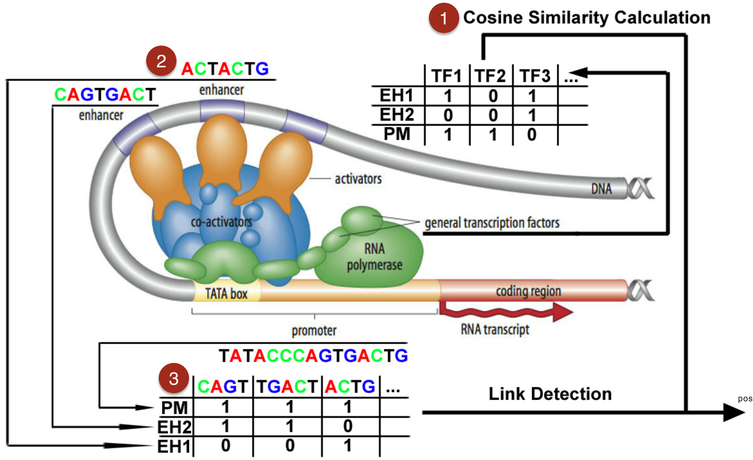
Predicting Transcription Factor Complexes Using Similarities in Genomic Motifs

Lichy Han, Maulik Kamdar, Daniel Kim
Biomedical Informatics Training Program, Stanford University, Stanford, CA

INTRODUCTION

- Gene regulation is an interplay of accessible DNA sequences with regulatory proteins, called transcription factors (TFs)
- TFs bind to regulatory sequences called enhancers and promoters which control gene transcription and are important for cell function
- Gene regulatory networks have been constructed in yeast to extensively study TF interactions, but contain missing links
- Community detection allows for identification of groups of promoters and enhancers with similar genomic motifs that define regulatory binding

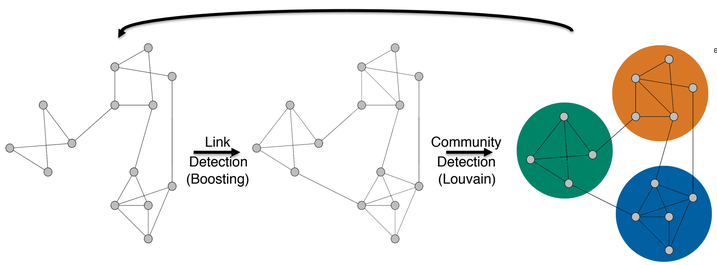
METHODS



1. We extracted data from chromatin immunoprecipitation experiments in ENCODE from a lymphoblastoid cell line and identified TFs that bind to each of the 124,530 promoter and enhancer regions. We calculated the cosine distance between each pair of enhancers and promoters and used high scoring pairs as our initial edges.

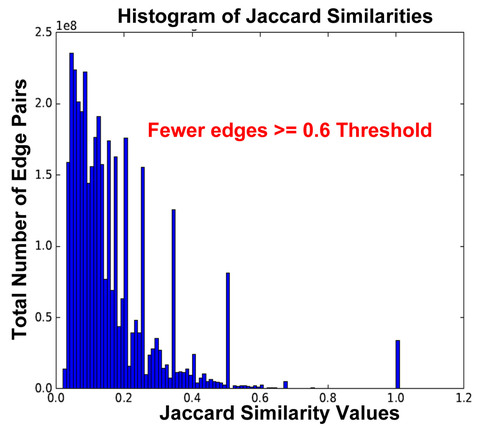
2. For each enhancer and promoter element, we broke down the sequence into *k*-mers as features for our nodes, for k=[4,5].

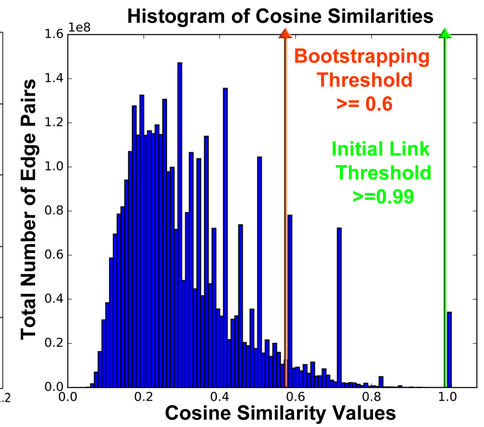
3. The presence of any particular *k*-mer is marked with a 1 in the corresponding vector index. Each edge is featurized by considering the shared *k*-mers between the two nodes.

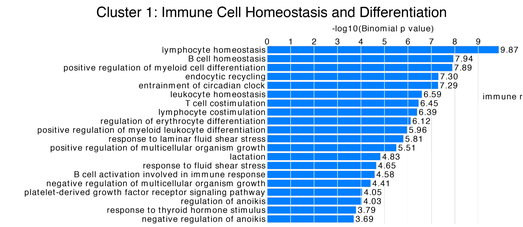


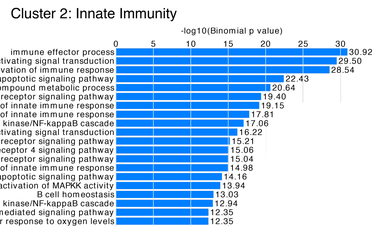
Iterative link detection and community detection using confidence rated AdaBoost and the Louvain method.

