Knowledge Discovery Using Big Data in Biomedical Systems

Sarath Chandra Janga, Dongxiao Zhu, Jake Y. Chen, and Mohammed J. Zaki

THE 13th International Workshop on Data Mining in Bioinformatics (BIOKDD'14) was organized in conjunction with the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining on August 24, 2014 in New York, USA. It brought together international researchers in the interacting disciplines of data mining, systems biology, and bioinformatics at the Bloomberg Headquarters venue. The goal of this workshop is to encourage Knowledge Discovery and Data mining (KDD) researchers to take on the numerous challenges that Bioinformatics offers. This year, the workshop featured the theme of "Knowledge discovery using big data in biological/biomedical systems".

In the last few years, there has been a rapid development in various high-throughput technologies, which has led to the accumulation of a large amount of data from different areas of molecular and cellular biology. These developments, together with increasing interest in the community for gaining a systems-wide understanding of the cellular machinery, have provided us unprecedented insights into the structure, organization, and dynamics of various major cellular processes such as transcription, translation, degradation, replication, metabolism, etc. Likewise, efforts to understand the interaction of the cell with external environment have generated global phenotypic maps such as those due to small-molecule perturbations and human microbiomes, which provide us with unparalleled information on the wide variety of microbes that interact with the host's tissues and play an important role in health and disease of an individual. Despite the growing amount of data representing each of these processes it should be admitted that none of these cellular processes work in isolation but rather form an integrated network of different wiring diagrams, which is responsible for the observed behavior of the cell within the context of its environment. While there is mounting evidence from several recent studies that each of these networks of associations associated with a particular cellular process can be studied in detail to provide meaningful insights into how they contribute to the functioning of the cell, as well as to identify the factors that constrain their structure and how they influence the genomes on which they are encoded, it is clear that an open challenge of contemporary biology is to integrate these diverse cellular programs to first understand and model in quantitative terms the topological and dynamic properties of such a unified cellular network, and then to exploit them for the therapeutic benefit of mankind. This field of integrative systems biology, generating large and disparate kinds of biological datasets, is full of opportunities for applying computational and statistical approaches, especially from data mining and machine learning. The goal here is to build accurate predictive or descriptive models of biological processes and diseases, and in integrating data and knowledge-bases from diverse sources to provide experimentally testable hypotheses. These approaches have already revolutionized new age biology by enabling novel discoveries from basic biology to complex disease contexts, as well as in the development of therapeutics. Data mining will continue to play an essential role in understanding these fundamental problems and in the development of novel therapeutic/diagnostic solutions in post-genomic medicine.

Papers for this special section were selected from the BIOKDD'14 workshop. To meet the acceptance criteria for the *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (*TCBB*), each of the papers selected for presentation at the workshop underwent additional reviews by at least two reviewers managed by the *TCBB* editors. We are very grateful to the anonymous reviewers in helping us select the following papers for this special section. The first paper, "Divide and Conquer Approach to Contact Map Overlap Problem using 2D-Pattern Mining of Protein Contact Networks", by Suvarna Vani Koneru and Durga Bhavani S, presents an interesting divide and conquer approach for optimizing the running time for the contact map overlap (CMO) problem. Protein structure prediction and comparison is perhaps one of the most challenging problems in bioinformatics. It is typically modeled as CMO problem in which the similarity of two proteins being compared is measured by the amount of overlap between their corresponding protein contact maps. Protein contact map is a two-dimensional representation of the protein 3D structure. In this study, they propose a novel approach to the CMO problem, which involves finding matching regions between the two contact maps using an approximate 2D-pattern matching algorithm, and dynamic programming technique. These matched pairs of small contact maps are submitted in parallel to a fast heuristic CMO algorithm. The approach facilitates parallelization at this level since all the pairs of contact maps can be submitted to the algorithm in

- *S.C. Janga and J. Y. Chen are with the Department of Biohealth Informatics, School of Informatics and Computing, Indiana University—Purdue University, Indianapolis, IN 46202, and the Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN 46202.*
- *D. Zhu is with the Department of Computer Science, Wayne State University, Detroit, MI 48202.*
- *M. J. Zaki is with the Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY 12180-3590.*

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.2015.2454551

parallel. Then, a merge algorithm is used in order to obtain the overall alignment. The authors show that this algorithm along with achieving better running time can also obtain better overlap for certain protein folds.

The second paper, "Biclustering with Flexible Plaid Models to Unravel Interactions between Biological Processes" by Rui Henriques and Sara C. Madeira, presents an improved biclustering approach which can identify functional modules with associated or interacting genes. Biclusters are subspaces where a subset of rows exhibits a correlated pattern over a subset of columns. The plaid assumption considers the cumulative influence of the contributions from the genes involved in more than one biological process (bicluster) active at a specific condition. Biclusters under a plaid assumption are thus able to compose contributions from multiple biclusters on areas where their rows and columns overlap. Genes can participate in multiple biological processes at a time and thus their expression can be seen as a composition of the contributions from the active processes. Biclustering with a plaid assumption allows the modeling of interactions between transcriptional modules based on overlapping activity levels. The plaid model defines biclusters (subsets of genes with coherent behavior across subsets of conditions) assuming an additive composition of contributions in the areas where they overlap with other biclusters. The authors in this paper propose BiP (Biclustering using Plaid models), a biclustering algorithm with relaxations to allow expression levels to change in overlapping areas according to biologically meaningful assumptions. Such plaid models are biologically significant and can unravel meaningful and non-trivial functional interactions between biological processes associated with the putative regulatory modules. The third paper, "Unsupervised Structure Detection in Biomedical Data" by Julia E. Vogt, presents an intuitive method based on ranked neighborhood comparisons (RaNC) that detects structure in unsupervised data. The method is based on ordering objects in terms of similarity and on the mutual overlap of nearest neighbors. Since the approach is based on ranking of nearest neighbors they call it Ranked Neighborhood Comparison. One interesting aspect about the method is that it doesn't group data into strictly separated groups, but provides a network structure where the links between all objects remain preserved. Especially in biomedical data it is very likely that objects belong to more than just one group, e.g., genes might belong to more than one group by having more than one function. Since the approach doesn't cut the structure into strictly separated groups, it is able to preserve this important information. They also show that the method is robust against outliers. Many biomedical data sets are frequently abundant with outliers either due to experimental or measurement noise. Hence, robustness to outliers is a very useful feature as there is no need to detect and remove outliers in advance in order to avoid incorrect results. As outliers have a higher distance to many data points, these points result in singletons in the graph, and they will not impair finding the underlying structure.

