Guest Editorial for Selected Papers From BIOKDD 2019

Da Yan, Sharma Thankachan, and Jake Y. Chen

BIOKDD 2019 Overview. The International Workshop on Data Mining in Bioinformatics (BIOKDD), held in conjunction with the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining for 18 years, has successfully established an annual forum for researchers and practitioners to present and discuss advances in data mining techniques that primarily target biological data.

The 18th BIOKDD workshop (BIOKDD 2019) was held on August 5, 2019 in Anchorage, Alaska, and was part of SIGKDD 2019 Health Day (a special theme day initiated in 2018) including 3 workshops and BIOKDD 2019 was one of them. BIOKDD 2019 accepted 10 submissions in total.

The morning session is the Health Day Plenary Session with 3 keynote talks and 1 panel, including a BIOKDD keynote delivered by Dr. Ananth Kalyanaraman on "Scalable Structure Discovery for Life Science Applications using Topological Data Analysis", as well as another keynote delivered by our PC member Dr. Tae Hyun Hwang on "Lessons learned for deploying machine learning and AI solutions in the healthcare system: current status, challenges, and opportunities."

There are two afternoon sessions each presenting 5 papers accepted by BIOKDD 2019, followed by a Health Day Poster Session, and then the KDD Poster Session in the evening.

Special Issue Overview. This special issue of *TCBB* features the extended versions of 3 quality papers presented in BIOKDD 2019. Each invited paper was reviewed by at least 2 additional reviewers invited by the *TCBB* guest editors and the workshop reviews were shared with them. The papers also went through 1 to 2 rounds of revisions.

The first invited paper, "A pipeline for integrated theory and data-driven modeling of biomedical data," [1] by Vineet Raghu, Xiaoyu Ge, Arun Balajiee, Daniel J. Shirer, Isha Das, Panayiotis Benos, and Panos K. Chrysanthis [1] presented a pipeline for knowledge discovery from integrated genomic and clinical data. The study emphasized the importance of reinforce high-throughput genome measurements with phenotypic, environmental, and behavioral data from individuals, to understand causes of disease and the effects of medical interventions. The pipeline begins with a novel variable selection method, and uses a probabilistic graphical model to understand the relationships between features in the data. The work demonstrated how this pipeline can improve breast cancer outcome prediction models, and can provide a biologically interpretable view of sequencing data.

The second invited paper, "Biomedical Knowledge Graphs Construction from Conditional Statements," [2] by Tianwen Jiang, Qingkai Zeng, Tong Zhao, Bing Qin, Ting Liu, Nitesh Chawla, and Meng Jiang [2] indicated that existing biomedical knowledge graphs (BioKGs) only focus on factual knowledge and ignore the conditions for the facts being valid. To fill the gap, the study considered both facts and their conditions in biomedical statements, and proposed a three-layered information-lossless representation of BioKG where the first, second, and their layers refer to biomedical concepts/attributes, biomedical fact and condition tuples, and biomedical statements, respectively. BioKG construction was solved as a sequence labeling problem based on a novel designed tag schema, and a Multi-Input Multi-Output sequence labeling model (MIMO) was designed to learn from multiple input signals and to generate proper number of multiple output sequences for tuple extraction. The authors showed that the BioKGs constructed provide a good understanding of the biomedical statements.

The third invited paper, "Revisiting Parameter Estimation in Biological Networks: Influence of Symmetries," by Jithin K. Sreedharan, Krzysztof Turowski, and Wojciech Szpankowski [3] showed that existing parameter estimation techniques for biological graph models overlook the critical property of graph symmetry (also known formally as graph automorphisms), thus the estimated parameters give statistically insignificant results concerning the observed biological network. To demonstrate this observation and to develop more accurate estimation procedures, the authors focus on the biologically inspired duplication-divergence model, and the up-to-date data of protein-protein interactions of seven species including human and yeast. Using exact recurrence relations of some prominent graph statistics, a parameter estimation technique was devised that provides the right order of symmetries and that uses phylogenetically old proteins as the choice of seed graph nodes. The results were found consistent with the ones obtained from maximum likelihood estimation (MLE), but

