Editorial

Chi Zhang

Department of Medical and Molecular Genetics,
School of Medicine, Indiana University,
Indianapolis, 46202, USA
Email: czhang87@iu.edu

Lai Wei

Department of Biomedical Informatics,
The Ohio State University,
Columbus, 43210, USA
Email: Lai.Wei@osumc.edu

Ewy Mathé

Division of Preclinical Innovation,
National Center for Advancing Translational Sciences,
NIH, 9800 Medical Center Dr,
Rockville, MD, 20892, USA
Email: ewy.mathe@nih.gov

Maciej Pietrzak

Department of Biomedical Informatics,
The Ohio State University,
Columbus, 43210, USA
Email: Maciej.Pietrzak@osumc.edu

Abstract: Since 2012, The International Association for Intelligent Biology and Medicine has been organising International Conferences on Intelligent Biology and Medicine (ICIBM). In 2019, ICIBM was hosted by The Ohio State University in Columbus, Ohio, USA. The meeting included 12 scientific sessions, four keynote speakers, four eminent scholar speakers, five tutorials and workshops and a poster session. Of 105 original research manuscripts submitted to the conference, we selected 10 for publication in this special issue of the *International Journal of Computational Biology and Drug Design*.

Biographical notes: Chi Zhang focuses on computational modeling of cancer micro-environment including the level of hypoxia, oxidative stress, acidity and dysregulation of extracellular matrix as well as altered immune responses in cancer by using large scale omics data. He is also interested in developing novel computation methods to integrate multiple tissue level and single cell omics data types to understand the mechanism of cancer initiation, progression and metastasis. In addition, his research application also includes inference of intra-tumour heterogeneity and reprogrammed metabolism, and prediction of gain or loss of functions led by a certain mutation or collective effect of multiple mutations.

Lai Wei, PhD, MS, is a Research Assistant Professor within the Department of Biomedical Informatics and served as the program leader of the clinical trial

unit within the Center for Biostatistics at The Ohio State University. She has over 13 years of experience on collaborative research and her areas of expertise including designing phase I/II/III clinical trials, group sequential trials, adaptive clinical trials, adaptive laboratory experiment, and sample size re-estimation. She is also collaborating with bioinformaticians on drug repurposing research and utilising pharmacokinetics and toxicity data in translational drug interaction knowledge base by leveraging her clinical trial expertise.

Ewy Mathé is the Director of Informatics in the Division of Preclinical Innovation at NCATS. She leads a diverse team of experts in bioinformatics, cheminformatics, data science, and software development that empower translational scientists to make meaningful data-driven decisions in their research. Her personal research interests are to develop methods and frameworks to guide analysis, integration, and interpretation of metabolomics and multi-omics data to identify valid biomarkers and therapeutic targets for the diagnosis, prognosis, and treatment of various diseases. She is applying these computational approaches in collaborative translational research projects define disease-specific molecular mechanisms and phenotypes.

Maciej Pietrzak is a Research Assistant Professor in the Department of Biomedical Informatics at The Ohio State University. He has an extensive background in molecular and cellular biology combined with expertise in bioinformatics and computational biology. His work in computational biology and bioinformatics fields is focused on applying computational methods and novel mathematical approaches to address systems biology and functional genomics questions. Previously, he studied cellular processes such as cell proliferation and programmed cell death, on a molecular level, identifying mechanisms and signalling pathways shaping cellular responses to environmental stimuli.

1 Introduction

The *International Conference on The Intelligent Biology and Medicine (ICIBM)* is the official conference of the International Association for Intelligent Biology and Medicine (IAIBM). In June 9–11 2019 the conference was co-hosted by IAIBM and the Department of Biomedical Informatics at The Ohio State University in Columbus, OH. The meeting brought together 164 researchers representing various areas of bioinformatics, computational and systems biology, who presented over 100 conference papers. The details of the conference, its organisation and achievements were summarised in Guo et al. (2020), Zhao et al. (2019), Mathé et al. (2019) and Ning et al. (2020).

