SCIENCE CHINA
Information Sciences

• LETTER •

August 2024, Vol. 67, Iss. 8, 189104:1–189104:2
https://doi.org/10.1007/s11432-024-4098-3

BioKG-CMI: a multi-source feature fusion model based on biological knowledge graph for predicting circRNA-miRNA interactions

Mengmeng WEI[1†], Lei WANG[1,5*†], Yang LI[4], Zhengwei LI[1,5], Bowei ZHAO[3], Xiaorui SU[3], Yu WEI[6] & Zhuhong YOU[2*]

[1] *School of Computer Science and Technology, China University of Mining and Technology, Xuzhou* 221116*, China;*
[2] *School of Computer Science, Northwestern Polytechnical University, Xi'an* 710129*, China;*
[3] *Xinjiang Technical Institute of Physics & Chemistry, Chinese Academy of Sciences, Urumqi* 830011*, China;*
[4] *School of Computer Science and Information Engineering, Hefei University of Technology, Hefei* 230601*, China;*
[5] *School of Information Science and Engineering, Zaozhuang University, Zaozhuang* 277100*, China;*
[6] *School of Information Engineering, Xijing University, Xi'an* 710123*, China*

Received 4 January 2024/Revised 12 April 2024/Accepted 19 June 2024/Published online 22 July 2024

Increasing experimental evidence has shown that circRNA has the potential to serve as a biomarker for disease diagnosis and prognosis, especially in cancer [1]. circRNA actively participates in numerous pathological processes by serving as an miRNA sponge. Consequently, the precise prediction of circRNA-miRNA interactions (CMIs) is crucial for narrowing down the scope of biological experiments and expediting the research and development of disease treatments.

Due to the time-consuming and costly nature of biological experiments, there is a pressing need for computational methods to expedite the research process. For example, Luo et al. [2] proposed a method for representing large-scale weighted networks, which was applied to describe weighted biological molecular networks. Bi et al. [3] introduced a feature analysis method based on nonnegative AutoEncoder, which can perform representation learning on missing biological data. He et al. [4] proposed a variant of the random forest algorithm, which selects features for modeling from all subsequence sets and outperforms state-of-the-art classification algorithms in terms of computational speed and accuracy.

These methods have effectively propelled research in CMI prediction and achieved commendable results. However, there are still some factors that have been overlooked. (1) Graph representation learning is a complex and challenging task, and traditional graph embedding algorithms may not fully consider the heterogeneity of entities and relationships. (2) A reasonable negative sample generation strategy can effectively utilize biological logic and enhance the prediction performance of the model. (3) The sequences of circRNAs and miRNAs harbor a wealth of biological information, but this information remains inadequately explored at present.

In this study, we propose a model named BioKG-CMI based on a biological knowledge graph that employs multi-source features to predict CMIs. The flowchart of BioKG-CMI is shown in Figure 1. Initially, BioKG-CMI performs subcellular localization by utilizing the sequence information of circRNAs and miRNAs, generating negative samples accordingly. Subsequently, we construct a biological knowledge graph containing known relationships between circRNAs and miRNAs. The DisMult algorithm is used to learn feature representations of entities and relationships in graphs. Then, the spatial proximity between nodes of the same type is calculated, and the bidirectional encoder representations from transformers (BERT) is used to learn the representation of sequence features. Finally, these features are fused and an AdaBoost classifier is used to predict potential CMIs. The results indicate that the prediction performance of the model can be effectively improved by generating negative samples through subcellular localization and adopting a multi-feature fusion strategy.

Dataset. In the experiments, we primarily utilize the CMI-9905 dataset to validate the performance of BioKG-CMI, and the CMI-9589 dataset to assess the model's generalization ability. The subcellular localization information is used to generate negative samples. For more detailed information about the datasets, please refer to Appendix A.

BioKG-CMI algorithm. BioKG-CMI introduced the pretraining algorithm BERT to learn the sequence representations of circRNAs and miRNAs. Neighboring molecules of the same type typically share similar functionalities. Based on this principle, we employ Word Mover's distance to capture spatial proximity among nodes of the same type. To capture the representation of entities and relationships in the biological knowledge graph accurately, BioKG-CMI adopts DisMult to learn low-dimensional vectors of entities based on neural networks. Please refer to Appendix B for details of the BioKG-CMI algorithm.

Results. In this study, we plotted receiver operating characteristic (ROC) curves and precision-recall (PR) curves, and calculated the area under ROC curve (AUC) and the

[*] Corresponding author (email: leiwang@cumt.edu.cn, zhuhongyou@nwpu.edu.cn)
[†] Wei M M and Wang L have the same contribution to this work.

