Saurabh Sinha, Ph.D., Georgia Tech

Novel Statistical and Machine Learning Methods for Analysis of Single Cell Data

Abstract:

With rapid advances in single cell technologies, new challenges and opportunities have emerged for discovering molecular mechanisms underlying biological processes. This talk will present two such challenges and the development of statistical and machine learning techniques to address them.

First, I will present recent work on analysis of spatial transcriptomics data at the sub-cellular resolution. Our newly developed toolkit called InSTAnT detects gene pairs with unusual patterns of co-localization within and across cells, using novel statistical tests. It then uses a probabilistic graphical model to discover special cases where an intra-cellular co-localization signal exhibits a non-random spatial pattern at the inter-cellular level. It also uses frequent subgraph mining algorithms to recover modules of co-localizing genes. Intra-cellular spatial patterns discovered by InSTAnT have diverse biological interpretations and provide a rich compendium of testable hypotheses regarding molecular functions.

Second, I will talk about the long-standing problem of reconstructing gene regulatory networks (GRNs), a powerful conceptual framework to describe mechanisms of gene expression changes accompanying biological processes. We are particularly interested in inference of GRN changes under two biological conditions. We present a causal inference approach to this task. We define the desired differential relationship as an estimand of "do calculus" and develop a computational tool called CIMLA to estimate this quantity. CIMLA uses non-linear models such as random forests and neural networks to model gene expression as a function of candidate regulators, employs "SHAPley Additive exPlanations" (SHAP) scores to assess each regulator's influence on a gene in each sample, and aggregates inter-group differences of this influence across all samples. We employed CIMLA to analyze a previously published single-cell dataset from subjects with and without Alzheimer's disease (AD).