Data Mining in Bioinformatics: Selected Papers from BIOKDD

Stefano Lonardi and Jake Chen

BIOINFORMATICS is the science of managing, mining, and interpreting information from observations of biological processes. Various genome projects have contributed to an exponential growth in DNA and protein sequence databases. In the postgenome era, advances in high-throughput technology such as microarrays and mass spectrometry have further created the fields of functional genomics and proteomics, in which one can monitor quantitatively the presence of multiple genes, proteins, metabolites, and compounds in a given biological state.

Data mining approaches are ideally suited for bioinformatics. The ongoing influx of these data, the presence of biological signals despite high data noises, and the gap between data collection and knowledge extraction have collectively created new and exciting opportunities for data mining researchers in this field. The extensive availability of open-access biological databases has created both challenges and opportunities for developing novel knowledge discovery and data mining methods specific to molecular biology. While tremendous progress has been made over the years, many of the fundamental problems in bioinformatics, such as protein structure prediction, protein-ligand interactions, molecular pathway mapping, and gene regulatory network modeling, are still open. Many of the current research problems are comprehensively covered in a recent book by Chen and Lonardi [1]. Data mining will play essential roles in understanding these fundamental problems and developing novel therapeutic/diagnostic solutions in post-genome medicine.

To provide avenues to data mining researchers active in bioinformatics, we organized the 2008 International Workshops on Data Mining in Bioinformatics (BIOKDD), held 24-27 August in Las Vegas, NV (http://bio.informatics.iupui.edu/biokdd08/). In this issue of *TCBB*, we present two extended papers chosen from nine peer-reviewed papers originally presented at the workshop. The two papers were among nine top ranked papers with the highest peer review scores from more than 30 papers received at the BIOKDD workshop. They have been further refereed according to the *TCBB* manuscript submission standards: each invited paper was reviewed by three external referees and went through one or two rounds of revisions. In the end, two out of the four invited papers were included for publication.

The two contributions appearing in this special section are "Molecular Function Prediction Using Neighborhood Features" by Petko Bogdanov and Ambuj K. Singh and "GPD: A Graph Pattern Diffusion Kernel for Accurate Graph Classification with Applications in Cheminformatics" by Aaron Smalter, Jun Huan, Yi Jia, and Gerald Lushington. Bogdanov and Singh describe a framework for predicting functional annotation in molecular interaction networks. Contrary to the popular assumption that genes with similar functions are topologically close in interaction networks, the authors held a different view—genes with similar functions share similar annotation patterns in their network neighborhood independent of their distance in the network. This led the authors to classify and subsequently infer gene functions based on network neighborhood. Their approach appears to be more robust against noise and to be more adaptable be integrated with homology information. Smalter et al. describe a new computational framework for chemical structure comparisons and database search, based on the new notion of graph pattern diffusion kernel. Their approach allows the discovery and labeling of frequent graph patterns in a database of chemical structures. It represents a significant novel development for the interdisciplinary research community that aims to support chemical biology.

In closing, we would like to thank all of the contributing authors, the reviewers, and the staff of *TCBB* for the support offered during this project.

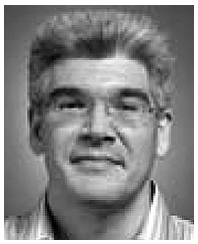
Stefano Lonardi
Jake Chen
Guest Editors

REFERENCES

[1] J. Chen and S. Lonardi, *Biological Data Mining.* Chapman & Hall/CRC, 2010.

- *S. Lonardi is with the Department of Computer Science and Engineering, University of California at Riverside, Riverside, CA 92521. E-mail: stelo@cs.ucr.edu.*
- *J. Chen is with the School of Informatics, Indiana University–Purdue University, Indianapolis, IN 46202. E-mail: jakechen@iupui.edu.*

For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org.



Stefano Lonardi is an associate professor of computer science and engineering at the University of California, Riverside. He received his "Laurea cum laude" from the University of Pisa in 1994 and the PhD degree in 2001 from the Department of Computer Sciences, Purdue University, West Lafayette, IN. He also holds a doctorate degree in electrical and information engineering from the University of Padua (1999). During the summer of 1999, he was an intern at Celera Genomics, Department of Informatics Research, Rockville, MD. His recent research interest includes computational molecular biology, data compression, and data mining. He has published more than 35 papers in major theoretical computer science and computational biology journals and has about 45 publications in referred international conferences. In the year 2005, he received the CAREER award from the US National Science Foundation.

Jake Chen is the founding director of the Indiana Center for Systems Biology and Personalized Medicine since 2007, a faculty member of Purdue University School of Science, Indiana University School of Informatics, and a member of the Indiana University Cancer Center, Center for Computational Biology and Bioinformatics, and Center for Biocomputing since 2004. He is the central Indiana chapter chair of the IEEE Engineering in Biology and Medicine Society, a senior member of both the IEEE and the ACM, and an associate editor of BMC Systems Biology. Prior to joining Academia, he spent six years in the biopharmaceutical industry in the Silicon Valley and Utah doing research in bioinformatics, functional genomics, and proteomics. His research interests spans over biological knowledge discovery, biological information visualization, and translational bioinformatics. He has published more than 70 peer-reviewed publications, including two recent edited books, *Biological Database Modeling* and *Biological Data Mining*.