LEARNING PREDICTIVE MODELS OF GENE REGULATION

Christina Leslie

Memorial Sloan-Kettering Cancer Center
New York City, NY

Studying the behavior of gene regulatory networks by learning from high-throughput genomic data has become one of the central problems in computational systems biology. Most work in this area focuses on learning structure from data -- e.g. finding clusters or modules of potentially co-regulated genes, or building a graph of putative regulatory "edges" between genes -- and generating qualitative hypotheses about regulatory networks.

Instead of adopting the structure learning viewpoint, our focus is to build predictive models of gene regulation that allow us both to make accurate quantitative predictions on new or held-out experiments (test data) and to capture mechanistic information about transcriptional regulation. Our algorithm, called MEDUSA, integrates promoter sequence, mRNA expression, and transcription factor occupancy data to learn gene regulatory programs that predict the differential expression of target genes. MEDUSA does not rely on clustering or correlation of expression profiles to infer regulatory relationships. Instead, the algorithm learns to predict up/down expression of target genes by identifying condition-specific regulators and discovering regulatory motifs that may mediate their regulation of targets. We use boosting, a technique from machine learning, to help avoid overfitting as the algorithm searches through the high dimensional space of potential regulators and sequence motifs. We will describe results of a recent gene expression study of hypoxia in yeast, in collaboration with the lab of Li Zhang. We used MEDUSA to propose the first global model of the oxygen and heme regulatory network, including new putative context-specific regulators. We then performed biochemical experiments to confirm that regulators identified by MEDUSA indeed play a causal role in oxygen regulation.