Multiscale and Multimodal Analysis for Computational Biology

Xin Gao, Jake Y. Chen, and Mohammed J. Zaki

The 16th International Workshop on Data Mining in Bioinformatics (BIOKDD17) was organized in conjunction with the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining on August 14, 2017 in Halifax, Canada. It brought together international researchers in the interacting disciplines of data mining, medical informatics, and bioinformatics at the World Trade and Convention Centre venue. The goal of this workshop is to encourage Knowledge Discovery and Data mining (KDD) researchers to take on the numerous challenges that Bioinformatics offers. This year, the workshop featured the theme of “Multiscale and multimodal analysis for computational biology”.

Bioinformatics is the science of managing, mining, and interpreting information from biological data. Various genome projects have contributed to an exponential growth in DNA and protein sequence databases. Rapid advances in high-throughput technologies, such as microarrays, mass spectrometry and new/next-generation sequencing, can monitor quantitatively the presence or activity of thousands of genes, RNAs, proteins, metabolites, and compounds in a given biological state. The ongoing influx of these data, the pressing need to address complex biomedical challenges, and the gap between the two have collectively created exciting opportunities for data mining researchers.

While tremendous progress has been made over the years, many of the fundamental problems in bioinformatics, such as protein structure prediction, gene-environment interaction, and regulatory network mapping, have not been convincingly addressed. Besides these, new technologies such as next-generation sequencing are now producing massive amounts of sequence data; managing, mining and compressing these data raise challenging issues. Finally, there is a pressing need to use these data coupled with efficient and effective computational techniques to build models of complex biological processes and disease phenotypes. Data mining plays an essential role in addressing these fundamental problems and in the development of novel therapeutic/diagnostic/prognostic solutions in the post-genomics era of medicine.

This year, the workshop featured the theme of “Multiscale and multimodal analysis for computational biology”. The focus is on the use of data mining and machine learning approaches for the analysis of the large amount of heterogeneous complex biological and medical data being generated. The direction of deep learning methods was also particularly encouraged. The goal here is to build accurate predictive or descriptive models from these data enabling novel discoveries in basic biology and medicine.

Papers for this special section were selected from the BIOKDD17 workshop. In addition, two papers were invited to the special issue. To meet the acceptance criteria for the IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), each of the papers selected underwent additional reviews by at least one reviewer managed by the TCBB guest editors. We are very grateful to the anonymous reviewers in helping us select the following papers for this special section.

The first paper, “Using machine learning to improve the prediction of functional outcome in ischemic stroke patients”, by Miguel Monteiro, Ana Catarina Fonseca, Ana Teresa Freitas, Teresa Pinho e Melo, Alexandre P. Francisco, José M. Ferro, and Arlindo L. Oliveira [1], explores the application of machine learning techniques in predicting the functional outcome of ischemic stroke patients. Ischemic stroke is a leading cause of disability and death worldwide among adults. Despite advances in treatment, around one-third of surviving patients still live with long-term disability. In this study, they first show that only using features available at admission, different machine learning techniques have only marginal improvements over the baseline. However, when progressively including features available at further time points, the AUC can be significantly increased.

The second paper, “ANTENNA, a multi-rank, multi-layered recommender system for inferring reliable drug-gene-disease associations: repurposing diazoxide as a targeted anti-cancer therapy”, by Annie Wang, Hansaim Lim, Shu-Yuan Cheng, and Lei Xie [2], proposes an interesting multi-rank, multi-layered recommender system, ANTENNA, to integrate and mine large-scale chemical genomics and disease association data for the prediction of novel drug-gene-disease associations. ANTENNA integrates a novel tri-factorization based dual-regularized weighted and imputed one class collaborative filtering al-
algorithm with a statistical framework that is based on random walk with restart and can assess the reliability of a specific prediction. Using their method, they discover that FDA-approved drug diazoxide can inhibit multiple kinase genes whose malfunction is responsible for many diseases including cancer, and can kill triple negative breast cancer cells effectively at a low concentration.

The third paper, “Predicting hospital readmission via cost-sensitive deep learning”, by Haishuai Wang, Zhicheng Cui, Yixin Chen, Michael Avidan, Arbi Ben Abdallah, and Alexander Kronzer [3], combines the power of convolutional neural networks (CNN) and feature embedding to predict hospital readmission. They apply CNN to automatically learn features from time series of vital signs, and embed categorical features to encode feature vectors with heterogeneous clinical features, such as demographics, hospitalization history, vital signs and laboratory tests. Both CNN features and embedding features are fed into a multilayer perception for prediction. They validate the proposed method on real medical datasets from Barnes-Jewish Hospital. The 30-day readmission prediction accuracy is significantly higher than all the baseline methods.

The fourth paper, “Bioinformatic workflow extraction from texts: combining machine learning and ontologies”, by Ahmed Halioui, Petko Valtchev, and Abdoulaye Baniré Diallo [4], proposes an ontology-based extraction framework to acquire semantically rich workflows from texts. Their method extends the classic NLP techniques to extract and disambiguate workflow tasks using a dedicated word sense disambiguation method for bioinformatics tools. To validate their method, they apply the method to phylogenetic analyses and show promising results in the identification of semantically enriched bioinformatic workflows from scientific texts.

