BigLS 2015 ACM International Workshop on Big Data in Life Sciences

September 9, 2015

Held with the ACM Conference on Bioinformatics, Computational Biology and Health Informatics Atlanta, GA

TECHNICAL PROGRAM

8:25am - 10:30am	INVITED TALKS
8:25am - 8:30am	Opening remarks, Ananth Kalyanaraman and Jaroslaw Zola
8:30am - 9:15am	"Building Scalable Health Analytic Platform: Computational Phenotyping and Cloud-based Predictive Modeling" Jimeng Sun, School of CSE, Georgia Tech
9:15am - 10am	"Big Data Phylogenetics" Suzanne J. Matthews, US Military Academy
10am - 10:30am	Coffee break
10:30am-12pm	KEYNOTE ADDRESS
	"Machine Learning Framework for Classification in Medicine and Biology" Eva K. Lee, Director of the Center for Operations Research in Medicine and HealthCare
12pm-1:30pm	Lunch
1:30pm-2:15pm	INVITED TALK
	"GHOST of Viral Molecular Surveillance" Yury Khudyakov, Chief of Molecular Epidemiology and Bioinformatics Lab, Center for Disease Control and Prevention
2:15pm-3:15pm	PEER REVIEWED PAPERS
2:15pm-2:45pm	"Scalable Multipartite Subgraph Enumeration for Integrative Analysis of Heterogeneous Experimental Functional Genomics Data" C. Phillips, K. Wang, J. Bubier, E. Baker, E. Chesler, M. Langston
2:45pm-3:15pm	"A Multi-Agent System with Reinforcement Learning Agents for Biomedical Text Mining" M. Camara, O. Bonham-Carter, J. Jumadinova
3:15pm-5pm	STUDENT POSTERS AND MENTORING SESSION
3:15pm-4pm	Student poster presentations
4pm-5pm	Mentoring interaction between faculty/researchers and students

KEYNOTE ADDRESS AND INVITED TALK ABSTRACTS

Keynote address

Prof. Eva K. Lee, Director of the NSF-Whitaker Center for Operations Research in Medicine and HealthCare, Co-Director of the NSF I/UCRC Center for Health Organization Transformation Distinguished Scholar in Health System, Health System Institute, Georgia Tech/Emory University Professor, School of Industrial and Systems Engineering
Georgia Institute of Technology

Machine Learning Framework for Classification in Medicine and Biology

Systems modeling and quantitative analysis of large amounts of complex clinical and biological data may help to identify discriminatory patterns that can uncover health risks, detect early disease formation, monitor treatment and prognosis, and predict treatment outcome. In this talk, we describe a machine-learning framework for classification in medicine and biology. It consists of a pattern recognition module, a feature selection module, and a classification modeler and solver. The pattern recognition module involves automatic image analysis, genomic pattern recognition, and spectrum pattern extractions. The feature selection module consists of a combinatorial selection algorithm where discriminatory patterns are extracted from among a large set of pattern attributes. These modules are wrapped around the classification modeler and solver into a machine learning framework. The classification modeler and solver consist of novel optimization-based predictive models that maximize the correct classification while constraining the inter-group misclassifications. The classification/ predictive models 1) have the ability to classify any number of distinct groups; 2) allow incorporation of heterogeneous, and continuous/time-dependent types of attributes as input; 3) utilize a high-dimensional data transformation that minimizes noise and errors in biological and clinical data; 4) incorporate a reserved-judgement region that provides a safeguard against over-training; and 5) have successive multi-stage classification capability. Successful applications of our model to developing rules for gene silencing in cancer cells, predicting the immunity of vaccines, identifying the cognitive status of individuals, and predicting metabolite concentrations in humans will be discussed.

Invited Talk #1

Dr. Jimeng Sun, Associate Professor, School of Computational Science and Engineering, College of Computing, Georgia Institute of Technology

Building Scalable Health Analytic Platform: Computational Phenotyping and Cloud-based Predictive Modeling

As the adoption of electronic health records (EHRs) has grown, EHRs are now composed of a diverse array of data, including structured information (e.g., diagnoses, medications, and lab results), and unstructured clinical progress notes. Two unique challenges need to be addressed in order to utilize EHR data in clinical research and practice:

1) Computational Phenotyping: How to turn complex and messy EHR data into meaningful clinical

concepts or phenotypes?

2) Scalable predictive modeling: How to efficiently construct and validate clinical predictive models from EHR?

In this talk, we discuss our approaches to these challenges. For computational phenotyping, we present EHR data as inter-connected high-order relations i.e. tensors (e.g. tuples of patient-medication-diagnosis, patient-lab, and patient-symptoms), and then develop expert-guided sparse nonnegative tensor factorization for extracting multiple phenotype candidates from EHR data. Most of the phenotype candidates are considered clinically meaningful and with predictive power.

For predictive modeling, we introduce CloudAtlas, a cloud-based parallel predictive modeling system using big data infrastructure including Hadoop and Spark. Besides parallel model building, CloudAtlas can accurately estimate the running time and cost for a predictive modeling workflow then provisions the proper cluster on demand in the cloud. In particular, we demonstrate that CloudAtlas can achieve 40x speedup plus 40% cost saving compared to traditional sequential execution on large EHR datasets.

Invited Talk #2

Dr. Suzanne Matthews, Assistant Professor, US Military Academy

Big Data Phylogenetics

Reproducibility is a hot-button issue in the biological community. In this talk, we discuss the (lack of) preservation of the tree collections produced by phylogenetic search. Modern phylogenetic analyses produce tens to hundreds of thousands of equally weighted output trees. Scientists commonly summarize these tree collections into a single tree, discarding the output collections in the process. While recent phylogenetic reproducibility studies focus on the preservation of sequence alignments and final trees, output tree collections are largely ignored. We make the case for preserving tree collections by discussing their utility and applications, assessing the scale of data lost, reducing their storage requirements, and expediting their analysis.

Invited Talk #3

Dr. Yury Khudyakov, Molecular Epidemiology and Bioinformatics Laboratory, Laboratory Branch, Division of Viral Hepatitis, Centers for Disease Control and Prevention (CDC)

GHOST of Viral Molecular Surveillance

Viral hepatitis, a major health problem worldwide, is caused by infections with 5 viruses, all belonging to different viral families. The efficient molecular surveillance of viral hepatitis is needed to track infections and devise strategies for timely interventions to interrupt viral transmissions and reduce morbidity and mortality caused by viral infections in human population. Molecular surveillance must be massive to be efficient. Although much is being said about "Big Data" that can be generated by next-generation sequencing (NGS) technology, with various computational techniques being developed to handle the

data deluge, molecular surveillance cannot become sufficiently efficient via a mere NGS application. It is true that NGS may sample 105-106 intra-host viral variants from each infected individual and, thus, sequencing from only 1,000-10,000 individuals may easily overwhelm research capacity of many laboratories, thwarting global molecular surveillance efforts. However, considering that the estimated 170 million people are currently infected with hepatitis C virus alone, such analysis is not even close in scale to the needs of surveillance. It must be organized into a "structured crowd-sourcing" system for massive data gathering and analysis, which can be achieved using cyber-molecular assays and a special laboratory pipeline for cost-effective molecular testing by NGS. I'll describe one of the possible solutions to the problem of massive surveillance using Global Hepatitis Outbreak and Surveillance Technology (GHOST), which we are developing.