

Data Mining and Its Applications in Bioinformatics: Techniques and Methods

Xiaohua Hu

College of Information Science and Technology
Drexel University
Philadelphia, PA, 19104, USA
e-mail: xh29@drexel.edu

Abstract—In this talk, I will discuss some of the latest data mining techniques and methods and their applications in bioinformatics study, focusing on data integration, text mining and graph-based data mining in bioinformatics research. In data integration, I will present a semantic-based approach for multi source bioinformatics data integration. In our approach, a metamodel is utilized to represent the master search schema, and an effective interface extraction algorithm based on the hierarchical structure of the web and pattern is developed to capture the rich semantic relationships of the online bioinformatics data sources. Our final goal is to develop a meta-search interface for biologists as a single point of access to multiple online bioinformatics databases. In text mining, some of the challenging issues in mining and searching the biomedical literature are addressed, and I will present a unified architecture Bio-SET-DM (Biomedical Literature Searching, Extraction and Text Data Mining), discuss some novel algorithms such as semantic-based language model for literature retrieval, semi-supervised pattern learning for information extraction of biological relationships from biomedical literature. In the third part, graph-based data mining, the focus is on graph-based mining in biological networks. I will discuss how to apply graph-based mining techniques and algorithms in the analysis of modular and hierarchical structure of biological networks, how to identify and evaluate the subnetworks from complicated biological networks, and present the experimental results. To put these pieces together, a unified framework is introduced to integrate the three parts (data integration, text mining and graph-based data mining) in the bioinformatics data mining procedure.

SHORT BIOGRAPHY

Xiaohua (Tony) Hu is a full professor and the founding director of the [data mining and bioinformatics lab](#) at the College of Information Science and Technology. He is also serving as the IEEE Computer Society Bioinformatics and Biomedicine Steering Committee Chair, the IEEE Computer Society Granular Computing Steering Committee Co-Chair, and the Co-Director of the NSF Industry/University Cooperative Research Center (I/U CRC) on Visual and Decision Informatics. He founded the [International Journal of Data Mining and Bioinformatics](#) (SCI indexed) in 2006, [International Journal of Granular Computing, Rough Sets and Intelligent Systems](#) in 2008. Earlier, he worked as a research scientist in the world-leading R&D centers such as Nortel Research Center, and Verizon Lab (the former GTE labs). In 2001, he founded the DMW Software in Silicon Valley, California. He has a lot of experience and expertise

to convert original ideas into research prototypes, and eventually into commercial products, many of his research ideas have been integrated into commercial products and applications in data mining fraud detection, database marketing.

Tony's current research interests are in data/text/web mining, bioinformatics, information retrieval and information extraction, social network analysis, healthcare informatics, rough set theory and application. He has published more than 200 peer-reviewed research papers in various journals, conferences and books such as various IEEE/ACM Transactions (IEEE/ACM TCBB, IEEE TFS, IEEE TDKE, IEEE TITB, IEEE SMC, IEEE Computer, IEEE Intelligent Systems), JIS, KAIS, CI, DKE, IJBRA, SIG KDD, IEEE ICDM, IEEE ICDE, SIGIR, ACM CIKM, IEEE BIBE, IEEE CICBC etc, co-edited 16 books/proceedings. He has received a few prestigious awards including the [2005 National Science Foundation \(NSF\) Career award](#), the [best paper award](#) at the 2007 International Conference on Artificial Intelligence, the [best paper award](#) at the 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, [the 2010 IEEE Granular Computing Outstanding Contribution Awards](#), the [2007 IEEE Bioinformatics and Bioengineering Outstanding Contribution Award](#), the [2006 IEEE Granular Computing Outstanding Service Award](#), and the [2001 IEEE Data Mining Outstanding Service Award](#). He has also served as a program co-chair/conference co-chair of 14 international conferences/workshops and a program committee member in more than 80 international conferences in the above areas. He is the founding editor-in-chief of the [International Journal of Data Mining and Bioinformatics](#) (SCI indexed), [International Journal of Granular Computing, Rough Sets and Intelligent Systems](#), an associate editor/editorial board member of four international journals ([KAIS](#), [IJDWM](#), [JJSOI](#) and [JCIB](#)). His research projects are funded by the National Science Foundation (NSF), [US Dept. of Education](#), the [PA Dept. of Health](#), the Natural Science Foundation of China (NSFC).

Tony received his Ph.D. in Computer Science from [University of Regina](#), Canada 1995, M.Sc. in Computer Science from [Simon Fraser University](#), Canada in 1992, M.Eng. in Computer Engineering from [Institute of Computing Technology, Chinese Academy of Science](#) in 1988 and B.Sc. (Software) from [Wuhan University](#) in 1985.