In this *Special Issue of International Journal of Computational Biology and Drug Design*, we present 10 papers selected for publication that cover various areas of computational biology and bioinformatics including: novel methods in genomics and transcriptomics, database development and translational science. These computational resources utilised novel modelling considerations to address several unsolved challenges and offer new capabilities in medical genomics and health informatics research.

2 Methods in genomics and transcriptomics data analysis

Luo et al. (2020) developed an artificial neural network based discriminator, named Skyhawk, to automatically review candidate genetic variations that may associated with

certain disease. Skyhawk, implemented in Python and Tensorflow, mimics how a genetic expert visually identifies genomics features based on the sequencing data alignment results. A repurposed network was utilised to generate a probability of each possible option for multiple categories including

- variant type
- alternative allele
- zygosity
- indel-length.

Skyhawk outputs a quality score to evaluate the confidence level of each variant. Real-world data based validation suggested Skyhawk is capable to review five million genome sequencing variations in 30 minutes, which could greatly reduce the workload on reviewing genetic variants.

Zucker and Coombes (2020) developed a capability to simulate both 'ground truth' and realistic CGH array and/or SNV data that could be utilised evaluate the performance of different copy number variation and genomic mutation analysis tools. A hierarchical consideration of the evolutionary path of cancer cells was implemented with an imputation approach of genetic mutations and copy number variations to simulate the CGH array or sequencing data of a tumour tissue with heterogeneous sub-clones. Population structure, sub-clone size, CNV length, mutations and errors can be adjusted by parameters. Data generated by the simulator were utilised to evaluate different segmentation and CNV detection methods. Their analysis demonstrated the simulator could generate realistic genetic variation data on the whole genome level that allow the robustness analysis of other methods.

Zhang and Kuo (2020) developed a novel computational analysis of the liver gene expression data of chimpanzee, human, non-tumour tissue and primary liver cancer tissue to mine the possible associations between tumour evolution and human evolution. Gene expression data were collected from public domain and normalised. A linear mixed model considering the species, probe, species-probe interaction, individual random effect, species within individual random effect and prober within individual random effect was utilised to model the expression variation of each gene in the collected data, whose statistical significance was assessed by ANOVA. Their analysis suggested a more rapid evolution rate of the liver cancer gene expression compared with human evolution gene expression. They also identified the human specific gene expression is more likely to be cancer specific.

Hayes and Mullins (2020) developed a new model of genomic structural variants calling by formulating the task into a minimum weight clique partition problem. Specifically, structural variants can be detected as the discordant alignments in the next generation sequencing, where each variant event can be reflected by a distinct cluster of reads. Under this consideration, an undirected graph was first constructed by treating each discordantly-mapped read pair as one vertex and assigning an edge to each pair of overlapped read pairs. Then structural variants were identified by iteratively finding maximal cliques in all connected components. The method was validated on synthetic and real world data. The experiments suggested the method is fast and accurate in detecting large structural variants.

Bartlett et al. (2020) developed a novel computational pipeline to predict colon cancer prognosis by using gene expression data. Univariate Cox regression model was first utilised to identify prognostic predictive transcriptional factors. A risk score was assigned to each sample to build a multi-variate model. The method was validated by cross validation and benchmarked on independent data sets.

3 Database development

Awasthi et al. (2020) developed the first comprehensive database of reported regulatory pseudogenes and a user-friendly app called PgenePapers (https://integrativeomics.shinyapps.io/PgenePapers/) which allows flexible database search and provides network visualisation. Pseudogenes play a pivotal role in gene regulation and disease progression as part of ceRNA networks. Realising that no database focusing on regulatory roles of pseudogenes yet, the authors extracted information about regulatory pseudogenes by analysing PubMed literature using natural language processing techniques followed by manual curation. The expression values of genes and pseudogenes for all 31 cancer types studied in TCGA were used to get the correlation between genes and pseudogenes. Based on this information, the authors reconstructed the regulatory networks involving pseudogenes and regulated genes (pseudogene-gene pairs) with disease and tissue specific annotations. They further extended the pseudogene-gene networks to include information on potential miRNAs and drugs targeting components of the networks, based on expression profiles, miRNA binding predictions and known FDA approved drugs. PgenePapers app can display the pseudogene-gene pairs with their functional categories, all the supporting text from literature, interactive visualisation of the pseudogene-gene association networks, and customised gene-pseudogene-miRNA-drug networks.

