IVYSPRING INTERNATIONAL PUBLISHER

International Journal of Biological Sciences

2018; 14(8): 807-810. doi: 10.7150/ijbs.27554

Editorial

Special issue on Computational Resources and Methods in Biological Sciences

Hao Lin[1,2], Shaoliang Peng[3], Jian Huang[1,2] ✉

1. Center for Informational Biology, University of Electronic Science and Technology of China (UESTC), Chengdu 611731, China
2. School of Life Science and Technology, University of Electronic Science and Technology of China (UESTC), Chengdu 610054, China
3. School of Computer Science, National University of Defense Technology, Changsha 410073, China

✉ Corresponding author: hj@uestc.edu.cn

Received: 2018.05.30; Accepted: 2018.06.03; Published: 2018.07.01

Abstract

This special issue covers a wide range of topics in computational biology, such as database construction, sequence analysis and function prediction with machine learning methods, disease-related diagnosis, drug–target and drug discovery, and electronic health record system construction.

Key words: Computational Resource, Computational Method

Computational biology has become an increasingly important research area in biological sciences in the recent decades. The new-generation technologies have become relatively cheaper and more popular in biolabs, and huge volumes of biological data output have consequently driven the biological field to a type of data science to a certain extent [1]. Computational resources and methods have since played a significant role in many biological studies [2]. Big data in biological sciences have been considered for curation enhancement, efficient analysis and precise interpretation, and accurate modeling and prediction. The special issue was focused on the recent development of computational resources and methods in biological sciences.

In this special issue, 20 studies have been published to elaborate the progresses and results of data collection and database construction, sequence analysis and classification, macromolecular structure analysis, disease-related analysis and diagnosis, and bioinformatics in biological techniques.

With the accumulation of biological data, the collection, storage, and management of these data have become the prerequisite of subsequent biological analysis. *Oryza sativa*, one of the model plants, has attracted much attention because it is not only produced as a major cereal, but also because it is easy to study genetically. Jiang *et al.* manually collected the yield-related genomics and proteomics data of rice from published databases and the literature, and then constructed a comprehensive database called RicyerDB to store these data [3]. The database comprises various yield-related information, such as genes, proteins, chromosomal locations, traits, protein–protein interaction networks, gene ontology, etc. All these data can be freely accessed at http://server.malab.cn/Ricyer/index.html. The database can also provide a relatively more convenient way to conduct yield-related research about rice. In the past ten years, an increasing number of scholars have focused on non-coding RNAs (ncRNAs) because this type of molecule plays an important role in various biological processes. Many databases about ncRNAs are currently built for different research purposes. To provide easy access and understanding of ncRNA resources for plant research, Liao *et al.* summarized 83 published ncRNA databases [4] and provided suggestions on how to use these resources and subsequently perform meaningful analysis on plant microRNAs (miRNAs) and LncRNAs.

Sequence analysis and type classification are classical topics in computational biology. On the basis of the principle that sequence determines structure and structure determines function, eight papers in the special issue have focused on different computational methods that can be used to determine the structures and functions of biological macromolecules based on their sequence information. Membrane proteins play extremely important roles in biological signal transfer and are important candidates for drug–target discovery, but their structures are different to solve experimentally. According to the sequence alignment principle, Wang *et al.* developed a segment alignment method to predict the fold of alpha-helical transmembrane proteins (αTMP) by referring to topological structures and evolutionary information [5]. The proposed method exhibited high alignment accuracy on a non-redundant dataset, and the RNA secondary structure is highly correlated with its function. Consequently, the prediction of the secondary structures of RNA has become a popular topic. However, an RNA sequence seldom provides sufficient information for RNA secondary structure prediction. Zhu *et al.* investigated the folding diversity of different RNA families and designed an RNA folding rule-based statistical learning model to predict RNA secondary structures [6].

