Quantitative Biology 2017, 5(4): 277–279
DOI 10.1007/s40484-017-0128-z

EDITORIAL

Special collection of bioinformatics in the era of precision medicine

Zhaohui S. Qin*

Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA 30322, USA
Department of Biomedical Informatics, Emory University School of Medicine, Atlanta, GA 30322, USA
* Correspondence: zhaohui.qin@emory.edu

Received October 23, 2017

Precision medicine advocates for the practice of customized disease treatment and prevention such that all clinical decisions are made based on the characteristics of individual patients. The precision medicine framework has been enthusiastically endorsed by the health care community and is set to have a profound impact on the health care practice. However, adopting the precision medicine ideology in the clinics requires solving a series of technical challenges, particularly in informatics and data analytics. To better serve this newly emerging area of research, we present this special issue to showcase the latest research developments in bioinformatics which has been widely regarded as a major stakeholder in the advancement of precision medicine. The ten papers appear in special collection cover a wide range of topics and can be broadly categorized into the following three main areas: statistical method development; informatics method development and algorithms for translational bioinformatics.

STATISTICAL METHOD DEVELOPMENT

Random Forest (RF) has been widely used as a powerful over-the-counter classification tool in many bioinformatics applications. However, too many features will reduce the power of RF. Liu and Zhao proposed a novel method named variable importance-weighted RF to overcome this problem. Instead of sampling features with equal probability at each node, Liu and Zhao proposed to sample features according to their variable importance scores, and then selected the best split from the randomly selected features. The goal of their new method is to better utilize more informative features without completely ignoring ones that are less informative. The authors have confirmed that their new method is able to improve the prediction accuracy in the presence of weak signals and large noises.

Mendelian randomization(MR) is a popular method in epidemiology that uses genetic data as the instrumental variable to conduct causal inference. MR has been increasingly used in biomedical research. In a paper from Zhang and Ghosh, the authors explored the feasibility of using least-squares kernel machines (LSKM) in MR studies. Their results suggested that LSKM based on genotype score or genotype can be used effectively in two-stage least square estimator (TSLS). The authors speculated that their new approach will provide higher power when the actual correlation between exposure and genetic instrumental variables is nonlinear.

Pimentel[1)] *et al*. introduceed a new biclustering algorithm, named SCCA-BC, or sparse canonical correlation analysis bi-clustering, to discover subsets of genomic expression features from large scale gene expression datasets. Due to the plasticity of the transcriptome under different conditions, bi-clustering is an important tool for analyzing a large set of gene expression datasets. SCCA-BC is an extension of the SCCA method coupled

[1)] The paper will be published in next issue.

with repeated random partitioning and subsampling of the expression data set to capture strong group interactions, or conditional dependencies between expression features.

INFORMATICS METHOD DEVELOPMENT

Zhang[1)] *et al*. developed a nonparametric approach, namely MRHCA, or mutual rank-based hub and co-expression analysis, to identify hub genes and modules in a large co-expression network with moderate computational and memory cost. The authors demonstrated that MRHCA has several desirable utilities including the capability of dealing with large association networks, assessing statistical significance for hubs and module sizes, identifying co-expression modules with low associations, detecting small and significant modules and allowing genes to be present in more than one modules. The authors have validated the utility of MRHCA suing simulated datasets. When applying MRHCA and several differential co-expression analysis methods on real data from *E. coli* and TCGA data, the authors were able to identify significant condition specific activated genes in *E. coli* and distinct gene expression regulatory mechanisms between the cancer types with high copy number variation and small somatic mutations.

The paper from Zheng and He is on the topic of ontology, an important branch of bio informatics research. The authors focused on the specific vaccine host response domain, and demonstrated that the community-based Ontology of Biological and Clinical Statistics (OBCS), combined with the domain-specific Vaccine Ontology (VO), can be utilized to represent experimental data and metadata of host immune responses to immunizations with yellow fever vaccine. Their ontology-based meta-analysis results suggested that different experimental results may be explained by different experimental assays and conditions, sample variations, and data analysis methods.

Chen *et al*. took advantage of the Cap-seq data to determine the transcription start sites (TSSs) and to study the transcriptional regulation of intergenic miRNAs in *C. elegans* and mouse. In both species, the authors were able to successfully identify a class of special pre-miRNAs whose 5′ ends are capped, and are most probably generated directly by transcription. Additionally, Chen *et al*. distinguished another class of special pre-miRNAs that are 5′-capped but are also part of longer primary miRNAs, suggesting they likely have more than one transcription mechanism.

ALGORITHMS FOR TRANSLATIONAL BIOINFORMATICS

Vasudevaraja *et al*. developed a precision medicine style target-drug selection method using a computational method based on genetic interaction. Their computation method named PMTDS utilize genetic interaction networks to select the optimum targets and associated drugs for the treatment of cancer. The PMTDS system includes three major components: a personalized medicine knowledge base that is cancer type specific, a genetic interaction network-based algorithm and a single patient molecular profile. The author believed that the newly developed PMTDS system provides an accurate and reliable source for off-label drug selection for precision cancer medicine.

Malik and Zhao provided an association rule mining-based approach for identifying integrated markers through mutual information based statistically significant feature extraction, and applied it to acute myeloid leukemia (AML), prostate carcinoma (PC) gene expression and methylation profiles. The authors run a Jarque-Bera normality test to separate the whole dataset into normal and non-normal parts. Then the authors applied statistical tests on top-ranked genes in each group to identify biomarkers and validate them using multiple bioinformatics approaches. Through this novel analyses, the authors identified several novel biomarkers for these two cancer types. The authors noted that their novel approach can be applied to other complex diseases.

Cao *et al*. attempted to answer the question whether cancer and normal tissue cells execute the same transcription regulation programs and why. By conducting an integrated and comprehensive analyses of gene expression and DNA-methylation data, Cao *et al*. made key discoveries that a set of genes are consistently regulated by DNA methylation in cancer, especially genes related to stress response. And for most cancer types, the involvement of DNA methylation in transcription regulation of genes increases as cancer advances. Overall, they found the transcription regulation program in cancer and normal cells are fundamentally different.

Finally, Wang *et al*. presented a review on transcriptome assembly strategies for precision medicine. The survey is focused on the recent methodology development of transcriptome assembly approach using RNA-

[1)] The paper will be published in next issue.

Seq. RNA-Seq plays an important role in precision medicine since it can provide valuable and timely information to clinicians that may lead to improved clinical diagnosis, prognosis and treatment plan. Among all the analysis strategies designed for RNA-seq data, transcription assembly is a very challenging problem but it is able to provide key insights that no other assay can supply. The detailed and comprehensive review will help researchers to quickly understand the contemporary state-of-art in this important research area.

ACKNOWLEDGMENTS

I want to extend my sincere appreciation to all the authors, reviewers who kindly agree to lend their time and effort to this important endeavor in a timely manner. I felt fortunate to have the opportunity to work so many prominent researchers in the bioinformatics field. This special issue would not be possible without the hard work of the managing editor. Dr. Huaying Liu which I greatly appreciate. Finally, I want to thank the editors and editorial board for their support and guidance.

