Editorial

Zhongming Zhao*

Departments of Biomedical Informatics
Psychiatry, and Cancer Biology,
Vanderbilt University Medical Centre,
Nashville, TN 37232, USA
Email: zhongming.zhao@vanderbilt.edu
*Corresponding author

Hua Xu

School of Biomedical Informatics,
The University of Texas Health Science Center at Houston,
Houston, TX 77030, USA
Email: hua.xu@uth.tmc.edu

Bing Zhang

Departments of Biomedical Informatics and Cancer Biology,
Vanderbilt University Medical Centre,
Nashville, TN 37232, USA
Email: bing.zhang@vanderbilt.edu

Yufei Huang

Department of Electrical and Computer Engineering,
The University of Texas at San Antonio,
San Antonio, TX 78249, USA
Email: yufei.huang@utsa.edu

Biographical notes: 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 an Ingram Associate Professor of Cancer Research and an Associate Professor in the Departments of Biomedical Informatics, Psychiatry, and Cancer Biology at the Vanderbilt University Medical Center. His research interests include bioinformatics and systems biology approaches to studying complex diseases, personalised medicine, pharmacogenomics, and biomedical informatics.

Hua Xu received his PhD in Biomedical Informatics from Columbia University at New York City in 2008. Currently he is an Associate Professor in the School of Biomedical Informatics at the University of Texas Health Science Center at Houston. His research interests include clinical natural language processing, biomedical text mining, and healthcare data mining.

Bing Zhang received his PhD degree in Molecular Genetics from the Chinese Academy of Sciences, Shanghai, China in 1999. Currently, he is an Assistant Professor in the Departments of Biomedical Informatics and Cancer Biology at the Vanderbilt University Medical Center. His research focuses on the development and application of integrative bioinformatics approaches to the study of complex diseases.

Yufei Huang received his PhD degree in Electrical Engineering from the State University of New York at Stony Brook in 2001. Since 2002, he has been with the Department of Electrical and Computer Engineering at the University of Texas at San Antonio, where he is now a Professor. His current research focus is on computational biology and computational neuro-ergonomics, including high throughput biomedical data modelling and integration, genomic regulatory networks, microRNA target and functional prediction, EEG data modelling and characterisation.

The International Conference on Intelligent Biology and Medicine 2012 (ICIBM 2012) was held on 22–24 April 2012 at the Holiday Inn at Vanderbilt in Nashville, Tennessee. Our program included a Workshop on Next Generation Sequencing, a Tutorial of Proteome Informatics, 3 keynote speeches, 6 presentation sessions, and a poster session. More specifically, the six regular scientific sessions were: Genomics, Systems Biology I and II, Algorithms and Methods, Intelligent Computing, and Applications and Tools. These six regular sessions presented 40 manuscripts about recent innovative work in the areas of bioinformatics, systems biology, medical informatics, and intelligent computing. The presenters were selected through a rigorous review process, and their work stood out among the submissions as novel and significant. The details are available on the conference website (http://bioinfo.mc.vanderbilt.edu/icibm/).

We received numerous, highly competitive manuscripts. After first round of review, the selected research manuscripts were invited for further expansion and review. This special issue includes 10 original articles that describe recent advances in genomics, systems biology and intelligent computing. The other selected papers were included in two supplemental issues to BMC Genomics (http://www.biomedcentral.com/bmcgenomics/supplements/13/S8) and BMC Systems Biology (http://www.biomedcentral.com/bmcsystbiol/supplements/6/S3). Below, we briefly summarise the 10 articles in this special issue.

Systematic review of biomedical literature is critical for developing evidence-based medicine and improving effectiveness of medical diagnosis and treatment. Current manual review approach is extremely labour-intensive; thus, automatic approaches such as natural language processing (NLP) are in strong demand. Jonnalagadda and Petitti proposed the use of distributional semantics and user feedback as an approach to reduce workload and demonstrated its effectiveness through their evaluation on a set of 15 published drug systematic reviews.