Many studies have performed prediction by directly using sequence information [7, 8, 9]. Meiotic recombination is an important genome evolution mode, but it is not randomly distributed in a genome. Thus, to identify the region with high DNA recombination frequency, Yang *et al.* designed a support vector machine-based method to discriminate recombination hotspot from recombination coldspot by optimizing the key hexamer features via binomial distribution [10]. A comparison with other published methods demonstrated that the proposed predictor was superior to its counterparts in terms of various evaluation indexes. Based on the method, the authors established a free webserver called iRSpot-Pse6NC, which can be freely accessed from http://lin-group.cn/server/iRSpot-Pse6NC. Several proteins with special functions have also attracted the attention of scholars. Hormone-binding proteins (HBPs) can interact with hormones, but their functions are currently unclear. Thus, a bioinformatics method was designed to correctly identify HBPs and provide vital clues about their functions. Tang *et al.* collected a high-quality benchmark dataset and developed a powerful predictor named HBPred (http://lin-group.cn/server/HBPred) to assist in the recognition of HBPs by their optimal dipeptide composition features [11]. Matrix metalloprotease (MMP) is another kind of function-special protein related to metabolism. Song *et al.* presented a random forest-based method to identify peptides that bind with MMP by referring to the physicochemical properties of residues [12]. As a result, 1300 known peptide motifs that bind with 7 MMPs were identified, and these results provided vital clues and insights into candidate inhibitor drugs. Some proteins with special properties, such as the self-interacting proteins (SIPs), also play important roles in cellular regulation. A predictive method was proposed to recognize SIPs based on sequence evolution information [13], and deep neural network was utilized to perform feature selection. The cross-validations showed that the proposed method can produce highly accurate results. Wang *et al.* verified the performance of their model by comparing the method with other existing techniques [13]. Almost all biological processes are influenced by protein post-translational modifications (PTMs). Accordingly, methods for bioinformatics were developed to predict PTM sites in proteins. Liu *et al.* developed a powerful webserver called PTM-ssMP to identify PTM sites by using protein sequence information [14]. The webserver provides nine kinds of PTM prediction and is freely available at http://bioinformatics.ustc.edu.cn/PTM-ssMP/index/. All users can easily obtain the research results from the published web servers without having to refer to detailed mathematical equations.

CRISPR is a family of DNA sequences in bacteria and archaea, and it can be used as defense against phage attacks. At present, CRISPR has become the most popular biological technique for precise genome editing. However, the research community generally lacks user-friendly avenues to derive CRISPR-related information from high-throughput sequencing data. You *et al.* developed a tool called CRISPRMatch that can automatically process the high-throughput genome-editing data of CRISPR [15]. The CRISPRMatch toolkit and the original code can be freely obtained from https://github.com/zhangtaolab/CRISPRMatch.

Computational methods can be used for drug–target discovery, drug design, disease diagnosis, and health recording, which altogether can guide disease treatment. Ten studies performed intensive research on these subtopics.

Cancer is the most common complex genetic disease. Scholars have exerted considerable effort to investigate the nosogenesis of cancer and design suitable drugs. miRNAs play important roles in multiple biological processes and have great potentials as drug candidates. Based on miRNA data about target genes and tissue specificity of diseases, as well as information from the Food and Drug Administration, Yu *et al.* constructed a drug-miRNA-disease network and developed a method called miTS to predict the potential treatment of diseases [16]. The method was applied on breast cancer cases, and several new potential drugs for breast cancer treatment were found. The method provided a convenient way to determine new drug candidates to treat complex diseases. The identification of differentially expressed genes (DEGs) not only helps in understanding the nosogenesis of cancer, but it also provides vital clues for cancer treatment. However, the detection of weakly differential expression signals is difficult to conduct between two phenotypes with limited sample sizes. Cai *et al.* improved their previous method RankComp to detect DEGs between two phenotypes without the need of batch effect adjustments [17]. The method with breast cancer drug-response data was applied to detect weak differential expression signals. Gastric cancer is one of the most common causes of cancer-related death in the world, and investigating its key genes can provide important clues for drug–target discovery. Zeng *et al.* constructed a co-expression network of gastric cancers and performed a series of enrichment analysis to detect the key genes [18]. Their results found that most of the identified key genes were unique, which suggest that key genes can be regarded as the potential biomarker of a subtype, and each subtype of a gastric cancer may differ by occurrence and developmental mechanism. Gene fusion plays an important role in cancer, but the current methods on tumor DNA composition has low sensitivity. Chen *et al.* developed a highly efficient tool called GeneFuse to accurately detect and visualize gene fusions [19]. The tool can be downloaded from https://github.com/OpenGene/GeneFuse. Cholangiocarcinoma is a rare and often fatal cancer. In the era of big data, translation bioinformatics can be enhanced with new methods and knowledge on cancer research. Qian *et al.* reviewed the development of various computational methods applied in cholangiocarcinoma, which ranged from biological data classification, knowledge discovery and annotation, and clinical application [20]. They also discussed future opportunities and challenges, including data collection and precise diagnosis for cholangiocarcinoma treatment.

Autism spectrum disorder (ASD) is a complex neurodevelopmental disease. However, the molecular mechanism of ASD is unclear. Huang *et al.* analyzed the RNA-sequence data from corpus callosum of six patients with ASD and six normal individuals [21]. The co-expression module analysis for ASD-associated genes showed that the detected DEGs and ncRNAs were significantly involved in nervous system, sensory system, immune system, and organ development. These findings can help uncover core dysfunctional modules and understand the molecular mechanism of ASD.

