Clustering of Pathogenic Genes in Human Co-regulatory Network

Michael Colavita
Mentor: Soheil Feizi
Fifth Annual MIT PRIMES Conference
May 17, 2015

Topics

- **Background**
  - Genetic Background
  - Regulatory Networks
  - The Human Regulatory Network
  - Co-regulatory Networks
- Modularity
  - Purpose and Methods
  - Implementation
  - Results
- Clustering Algorithm
  - Goals
  - Algorithmic Basis
  - Initial Method and Progress

Genetic Background

- Genes have regulatory effects on each other
  - Upregulation
  - Downregulation
- **Transcription factors** have regulatory effects
- **Target genes** do not affect other genes
- Both can be subject to regulation by other genes

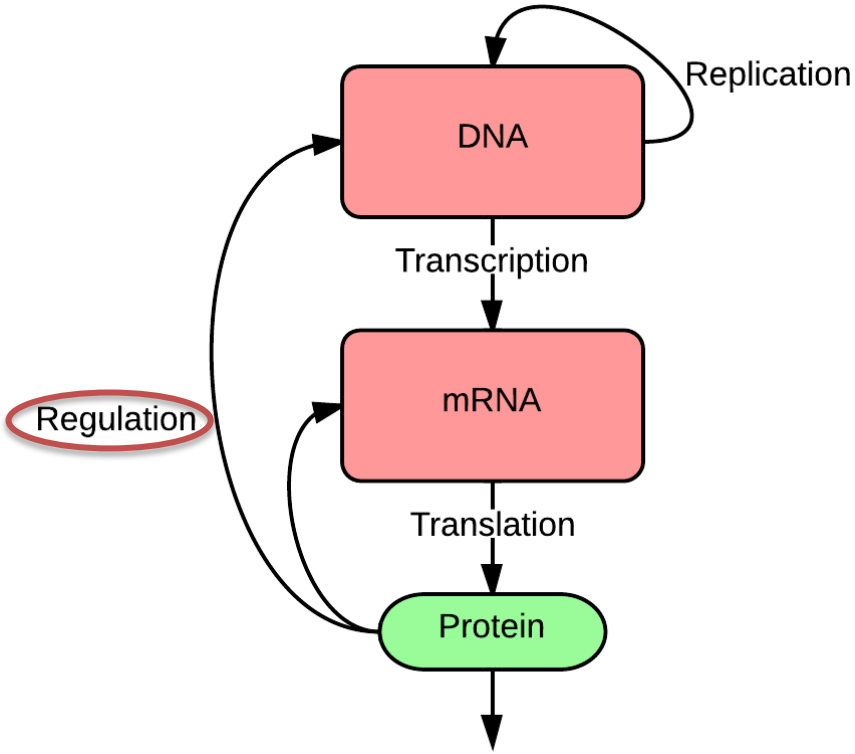


Figure: The central dogma of molecular biology including gene regulation

Genetic Regulatory Networks

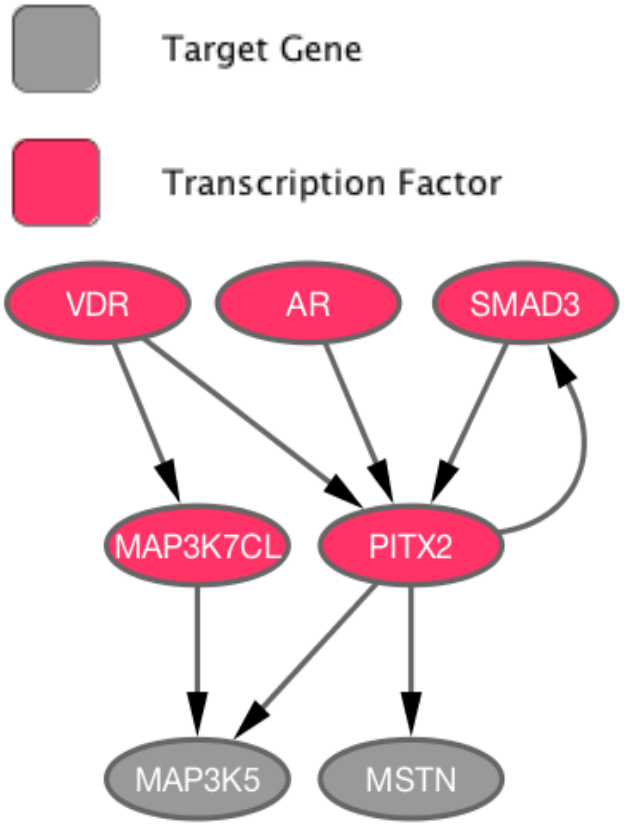


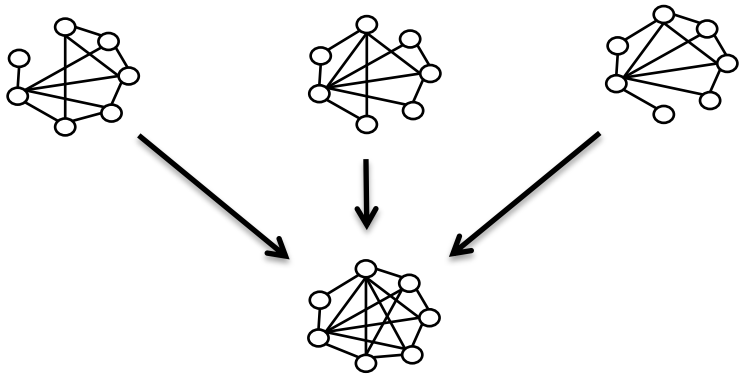
Figure: A small section of the human regulatory network

- Method of storing regulatory information in a computationally accessible format
  - Captures regulatory dynamics of a genome
  - Allows for the use of algorithms from the field of graph theory
- Nodes represent genes