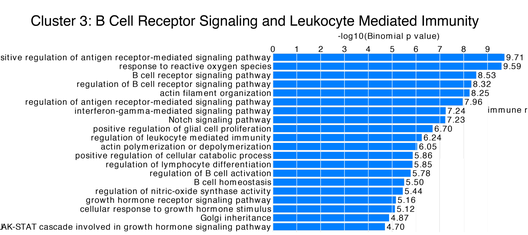
RESULTS

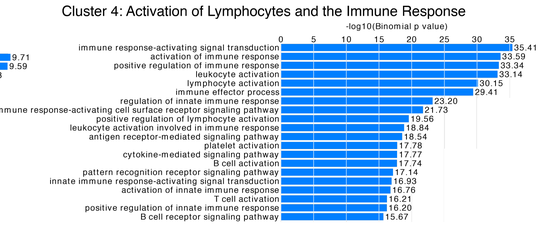


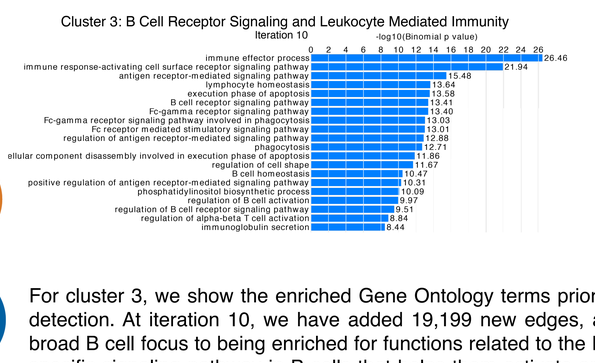










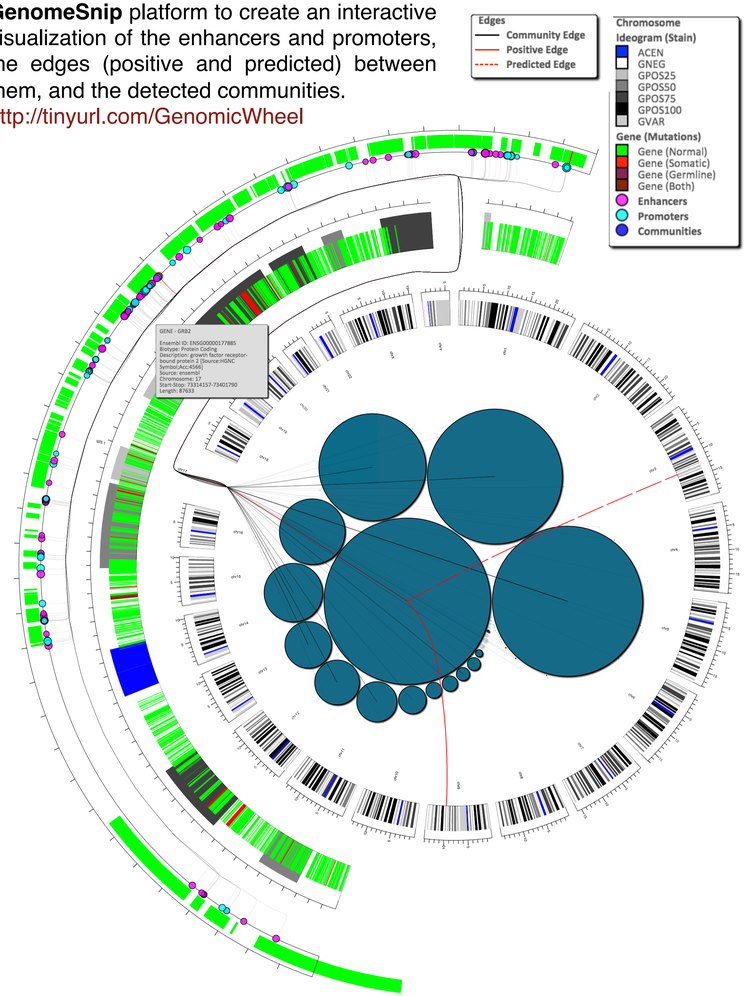


Four prominent clusters emerged from our analysis prior to iterative link detection, consisting of 3248, 4558, 1990, and 4903 regions, respectively. We identified genes close to our promoters and enhancers using Genomic Regions Enrichment of Annotations Tool (**GREAT**). We then looked for enriched functional Gene Ontology terms associated with the genes located close to the enhancers and promoters in each cluster. Each of the 4 largest clusters reflects a different aspect of the immune system.

For cluster 3, we show the enriched Gene Ontology terms prior to link detection, and after 10 iterations of link detection. At iteration 10, we have added 19,199 new edges, and we can see the cluster has shifted from a broad B cell focus to being enriched for functions related to the Fc-gamma receptor signaling pathway. This is a specific signaling pathway in B cells that helps them activate and respond to foreign antigens, and reflects the changing nature of the clusters to increasingly specific cell functions.

VISUALIZATION

GenomeSnip platform to create an interactive visualization of the enhancers and promoters, the edges (positive and predicted) between them, and the detected communities.
http://tinyurl.com/GenomicWheel



CONCLUSIONS

- Transcription factors with high motif similarity form clusters that correspond to distinct immunological functions
- Functions of communities can achieve greater specificity after edge refinement using our supervised machine learning model
- We created an interactive visualization for researchers to use to thoroughly explore the edges and clusters in relation to the genome

ACKNOWLEDGEMENTS

We would like to thank the CS224W teaching staff for their valuable time and input.