Cancers, age-related macular degeneration, choroidal neovascularization, and many other diseases are closely related to pathological angiogenesis, which can be treated with blockers of VEGF signaling pathway. According to previous knowledge, the peptide HRH can target vascular endothelial growth factor receptors and exhibit a dose-dependent angiogenesis-suppressing effect [22]. Ning *et al.* designed a peptibody with an antiangiogenic effect by fusing the HRH peptide to a human IgG1 Fc fragment, and they obtained a new peptibody called PbHRH [23]. A series of molecular dynamics simulations analysis was also performed, and they found that PbHRH can prevent the native VEGF-A binding site of VEGFR-1 D2 from directly binding its HRH functional domain. The designed peptibody may indirectly help improve the pharmacokinetic profile and bioavailability of HRH.

Microorganisms that reside in the human body can influence human health, and the abnormal distribution of microorganisms can cause various diseases. Wu *et al.* developed a computational method called PRWHMDA to predict human micro-disease associations [24]. The model was also applied to asthma, inflammatory bowel disease, and type 1 diabetes cases. The authors experimentally found and validated the existence of disease-associated microbes, which demonstrates that the method can be used to detect disease-related microbes.

Pathophysiology signals provide key clues for disease diagnoses. Sputum sounds can be used to estimate the status of sputum deposition in a respiratory system. Shi *et al.* designed an artificial neural network-based method to recognize sputum sounds by using wavelet transform, and their method achieved highly accurate results [25]. The proposed method can contribute to the efficiency of intensive care unit (ICU) staff in achieving timely clearance of secretion among patients with mechanical ventilation. The era of big data has rendered it highly important for online diagnostic services to store, process, and transmit electronic health records (EHRs) of patients across several healthcare platforms, even among clinical researchers. Yang *et al.* applied the MicroService Architecture to construct an EHR system that can enable clinical researchers and governmental or non-governmental organizations to conveniently and directly obtain clinical-related data [26].

In summary, this special issue covers a wide range of topics in computational biology, such as database construction, sequence analysis and function prediction with machine learning methods, disease-related diagnosis, drug–target and drug discovery, and electronic health record system construction. The databases, methods, tools and discoveries published in this special issue can offer important knowledge and ways for the life sciences. Moreover, this special issue can promote the development of computational biology and bioinformatics-related fields.

