Data Mining for Biomedical Informatics Workshop Schedule

Section 1 (8:30 AM – 10:00 AM)

8:30 AM - 8:40 AM
Welcome comments

8:40 AM - 9:00 AM
The L2 Discrepancy Framework to Mine High-Throughput Screening Data for Targeted Drug Discovery: Application to AIDS Antiviral Activity Data of the National Cancer Institute
F. El Khettabi and P. Kyriakidis

9:00 AM - 9:20 AM
Methods for Effective Scaffold Hopping in Chemical Compounds
N. Wale, G. Karypis, and I. Watson

9:40 AM - 10:00 AM
A Statistical Model for Functional Characterization of Regulatory Pathways
J. Pandey, M. Koyuturk, W. Szpankowski, and A. Grama

10:00 AM – 10:30 AM
Coffee Break

10:30 AM – 11:15 AM
Keynote Talk
Inferencing Across Clinical and Genomic Data: Mining Madness, Statistical Folly, and the Joy of Systems Biology
Christopher G. Chute, Mayo Clinic, College of Medicine

The advent of the human genome and its integration into clinical medicine poses one of the great challenges for biomedicine into the 21st century. Colloquially referenced as translational medicine, the challenges from data management, data representation, and inferencing perspectives are formidable. Recent technology is able to generate over one million SNPs per specimen per analytic
work frame. Correlating this volume of data with hundreds or thousands of phenotypic expression characteristics causes traditional stochastic models and statistical methods to collapse. The historical work around to the “statistical fishing expedition” problem has been hypothesis driven research. However, in a data discovery model, hypothesis generation has almost equal importance in this brave new world. One mechanism is to leverage the knowledge resources implicit (explicitly not yet explicit) in systems biology and dynamic pathway networks of metabolic systems, sub-cellular physiology, and the new integrated biology. This talk will outline some of the data representation issues, frameworks for biological workflow analysis, and introduce the promise of systems biology.

Dr. Chute is Professor and Chair of Biomedical Informatics at Mayo Clinic College of Medicne. He received his undergraduate and medical training at Brown University, internal medicine residency at Dartmouth, and doctoral training in Epidemiology at Harvard. He is a Fellow of the American College of Physicians, the American College of Epidemiology, and the American College of Medical Informatics.

As a career scientist at Mayo, Dr. Chute's NIH and AHCPR/AHRQ funded research in medical concept representation, clinical information retrieval, and patient data repositories have been widely published. He is Vice-chair of the ANSI Health Information Standards Board,
Convener of Healthcare Concept Representation WG3 within the ISO Health Informatics Technical Committee, chair-elect of the US delegation to ISO TC215 for Health Informatics, co-chair of the HL7 Terminology Committee and a past member of the NIH Medical Informatics Study Section.
He has chaired International Medical Informatics Association WG6 on Medical Concept Representation since 1994.

<u>Section 2</u> (11:15 AM – 12:00 PM)

11:15 AM - 11:40 AM
Discovery of Principles of Nature From Matrix and Tensor Modeling of DNA Microarray Data
Orly Alter

11:40 AM - 12:00 PM
Comparative Study of Various Genomic Data Sets for Protein Function Prediction and Enhancements Using Association Analysis
R. Gupta, T. Garg, G. Pandey, M. Steinbach, and V. Kumar

12:00 PM – 1:15 PM
Lunch Break

1:15 PM – 2:00 PM
Keynote Talk
DNA sequence variation around the genome and around the world
Kenneth K. Kidd, Ph.D., Prof. of Genetics, Psychiatry, and Ecology & Evolutionary Biology
Yale University

Even before the explosion of genetic data on humans in the past two and a half decades, two facts were clear: considerable normal genetic variation exists in all human populations and the frequencies of the different variants (alleles) vary from population to population. The molecular data on DNA sequence variation has confirmed those in spades and allowed us to begin to quantify many aspects of variation on a genome-and species-wide basis. Numerous interesting questions exist and many are analytically and/or computationally challenging.

However, the data often exist in diverse locations and are often fragmentary, e.g., different populations studied for different genetic variants precluding any direct comparison between populations. Some resources are available and others will be available soon. The HapMap has data on the largest number of SNPs but is limited to only four populations that cannot fully represent global genetic variation. Data sets on larger numbers of populations have much more limited data. However, enough data exist now that the global pattern of human population similarity is
becoming clear and will be presented along with an overview of online resources currently available.

Biography:
Kenneth K. Kidd received his Ph.D. in Genetics from the University of Wisconsin in 1969. His early training included Drosophila genetics, classical immunogenetics, and population genetics. During his post-doctoral studies in Italy and at Stanford University, he established his reputation in human population genetics. He joined the Genetics faculty at Yale University School of Medicine in 1973 where he has remained and is currently Professor of Genetics, Psychiatry, and Ecology and Evolutionary Biology. At Yale he has pursued research in many areas of human genetics, including medical genetics (studies of neuropsychiatric disorders and simple Mendelian disorders), gene mapping (both physical and genetic), database design for modern genetic data, and a variety of molecular methodologies. More recently, his long-standing interest in human population genetics has been combined with his laboratory's expertise in molecular technology to examine human genome diversity at the DNA level. He is also responsible for ALFRED, the ALlele FREquency Database, a web accessible compilation of allele frequency data for DNA polymorphisms on anthropologically defined human populations.

During his career, Dr. Kidd has published more than 450 scientific articles in a broad range of subjects including population genetics, cancer and neuropsychiatric genetics, gene mapping, molecular methodology, genetic databases, and human diversity. He is one of the co-authors of a paper selected as the best biomedical paper of the year by The Lancet, a leading British medical journal. This and other publications by Dr. Kidd can be found on his web site http://info.med.yale.edu/genetics/kkidd. He is a certified Medical Geneticist by the American Board of Medical Genetics. He has served on several U.S. Government Review and Advisory Committees/Panels, on several editorial boards, and helped organize several international conferences. He is a member of several professional societies and a Fellow of the American Association for the Advancement of Science. Among his other awards, he has been recognized by the U.S. Federal Bureau of Investigation and the National Institute of Justice for his contributions toward acceptance of DNA methodologies in the courts. He recently served on national advisory panels for DNA identification of victims of the World Trade Center attack and victims of Katrina.

Section 3 (2:00 PM – 3:00 PM)

2:00 PM - 2:20 PM

Extracting tagging SNPs from Genome-wide Datasets
A. Javed and P. Paschou

2:20 PM - 2:40 PM
Multiscale Analysis of Data Sets with Diffusion Wavelets
M. Maggioni and R.R. Coifman

2:40 PM - 3:00 PM
Exploration of high dimensional biomedical datasets with
low-distortion embeddings
F. Meyer and X. Shen

3:00 PM – 3:30 PM
Coffee Break

Section 4 (3:30 PM – 5:00 PM)

3:30 PM - 3:50 PM
Mining Edge-disjoint Patterns in Graph-relational Data
C. Besemann and A. Denton

3:50 PM - 4:10 PM
Generalized Replicator Dynamics for Efficient Phylogenetic Inference
W. Li and Y. Liu

4:10 PM -4:30 PM
Finding Differentially Expressed Genes Through Noise Elimination
A. Denton and A. Kar

4:30 PM -4:50 PM
Computing Overlapping Protein Interaction Modules
C. Ding, C. Wang, and S. Holbrook