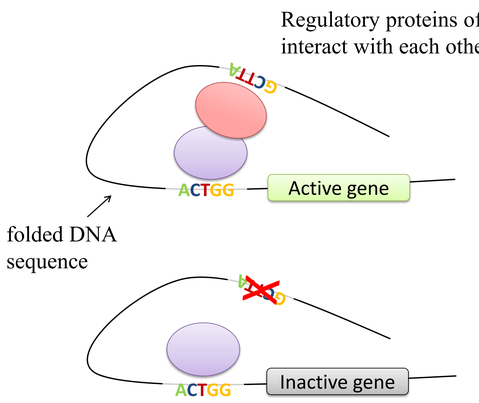
An integrative model of gene expression and epigenetic variation across human cell types

Sofia Kyriazopoulou-Panagiotopoulou[1], Marco Cusumano-Towner[1], Serafim Batzoglou[1], Anshul Kundaje[2]

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.



Regulatory proteins bind to **DNA sequence motifs and can activate their target genes.**

In some cell types, some regions of the genome might be **inaccessible** to binding.

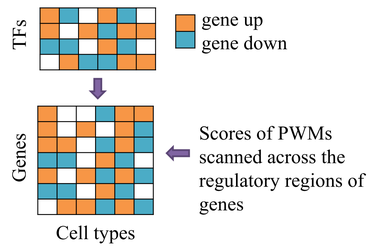
Goal: Build machine learning models for predicting cell-type-specific gene expression based on the expression of regulatory proteins and the sequence composition of the regulatory regions of genes.

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

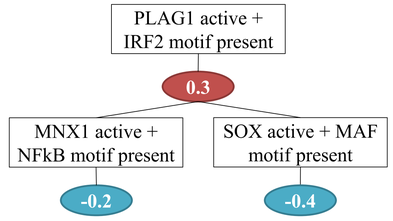
A challenging machine learning and data integration problem: Experimental data are **scarce, noisy, diverse, hard to integrate**. They contain many **correlated or interacting features**. The learned models should be **interpretable**, providing us with biological insights.

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



LogitBoost with weights on candidate rules



Inputs

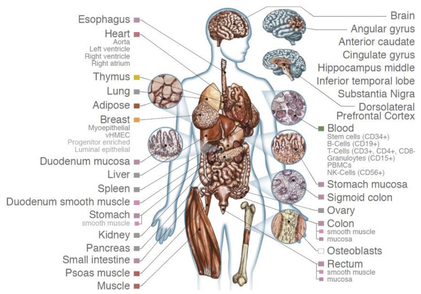
- Up/down-regulation of genes and regulators across cell types.
- Sequence motifs + cell-type-specific accessibility of these motifs.

- **Protein-protein interactions**, weighted by their confidence.
- Used to **weight** candidate rules during learning.

Output

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

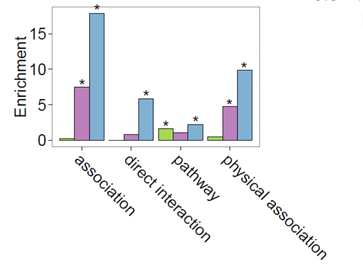
Model evaluation

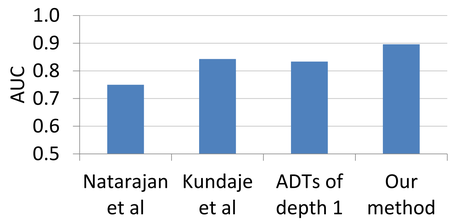


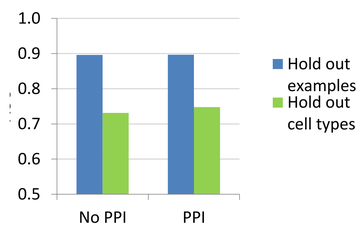
ROADMAP epigenomics PROJECT

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

Enrichment of regulator pairs in neighboring tree nodes



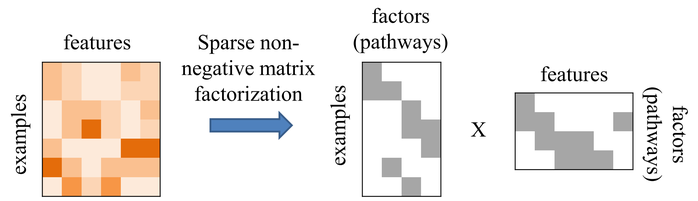




Module networks
Our method - no protein interaction information
Our method - using protein interaction information
⋆ p-value < 0.05

Biological significance

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



Our contributions

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

- Freund Y. and Mason L. (1999) The alternating decision tree learning algorithm. *ICML*.
- Kundaje A., Xin X., *et al*. (2008) A predictive model of the oxygen and heme regulatory network in yeast. *PLoS Computational Biology*, 4(11):e1000224.
- Mairal J., Bach F., *et al*. (2010) Online learning for matrix factorization and sparse coding. *ICML*.
- Jalali A., Ravikumar, P. D., et al. (2010) A dirty model for multi-task learning. *NIPS*.
- BioGRID, Cytoscape, GeneMANIA.