Mathematical Biosciences and Engineering

MBE, 17(4): 2822–2824.
DOI: 10.3934/mbe.2020156
Received: 04 March 2019
Accepted: 17 March 2020
Published: 19 March 2020

http://www.aimspress.com/journal/MBE

Editorial

Towards Machine Learning in Molecular Biology

Juexin Wang[2] and Yan Wang [1,*]

[1] Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education, College of Computer Science and Technology, Jilin University, China
[2] Department of Electrical Engineering and Computer Science, and Bond Life Science Center, University of Missouri, USA

* **Correspondence:** Email: wy6868@jlu.edu.cn.

We are pleased to present the edition in Mathematical Biosciences and Engineering of a Special Issue that highlights machine learning in molecular biology. Our aim is to report latest developments both in computational methods and analysis expanding the existed biological knowledge in molecular biological systems. We feature both web-based resources, which provide easy access to users, downloadable tools of particular use for in-house processing, and the inclusion into pipelines being developed in the laboratory.

In this special issue, Zhu *et al.* [1] developed a new approach to computationally reconstruct the 3D structure of the X-chromosome during XCI, in which the chain of DNA beads representing a chromosome is stored and simulated inside a 3D cubic lattice. They first generated the 3D structures of the X-chromosome before and after XCI by applying simulated annealing and Metropolis-Hastings simulations. Then, Xist localization intensities on the X-chromosome (RAP data) are used to model the traveling speeds or acceleration between all bead pairs during the process of XCI. With their approach, the 3D structures of the X-chromosome at 3 hours, 6 hours, and 24 hours after the start of the Xist expression, which initiates the XCI process, have been reconstructed.

Long noncoding RNAs (lncRNA) play important roles in gene expression regulation in diverse biological contexts. While lncRNA-gene interactions are closely related to the occurrence and development of cancers, the new target genes could be detected from known lncRNA regulated genes. Lu *et al.* [2] developed a method by using a biclustering approach for elucidating lncRNA-gene interactions, which allows for the identification of particular expression patterns across multiple datasets, indicating networks of lncRNA and gene interactions. Their method was applied and evaluated on the breast cancer RNA-seq datasets along with a set of known lncRNA regulated genes. Their method provides useful information for future studies on lncRNAs.

RNA modification site prediction offers an insight into diverse cellular processing in the regulation

of organisms. Deep learning can detect optimal feature patterns to represent input data other than feature engineering from traditional machine learning methods. Sun *et al.* [3] developed DeepMRMP (Multiple Types RNA Modification Sites Predictor), a predictor for multiple types of RNA modifications method, which is based on the bidirectional Gated Recurrent Unit (BGRU) and transfer learning. Using multiple RNA site modification data and correlation among them, DeepMRMP build predictor for different types of RNA modification sites. DeepMRMP identifies N1-methyladenosine (m1A), pseudouridine (Ψ), 5-methylcytosine (m5C) modification sites through 10-fold cross-validation of the RNA sequences of H. sapiens, M. musculus and S. cerevisiae,

In biomedical research, near infrared spectroscopy (NIRS) is widely applied to analysis of active ingredients in medicinal fungi. Huang *et al.* [4] introduced an autonomous feature extraction method to model original NIRS vectors using attention based residual network (ABRN). Attention module in ABRN is employed to enhance feature wave bands and to decay noise. Different from traditional NIRS analysis methods, ABRN does not require any preprocessing of artificial feature selections which rely on expert experience. Comparing with other methods on various benchmarks and measurements, ABRN has better performance in autonomously extracting feature wave bands from original NIRS vectors, which can decrease the loss of tiny feature peaks.

Selectively and non-covalently interact with hormone, the soluble carrier hormone binding protein (HBP) plays an important role in the growth of human and other animals. Since experimental methods are still labor intensive and cost ineffective to identify HBP, it's necessary to develop computational methods to accurately and efficiently identify HBP. In Tan *et al.*'s paper [5], a machine learning-based method named as HBPred2.0 was proposed to identify HBP, in which the samples were encoded by using the optimal tripeptide composition obtained based on the binomial distribution method. The proposed method yielded an overall accuracy of 97.15% in the 5-fold cross-validation test. A user-friendly webserver is also provided.

Sun *et al.* [6] propose novel machine learning methods for recognition cancer biomarkers in saliva. As cancer tissues can make disease-specific changes in some salivary proteins through some mediators in the pathogenesis of systemic diseases, effectively identify these salivary proteins as potential markers is one of the challenging issues. With the proposed approach, salivary secreted proteins are recognized which are considered as candidate biomarkers of cancers. SVC-KM method is used to cluster the remaining proteins, and select negative samples from each cluster in proportion. Experimental results show the proposed methods can improve the accuracy of recognition by solving the problems of unbalanced sample size and uneven distribution in training set. They analyze the gene expression data of three types of cancer, and predict that 33 genes will appear in saliva after they are translated into proteins. This study provides a computational tool to help biologists and researchers reduce the number of candidate proteins and the cost of research in saliva diagnosis.

We hope that the readers will find this Special Issue helpful in identifying tools and analysis to help them in their study of particular molecular biological problems. In addition, this Issue is also providing an insight into current developments in bioinformatics where the articles describe the strategies being employed to exploring and interpreting sophisticate biological mechanisms, inferring underling relationships and interactions, predicting consequences from disturbance and building hypothesis in molecular biological systems.

Last but not least, we thank all the authors contributing to this special issue, and editor May Zhao's help and excellent work.

References

1. H. Zhu, N. Wang, J. Z. Sun, R. B. Pandey, Z. Wang, Inferring the three-dimensional structures of the X-chromosome during X-inactivation, *Math. Biosci. Eng.,* **16** (2019), 7384–7404.
2. S. J. Lu, J. Xie, Y. Li, B. Yu, Q. Ma, B. Q. Liu, Identification of lncRNAs-gene interactions in transcription regulation based on co-expression analysis of RNA-seq data, *Math. Biosci. Eng.,* **16** (2019), 7112–7125.
3. P. P. Sun, Y. B. Chen, B. Liu, Y. X. Gao, Y. Han, F. He, et al., DeepMRMP: A new predictor for multiple types of RNA modification sites using deep learning, *Math. Biosci. Eng.,* **16** (2019), 6231–6241.
4. L. Huang, S. Y. Guo, Y. Wang, S. Wang, Q. B. Chu, L. Li, et al., Attention based residual network for medicinal fungi near infrared spectroscopy analysis, *Math. Biosci. Eng.,* **16** (2019), 3003–3017.
5. J. X. Tan, S. H. Li, Z. M. Zhang, C. X. Chen, W. Chen, H. A. Tang, et al., Identification of hormone binding proteins based on machine learning methods, *Math. Biosci. Eng.,* **16** (2019), 2466–2480.
6. Y. Sun, W. Du, L. L. Yang, M. Dai, Z. Y. Dou, Y. X. Wang, et al., Computational methods for recognition of cancer protein markers in saliva, *Math. Biosci. Eng.,* **17** (2020), 2453–2469.

AIMS Press