Similarly, development of automatic approaches to extract pharmacogenomics (PGx)-specific drug-gene relationships from knowledgebase has been an active research area. Such approaches are often based on statistical modelling, computational algorithms, and NLP techniques. Xu and Wang developed a bootstrapping PGx gene prioritisation learning technique to rank 33,310 human genes and 2388 PharmGKB genes on their relevance to drug response, using 20 million MEDLINE abstracts. Their evaluation

indicated that their algorithm could achieve similar performance in ranking genes in PharmGKB, the largest manually curated resource for variation on genes related to drug response.

Adefioye et al. presented a multi-view spectral clustering framework by leveraging the inherent consistency among multi-view data and integration. Their clustering method is based on tensor decomposition. The authors applied their clustering algorithms to chemical compounds and demonstrated that their algorithms had better performance than spectral clustering of single-view data.

Gene co-expression from microarray data provides abundant biological information under a set of related conditions, e.g. development or disease progression. Network analysis of such data has been proven effective on many applications such as disease biomarkers and functional regulations. In this special issue, Kalluru et al. proposed a method for identifying condition specific co-expression networks by combining quasi-clique mining algorithm with the Expected Conditional F-statistic. They applied the new method to a breast cancer microarray dataset and identified the dynamic differential transcriptional programmes between the basal and non-basal subtypes of breast cancers.

In another work from the same group, the authors introduced a workflow for identifying gene co-expression networks that are associated with colon cancer metastasis (Zhang et al.). Their workflow includes dense network discovery from weighted gene co-expression network followed by network activity analysis. The authors reported several genomic loci as potential genetic aberrations related to colon cancer malignancy.

Cheng et al. also performed research on gene-gene interactions from microarray data. To tackle the problem of gene expression similarity (or distance) measure in microarray datasets, the authors proposed a Kernel correlation coefficient method and evaluated the performance at the biological network level on a public dataset of 113 expression measurements of the 6229 yeast genes.

Berry et al. introduced a software tool, PolyLens, a Java-based, integral visual analytical toolkit that can systematically process population genomic data, visualise geographic distributions of genealogical lineages, and display allele distribution patterns. PolyLens can be used to visualise specific DNA sequences within each individual and their related locations. A data mining method, Non-negative Tensor Factorisation (NTF), was implemented to derive hidden patterns from the data set and to generate clusters.

Major histocompatibility complex class II (MHC II) molecules bind peptides derived from antigens and form MHC II-peptide complexes. To explore the potential correlation between MHC II peptides and B cell epitopes, Li et al. specifically examined the similarity between the segments in B cell conformational epitopes and MHC II peptides based on residual composition, sequence identity and similarity. The authors suggested that the epitope segments have a higher similarity to the known MHC II peptides, thus, playing a critical role of MHC II molecule in picking up protein epitopes

Many algorithms have been developed for sequence similarity analysis due to the exponential growth of sequence data during the last three decades. In this special issue, Barton et al. introduced algorithms for the extreme similarity sequencing problem, aiming to dramatically improve sequence data storage and analysis in the era of high throughput sequencing being feasible in a typical lab. Specifically, the authors presented an asymptotically fast $O(n + occ \log occ)$ algorithm and a practical $O(nk/w)$ algorithm for the extreme similarity sequencing problem.

In the last paper, to address the protein sub-similarity matching problem, Cimen et al. proposed a novel matching and protein alignment algorithm method based on clustering and longest common subsequence (LCS) techniques. The authors tested their method in HIV-1 related proteins and relationship to meningitis. Specifically, the authors clustered dihedral angles of potential human immune system proteins and the dihedral angles of the meningitis outer membrane antigen and found their results promising.

We thank Rebecca Hiller Posey for her assistance in conference management and all the reviewers for judging the scientific merits of the manuscripts submitted to ICIBM 2012 including this special issue. We would like to acknowledge Drs. Kun Huang, Dongxiao Zhu, and David Tabb for organising the workshop and tutorial in ICIBM 2012. We thank the National Science Foundation (NSF grant 1141979) and Vanderbilt Center for Quantitative Sciences for financial support of ICIBM 2012.