Introduction

Yunyun Zhou

Department of Data Science,
The University of Mississippi Medical Center,
Jackson, MS 78249, USA
Email: yzhou.umc@gmail.com

Junqing Wang

Department of Surgery,
Ruijin Hospital,
Shanghai Jiao Tong University School of Medicine,
Shanghai, 20025, China
Email: wangjunqingmd@163.com

Xiaofeng Song

Department of Biomedical Engineering,
Nanjing University of Aeronautics and Astronautics,
Nanjing, 210016, China
Email: xfsong@nuaa.edu.cn

Zhongming Zhao

Center for Precision Health,
School of Biomedical Informatics,
The University of Texas Health Science Center
at Houston Houston, TX 77030, USA
Email: zhongming.zhao@uth.tmc.edu

Biographical notes: Yunyun Zhou received her PhD in the School of Electrical Engineering and Computer Science from Washington State University in 2012. Later, she worked as the computational biologist in University of Texas Southwestern Medical Center at Dallas. Currently, she is an Assistant Professor in the Department of Data Science, John D. Bower School of Population Health at the University of Mississippi Medical Center. Her research interests include the development of bioinformatics and biostatistics tools for systems biology data analysis, biomarker discovery for cancer clinical research, next-generation sequencing data analysis, comparative genomics and quantitative biomedical informatics.

Junqing Wang received his PhD from Shanghai Jiao Tong University School of Medicine, Shanghai, China in 2013. Currently, he is an Associate Chief Physician from the Department of Surgery, Ruijin Hospital, China. His research interests include using system biology approaches to study molecular cancer, integrative analysis and data mining for biomarkers discovery and cancer genomics studies for the functional investigations of microRNAs.

Xiaofeng Song, PhD, is a Professor and Principle investigator in the Department of Biomedical Engineering, Nanjing University of Aeronautics and Astronautics. His main interest is to study proteomics, genomics, and bioinformatics methods.

Zhongming Zhao received his PhD in Human and Molecular Genetics from The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences, Houston, Texas in 2000. He holds Chair Professor for Precision Health and is a professor in School of Biomedical Informatics. He is Founding Director of the Center for Precision Health at The University of Texas Health Science Center at Houston. He is also Founding President of The International Association for Intelligent Biology and Medicine (IAIBM). His research interests include bioinformatics and systems biology approaches to studying complex diseases, deep learning, precision medicine, and pharmacogenomics.

This special issue collects six papers submitted to *The IEEE International Conference on Bioinformatics and Biomedicine (IEEE BIBM 2017)*, which was held on November 13–16, 2017 in Kansas City, MO, USA. The IEEE BIBM 2017, which was built on the success of previous conferences, attracted more than 300 participants from many institutions around the world. IEEE BIBM 2017 provides a leading forum for disseminating the latest research in bioinformatics and health informatics. It brings together academic and industrial scientists from computer science, biology, chemistry, medicine, mathematics and statistics. The BIBM program included four keynote speeches, eight scientific sessions, three tutorials, nine highlight talks, and a poster session. The details of all presentations are available on the conference website (https://muii.missouri.edu/bibm2017/). The four keynote speakers, all world-renowned leaders in bioinformatics, genomics, systems biology and computational medicine, are Dr. Sanguthevar Rajasekaran from The University of Connecticut, Dr. Andrey Rzhetsky from the University of Chicago, Dr. Predrag Radivojac from Indiana University Bloomington, and Dr. Jun (Luke) Huan from the University of Kansas. Twenty four scientific sessions included the presentations selected from the rigorous review process handled by a program committee of more than 90 experts in the field based on their scientific merit and technical quality. The examples of these sessions are biological data mining, visualisation, high-performance computing, clinical decision support and informatics, AI and machine learning methods in biomedical informatics. Twenty four workshops cover current hot research topics such as deep learning in bioinformatics, network-based data integration and analysis, as well as health informatics and data science. These tutorials provided a wealth of information on these cutting-edge techniques and are well appreciated by the conference participants.

Thanks to the high-throughput next generation deep sequencing technologies, biologists now can observe and measure thousands of molecules in cells simultaneously. This wealth of data provides an unprecedented challenge and opportunity to construct predictive and mechanistic models for the complex molecular system. Machine learning and statistical modelling will continue playing critical roles in this field of systems biology, and the demand for more efficient and accurate machine algorithms is still urgent. The six original papers selected in this special issue describe some recent

development of machine learning and statistical methods in systems biology, reflecting the rapid advances in these topics. Among these manuscripts, most of them are related to identifying genetic biomarkers from heterogeneous tumours through developing novel frameworks for integrative analysis.