- Directed edges indicate **upregulatory** effects
  - Edge weights indicate strength of regulatory activity

The Human Regulatory Network

- Primary dataset used for pulling regulation data

- Created by combining datasets into a unified network
  - Co-expression network
  - Motif network
  - ChIP network



- 2757 **transcription factors**
- 16464 **target genes**
- ~1,000,000 regulatory relationships (cutoff = .95)

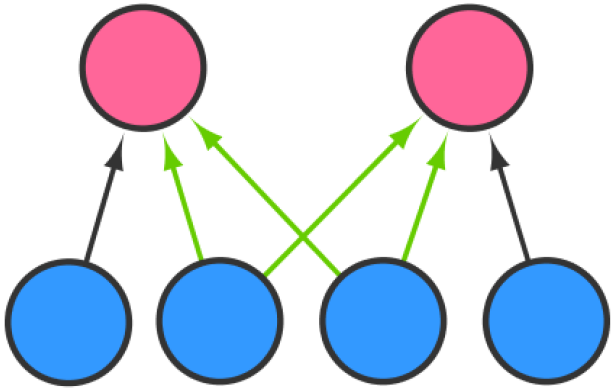
Co-regulatory Networks

- Capture different relationships than regulatory networks

- Nodes still represent genes; edges represent similar regulatory profiles

$$\frac{|R_a \text{ Ç } R_b|}{|R_a \text{ È } R_b|} \text{ 3 } C$$

- **Undirected** network
  - Clustering is better defined



Topics

- Background
  - Genetic Background
  - Regulatory Networks
  - The Human Regulatory Network
  - Co-regulatory Networks
- **Modularity**
  - Purpose and Methods
  - Implementation
  - Results
- Clustering Algorithm
  - Goals
  - Algorithmic Basis
  - Initial Methodology and Progress

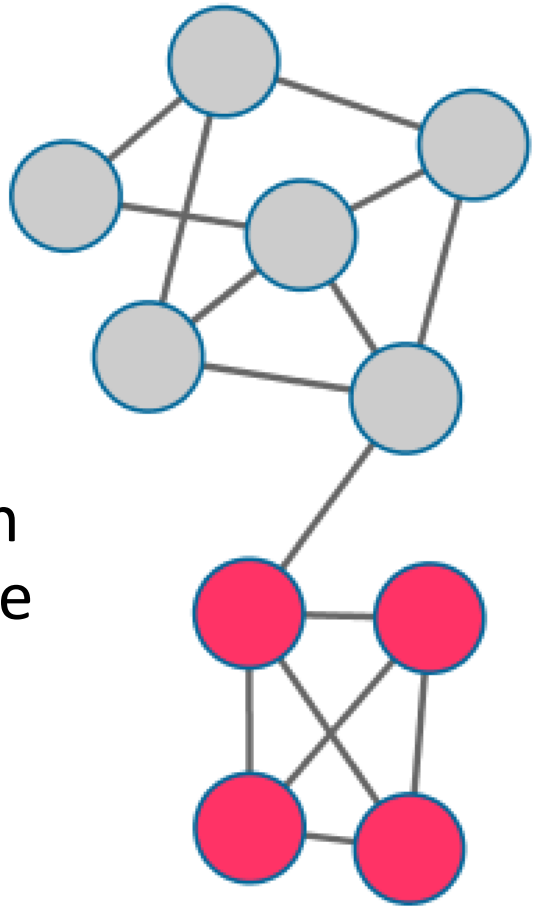
Motivations for Analysis

- **Pathogenic genes** are associated with a specific genetic disease (dbGaP)

