Editorial

Sunny Sun and Yi-Ping Phoebe Chen

THE special issue of *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* is a collection of ten papers presented at the 19th Asia Pacific Bioinformatics Conference (APBC2021), which was a virtual conference held in Tainan city, Taiwan, from February 3 to 5 in 2021. The APBC2021 brought together academia and industry to share knowledge and experiences in bioinformatics and artificial intelligence, which covers the greatest diversity and complexity of data flows generated from various high-throughput technologies. We are grateful and proud to present the 3-day program that includes three outstanding keynotes, 63 excellent oral presentations and 3 industrial talks. As the first virtual conference in the APBC conference series, APBC2021 has drew over 120 attendees, speakers and moderators, from all over the world and in average 50-60 people online watching the conference concurrently. The maximum number of participants once reaches 80!

Year 2020 is a very special year, the global COVID-19 pandemic has brought so many changes and challenges to our working lives. We would like to thank the program committee members and external researchers who reviewed the selected papers. Special thanks also go to the editorial staff of the *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. None of these would be possible without their commitment and dedication in these unprecedented times. In the current special section, the selected papers cover comprehensive topics that include Medical informatics, application of artificial intelligence in clinical data analysis, Algorithms and Data Mining, Genomics and epigenomics, and Protein structures and functions. A brief introduction is as followings.

"Resting State Functional Connectivity Analysis During General Anesthesia: A High-Density EEG Study" by Hui Bi *et al.* present a novel sparse representation (SR) method and compare to the traditional coherence analysis (CA) method to analyze Functional connectivity (FC). They report calculated network parameters, average clustering coefficient (ACC) and average shortest path length (ASPL) before and after anesthesia. Their results showed SR method find more significant FC differences in frontal and occipital cortices, and whole brain network ($p<0.05$). In contrast, CA can hardly obtain consistent ASPL in the whole brain network ($p>0.05$). Further, ASPL calculated by SR for whole brain connections in all of three anesthesia groups increased, which can be a unified EEG biomarker of general anesthetics-induced loss of consciousness (LOC). The study concluded that FC based on SR analysis has better performance in distinguishing anesthetic-induced LOC from awake state.

"Shaking the β-Bulges" by Pierrick Craveur *et al.* investigate the dynamical behavior of β-bulges using the largest known set of protein molecular dynamics simulations. They observed that more than 50% of the existing β-bulges in protein crystal structures remained stable during dynamics while more than1/6th were not stable at all and disappeared entirely. Surprisingly, 1.1% of β-bulges that appeared remained stable. The most common β-bulges' subtypes are the smallest insertion in β-strands (namely AC and AG); they are found as stable as the whole β-bulges dataset. Low occurring types (namely PC and AS), that have the largest insertions, are significantly more stable than expected. Thus, this pioneer study allowed to precisely quantify the stability of the β-bulges, demonstrating their structural robustness, with few unexpected cases raising structural questions.

"A New Family of Similarity Measures for Scoring Confidence of Protein Interactions Using Gene Ontology" by Madhusudan Paul and Ashish Anand utilize Gene Ontology (GO) to introduce a new set of specificity measures: Relative Depth Specificity (RDS), Relative Node-based Specificity (RNS), and Relative Edge-based Specificity (RES), leading to a new family of similarity measures to obtain a confidence score for each protein-protein interaction (PPI). Using four different benchmarks, they evaluated the new measures and show that all the three measures are quite effective. Notably, RNS and RES more effectively distinguish true PPIs from false positives than the existing alternatives. RES also shows a robust set-discriminating power and can be useful for protein functional clustering as well.

"A Clinical Dataset and Various Baselines for Chromosome Instance Segmentation," by Runhua Huang *et al.* construct a clinical dataset for deep learning-based chromosome instance segmentation models and propose a chromosome instance segmentation framework and implement multiple baselines for the proposed framework based on various instance segmentation models. Experiments evaluated on the clinical dataset show that the best baseline of the proposed framework based on the Mask-RCNN model yields an outstanding outcomes that exceed results reported in current chromosome instance segmentation methods.

"SPP-CPI: Predicting Compound–Protein Interactions Based On Neural Networks" by Ying Qian *et al.* present a

- *Sunny Sun is with the Institute of Molecular Medicine, College of Medicine, National Cheng Kung University, Tainan 701, Taiwan. E-mail: hssun@mail.ncku.edu.tw.*
- *Yi-Ping Phoebe Chen is with the Department of Computer Science and Information Technology, La Trobe University, Melbourne Bundoora, VIC 3086, Australia.. E-mail: phoebe.chen@latrobe.edu.au.*

Digital Object Identifier no. 10.1109/TCBB.2021.3089195

unique method that uses SPP (Spatial pyramid pooling)-net to extract compound features and extracts protein features through the natural language processing method (doc2vec) to obtain sequence semantic information, thus it represents a compound with a distance matrix. The authors evaluated their method on three benchmark datasets—human, C. elegans, and DUDE—and the experimental results demonstrate that the proposed model presents competitive performance against state-of-the-art predictors. They also carried out drug–drug interaction (DDI) experiments to verify the strong potential of distance matrix as molecular characteristics. The source code and datasets are available at https://github.com/lxlsu/SPP_CPI.