Wang et al. (2020) developed an online tool, LCLE (https://datasciences123.shinyapps.io/LCLE), which is able to systematically analyse gene expression data to identify more comprehensive relationships among lncRNAs and protein-coding genes (PCGs) from five different distances metrics. In the network analysis, determining which distance to use is probably one of the most difficult decisions. The authors noticed that the non-linear dependent correlations among genes have been ignored in most of the currently available co-expression network analysis method. They demonstrated that the selection of an appropriate distance method could help to identify novel important genes from networks. LCLE allows users to visualise figures, and download tables analysed from publically available RNAseq data such as TCGA and GTEx or upload their own data for analysis. Their web portal will benefit for biologists or clinicians without programming background in identifying novel co-regulation relations for lncRNAs and PCGs.

Zhang et al. (2020) presented a Platform for Analysis of Time-course High-dimensional data (PATH) (https://ouyanglab.shinyapps.io/PATH/) with applications in genomics research. They implemented the visualisation functions for PATH based on R libraries including gplots, RColorBrewer, and ggplot2. PATH web application provides a user-friendly interface with interactive data visualisation, dimension reduction, pattern discovery, feature selection based on the Principal Trend Analysis (PTA). Furthermore, the web application enables interactive and integrative analysis of time-course high-dimensional data based on the Joint Principal Trend Analysis (JPTA). The authors also compared PTA with traditional time-course data analysis methods in this paper. Simulation and real data examples demonstrated the functionalities of PATH for feature

selection and dynamic pattern discovery. The PATH web server provides effective and convenient functions for interactive analysis and visualisation of time-course high-dimensional data, with the potential to be broadly applied to genomics research.

4 Translational science

Srinivasan et al. (2020) developed a new predictive model of the re-admission of diabetic patients based on electronic medical record data. Around 9.4% of the US population had diabetic conditions while an estimated 33.9% of the adult population has a pre-diabetic condition. A good manage of patients' health condition is critical in the treatment of this chronic disease. A key question is to understand what features is associated with the re-admission event. In this study, 130-US hospital dataset containing patient records spanning 10 years from 1999–2008 were collected and manually curated. A random forest classifier with hyperparameter optimisation was applied to predict the readmission event based on other clinical data. The study revealed that using the attributes including number of inpatient visits, discharge disposition, admission type, and number of laboratory tests achieves a maximum AUC of 0.684 and a precision and recall of 46% and 60%. The model as well as the predictive features could help the patient management and facilitate further studies to improve the health condition of diabetic patients.

Pietrzak et al. (2020) developed new analytical approach for the imunnome data analysis. Immune profiling by flow cytometry using multi-colour immunome panels allows to assess multiple surface markers that carry information about the cell type, function and activation status. Proposed algorithms, that use information theoretical approach, allow to detect both large-scale changes as well as consistent differences in less abundant features and the relation between various subsets of immunome. Presented approach was tested using both simulated data and real-world data of Hairy Cell Leukemia (HCL), obtained using 5-colour immunome panel. Synthetic data based sensitivity and specificity test demonstrated high performance of this new algorithm. The analysis of HCL immunome data revealed subtle differences between case and control immunomes with low false discovery rate. Communicated approach may reveal the mechanisms of the responses of immune system to cancer treatment and the mechanisms of autoimmune diseases.

5 Summary

International Conferences on Intelligent Biology and Medicine 2019 brought together both senior and young investigators across the world, who are studying various areas of bioinformatics, biomedical informatics, computational and systems biology. This meeting was an excellent platform that allowed them to exchange the ideas in friendly and highly interactive environment. The researchers presented over 100 papers of which, 10 manuscripts communicating innovative bioinformatics research were selected for this special issue of IJCBDD. We hope that presented works will inspire novel, innovative research addressing key bioinformatics and systems biology questions.

Acknowledgement

We thank the National Science Foundation (NSF grant IIS-1902617) for the financial support of ICIBM 2019. This paper has not received sponsorship for publication.