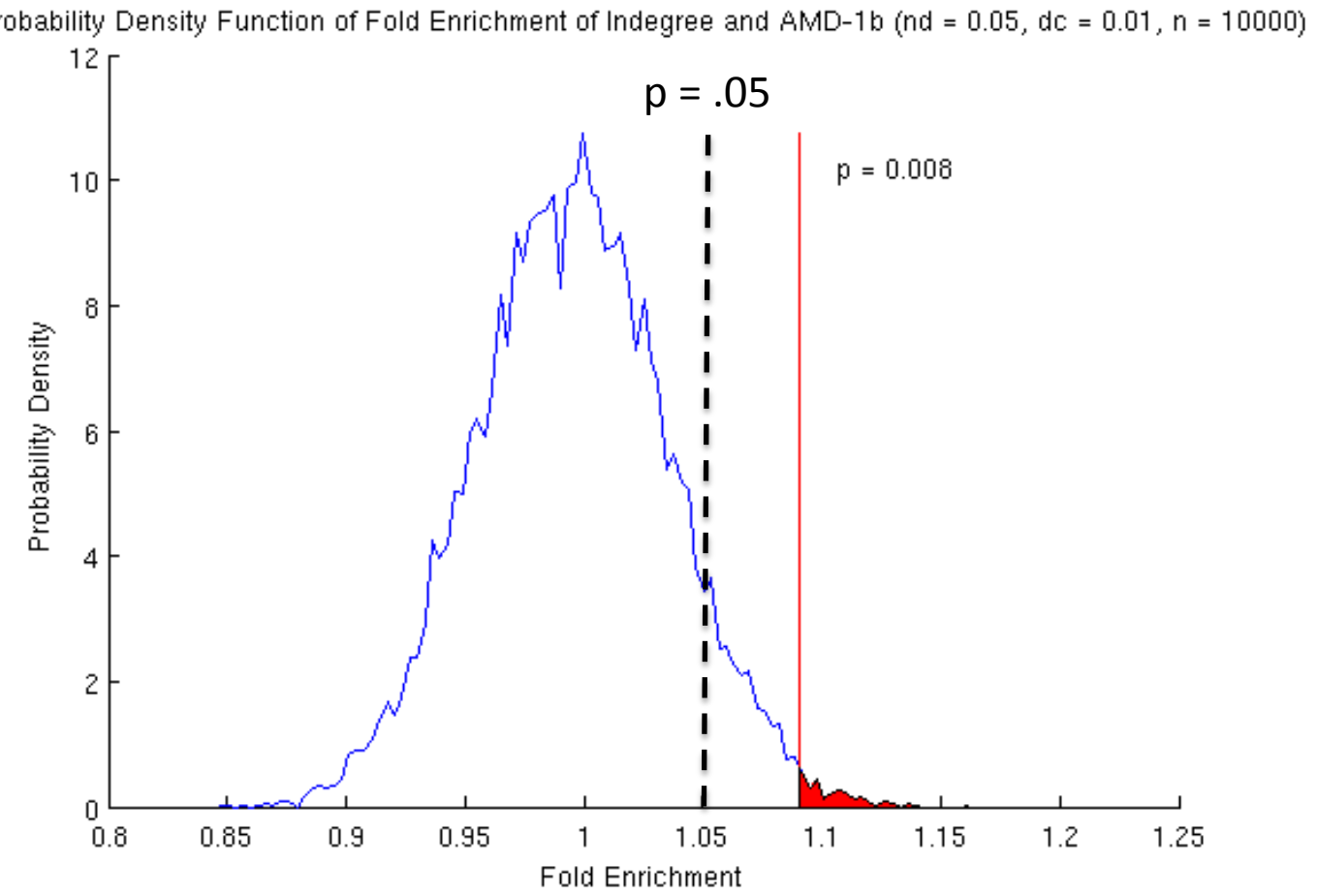
- Search for differences and patterns in how pathogenic genes are regulated
  - Understanding the basis of genetic diseases
  - Applications to gene therapy

Preliminary Analysis: Modularity

- Method of examining the types of connections in the network:
  - **Non-pathogenic - Non-pathogenic**
  - **Pathogenic – Pathogenic**
  - Non-pathogenic - Pathogenic
- How does the number of edges between nodes of the same classification compare to the expected value (null model)?
  - Assortative (preference for same classification)
  - Disassortative (preference for different)



Hypothesis Test and P-value Example



Modularity Testing

- Analyzed **45 diseases** across the network of 19,221 genes
  - MATLAB for parallel operations

- Possible outcomes:
  - Insignificant ($p > 0.05$)
  - **Assortative** (modular)
  - Disassortative

