Introduction to the Special Issue on Data Mining for Health Informatics

Raymond T. Ng
Department of Computer Science
The University of British Columbia
Vancouver, BC, Canada
rng@cs.ubc.ca

Jian Pei
School of Computing Science
Simon Fraser University
Burnaby, BC, Canada
jpei@cs.sfu.ca

One of the holy grails of medical research in the next decade is what is often called "personalized" medicine. The goal is to individualize therapy and treatment options, and even prevention strategies. This movement is fueled by the rapid advances made in high-throughput biotechnologies. Prime examples include SNP (Single Nucleotide Polymorphisms) chips and CGH (Comparative Genomic Hybridization) arrays for DNA profiling, oligonucleotide arrays for mRNA expression, and advanced mass spectrometry for peptide/protein and metabolite quantitation.

Regarding these technologies, the good news is that they have become relatively inexpensive (e.g., hundreds of dollars per sample), making them widely accessible to researchers. However, the bad news to many medical researchers is that the amount of data collected by these devices is phenomenal. For instance, the number of "units" on a CGH array is over 20,000, and there are now 100K SNP chips being used. Oligonucleotide arrays can be of a resolution an order of magnitude higher. Thus, as a simple illustration, a medical study with, say, 200 patients and 5 samples per patient can lead to a data set with hundreds of millions of data points. Data mining techniques, hence, become attractive for many medical and health studies. Broadly defined, there are four general objectives for the data mining activities in health informatics:

- Diagnostics: to determine whether a patient is suffering from a certain medical condition. For instance, early stage lung and oral cancers are very hard to diagnose by conventional means; genomic signatures can be used to provide more timely and perhaps more accurate diagnosis.
- Prognostics: to predict how well a patient would recover or how the medical condition would progress over time. For instance, biomarkers have been identified to predict how well a transplanted organ would be tolerated in the recipient's body.
- Treatment optimization: to predict the response to treatment or therapies. For example, for certain cancer types, biomarkers can be used to predict whether a certain chemotheraphy regimen would be effective or not. Pharmaco-genomics is a very active area of research on understanding how pharmaceuticals and medications can affect the genomic profile of a patient.
- Understanding of disease mechanisms: to provide new insights into how a certain medical condition is triggered. For example, it is an active area of research on finding out how signalling pathways interact during a viral infection.

This special issue focuses on health informatics. The objectives shown above serve to distinguish health informatics from bioinformatics, which focuses more on general biology. Although bioinformatics has received tremendous interest from data mining researchers, data mining for health informatics is relatively new for data mining community. To promote real problems and challenges in health informatics to data mining community, we organized this special issue.

The call-for-papers of this special issue received much attention. From a good number of submissions, we are lucky to have the four very interesting articles in this special issue. All of them address interesting, important and diverse issues in health informatics research and practice.

Analysis of gene expression data has been established as one of the major topics in health informatics. The first paper written by Donny Soh, Difeng Dong, Yike Guo, and Limsoon Wong, titled "*Enabling more sophisticated gene expression analysis for understanding diseases and optimizing treatments*" reviews the state-of-the-art progress of gene expression data analysis, and identifies three important challenges that the current methods cannot meet satisfactorily. The authors report their on-going project towards an advanced integrated system for gene expression profile analysis, and illustrate their ideas on a logical model and reasoning system for biomolecular events.

In in-silico drug discovery, small biomolecular classification and structural similarity search are important tools. The second paper, "*Novel approaches for small biomolecule classification and structural similarity search*" by Emre Karakoc, S. Cenk Sahinalp, and Artem Cherkasov, presents the design of the optimal weighted Minkowski distance for k-nearest-neighbor search on biomolecular structures. New data structures and algorithms are developed. The novel approach outperforms the traditional methods according to their empirical study.

Interdisciplinary collaboration is crucial in health informatics research. The third paper, "*Drug exposure side effects from mining pregnancy data*" by Yu Chen, Lars Henning Pedersen, Wesley W. Chu, and Jorn Olsen, demonstrates a successful interdisciplinary collaborative research project on mining possible side effects due to explosure to multiple drugs at different duration of pregnancy. Different from the

first two papers, this paper focuses on clinical data instead of genomic data.

Last but not least, the fourth paper "*Automatic in vivo microscopy video mining for leukocytes*" by Chengcui Zhang, Wei-Bang Chen, Lin Yang, and Xin Chen provides an interesting example of applying video mining approaches to leukocytes recognition. Blood is one of the most important fluids in the human body. In general, tracking the flow of a body fluid has many diagnostic applications.

A few experts in the field provided timely and careful reviews of the submissions. We are grateful to the reviewers for their timely and constructive reviews. The reviewers in alphabetical order include Jack Nansheng Chen, Ben Good, David Lowe, Zengzeng Xing, Hui Xiong, Mohammed Zaki. Without their help, this special issue will not be possible. We also thank the authors who responded to the call-for-papers and submitted their articles to this special issue.