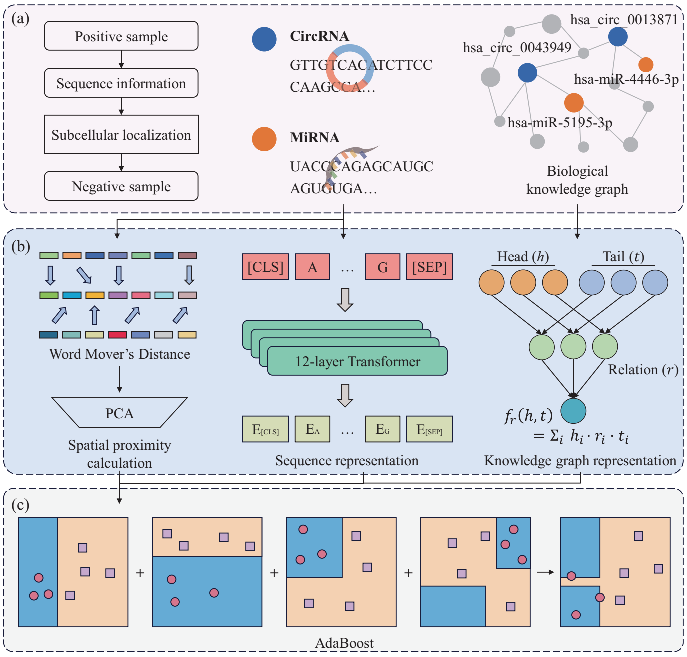

Figure 1 (Color online) Flowchart of BioKG-CMI. (a) Constructing a biological knowledge graph and generating negative samples; (b) learning multi-source features of circRNAs and miRNAs; (c) features are fused, and CMIs are predicted using AdaBoost.

area under PR curve (AUPR) as standard evaluation indicators. Besides, the accuracy rate, precision rate, specificity, etc. are also employed as indicators to validate the model. To comprehensively assess the predictive performance of the model, we first conducted validation using two benchmark datasets, CMI-9589 and CMI-9905 (Tables C1 and C2, Figures C1 and C2). The implementation of a judicious strategy for generating negative samples is pivotal in enhancing model's performance, so we compare the performance of BioKG-CMI under various strategies for generating negative samples (Figure C3). Secondly, to assess the extent to which different types of features contribute to model performance, we performed ablation experiments (Table C3). Then, we compared DisMult with other graph embedding methods and further explored the impact of knowledge graph features with different embedding dimensions on model performance (Figure C4, Table C4). Next, we also conducted comparative experiments with different classifiers (Table C5, Figure C5). Subsequently, to demonstrate the exceptional performance of BioKG-CMI, we compare it with other cutting-edge models (Tables C6 and C7). Finally, case studies related to diseases have demonstrated the practicality of the model (Table C8).

Conclusion. This study proposes a model named BioKG-CMI to predict CMIs based on a biological knowledge graph. Faced with limited data, we employ subcellular localization to generate negative samples that align more closely with biological logic. To mine semantic information in circRNA and miRNA sequences, we introduce the pre-trained model BERT to learn sequence feature representation. Guided by the hypothesis that adjacent molecules have similar functions, we calculate spatial proximity between nodes of the same class. The DisMult algorithm is applied to extract the potential logical rules of the knowledge graph and learn entity and relationship representations. Subsequently, the integration of multi-feature successfully addresses the challenge of expressing the complex biological knowledge graph and overcoming the limitation of single-feature inadequacy. Multiple comparative experiments and case studies demonstrate the robustness of the proposed model.

Discussion. However, due to limitations in the dataset, the full potential of BioKG-CMI remains to be explored. Additionally, by integrating advanced graph representation methods into the model, more representative features can be extracted [5]. In the future, we are committed to collecting more relevant data to improve the predictive capability and applicability of BioKG-CMI further.

Acknowledgements This work was supported by Natural Science Foundation of Shandong (Grant No. ZR2022LZL003), Natural Science Foundation of Guangxi (Grant No. 2022JJD170019), National Natural Science Foundation of China (Grant Nos. 62172355, 61702444), and National Science Fund for Distinguished Young Scholars of China (Grant No. 62325308).

Supporting information Appendixes A–C, Figures C1–C5, and Tables C1–C8. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Dolgin E. Why rings of RNA could be the next blockbuster drug. Nature, 2023, 622: 22–24

2 Luo X, Wu H, Wang Z, et al. A novel approach to large-scale dynamically weighted directed network representation. IEEE Trans Pattern Anal Mach Intell, 2021, 44: 9756–9773

3 Bi F, He T, Luo X. A fast nonnegative autoencoder-based approach to latent feature analysis on high-dimensional and incomplete data. IEEE Trans Serv Comput, 2024, 17: 733–746

4 He Z, Wang J, Jiang M, et al. Random subsequence forests. Inf Sci, 2024, 667: 120478

5 He T, Ong Y S, Bai L. Learning conjoint attentions for graph neural nets. In: Proceedings of Advances in Neural Information Processing Systems, 2021. 34: 2641–2653