- *Da Yan is with the Department of Computer Science, University of Alabama at Birmingham (UAB), Birmingham, AL 35294 USA. E-mail: yanda@uab.edu.*
- *Sharma Thankachan is with the Department of Computer Science, University of Central Florida, Orlando, FL 32816 USA. E-mail: sharma.thankachan@ucf.edu.*
- *Jake Y. Chen is with the UAB Informatics Institute, Genetics, Computer Science, and Informatics Section of the UAB Department of Genetics, University of Alabama at Birmingham, Birmingham, AL 35294 USA. E-mail: jakechen@uab.edu.*

Digital Object Identifier no. 10.1109/TCBB.2021.3067071

the proposed methods were found to be significantly faster than the MLE approach.

Acknowledgments

As guest editors of this special issue, we would like to thank the contributing authors, BIOKDD 2019 program committee, the TCBB reviewers who reviewed papers in this special issue, and the TCBB staff for the support to make this special issue possible.

References

[1] V. Raghu *et al.*, "A pipeline for integrated theory and data-driven modeling of biomedical data," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 18, no. 3, pp. 811–822, May/June 2021.
[2] T. Jiang *et al.*, "Biomedical knowledge graphs construction from conditional statements," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 18, no. 3, pp. 823–835, May/June 2021.
[3] J. Sreedharan, K. Turowski, and W. Szpankowski, "Revisiting parameter estimation in biological networks: Influence of symmetries," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 18, no. 3, pp. 836–849, May/June 2021.

Da Yan is an assistant professor with the Department of Computer Science, University of Alabama at Birmingham (UAB), Birmingham, Alabama. Before joining UAB, he was an established data scientist in Hong Kong, and was the sole winner of Hong Kong 2015 Young Scientist Award in Physical/Mathematical Science. His research expertise lies in developing scalable systems and algorithms for Big Data analytics, with experience in Data Science projects on bioinformatics. He frequently publishes in conferences such as SIGMOD, VLDB, ICDE, SIGKDD, AAAI, and he also regularly serves in the program committee of conferences such as SIGKDD 2020–2021, SIGMOD 2019–2021, VLDB 2018 and 2021, IJCAI 2017 and 2021 (SPC), 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*. He developed a series of systems for Big Data analytics, which are of high impact and are now being used by many researchers and companies. As an expert in Big Data and Data Science, He has been invited to publish surveys and books as the first author in prestigious venues such as Foundations and Trends® in Databases and Springer Briefs in Computer Science. He has organized workshops including BIOKDD 2018–2020 with SIGKDD, and DMBIH 2019 with ICDM.

Sharma Thankachan received the BTech degree in electrical and electronics engineering from the National Institute of Technology Calicut, India, in 2006, and the PhD degree in computer science from Louisiana State University, Baton Rouge, Louisiana, in 2014. He is an assistant professor with the Department of Computer Science, University of Central Florida, Orlando, Florida. His research interests include string algorithms, computational biology, and succinct data structures.

Jake Y. Chen received the BS degree in biochemistry and molecular biology from Peking University, China, and both MS and PhD degrees in computer science and engineering from the University of Minnesota, Minneapolis, Minnesota. He is a professor of genetics and computer science, chief bioinformatics officer of the newly established Informatics Institute, and head of the Informatics Section of the Genetics Department, University of Alabama at Birmingham, Birmingham, Alabama. He has more than 20 years of bioinformatics R&D experience, including biological data mining, computational systems biology, and translational bioinformatics, with more than 150 peer-reviewed publications. His research focuses on building quantitative biomolecular systems models from genomic and clinical big data, thus helping understand, simulate, and predict complex disease biology outcomes. Prior to joining UAB, he holds tenured faculty positions at Indiana University School of Informatics and Computing and at Purdue University Computer and Information Science Department. He is also an entrepreneur who created several startup companies to make emerging biomedical data easy to interpret and use by growing Medicine 2.0 stakeholders.