High-resolution sequence and chromatin signatures predict transcription factor binding in the human genome

Christina Leslie, Memorial Sloan-Kettering Cancer Center

Abstract:

Gene regulatory programs are orchestrated by proteins called transcription factors (TFs), which coordinate expression of target genes both through direct binding to genomic DNA and through interaction with cofactors. Accurately modeling the DNA sequence preferences of TFs and predicting their genomic binding sites are key problems in regulatory genomics. These efforts have long been frustrated by the limited availability and accuracy of TF binding site motifs. Today, protein binding microarray (PBM) experiments and chromatin immunoprecipitation followed by sequencing (ChIP-seq) experiments are generating unprecedented high-resolution data on in vitro and in vivo TF binding. Moreover, genome-wide data on the cell-type specific chromatin state, including ChIP-seq experiments that profile histone modifications associated with active or inactive transcriptional states, provide additional information for predicting the genomic binding locations of TFs.

We will present a flexible new discriminative framework for representing and learning TF binding preferences using these massive data sets. We will first describe in vitro models of TF-DNA sequence affinities, where we train support vector regression (SVR) models with a novel string kernel on PBM data to learn the mapping from probe sequences to binding intensities. In a large data set of over 180 yeast and mouse TF binding experiments, our SVR models better predicted in vitro binding than popular motif discovery approaches or methods based on enrichment of k-mer patterns.

We will then show how to train kernel-based SVM models directly on TF ChIP-seq data to learn in vivo TF sequence models and investigate the cell-type specificity of TF binding profiles. In a large-scale evaluation on 184 TF ChIP-seq experiments from the ENCODE project –comprising data from 64 TFs across two human hematopoietic cell lines – we confirmed that our discriminative sequence models significantly outperform existing motif discovery algorithms. Moreover, using histone ChIP-seq data and DNase-seq data from the same cell types, we trained discriminative chromatin models to capture the spatial organization of multiple histone modification modifications or DNase sensitivity. For most TFs, combining the sequence and DNase signatures trained in one cell type allowed accurate binding predictions in the second cell type, indicating that changes in chromatin accessibility alone can account for cell-specific binding in these cases. However, we also identified a number of TFs with strongly cell-type specific binding profiles that TFs display distinct sequence signatures between cell types, showing that the DNA sequences recognized by TFs can indeed depend on cellular context , e.g. due to differences in usage of cofactors or composition of the TF complex.

This work establishes effective new techniques for analyzing next generation sequencing data sets to study the role of chromatin and sequence in TF binding in the human genome.