To conclude, the authors thank the authors and the reviewers for their contribution to this special section of the *TCBB* journal. They also thank Prof. Ying Xu and Prof. Dong Xu for their support as editors and assistance from the editorial staff at *TCBB* for making this special section possible.

Sarath Chandra Janga received the PhD degree from the MRC Laboratory of Molecular Biology & University of Cambridge in 2010. He is currently an assistant professor of informatics in the Department of Biohealth Informatics at the School of Informatics and Computing, Indiana University-Purdue University at Indianapolis and a faculty member in the Center for Computational Biology and Bioinformatics at the Indiana University School of Medicine. His research interests include understanding the design principles and constraints imposed on gene regulatory systems within the broader field of computational and systems biology his lab works on. He has published more than 60 research publications on various aspects of basic and applied translational research in the fields of computational, molecular and systems biology.

Dongxiao Zhu received the PhD degree from the University of Michigan in 2006. He is currently an assistant professor in the Department of Computer Science, Wayne State University. From 2008 to 2011, he was an assistant professor in the Department of Computer Science, University of New Orleans. From 2006 to 2008, he was at Stowers Institute for Medical Research as a biostatistician. His research interests have been in areas of genomic data science and machine learning. He has published more than 40 peer-reviewed publications and numerous book chapters and he served on several editorial boards of bioinformatics journals. His research has been supported by NIH, the US National Science Foundation (NSF), and private agencies and he has served on multiple NIH and NSF grant review panels. He has advised numerous students at undergraduate, graduate, and postdoctoral levels.

Jake Chen received the BS degree in biochemistry and molecular biology from Peking University of China, and both the MS and PhD degrees in computer science and engineering from the University of Minnesota at Twin Cities. He is currently an associate professor of bioinformatics and computer science at Indiana University - Purdue University Indianapolis (IUPUI), the founding director of the Indiana Center for Systems Biology and Personalized Medicine, and the founding director of the Zhejiang Institute of Biopharmaceutical Informatics and Technologies in China. He currently serves on the editorial boards of several journals including *BMC Systems Biology*, the *IEEE Journal of Biomedical and Health Informatics, Personalized Medicine,* and *Network Biology*. Prior to joining academia in 2004, he worked for six years as a bioinformatics computer scientist in the Silicon Valley and Salt Lake City to develop late-breaking DNA microarray and proteomics technologies. His primary research interest is in developing computational systems biology techniques and models that lead to future predictive and personalized medicine. At IUPUI, he has published more than 100 peer-reviewed research papers and filed several key patents in systems pharmacology—the application of systems biology techniques to drug discovery. In 2011, he was selected by the National Academy to serve on an Institute of Medicine committee that advises FDA on food and drug regulatory systems harmonization matters in developing countries. He was cited by HealthTechTopia as one of the "17 Informatics Experts Worth Listening To" in 2011. In 2012, he received a "Cancer Systems Biology Grand Challenge Award" by Innocentive.com, for successfully predicting a cancer drug's molecular mechanisms of action among the site's 250,000 community of scientists worldwide. In both 2013 and 2014, he was recognized as an "Indiana Technology Educator of the Year" MIRA Award finalist, for his contribution to informatics research, education, and their impact to the state of Indiana. He is the founder of Medeolinx, LLC and Medeolinx Software, Ltd., companies aiming to advance innovative informatics applications in drug development.

Mohammed J. Zaki received the PhD degree in computer science from the University of Rochester in 1998. He is a professor of computer science at RPI. His research interests focus on developing novel data mining techniques, especially for applications in bioinformatics and social networks. He has over 225 publications, including the textbook, *Data Mining and Analysis: Fundamental Concepts and Algorithms*, Cambridge University Press, 2014. He is an area editor for *Statistical Analysis and Data Mining*, and an associate editor for *Data Mining and Knowledge Discovery*, the *ACM Transactions on Knowledge Discovery from Data*, and *Social Networks and Mining*. He was the program co-chair for SDM'08, SIGKDD'09, PAKDD'10, BIBM'11, CIKM'12, ICDM'12, BigData'15. He is currently serving on the Board of Directors for ACM SIGKDD. He is a recipient of the US NSF CAREER Award and the DOE Early Career Award. He is a senior member of the IEEE, and an ACM Distinguished Scientist. His research is supported in part by the US National Science Foundation (NSF), NIH, DOE, Google, HP, and Nvidia.