Guest Editorial for Special Section on Big Data Computing and Processing in Computational Biology and Bioinformatics

Chao Wang, Hong Yu, Aili Wang, and Kai Xia

BIG data has emerged as an important application field which has shown its huge impact in different scientific research domains. In particular, the big data bioinformatics applications such as DNA sequence analysis have posed significant challenges to the state-of-the-art processing and computing systems. With the growing explosive data scale, the collection, storage, retrieval, processing, scheduling, and visualization are key big data issues to be tackled. Up to now, many researchers have been seeking high-level parallelism using novel big data computing architectures and processing mechanisms.

As an emerging research area, a significant number of state-of-the-art research works was conducted in the past few years. However, there are still many open issues that need to be addressed, for instance:

1) Big data analysis methodologies for bioinformatics and computational biology, including data mining algorithms and processing techniques, big data acquisition, collection, and storage. Meanwhile, the cutting edge big data processing mechanisms, including data sorting, retrieval, classification, scheduling, and visualization for big data applications, have been focused during the past few years.
2) Emerging data-intensive bioinformatics applications, such as DNA/RNA word analysis, segmentation, functional genomics, evolutionary genomics, and comparative genomics.

This present special section of the *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (*TCBB*) aims at introducing and demonstrating the latest research activities on innovative ideas and solutions in all aspects around data intensive computing in system and application domains. In total, 16 papers were received. Each paper was reviewed by three experts in this area during the review process. After undergoing three rounds of revisions, five papers were finally accepted for this special section. The main contributions of the five papers are briefly introduced as follows:

In the paper titled “Expectation Maximization of Frequent Patterns, a Specific, Local, Pattern-Based Biclustering Algorithm for Biological Datasets”, Erin Jessica Moore and Thirmachos Bourlai propose an algorithm that can analyze datasets with a large attribute set at different densities. The algorithm produces biclusters with low root mean squared error and false positive rate. In particular, the algorithm is a hybrid, axis-parallel, pattern-based algorithm with a variable confidence threshold, and the novel use of local density comparisons versus the standard global threshold. Moreover, the authors also introduce a framework to ease comparison with other algorithms, and compare to both binary and general biclustering algorithms.

In the paper entitled “Improve Glioblastoma Multiforme Prognosis Prediction by Using Feature Selection and Multiple Kernel Learning”, Ya Zhang, Ao Li, Chen Peng, and Minghui Wang have improved prognosis prediction accuracy by taking advantage of the minimum redundancy feature selection method and the multiple kernel machine learning method. The authors build an integrated model which could predict prognosis with high accuracy.

In the paper entitled “Inferring Disease Associated Phosphorylation Sites via Random Walk on Multi-Layer Heterogeneous Network”, Xiaoyi Xu and Minghui Wang propose a multi-layer heterogeneous network model to make use of the kinase information to infer disease-phosphorylation site relationship. Random walk mechanism on the heterogeneous network is implemented. Experimental results reveal that multi-layer heterogeneous network model with kinase layer is superior in inferring disease-phosphorylation site relationship when comparing with existing random walk model and common used classification methods.

In the paper entitled “MrBayes tgMC^3++: A High Performance and Resource-Efficient GPU-Oriented Phylogenetic Analysis Method”, Cheng Ling, Tsuyoshi Hamada, Jingyang Gao, Guoguang Zhao, Donghong Sun, and Weifeng Shi present a high performance and resource-efficient method for GPU-oriented parallelization of likelihood estimations. Instead of having to rely on empirical programming, the proposed novel decomposition storage model implements high performance data transfers implicitly. In terms of performance improvement, a speedup

- *C. Wang is with the School of Computer Science, University of Science and Technology of China, Hefei 230000, China. E-mail: cswang@ustc.edu.cn.*
- *H. Yu is with the National Center for Plant Gene Research, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100000, China. E-mail: hyu@genetics.ac.cn.*
- *A. Wang is with the School of Software Engineering, University of Science and Technology of China, Hefei 230000, China. E-mail: wangal@ustc.edu.cn.*
- *K. Xia is with the Department of Psychiatry, University of North Carolina, Chapel Hill, NC 27514. E-mail: kxia@med.unc.edu.*

For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TCBB.2016.2581460

factor of up to 178 can be achieved on the analysis of simulated datasets by 4 Tesla K40 cards.

In the paper entitled "Mining Contiguous Sequential Generators in Biological Sequences", Jingsong Zhang, Yinglin Wang, Chao Zhang, and Yongyong Shi present ConSgen, an efficient algorithm for discovering contiguous sequential generators. ConSgen adopts the n-gram model to generate potential frequent subsequences and leverages several pruning techniques to prune the unpromising parts of search space. The authors developed a new machine learning approach by taking advantages of both mRMR feature selection method and MKL classification method.

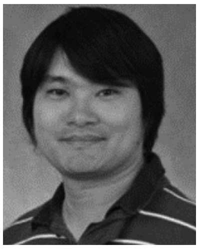
The guest editors are very grateful to the authors of this special section and to the reviewers for their tremendous service by critically reviewing the submitted papers. The editors would like also to thank the editor-in-chief of *TCBB* Prof. Ying Xu, the associate editor-in-chief of *TCBB* Prof. Dong Xu, and especially Joyce Arnold for the editorial assistance and excellent cooperative collaboration to produce this scientific work. The authors hope that the readers will share our excitement to the section papers on big data computing and processing in computational biology and bioinformatics and will find it useful.

Chao Wang
Hong Yu
Aili Wang
Kai Xia
Guest Editors

Chao Wang received the BS and PhD degrees from the University of Science and Technology of China, in 2006 and 2011, respectively, both in computer science. He is with the Embedded System Lab in the Suzhou Institute at the University of Science and Technology of China, Suzhou, China. His research interests include big data and reconfigurable computing. He is now an editorial board member of *IET CDT MICPRO* and *IJHPSA*. He also served as the publicity cochair of ISPA 2014 and HiPEAC 2015. He is a member of the IEEE, ACM, and CCF.

Hong Yu received the BS degree from Tsinghua University in 2006 and the PhD degree in bioinformatics from the Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, in 2011. He is an assistant professor with the National Center for Plant Gene Research, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences. He has published more than 10 international journal and conference articles in the research areas of bioinformatics on scientific journals, including *Nature*, *Genome Research*, *Bioinformatics*, and *BMC Systems Biology*.

Aili Wang is a lecturer in the School of Software Engineering, University of Science and Technology of China. She serves as the guest editor of *Applied Soft Computing* and the *International Journal of Parallel Programming*. She is also a reviewer for the *International Journal of Electronics*. She has published 12 international journal and conference articles in the areas of software engineering, operating systems, and distributed computing systems.



Kai Xia received the PhD degree in bioinformatics from the Institute of Genetics and Developmental Biology at the Chinese Academy of Sciences. He is a research assistant professor with UNC Hospitals - Chapel Hill. He has been a postdoctoral researcher with Lineberger Comprehensive Cancer Center and the Department of Biostatistics, UNC-Chapel Hill. His research focuses on Genome wide association study (GWAS) of early brain development in humans, meta-analysis of eQTL studies with independent and correlated subjects, development of efficient tools for GWAS with twin/correlated subjects, and modular regulation in protein-protein interaction network.