The fifth paper, “Super gene set’ casual relationship discovery from functional genomics data”, by Zongliang Yue, Michael T. Neylon, Thanh Nguyen, Timothy Ratliff, and Jake J. Chen [5], presents a powerful framework to identify stimulatory and inhibitory regulatory relationships among super gene sets, including pathways, annotated lists, and gene signatures (PAG). They extend their previous work on identifying PAG-to-PAG relationships by further requiring them to be significantly enriched with gene-to-gene co-expression pairs across the two PAGs involved. Comprehensive experiments on a functional genomics benchmark dataset from the GEO database and a myeloid-derived suppressor cells dataset demonstrate the effectiveness of the proposed method.

The last paper, “A slice-based 13C-detected NMR spin system forming and resonance assignment method”, by M. Alazmi, A. Abbas, X. Guo, M. Fan, L. Li, and X. Gao [6], proposes the first spin system forming method for 13C-detected NMR spectra. 13C-detected spectra provide possible solutions to solve larger proteins by NMR, owing to the fact that such spectra have less overlapping signals and better connectivity than the 1H-detected spectra. They develop a slice picking method and use the correlated information in slices from different spectra to form spin systems and assign resonance. They feed the assigned chemical shifts to CS-ROSETTA and accurately determine the 3D structure of Ubiquitin.

To conclude, we thank the authors and the reviewers for their contribution to this special section of the TCBB journal. We also thank Prof. Dong Xu and Prof. Sun Kim for their support as the associate editor-in-chief and assistance from the editorial staff at TCBB for making this special issue possible.

REFERENCES


Xin Gao is an associate professor, and group leader of Structural and Functional Bioinformatics (SFB) group in Computational Bioscience Research Center, and in Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST). He received the BS and PhD degrees in computer science from Tsinghua University, China, and University of Waterloo, Canada, respectively. Before joining KAUST, he was a Lane Fellow in Lane Center for Computational Biology at Carnegie Mellon University, US. His research interests include bioinformatics, computational biology, machine learning and algorithms. His group works on building computational models, developing machine learning techniques, and designing efficient and effective algorithms to solve key open problems along the path from genome-scale sequence analysis, to protein structure prediction/determination, to function annotation, and to understanding and controlling molecule behaviors in complex biological systems. He has published more than 140 articles in leading journals and conferences in the fields of bioinformatics and machine learning.
Jake Y. Chen Ph.D., is the Chief Bioinformatics Officer and Associate Director of the Informatics Institute, Chief of the Informatics Section of the Genetics Department, and Professor of Genetics and Computer Science of the University of Alabama at Birmingham. He holds a BS degree in Biochemistry and Molecular Biology from Peking University of China and MS/PhD degrees in Computer Science and Engineering from the University of Minnesota at Twin Cities.

In the past two decades, Dr. Chen has been studying "translational bioinformatics" - the science of collecting, representing, storing, retrieving and processing data and knowledge for the improvement of human health. His research interest focuses on systems biology, data mining, advanced visual analytics for therapeutic discovery and clinical decision-making applications. He has published more than 130 peer-reviewed scientific papers and edited two bioinformatics books, "Biological Database Modeling" and "Biological Data Mining". He has organized many Bioinformatics conferences, among which the BIOKDD workshop at the Annual ACM Knowledge Discovery and Data Mining Conference have been running consecutively for nearly two decades. He is a frequent grant review panelist for NIH, NSF, and DOD in bioinformatics. He was named a MIRA Award finalist as "Indiana's Technology Educator of the Year" for his contribution to bioinformatics research, education, and entrepreneurship in the state of Indiana.

Prior to joining UAB, Dr. Chen was the founding director of the Indiana Center for Systems Biology and Personalized Medicine at Indiana University, a tenured bioinformatics faculty at the Indiana University School of Informatics, and a tenured faculty of computer science at Purdue University Department of Computer and Information Science at IUPUI. Through his international collaboration efforts since 2012, he also led to create the Center for Biomedical Big Data and Applications at the First Affiliated Hospital of Wenzhou Medical University-recognized among the best IT-intensive large hospitals in China.

Mohammed J. Zaki received the PhD degree in computer science from the University of Rochester in 1998. He is a professor of computer science at RPI. His research interests focus on developing novel data mining techniques, especially for applications in bioinformatics and social networks. He has over 225 publications, including the textbook, Data Mining and Analysis: Fundamental Concepts and Algorithms, Cambridge University Press, 2014. He is an area editor for Statistical Analysis and Data Mining, and an associate editor for Data Mining and Knowledge Discovery, the ACM Transactions on Knowledge Discovery from Data, and Social Networks and Mining. He was the program co-chair for SDM08, SIGKDD09, PAKDD10, BIBM11, CIKM12, ICDM12, BigData15. He is currently serving on the Board of Directors for ACM SIGKDD. He is a recipient of the US NSF CAREER Award and the DOE Early Career Award. He is an ACM Distinguished Scientist and a Fellow of the IEEE. His research is supported in part by the US National Science Foundation (NSF), NIH, DOE, Google, HP, and Nvidia.