Guest Editorial: Special Section on Biological Data Mining and Its Applications in Healthcare

Fei Wang, Xiao-Li Li, Jason T. L. Wang, and See-Kiong Ng

Abstract—Biologists are stepping up their efforts in understanding the biological processes that underlie disease pathways in the clinical contexts. This has resulted in a flood of biological and clinical data—genomic sequences, DNA microarrays, protein interactions, biomedical images, disease pathways, etc. The rapid adoption of Electronic Health Records (EHRs) across healthcare systems, coupled with the capability of linking EHRs to research biorepositories, provides a unique opportunity for conducting large-scale Precision Medicine research. As a result, data mining techniques, for knowledge discovery and deriving data driven insights from various data sources, are increasingly important in modern biology and healthcare. The purpose of this special section is to bring together the researchers in bioinformatics, healthcare informatics, and data mining to share about their current research, and their visions on future directions.

Index Terms—Biological data mining, bioinformatics, healthcare

1 Introduction

Today, biologists are stepping up their efforts in understanding the biological processes that underlie disease pathways in the clinical contexts. Recent advances in biotechnology have resulted in a flood of biological data such as genomic sequences, DNA microarrays, and protein interactions. At the same time, the healthcare industry has begun to embrace powerful big data technologies, leading to a dramatic shift toward data-driven healthcare management and decision making. While these data are useful for knowledge discovery and decision making, we are in a situation where our capability of generating biomedical data has greatly surpassed our abilities to mine and analyse them. As a result, large amounts of complex clinical data about patients, hospital resources, disease diagnosis and electronic patient records have been generated. However, we are faced with the following challenges: how to properly handle noisy and incomplete data (e.g., protein interactions have high false positive and false negative rates), how to efficiently process computationally-intensive tasks (e.g., large scale graph mining), and how to integrate heterogeneous data sources (e.g., linking genomic data with clinical databases).

Data mining techniques, designed to extract useful information from large databases or data warehouses, have started to demonstrate their huge potentials in solving the aforementioned challenges and could be the next technical innovation that enables biologists and medical researchers to gain insightful observations and make groundbreaking discoveries in molecular biology as well as the pharmaceutical and clinical domains. Therefore, there are unprecedented opportunities for data mining researchers to contribute to this meaningful scientific pursuit together with the biologists and clinical scientists.

- *F. Wang is with the Division of Health Informatics, Department of Healthcare Policy and Research, Weill Cornell Medical College, Cornell University, Ithaca, NY 14850. E-mail: feiwang03@gmail.com.*
- *X.-L. Li is with the Institute for Infocomm Research, Singapore 138632. E-mail: xlli@i2r.a-star.edu.sg.*
- *J.T.L. Wang is with the Department of Computer Science, New Jersey Institute of Technology, Newark, NJ 07102. E-mail: wangj@njit.edu.*
- *S.-K. Ng is with the Institute of Data Science, National University of Singapore, Singapore 117417. E-mail: seekiong@nus.edu.sg.*

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.2016.2612558

2 This Special Section

We started our workshop series on "Biological Data Mining and its Applications in Healthcare (BioDM)" in 2009, and it has since been held in conjunction with the IEEE International Conference on Data Mining (ICDM) each year. This special section provides a leading focused forum for timely, in-depth presentation of recent advances in algorithms, theory, and applications on data mining technologies for biological and clinical data. The papers have been selected through a rigorous reviewing process, addressing classic problems in healthcare, such as healthcare quality management, biomedical annotations, clinical outcome prediction, and military acute concussion evaluation, as well as tasks in biological studies such as protein sequence analysis, protein function prediction, genetic selection. Overall, the special section includes two lines of work.

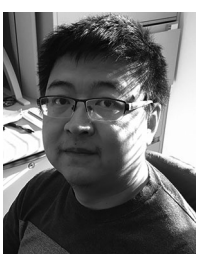
Biological Data Mining. The paper "From Protein Sequence to Protein Function via Multi-Label Linear Discriminant Analysis" written by Huang et al. concerns the problems of protein function prediction and solves it with a novel Multi-Label Linear Discriminant Analysis (MLDA) approach. In the paper "Towards Unsupervised Gene Selection: A Matrix Factorization Framework," Jianqiang Li and Fei Wang propose a novel unsupervised two-stage coarse-fine gene selection method that filters out redundant genes with a k-means algorithm and then selects the most representative genes using matrix factorization techniques. The paper "Efficient Approach to Correct Read Alignment for Pseudogene Abundance Estimates" by Chelsea J.-T. Ju, Zhuangtian Zhao, and Wei Wang proposes an extension of the PseudoLasso to learn the read alignment behaviors for RNA-sequencing, which is the leading technology to quantify expression of thousands of genes.

