Guest Editorial for Selected Papers From BIOKDD 2021

Da Yan, Zhaohui S. Qin, Debswapna Bhattacharya, and Jake Y. Chen

THE 20th International Workshop on Data Mining in Bioinformatics (BIOKDD 2021) was held virtually on August 15, 2021 due to the COVID-19 pandemic. BIOKDD 2021 featured the special theme of "Artificial Intelligence in Medicine" which particularly welcomed paper submissions and invited talks related to the use of machine learning and data mining techniques for the analysis of large amounts of heterogeneous complex biological and medical data, with a particular focus on deep learning methods that see fast advancement and wider adoption in Bioinformatics. As a whole-day workshop, BIOKDD 2021 accepted 9 submissions for oral presentation, and has 7 additional invited talks by domain experts. These presentations were divided into 4 sessions: (1) Structural Bioinformatics, (2) Clinical Informatics, (3) Bioinformatics, and (4) Network Biology and Machine Learning.

This special section of *TCBB* features the extended versions of 5 quality papers presented in BIOKDD 2021. Each of the 5 invited papers was reviewed by 3 reviewers invited by the *TCBB* guest editors, and the BIOKDD workshop reviews were shared with the TCBB reviewers. The papers also went through 1 to 2 rounds of revisions.

The first invited paper, "A Knowledge Graph-Enhanced Tensor Factorisation Model for Discovering Drug Target," by Cheng Ye, Rowan Swiers, Stephen Bonner, and Ian Barrett explores the use of machine learning classification algorithms and both matrix and tensor factorisation techniques to predict the clinical outcomes of unseen gene target-disease pairs. A 3D data tensor was created consisting of 1048 gene targets, 860 diseases and 230,011 evidence attributes and clinical outcomes connecting them, using data extracted from the Open Targets and PharmaProjects databases. The data is enriched with gene target representations learned from a drug discovery-oriented knowledge graph. Their results show that incorporating knowledge graph embeddings significantly improves the prediction accuracy and that training tensor factorisation alongside a dense neural network outperforms all other baselines.

The second invited paper, "MuCoMiD: A Multitask Graph Convolutional Learning Framework for miRNA-Disease Association Prediction," by Ngan Dong, Stefanie Mücke, and Megha Khosla studies the use of a multitask graph convolutional learning framework called MUCOMID for the problem of predicting miRNA-disease associations. Their approach allows automatic feature extraction while incorporating knowledge from five heterogeneous biological information sources in a multitask setting: associations between miRNAs/diseases and protein-coding genes (PCGs), interactions between protein-coding genes, miRNA family information, and disease ontology. Incorporating multiple sources of information helps compensate for the lack of information in any single source and, at the same time, enables the model to generate predictions for any new miRNA or disease. Their model can be employed in both transductive and inductive settings.

The third invited paper, "Heterogeneous Multi-Task Learning with Expert Diversity," by Raquel Aoki, Frederick Tung, and Gabriel L. Oliveira predicts multiple heterogeneous biological and medical targets simultaneously using multi-task learning (MTL). Their model, Multi-gate Mixture-of-Experts with Exclusivity (MMoEEx), optimizes multiple tasks with different characteristics by inducing more diversity among experts, thus creating representations more suitable for highly imbalanced and heterogenous MTL learning. This is realized with two mechanisms, exclusion and exclusivity, under which some experts only contribute to some tasks, while other experts are shared among all tasks. A two-step optimization approach inspired by MAML is also used to balance the tasks at the gradient level. The approach is validated on three MTL benchmark datasets, including UCI-Census-income dataset, Medical Information Mart for Intensive Care (MIMIC-III) and PubChem BioAssay (PCBA).

The fourth invited paper, "Biocode: A Data-Driven Procedure to Learn the Growth of Biological Networks," by Emre Sefer proposes Biocode, a framework to automatically discover novel biological growth models matching user-specified graph attributes in directed and undirected biological graphs. Such probabilistic biological network growth models have been utilized for tasks such as capturing mechanism and dynamics of biological growth activities, and capturing anomalies. Biocode designs a basic set of instructions which are common enough to model a number of well-known biological graph growth models such as Kronecker model, preferential attachment model, and duplication-based model. These instruction-wise representation are combined with a genetic

