Editorial

Jianhua Ruan

Department of Computer Science,
The University of Texas at San Antonio,
San Antonio, Texas 78249, USA
Email: jianhua.ruan@utsa.edu

Kun Huang

Department of Biomedical Informatics,
Ohio State University College of Medicine,
Columbus, OH 43210, USA
Email: kun.huang@osumc.edu

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

Yunlong Liu

Department of Medical and Molecular Genetics,
Indiana University School of Medicine,
Indianapolis, IN 46202, USA
Email: yunliu@iu.edu

Biographical notes: Jianhua Ruan received his PhD in Computer Science and Engineering from Washington University in St. Louis in 2007. He is currently an Associate Professor in the Department of Computer Science and Leader of Computational Systems Biology Core at The University of Texas at San Antonio. His research interests lie in broad areas of bioinformatics, computational systems biology, and data mining.

Kun Huang received his Bachelor's degrees in Biology and Computer Science in 1996 from Tsinghua University, in Beijing, China. From 1996 to 2004, he studied at the University of Illinois at Urbana-Champaign (UIUC), where he obtained his Masters' of Science in Physiology, Electrical Engineering, and Mathematics, as well as his PhD in Electrical and Computer Engineering. He is currently Professor in the Department of Biomedical Informatics and Director of the Division of Computational Biology and Bioinformatics at The Ohio State University. His research interests include computer vision, machine learning, medical imaging and computational biology.

Zhongming Zhao received his PhD in Human and Molecular Genetics from the Graduate School of the University of Texas Health Science Center at Houston and MD Anderson Cancer Center, Houston, Texas in 2000. Currently, he is the Dr. Doris L. Ross Professor in School of Biomedical Informatics and Founding Director of the Center for Precision Health at The University of Texas Health Science Center at Houston. Before coming to UTHealth in 2016, he was a Professor in the Departments of Biomedical Informatics, Psychiatry and Cancer Biology at Vanderbilt University School of Medicine. His research interests include bioinformatics and systems biology approaches to studying complex diseases, precision medicine, pharmacogenomics, and biomedical informatics.

Yunlong Liu received his PhD in Biomedical Engineering from Purdue University in 2004. He is currently an Associate Professor in the Department of Medical and Molecular Genetics at Indiana University School of Medicine. Using statistical and informatics approaches, his laboratory is actively developing computational methodologies and informatics pipelines in analysing data from next generation sequencing technology to study transcriptional, post-transcriptional and epigenetic regulation.

This special issue collects five papers submitted to *The International Conference on Intelligent Biology and Medicine 2015 (ICIBM 2015)*, which was held on 13–15 November, 2015 in Indianapolis, USA. The *ICIBM 2015*, which was built on the success of previous ICIBM conferences, attracted more than 120 participants from many institutions around the world. The ICIBM 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 (watson.compbio.iupui.edu/yunliu/icibm/). The four keynote speakers, all world-renowned leaders in bioinformatics, genomics, systems biology and computational medicine, are Dr. Jiajie Zhang from The University of Texas Health Science Center at Houston, Dr. Christopher Sanders from Harvard Medical School, Dr. Keith Dunker from Indiana University, and Dr. Sylvia Plevritis from Stanford University. Eight scientific sessions included the presentations selected from rigorous review process handled by a program committee of more than 90 experts in the field based on their scientific merit and technical quality. These sessions are the following: Session I: NGS Data Analysis I; Session II: Systems Biology; Session III: NGS Data Analysis II; Session IV: Integrative Genomics; Session V: Genomics and Genetics; Session VI: Epigenomics, Proteomics and Metabolomics; Session VII: Biomarker Discovery and Precision Medicine; and Session VIII: Pharmacogenomics and Systems Medicine. Three tutorial sessions cover current hot research topics:

- proteomics
- bioimaging informatics
- next-generation sequencing and data analysis.

These tutorials provided a wealth of information on these cutting-edge techniques and are well appreciated by the conference participants.

Thanks to the rapid advances in high-throughput experimental technologies, biologists now have the ability to observe and measure thousands of molecules in cell 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 learning algorithms is still urgent. The five 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. The other selected papers from ICIBM 2015 were included in two supplement issues to BMC Genomics (volume 17, supplement 7) and BMC Systems Biology (volume 10, supplement 3). Below, we briefly summarise the five papers in this special issue.