Clinical Data Mining. In addition, entity relationships, data representation, and pattern mining based on EHR data also receive extensive attentions. In the paper "Applications of Transductive Spectral Clustering Methods in a Military Medical Concussion Database," Peter B. Walker, Jacob N. Norris, Anna E. Tschiffely, Melissa L. Mehalick, Craig A. Cunningham, and Ian N. Davidson demonstrate the advantages of spectral graph methods in identifying the relationship between the administration of specific medications and reductions in traumatic brain injury symptomology from a big military medical database. In the paper "Modeling Healthcare Quality via Compact Representations of Electronic Health Records," Jelena Stojanovic, Djordje Gligorijevic, Vladan Radosavljevic, Nemanja Djuric, Mihajlo Grbovic, and Zoran Obradovic learn vector representations of patient conditions and clinical procedures in an unsupervised manner from a large-scale EHR database comprising more than 35 million hospitalizations. In the paper "Prognosis of Clinical Outcomes with Temporal Patterns and Experiences with One Class Feature Selection," Jelena Stojanovic, Djordje Gligorijevic, Vladan Radosavljevic, Nemanja Djuric, Mihajlo Grbovic, and Zoran Obradovic propose a framework that learns predictive models based on the temporal patterns in the clinical records that are prognostic markers and use these markers to train predictive models for eight clinical procedures. In another paper "Bi-convex Optimization to Learn Classifiers from Multiple Biomedical Annotations," Xin Wang and Jinbo Bi propose a general optimization approach for solving the label ambiguity problem in biomedical data.

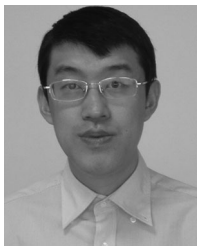
3 Conclusions and Future Directions

This special section contains seven high quality papers, which cover different aspects on biological data mining and its applications in healthcare. For future research, the following directions are very promising:

- *Integration of biological and clinical data*. The ability to combine biological data with clinical data effectively will be a key enabler for precision medicine research.
- *Privacy preservation*. Mining useful information from both biological and clinical data without leaking sensitive patient information is an important research problem.
- *Knowledge integration*. The biological and clinical knowledge are invaluable for healthcare. With limited patient samples, combining knowledge and data in the mining process is crucial.



Fei Wang is an assistant professor in the Division of Health Informatics, Department of Healthcare Policy and Research, Cornell University. His major research interests include data analytics and its applications in health informatics. He has published more than 150 papers on top data mining and medical informatics venues. His papers have received more than 3,300 citations so far. He won best student paper for ICDM 2015, best research paper nomination for ICDM 2010, Marco Romani Best paper nomination in AMIA TBI 2014, and his paper was selected as a best paper finalist in SDM 2011 and 2015. He is the vice chair of the KDD working group in AMIA. He is a senior member of the IEEE.



Xiao-Li Li is currently a department head (Data Analytics Department) and a senior scientist in the Institute for Infocomm Research, A*STAR, Singapore. He also holds adjunct associate professor positions with National University Singapore and Nanyang Technological University. He has published more than 150 peer-reviewed papers, including top tier data mining, machine learning, artificial intelligence, information retrieval conferences, such as KDD, ICDM, SDM, PKDD/ECML, PAKDD, ICDE, ICML, IJCAI, AAAI, UAI, ACL, SIGIR, EMNLP, CIKM, UbiCom, etc., as well as some top tier journals such as the *IEEE Transactions on Knowledge and Data Engineering* and the *IEEE Transactions on Reliability*. Some impactful bioinformatics papers were published in *PLOS Computational Biology, Bioinformatics, BMC Genomics, PLOS ONE, BMC Bioinformatics, Annals of the New York Academy of Sciences, Journal of Computational Biology*, as well as some top bioinformatics conferences such as ISMB/ECCB, CSB etc. He has served program committees/workshop chairs/session chairs for leading data analytics, artificial intelligence and bioinformatics related conferences, such as KDD, ICDM, SDM, ECML/PAKDD, PAKDD, ACML, WWW, IJCAI, AAAI, ACL, NAACL-HLT, CIKM, GIW, InCOB, BIBM, BIBE, BCB, ICSH etc.

Jason T. L. Wang is a professor of computer science in the New Jersey Institute of Technology and the director with the University's Data and Knowledge Engineering Laboratory. His research interests include databases and data mining, data science and analytics, big data, machine learning, and computational biomedicine. He has published more than 150 refereed papers and presented three SIGMOD software demos in these areas. His research has been supported by Novartis Pharmaceuticals Corporation, AT&T Foundation, Alfred P. Sloan Foundation, James S. McDonnell Foundation, Howard Hughes Medical Institute, U.S. Department of Defense (DoD), U.S. National Institutes of Health (NIH), and U.S. National Science Foundation (NSF). He has served on the program committees of more than 200 national and international conferences, was program co-chair of the 2001 Atlantic Symposium on Computational Biology, Genome Information Systems & Technology held at Duke University, program co-chair of the 1998 IEEE International Joint Symposia on Intelligence and Systems held at Rockville, Maryland, a founding chair of the ACM SIGKDD Workshop on Data Mining in Bioinformatics, a workshop co-chair of the 2010 ACM International Conference on Bioinformatics and Computational Biology, and a co-chair of the 2006 IEEE ICDM Workshop on Data Mining in Bioinformatics, and the 2011 IEEE ICDM Workshop on Biological Data Mining and its Applications in Healthcare. He is listed in Who's Who among Asian Americans, and Who's Who in Science and Engineering.

See-Kiong Ng is currently professor of practice in the Department of Computer Science at National University of Singapore's (NUS) School of Computing. He is the deputy director of the Institute of Data Science at NUS. Dr. Ng's diverse and cross-disciplinary research interests include bioinformatics, text mining, social network mining, and privacy-preserving data mining. His primary research objective is to unravel the underlying functional mechanisms of dynamic real world networks from protein interaction networks to social networks in order to understand the "biology" of complex systems through computational technologies and data-driven approaches.