An integrative model of gene expression and epigenetic variation across human cell types

Sofia Kyriazopoulou-Panagiotopoulou¹, Marco Cusumano-Towner¹, Serafim Batzoglou¹, Anshul Kundaje²

1. Department of Computer Science, Stanford University 2. Department of Genetics, Stanford University School of Medicine

Background

All cells in our body contain the same genetic material, but can perform very different functions. Why? Different genes are expressed in different cell types. What causes cell-type-specific gene expression? Cell-type-specific activity of regulatory proteins + cell-type-specific accessibility of regions where these proteins bind.

Goal: Build machine learning models for predicting cell-type-specific gene expression based on the expression of regulatory proteins and the sequence composition of the regulatory regions of genes.

Significance: The learned models will help us understand the mechanisms underlying disease and identify potential drug targets.

A challenging machine learning and data integration problem: Experimental data are scarce, noisy, diverse, hard to integrate. They contain many correlated or interacting features. The learned models should be interpretable, providing us with biological insights.

Our contributions: One of the most integrative models to date for modeling gene regulatory variation across cell-types. The learned models are both accurate and interpretable.

Modeling using boosting and Alternating Decision Trees

Input:
- Regulatory proteins bind to DNA sequence motifs and can activate their target genes.
- Up/down-regulation of genes and regulators across cell types.
- Sequence motifs + cell-type-specific accessibility of these motifs.

Output:
- An Alternating Decision Tree: Each regulatory rule is associated with a score.
- Paths define cellular contexts. Tree structure captures regulatory interactions.

Model evaluation

13K genes
56 human cell types
100s of thousands of interactions

Enrichment of regulator pairs in neighboring tree nodes

Biological significance

Our models discover cell-type-specific regulatory proteins and sequence motifs as well as regulatory pathways:

- The most integrative model to date for modeling gene regulatory variation across cell-types.
- Interpretable models that provide biological insights.

Our contributions:

- BioGRID, Cytoscape, GeneMANIA.