Thesis Defense

MACHINE LEARNING DEPARTMENT

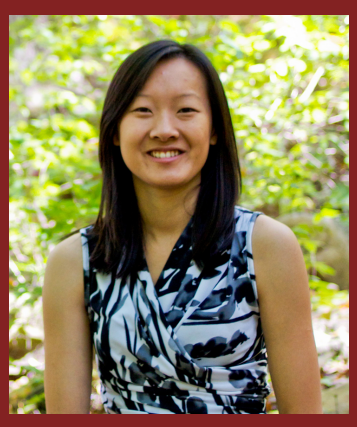
Statistical Approach for Functionally Validating Transcription Factor Bindings Using Population SNP and Gene Expression Data

Abstract:

Understanding transcriptional gene regulation is an important step to understanding how essential mechanisms are controlled in biological systems. Functional assays such as ChIP-seq and DNase I have been used to obtain a binding map of transcription factor (TF) binding sites on DNA and to determine the transcriptional regulatory network of TFs and their target genes. However, binding alone may not result in a change in target gene expression. The standard approach to identifying functional binding events is to perform artificial TF knockdown experiments and declare the differentially expressed genes as functionally validated target genes. Instead of artificial perturbation, in order to functionally validate the TF binding map, we propose to leverage the naturally-occurring genetic variations as the source of perturbations that vary gene expressions and to analyze population single nucleotide polymorphism (SNP) and gene expression data. Compared to the standard approach that perturbs TF concentration for a single TF at a time, our approach is potentially more powerful, because any aspects of the TF-target interaction, including TF concentration and TF binding affinity, can be perturbed by a large number of SNPs found across the genome.

In this thesis, we first introduce a statistical approach, based on conditional Gaussian Bayesian networks, that integrates population SNP and gene expression data with TF binding data to validate the TF binding map. We developed an efficient learning algorithm for learning the gene regulatory network by using TF binding data as prior knowledge, and selecting the TF-target interactions that are validated based on population SNP and gene-expression data. Given the estimated network, we perform inference on the estimated probabilistic graphical models to determine downstream genes that are differentially expressed due to the effect of the TF-target interactions.

We apply our method to learn transcriptional regulatory networks in lymphoblastoid cell lines (LCLs) and breast cancer tumours. First, we demonstrate our approach for validation of the TF binding map derived from ENCODE DNase I and ChIP-seq data from 71 TFs in LCLs, with SNP and gene expression data from the 1000 genomes and HapMap 3 projects respectively. We examined functional target genes that were validated under perturbation of TF concentration and TF binding affinity. Finally, we apply our method to perform TF binding map validation for ER and its coregulators which include 38 TFs obtained from Cistrome TF binding data, by using The Cancer Genome Atlas SNP and expression data from breast cancer tumors. We identified many previously known interactions between ER and its coregulators. We also found expression quantitative trait loci (eQTLs) in local binding regions of target genes that are potential super enhancers and eQTLs in coding regions that may affect the protein structure of important regulators.



PhD Candidate:

Jing Xiang

Thesis Committee:

Seyoung Kim (Chair)
Geoff Gordon
Carl Kingsford
Steffi Oesterreich (University of Pittsburgh)

Sept. 11, 2017
10:00am
6501 GHC

Link to draft document:
https://www.cs.cmu.edu/~jingx/docs/jing_xiang_thesis.pdf