The NIH human microbiome project (HMP) and other initiatives have produced massive -omics data from uncultured microbial community in environmental samples, providing a powerful lens to view the microbial world and its interaction with the environment. Recently there is growing interest in RNA-seq based metatranscriptomic sequencing of the complete set of transcripts from environmental samples, in order to monitor the dynamic change of the gene activities in these communities when responding to the changes in their environment. However, one of the strong challenges in meta-transcriptomics is to assemble the sequencing reads into transcripts without any reference genome, as the complex environmental samples may contain many previously uncharacterised microbial species. To address this important problem, Mohsen et al. developed a novel algorithm, DNPipe, based on expectation-maximisation (EM) and sampling approaches that utilise abundance information of transcript contigs, which can substantially improve the quality of metatranscriptome assembly, producing longer and more accurate transcripts compared to existing state-of-the-art approaches, demonstrating the advantage of this machine-learning based method.

In another paper, Jiang et al. proposed an empirical Bayes mixture model for eQTL mapping and identification of gene regulatory hotspots. The key components in their method include a modified two-sample t-statistic, a two-component negative binomial mixture model, and an expectation-maximisation (EM) algorithm to identify the parameters involved in the model. This is the first paper that used the joint empirical Bayes model to evaluate the false discovery rate of hotspot identification in the expression quantitative loci. Their experimental results show that many hotspots have annotated function as the binding sites of transcription factors and histone proteins.

The recently discovered R-loop structure in the genome is composed of a DNA/RNA hybrid and a displaced single-stranded DNA and has been shown to play important roles including genome instability and transcriptional regulation. R-loop can be identified by the protocol of DNA-RNA immuno-precipitation (DRIP) followed by next-generation sequencing (DRIP-seq). Liu et al. developed a suite of statistical tools, DRIPer, for investigating DRIP-seq data and perform correlation analysis and Kolmogorov-Smirnov tests to study associations with publicly available DRIP-seq data and ENCODE ChIP-seq, which enables biologists to quickly evaluate the functional relationship between R-loops and nearby protein binding sites and target gene expression as well as gene sets such as canonical biological pathways and Gene Ontology terms.

Biological networks such as protein-protein interaction (PPI) networks share many common topological properties with real-world networks (e.g., social networks and the internet). These properties include scale-free characteristics and high modularity.

It is also well known that a protein's topological property in the PPI network reflects the importance of its functional role, and proteins with similar topological properties may have similar functions. For example, hub nodes in the yeast protein-protein interaction network tend to have higher essentiality (i.e., more critical for survival).

Yu et al. performed network topological analysis using the well-known machine learning algorithm, PageRank, to identify proteins that are potentially influential (i.e., crucial for the connectivity) in the protein-protein interaction network in the malaria parasite *Plasmodium falciparum*. The proteins that were predicted by the algorithm to be most influential in the network are found to be involved in transcriptional regulation, signalling, proteolysis, and heat shock response and may play a role in a variety of fundamental processes ranging from genetic information processing, metabolism, transport, development, to virulence to the host. Functional characterisation of these proteins may open venues for novel therapeutics for effective malaria eradication.

In another study, Li et al. debated that topological properties alone are insufficient to predict gene essentiality, as PPI network is static and lacks specific biological context. To address the issue, they proposed a novel algorithm based on the theories of dynamic systems and kinetic modelling to predict the consequence of gene knockout by simulating the changes in systems dynamics of fundamental metabolic/physiological pathways in the cell. When their method was applied to the central carbon metabolism pathway in *Escherichia coli*, it could accurately predict gene essentialities that were subsequently validated experimentally. Accordingly, it outperformed the classical flux balance analysis method and their previous algorithm. Their method can be a useful tool for virtual knockout experiment for *in silico* investigations of complex biological systems, which may have many applications in basic and translational studies, including metabolic engineering and drug target selection.

We thank all the reviewers for judging the scientific merits of the manuscripts submitted to *ICIBM 2015* including this special issue. We are grateful to the local organisation committee members and volunteers for making *ICIBM 2015* a great venue for exchanging research results and ideas for fostering collaboration, and for training next-generation informaticians in biological and biomedical fields. We thank the Indiana University – Purdue University Indianapolis (IUPUI) for hosting *ICIBM 2015*.