References

1. He B, Jiang L, Duan Y, *et al*. Biopanning data bank 2018: hugging next generation phage display. *Database (Oxford)*. 2018; 2018 doi: 10.1093/database/bay032
2. He B, Chai G, Duan Y, *et al*. BDB: biopanning data bank. *Nucleic Acids Res*. 2016; 44(D1): D1127-1132 doi: 10.1093/nar/gkv1100
3. Jiang J, Xing F, Zeng X, *et al*. RicyerDB: A Database For Collecting Rice Yield-related Genes with Biological Analysis. *Int J Biol Sci*. 2018; 14(8): 965-970 doi: 10.7150/ijbs.23328
4. Liao P, Li S, Cui X, *et al*. A comprehensive review of web-based resources of non-coding RNAs for plant science research. *Int J Biol Sci*. 2018; 14(8): 819-832 doi: 10.7150/ijbs.24593 Review
5. Wang H, Wang J, Zhang L, *et al*. A Sequential Segment Based Alpha-Helical Transmembrane Protein Alignment Method. *Int J Biol Sci*. 2018; 14(8): 901-906 doi: 10.7150/ijbs.24327
6. Zhu Y, Xie Z, Li Y, *et al*. Research on folding diversity in statistical learning methods for RNA secondary structure prediction. *Int J Biol Sci*. 2018; 14(8): 622-632 doi: 10.7150/ijbs.24595
7. Li N, Kang J, Jiang L, *et al*. PSBinder: A Web Service for Predicting Polystyrene Surface-Binding Peptides. *Biomed Res Int*. 2017; 2017: 5761517 doi: 10.1155/2017/5761517
8. He B, Kang J, Ru B, *et al*. SABinder: A Web Service for Predicting Streptavidin-Binding Peptides. *Biomed Res Int*. 2016; 2016: 9175143 doi: 10.1155/2016/9175143
9. Kang J, Fang Y, Yao P, *et al*. NeuroPP: A Tool for the Prediction of Neuropeptide Precursors Based on Optimal Sequence Composition. *Interdiscip Sci*. 2018: doi: 10.1007/s12539-018-0287-2
10. Yang H, Qiu W-R, Liu G, *et al*. iRSpot-Pse6NC: identifying recombination spots in Saccharomyces cerevisiae by incorporating hexamer composition into general PseKNC. *Int J Biol Sci*. 2018; 14(8): 883-891 doi: 10.7150/ijbs.24616
11. Tang H, Zhao Y-W, Zou P, *et al*. HBPred: a tool to identify growth hormone-binding proteins. *Int J Biol Sci*. 2018; 14(8): 957-964 doi: 10.7150/ijbs.24174
12. Song J, Tang J, Guo F. Identification of Inhibitors of MMPS Enzymes via a Novel Computational Approach. *Int J Biol Sci*. 2018; 14(8): 863-871 doi: 10.7150/ijbs.24588
13. Wang Y-B, You Z-H, Li L-P, *et al*. Improving Prediction of Self-interacting Proteins Using Stacked Sparse Auto-Encoder with PSSM profiles. *Int J Biol Sci*. 2018; 14(8): 983-991 doi: 10.7150/ijbs.23817
14. Liu Y, Wang M, Xi J, *et al*. PTM-ssMP: A Web Server for Predicting Different Types of Post-translational Modification Sites Using Novel Site-specific Modification Profile. *Int J Biol Sci*. 2018; 14(8): 946-956 doi: 10.7150/ijbs.24121
15. You Q, Zhong Z, Ren Q, *et al*. CRISPRMatch: An Automatic Calculation and Visualization Tool for High-throughput CRISPR Genome-editing Data Analysis. *Int J Biol Sci*. 2018; 14(8): 858-862 doi: 10.7150/ijbs.24581
16. Yu L, Zhao J, Gao L. Predicting Potential Drugs for Breast Cancer based on miRNA and Tissue Specificity. *Int J Biol Sci*. 2018; 14(8): 971-982 doi: 10.7150/ijbs.23350
17. Cai H, Li X, Li J, *et al*. Identifying differentially expressed genes from cross-site integrated data based on relative expression orderings. *Int J Biol Sci*. 2018; 14(8): 892-900 doi: 10.7150/ijbs.24548
18. Zeng W, Rao N, Li Q, *et al*. Genome-wide Analyses on Single Disease Samples for Potential Biomarkers and Biological Features of Molecular Subtypes: A Case Study in Gastric Cancer. *Int J Biol Sci*. 2018; 14(8): 833-842 doi: 10.7150/ijbs.24816
19. Chen S, Liu M, Huang T, *et al*. GeneFuse: detection and visualization of target gene fusions from DNA sequencing data. *Int J Biol Sci*. 2018; 14(8): 843-848 doi: 10.7150/ijbs.24626
20. Qian F, Guo J, Jiang Z, *et al*. Translational Bioinformatics for Cholangiocarcinoma: Opportunities and Challenges. *Int J Biol Sci*. 2018; 14(8): 920-929 doi: 10.7150/ijbs.24622 Review
21. Huang Y, Chang Z, Li X, *et al*. Integrated multifactor analysis explores core dysfunctional modules in autism spectrum disorder. *Int J Biol Sci*. 2018; 14(8): 811-818 doi: 10.7150/ijbs.24624
22. Zhang Y, He B, Liu K, *et al*. A novel peptide specifically binding to VEGF receptor suppresses angiogenesis in vitro and in vivo. *Signal Transduct Target Ther*. 2017; 2: 17010 doi: 10.1038/sigtrans
23. Ning L, Li Z, Bai Z, *et al*. Computational Design of Antiangiogenic Peptibody by Fusing Human IgG1 Fc Fragment and HRH Peptide: Structural Modeling, Energetic Analysis, and Dynamics Simulation of Its Binding Potency to VEGF Receptor. *Int J Biol Sci*. 2018; 14(8): 930-937 doi: 10.7150/ijbs.24582
24. Wu C, Gao R, Zhang D, *et al*. PRWHMDA: Human Microbe-Disease Association Prediction by Random Walk on the Heterogeneous Network with PSO. *Int J Biol Sci*. 2018; 14(8): 849-857 doi: 10.7150/ijbs.24539
25. Shi Y, Wang G, Niu J, *et al*. Classification of sputum sounds using artificial neural network and wavelet transform. *Int J Biol Sci*. 2018; 14(8): 938-945 doi: 10.7150/ijbs.23855
26. Yang Y, Zu Q, Liu P, *et al*. MicroShare: Privacy-Preserved Medical Resource Sharing through MicroService Architecture. *Int J Biol Sci*. 2018; 14(8): 907-919 doi: 10.7150/ijbs.24617