References

Guo, Y., Ning, X., Mathé, E., Wang, K., Li, L., Zhang, C. and Zhao, Z. (2020) 'Innovating computational biology and intelligent medicine: ICIBM 2019 Special Issue', *Genes*., Vol. 11, No. 4, https://doi.org/10.3390/genes11040437

Mathé, E., Zhang, C., Wang, K., Ning, X., Guo, Y. and Zhao, Z. (2019) 'The international conference on intelligent biology and medicine 2019 (ICIBM 2019): conference summary and innovations in genomics', *BMC Genomics*, Vol. 20, Suppl. 12, p.1005, https://doi.org/10.1186/s12864-019-6326-5

Ning, X., Zhang, C., Wang, K., Zhao, Z. and Mathé, E. (2020) 'The international conference on intelligent biology and medicine 2019: computational methods for drug interactions', *BMC Medical Informatics and Decision Making*, Vol. 20, Suppl. 2, p.51, https://doi.org/10.1186/s12911-020-1051-1

Zhao, Z., Dai, Y., Zhang, C., Mathé, E., Wei, L. and Wang, K. (2019) 'The international conference on intelligent biology and medicine (ICIBM) 2019: bioinformatics methods and applications for human diseases', *BMC Bioinformatics*, Vol. 20, Suppl. 24, p.676, https://doi.org/10.1186/s12859-019-3240-4

Luo, R., Lam, T-W. and Schatz, M.C. (2020) 'Skyhawk: an artificial neural network-based discriminator for reviewing clinically significant genomic variants', *Int. J. Computational Biology and Drug Design*, Vol. 13, Nos. 5/6, pp.431–437.

Zucker, M.R. and Coombes, K.R. (2020) 'Generating simulated SNP array and sequencing data to assess genomic segmentation algorithms' *Int. J. Computational Biology and Drug Design*, Vol. 13, Nos. 5/6, pp.438–453.

Zhang, F. and Kuo, M.D. (2020) 'Rapid evolution of expression levels in hepatocellular carcinoma', *Int. J. Computational Biology and Drug Design*, Vol. 13, Nos. 5/6, pp.454–474.

Hayes, M. and Mullins, D. (2020) 'The minimum weight clique partition problem and its application to structural variant calling', *Int. J. Computational Biology and Drug Design*, Vol. 13, Nos. 5/6, pp.475–487.

Bartlett, B., Zhu, Y., Menor, M., Khadka, V.S., Zhang, J., Zheng, J., Jiang, B. and Deng, Y. (2020) 'Development of a RNA-Seq based prognostic signature for colon cancer', *Int. J. Computational Biology and Drug Design*, Vol. 13, Nos. 5/6, pp.488–503.

Awasthi, A., Tan, S., Johnson, T.S., Mo, X-M., Fan, K. and Zhang, Y. (2020) 'PgenePapers: a novel database and search tools of reported regulatory pseudogenes', *Int. J. Computational Biology and Drug Design*, Vol. 13, Nos. 5/6, pp.504–519.

Wang, X., Xu, K., Wang, J. and Zhou, Y. (2020) 'LCLE: a web portal for lncRNA network analysis in liver cancer', *Int. J. Computational Biology and Drug Design*, Vol. 13, Nos. 5/6, pp.520–528.

Zhang, Y., Chen, Y. and Ouyang, Z. (2020) 'PATH: An interactive web platform for analysis of time-course high-dimensional genomic data', *Int. J. Computational Biology and Drug Design*, Vol. 13, Nos. 5/6, pp.529–538.

Srinivasan, S.M., Paat, Y-F., Halls, P., Kalule, R. and Harvey, T.E. (2020) 'Predicting re-admission to hospital for diabetes treatment: a machine learning solution', *Int. J. Computational Biology and Drug Design*, Vol. 13, Nos. 5/6, pp.539–554.

Pietrzak, M., Lozanski, G., Grever, M., Andritsos, L., Blachly, J., Rogers, K. and Seweryn, M. (2020) 'On the analysis of the human immunome via an information theoretical approach', *Int. J. Computational Biology and Drug Design*, Vol. 13, Nos. 5/6, pp.555–581.