- *Da Yan is with the Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL 35294 USA. E-mail: yanda@uab.edu.*
- *Zhaohui S. Qin is with the Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA 30322 USA. E-mail: zhaohui.qin@emory.edu.*
- *Debswapna Bhattacharya is with the Department of Computer Science, Virginia Tech, Blacksburg, VA 24061 USA. E-mail: dbhattacharya@vt.edu.*
- *Jake Y. Chen is with the Informatics Institute, University of Alabama at Birmingham, Birmingham, AL 35294 USA. E-mail: jakechen@uab.edu.*

Date of current version 8 December 2022.
Digital Object Identifier no. 10.1109/TCBB.2022.3208759

algorithm based optimization procedure to encode models for various biological networks. The performance of Biocode has been evaluated in discovering models for biological collaboration networks, gene regulatory networks, and protein interaction networks with features such as assortativity, clustering coefficient, degree distribution closely match with the true ones in the corresponding real biological networks.

The fifth invited paper, "Finding Overlapping Rmaps via Clustering," by Kingshuk Mukherjee, Daniel Dole-Muinos, Massimiliano Rossi, Ayomide Ajayi, Mattia Prosperi, and Christina Boucher considers the context of optical mapping which is a method for creating high resolution restriction maps of an entire genome. Optical mapping first produces single molecule restriction maps, called Rmaps, which are assembled to generate genome wide optical maps. This work develops a method, called OMCLUST, for finding overlapping Rmaps that uses Gaussian mixture model clustering, and does not require any quantization. Their work demonstrates that OMCLUST not only achieves the highest precision and was more efficient than competing methods, but can also be integrated into the error correction methods to improve their quality of error correction. OMCLUST may serve as a filtering step for finding related Rmaps for error correction, assembly or other applications of optical mapping data.

Acknowledgments

As guest editors of this special section, we would like to thank the contributing authors, BIOKDD 2021 program committee, the TCBB reviewers who reviewed papers in this special section, and the TCBB staff for the support to make this special section possible.

Da Yan is an assistant professor of computer science with the University of Alabama at Birmingham (UAB). He was the sole winner of Hong Kong 2015 Young Scientist Award in Physical/Mathematical Science. His research interests include expertise lies in developing scalable systems and algorithms for Big Data analytics, with experience in Data Science projects on bioinformatics. He frequently serves in the program committee of conferences such as SIGMOD, VLDB, SIGKDD, ICDE, and AAAI, and serves as reviewers of journals such as *ACM Transactions on Database Systems*, *VLDB Journal*, *IEEE Transactions on Parallel and Distributed Systems*, and *IEEE Transactions on Knowledge and Data Engineering*, where he also frequently publishes his work. He has organized the BIOKDD workshops since 2018.

Zhaohui S. Qin received the PhD degree in statistics from the University of Michigan. He is a professor of biostatistics and bioinformatics with Emory University. His research interests include focused on developing statistical and machine learning methods to analyze data generated from high-throughput technologies, and on developing computer programs so that the methods can be easily adopted by the research community. He is also actively collaborating with scientists and clinicians to better understand complex human diseases using genetics, genomics and epigenomics.

Debswapna Bhattacharya is an associate professor with the Department of Computer Science, Virginia Tech. His research interests include the intersection of computational biology and applied machine learning. He is a recipient of the NSF CAREER Award and NIH Maximizing Investigators' Research Award (MIRA).

Jake Y. Chen is the chief bioinformatics officer with the Informatics Institute, University of Alabama at Birmingham and, a tenured professor of genetics, computer science, and biomedical engineering, the past President of the Midsouth Computational Biology and Bioinformatics Society. He has more than 25 years of R&D experience in biological data mining and systems biology with more than 190 peer-reviewed publications and more than 200 invited talks worldwide on bioinformatics methodologies and biomedical applications. At UAB, he leads the AI.MED laboratory (http://aimed-lab.org/) to advance multi-omics modeling and artificial intelligence application in medicine. He is an ACM distinguished scientist and an elected fellow of the American College of Medical Informatics (ACMI) and of the American Institute of Medical and Biological Engineering (AIMBE). He also serves on the editorial boards of *BMC Bioinformatics*, *Journal of American Medical Informatics Association (JAMIA)*, and *Frontiers in Artificial Intelligence and Big Data*. He was recognized as one of the "17 Informatics Experts Worth Listening To" by HealthTechTopia (2011), as a finalist for the "Indiana's Technology Educator of the Year" Award (2012-2014), and as one of the "Top 100 AI Leaders in Drug Discovery and Healthcare" by Deep Knowledge Analytics (2019).