"MultiTrans: An Algorithm for Path Extraction Through Mixed Integer Linear Programming for Transcriptome Assembly" by Jin Zhao *et al.* formulate the transcriptome assembly problem as path extraction on splicing graphs (or assembly graphs), and propose a novel algorithm MultiTrans for path extraction using mixed integer linear programming. MultiTrans takes into consideration coverage constraints on vertices and edges, the number of paths and the paired-end information simultaneously. They benchmarked MultiTrans against two state-of-the-art transcriptome assemblers, TransLiG and rnaSPAdes, and results show that MultiTrans generates more accurate transcripts compared to the existing tools. MultiTrans is freely available at https://github.com/jzbio/MultiTrans.

"GraphPlas: Refined Classification of Plasmid Sequences Using Assembly Graphs" by Anuradha Wickramarachchi and Yu Lin present a novel approach for plasmid recovery using coverage, composition and assembly graph topology. They evaluated GraphPlas on simulated and real short read assemblies with varying compositions of plasmids and chromosomes and the experiments show that GraphPlas is able to significantly improve accuracy in detecting plasmid and chromosomal contigs on top of popular state-of-the-art plasmid detection tools.

"Leveraging Sequential and Spatial Neighbors Information by Using CNNs Linked With GCNs for Paratope Prediction" by Shuai Lu *et al.* propose a method to identify which amino acid residues of an antibody directly interact with its associated antigen based on the features from sequence and structure. The algorithm uses convolution neural networks (CNNs) linked with graph convolution networks (GCNs) to make use of information from both sequential and spatial neighbors to understand more about the local environment of target amino acid residue. Furthermore, they process the antigen partner of an antibody by employing an attention layer.

"An Aggregation Method to Identify the RNA Meta-Stable Secondary Structure and its Functionally Interpretable Structure Ensemble" by Tzu-Hsien Yang proposes a novel method to identify the functionally interpretable structure ensemble of a given RNA sequence and provide the metastable structure, or the most frequently observed functional RNA cellular conformation, based on the ensemble. The proposed method outperformed existing tools on a yeast test set. The inferred functional aspects were then manually checked and demonstrated a micro-averaging F1 value of 0.92. Further, a biological example of the yeast ASH1-E1 element was discussed to articulate that these functional aspects can also suggest testable hypotheses. Then the proposed method was verified to be well applicable to other species through a human test set. Finally, the proposed method was demonstrated to show resistance to sequence length-dependent performance deterioration.

"An Extensive Examination of Discovering 5-Methylcytosine Sites in Genome-Wide DNA Promoters Using Machine Learning Based Approaches" by Trinh-Trung-Duong Nguyen *et al.* compare the effectiveness of the most popular and strong machine learning techniques namely XGBoost, Random Forest, Deep Forest, and Deep Feedforward Neural Network in predicting the 5mC sites of genome-wide DNA promoters. A feature extraction method based on k-mers embeddings learned from a language model were also applied. Overall, the performance of all the surveyed models surpassed deep learning models of the latest studies on the same dataset employing other encoding scheme. Furthermore, the best model achieved AUC scores of 0.962 on both cross-validation and independent test data. Thus they conclude that the current approach was efficient for identifying 5mC sites of promoters with high performance.

Sunny Sun received the PhD degree from the University of Wisconsin-Madison, USA. She is currently a professor and the director of molecular medicine with the National Cheng Kung University (NCKU), Taiwan. She has been involved in genomic researches for the past 20 years. Her research focuses on the application of computational strategy and advanced molecular biology techniques to identify genes responsible for various human diseases and study the regulation of gene expression in physiological and pathological conditions. She is also the director of Center for Genomic Medicine, NCKU, where she leads a team to provide services for clinical diagnosis and supports for clinical research.

Yi-Ping Phoebe Chen received the BInfTech degree (with first-class honors) and the PhD degree from the University of Queensland. She is currently a professor and the chair with the Department of Computer Science and Information Technology, La Trobe University, Melbourne, Australia. She is also the chief investigator with the ARC Centre of Excellence in Bioinformatics. She is currently working on knowledge discovery technologies and is especially interested in their application to genomics and biomedical science. She has been working in bioinformatics, health informatics, multimedia databases, query system, and systems biology, and has coauthored more than 250 research papers, with many published in top journals and conferences. Her research focuses to find the best solutions for mining, integrating, and analyzing complex data structure and functions for scientific and biomedical applications. She is the steering committee chair of the Asia-Pacific Bioinformatics Conference (founder) and the International Conference on Multimedia Modelling. She has been on the program committees of more than 100 international conferences, including top ranking conferences, such as ICDE, ICPR, ISMB, and CIKM.