SCHOOL OF INFORMATICS, COMPUTING, AND ENGINEERING

# CENTER FOR BIOINFORMATICS RESEARCH TALK

## Somali Chaterji

**Purdue University**

Tuesday, October 3, 2017

2:00 pm

IMU, Oak Room

## Predictive Algorithms & Cyberinfrastructures for Data-Driven Genomic Medicine

Abstract: Personalized medicine or genomic medicine emphasizes the way in which disease risk or its manifestation varies in every single individual. This is because the health of an individual is a function of both her genome, with its sequence variants, and of her epigenome. Every cell of the human body has the same genome. How then is a heart cell different from a liver cell or a normal cell different from a cancer cell? This is where the cell's epigenome offers a distinct "symphony" to each cell type and in varied spatio-temporal contexts. Driven by the exabytes of sequencing data being generated, there is an increasing need to analyze genomic big data to interpret and stratify disease and to personalize therapies. How can this genomical big data enable the strides of precision medicine? What kinds of algorithms can deal with the inherent heterogeneity, the noise, and the high-dimensionality of this kind of data? Are there recurrent kernels or motifs in these algorithms that can be identified and then be used to speed up the modeling of this data? Can these efforts result in high-precision genomic medicine and biomarkers, by revealing the underlying patterns of epigenomic marks?

In this talk, I will answer some of the above questions through two exemplar ML algorithmic suites—*Avishkar* and *Tiresias*. First, in our *Avishkar* suite, we uncover the non-canonical signatures of small regulatory RNA or (therapeutic) microRNA (miRNA) targets. Regulatory miRNA-mRNA interactions are known to control a vast swath of cellular processes. Using our support vector machine-based algorithms, we are able to predict both canonical and non-canonical miRNA targets in a unified manner. Then, I will present our recent approach called Tiresias that can uncover combinatorial regulatory effects whereby multiple miRNAs together regulate gene expression. Finally, I will present lessons on developing large-scale cyberinfrastructures for supporting genomics, and more specifically, metagenomics workloads, which are both dynamic and hard-to-predict. This is in the context of MG-RAST, the leading metagenomics portal and analysis pipeline. I will present *Rafiki*, our algorithm for identifying the optimal parameters for the large datastore needed for a global cyberinfrastructure of this scale.

Biography: Somali Chaterji is a Visiting Faculty at Purdue University, where she specializes in developing algorithms and statistical models in the area of computational genomics. She got her PhD in Biomedical Engineering from Purdue University, winning the Chorafas International Award (2010), College of Engineering Best Dissertation Award (2010), and the Future Faculty Fellowship Award (2009). She did her Post-doctoral Fellowship at the University of Texas at Austin in the Department of Biomedical Engineering, where her work was supported by an American Heart Association award. She won the best paper award at the ACM BCB conference in 2015. Dr. Chaterji is also a technology commercialization enthusiast and has been consulting for the IC2 Institute at the University of Texas at Austin since Spring 2014.

INDIANA UNIVERSITY

SCHOOL OF INFORMATICS, COMPUTING, AND ENGINEERING