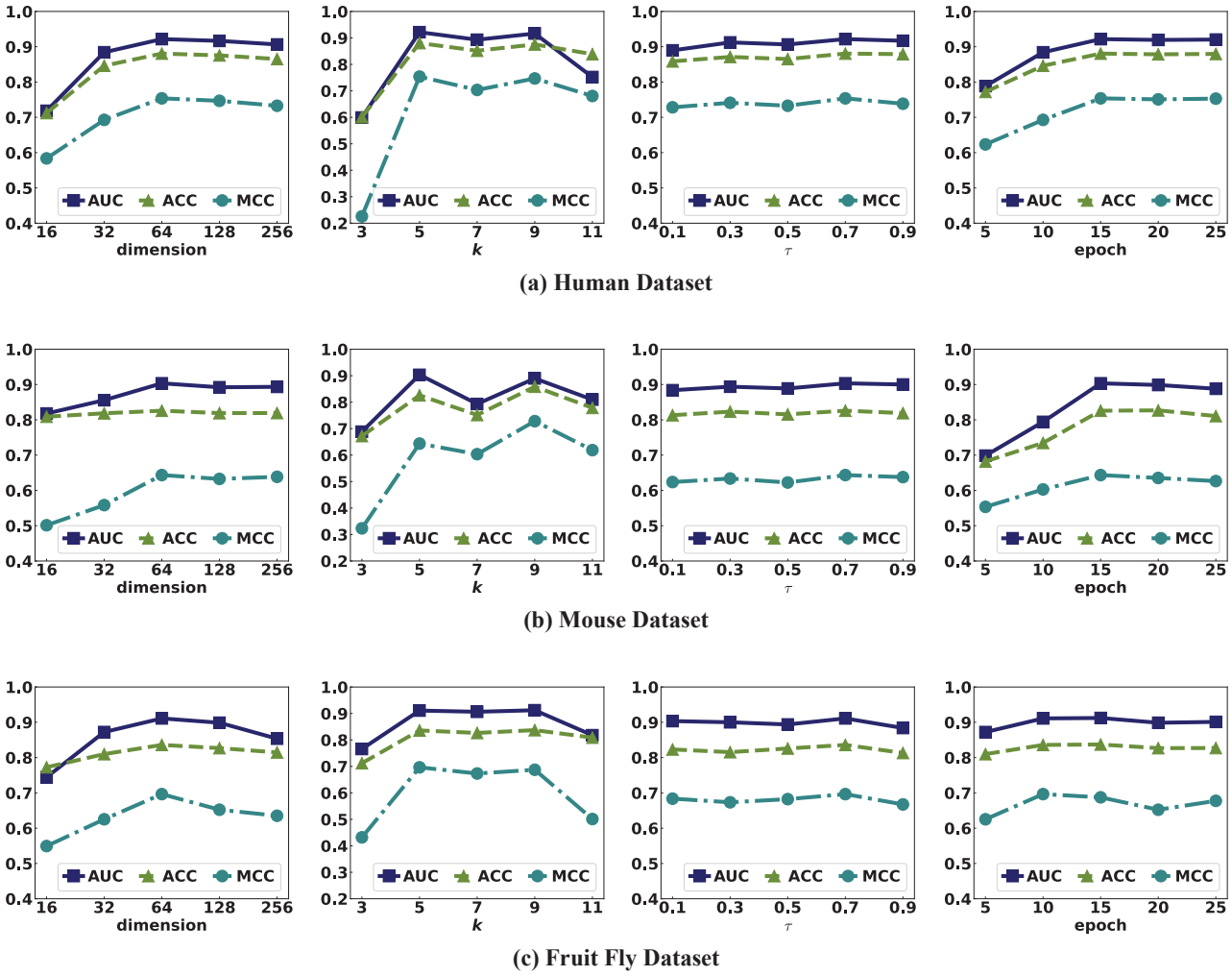

Fig. 1. The parameter sensitivity analysis.

as it is easy to ignore the important interaction signals. From the results, it can be observed that the performance fluctuation on different datasets varies with the increase of k. However, the performance of the model decreases on all three datasets when it exceeds $k = 9$.

3) Effect of τ: Next, we analyze the sensitivity of the temperature coefficients τ in two feature representations for fusing in the contrastive learning. This part observes the change of the model's performance by choosing them in the range of $\{0.1, 0.3, 0.5, 0.7, 0.9\}$. From the results in Fig. 1, it can be found that under the three evaluation metrics, the overall performance of the model on all datasets does not fluctuate greatly with the change of τ, and in general, it tends to be stable, which reveals that the model is insensitive to the hyper-parameter τ within the appropriate range, and it also indicates that the differentiation of similar samples has little effect on the fusion of feature representations in the contrastive learning.

4) Effect of epoch: Finally, we further analyze the sensitivity of the number of iterations (epoch) set for the training model, and this part observes the convergence speed of the model by choosing them in the range of $\{5, 10, 15, 20, 25\}$. From the experimental results in Fig. 1, it can be found that the convergence speed of the model is different on different datasets, which is also related to the data distribution in each dataset, but in general, the model basically reaches convergence when the number of iterations is 15, therefore, the number of training iterations of the model defaults to 15 in the setup of the experiments.

REFERENCES

[1] J. Jiang, C. J. Ju, J. Hao, M. Chen, and W. Wang, "JEDI: circular RNA prediction based on junction encoders and deep interaction among splice sites," *Bioinformatics*, vol. 37, no. Supplement, pp. 289–298, 2021.

[2] Y. He, Z. Shen, Q. Zhang, S. Wang, and D. Huang, "A survey on deep learning in DNA/RNA motif mining," *Briefings in Bioinformatics*, vol. 22, no. 4, p. bbaa229, 2021.

[3] Z. Shen, Y. Shao, W. Liu, Q. Zhang, and L. Yuan, "Prediction of back-splicing sites for circRNA formation based on convolutional neural networks," *BMC Genomics*, vol. 23, p. 581, 2022.