Biomarker discovery from complex 'omics' data has become an important method to identify the genetic targets for disease prevention, diagnosis, and treatment in precision medicine. However, due to the complexity of increasing number of NGS data generated, developing novel integrative analysis methods will help researchers to detect new knowledge and biological signals. Mallik and Zhao developed a rule mining framework for identifying new biomarkers using the shortest distance method on the sarcomas tissues' omics profiling including gene expression, DNA methylation, and protein-protein interaction. Their method is useful to extract novel biomarkers from 'omics' profiles of the data for the complex disease or cellular conditions.

Instead of identifying complex biomarkers from heterogeneous multi-omics profiling within the same study, there are other challenges in identifying biomarkers from the same type of omics data (i.e., transcriptomics) across different platforms and studies. For example, integrating gene expression profiles from both Microarray and RNAseq platforms has been widely used to increase the statistical power in identifying phenotype-specific biomarkers from the large sample size. However, there are limited studies which comparatively investigate the biomarker selection from the two platforms. Zhang et al., comparatively studied the performance of identified biomarkers using four popular feature selection algorithms (i.e., SVM) across different platforms and cohorts. The strengths and weakness of selected biomarkers across different experimental platforms and studies were discussed. This project showed the biomarker selection accuracy was influenced more by the characteristic of the data than by algorithms, platforms or datasets.

In addition, tumours contain genetically diverse subclonal populations of cells through successive waves of expansion and selection. Developing algorithms to analyse the extent of tumor heterogeneity is critically important in explaining cancer evolution history and identifying the somatic variations or aneuploidy events with subclonal frequency. Chu et al., developed a simulation tool, Pysubsim-tree, which could identify different aneuploidy events and somatic variations based on the given tumor evolution history of distinct subclonal genomes. This project is significant since it fills the gap that no existing ground truth methods which can benchmark the somatic variations or aneuploidy events at subclonal populations in the tumor.

Other than developing integrative analysis framework to identify biomarkers from the complex biological system, construction of data management system to extract relevant biological information and track the data provenance at specific data points is another type of hot topic. Data provenance means documenting the paths of the input data and data retrieve history during the experimental process from the beginning to end. Almeida et al. developed a tool, AProvBio, which enables the documentation of the storage of data provenance using the PROV-DM standard model in bioinformatics workflows. Three types of data provenance, which are prospect, retrospect, and the user-defined type, was implemented in the AProvBio workflow. Given how graphs conveniently express PROV-DM, Almeida et al., have designed the method for storing the data provenance in a graph database system and evaluated the AProvBio system through two real case experiments.

Integrative network-based method to identify biomarkers is another type of hot topic, which plays a significant role in characterising biological relationships. However, most

current available network modelling studies focus on 'static' network-based analysis while ignoring the time-series gene expression alterations. Cooper et al., used a temporal, network-based approach to identify and rank genes that exhibit variation in short-course gene expression. They used a Caenorhabditis elegans (*C. elegans*) gene correlation dataset as the experiment and compared their method with some traditional methods. They identified novel genes that are inherently different from differentially expressed genes in traditional ways, raising new questions about structural meaning in expression networks and how changes in expression.

In final project, Zheng et al., used a network enrichment analysis method to identify the interior warming mechanisms of Ginger. Ginger has been widely used as a traditional Chinese medicine (TCM) for its warming function in the stomach and small intestine. Through data mining, they extracted regulatory proteins database from publically available database and investigated the protein regulating networks from ginger's two bio-active compounds, 6-gingerol and 6-shaogaol. They identified five key metabolic processes on ATP, glycogen, coenzyme, glycerolipid and fatty acid and concluded that these five key metabolic processes could carry out the warming interior function in ginger.

We thank all the reviewers for judging the scientific merits of the manuscripts submitted to IEEE BIBM 2017 including this special issue. We are grateful to the local organising committee members and volunteers for making IEEE BIBM 2017 an excellent venue for exchanging research results and ideas for fostering collaboration, and for training next-generation bioinformaticians in biological and biomedical fields.