Modularity Results

- 12/45 (26.7%) diseases displayed statistically significant **assortativity** ($p < 0.05$)
  - Clopidogrel a, b, j, k, l ($p = 0.01$)
  - Cardiovascular disease risk
  - T1D
  - Type 1 Multiple Sclerosis
  - Psoriasis

- More connections between similarly classified genes than expected

Implications

- Suggests that the network contains **communities** of pathogenic and non-pathogenic genes
  - Potential for statistically significant clusters based on pathogenicity

- Significance of the co-regulatory structure
  - Suggests that pathogenic genes share common regulatory characteristics

Topics

- Background
  - Genetic Background
  - Regulatory Networks
  - The Human Regulatory Network
  - Co-regulatory Networks
- Modularity
  - Purpose and Methods
  - Implementation
  - Results
- **Clustering Algorithm**
  - Goals
  - Algorithmic Basis
  - Initial Methodology and Progress

Clustering

- Another point of interest for genetic diseases
- Typically based on connectivity
- Searching for cohesive regulatory units
  - Based on **modularity**
- Provides more information about how genes interact
  - Identifies patterns in regulatory profiles

Co-regulatory Clustering Goals

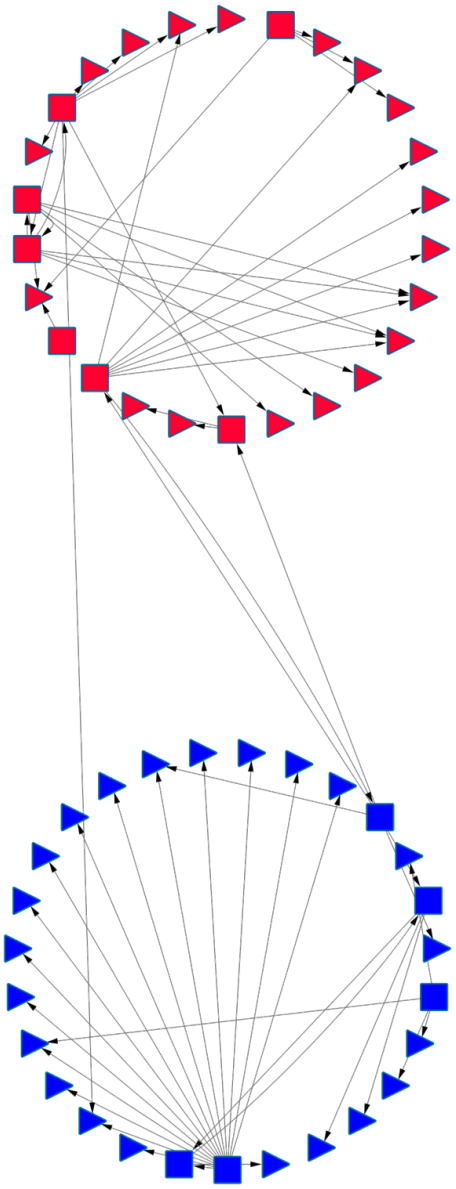
- Identify clusters by combining network structure with pathogenicity classification
  - Combine co-regulation with common genetic disease associations

- Clusters should indicate groups of pathogenic genes that share regulatory profiles
  - Indicates regulatory patterns that can lead to **genetic disease**

Algorithmic Basis: Spectral Partitioning

- **Goal**: divide a network into two groups such that the modularity is minimized

- **Method**: use the sign of values in the second eigenvector of the graph Laplacian to determine classification
  - Estimation stemming from a constraint relaxation



Algorithmic Basis: Spectral Clustering

- Similar to spectral partitioning, but produces *k* clusters

- **Basic Algorithm:**
  - Define a similarity matrix that quantifies the similarity between two vertices
  - Use the similarity matrix to produce a graph Laplacian
  - Use the values in the first *k* eigenvectors as input to the **k-means** algorithm

Current Hybrid Algorithm

- Construct a **similarity matrix** *S* capturing the structure of the network and the classifications of vertices
  - Assign a similarity value based on pure connectivity
  - Scale these values for each pairing using their classifications
    - Genes of the same type will have a higher similarity
    - Genes of different types will have a lower similarity
- **Spectral clustering** is applied to *S*

Future Goals

- Continue work on the clustering algorithm
  - Incorporate shortest path and other measures of distance into the similarity matrix
  - Refine similarity values for pairings
- Extend study to examine genetic disease information
  - Linear mixed model for genome-wide association studies

Thank You

- To **MIT PRIMES** for providing this research opportunity

- To my mentor **Soheil Feizi** for his support, suggestions, and assistance

- To **Professor Manolis Kellis** for suggesting the project