Using the Network Structure of Annotation Data to Gain Insights into Gene Interactions and the Organization of Biological Function

Michelle Girvan

in collaboration with:

Kimberly Glass,

Ed Ott,

Wolfgang Losert

Why statistical physicists are interested in network problems

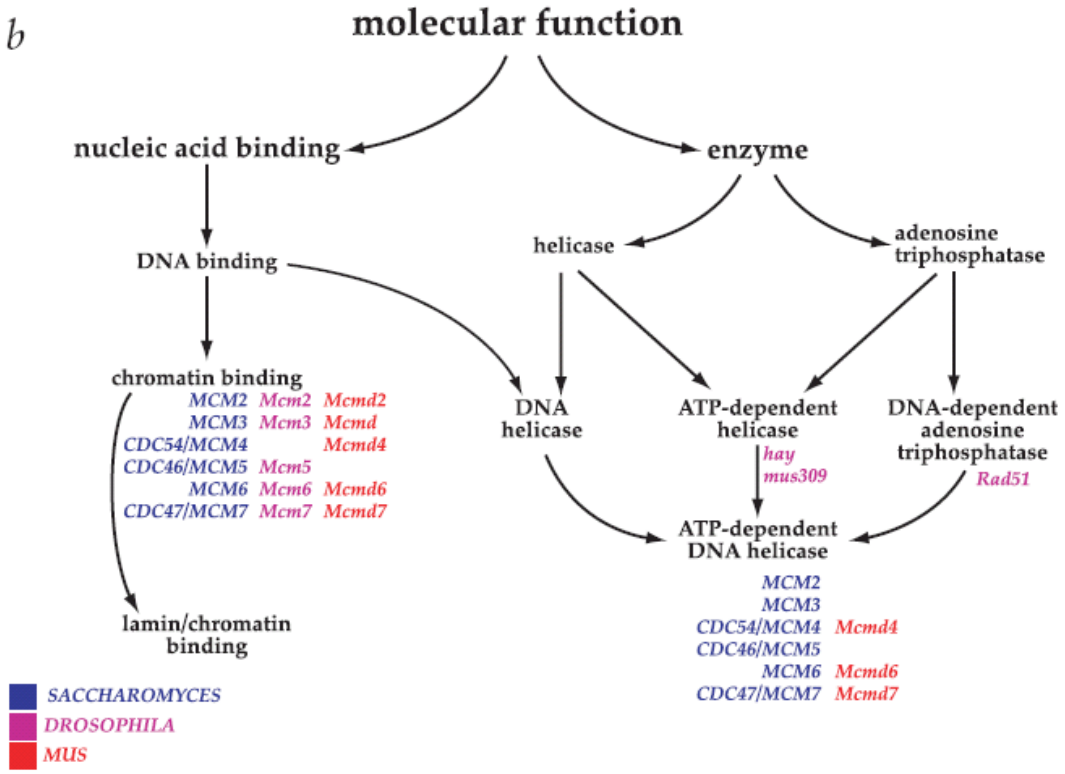
- Statistical physics is well-equipped to deal with networks that are highly regular (e.g. the lattice connections of atoms in a solid) or highly random (e.g. the interactions of gas molecules).
- Heterogeneous networks represent a new area in which to extend the tools of statistical physics.
- Statistical physicists have a long tradition of applying their approaches to many body problems in other fields: animal flocking, market behaviors, etc.

Why analyze the graph structure of gene annotations?

- Determine if there are undocumented, biologically meaningful relationships between terms.
- Understand large-scale functional relationships between genes.

Structure of the Gene Ontology

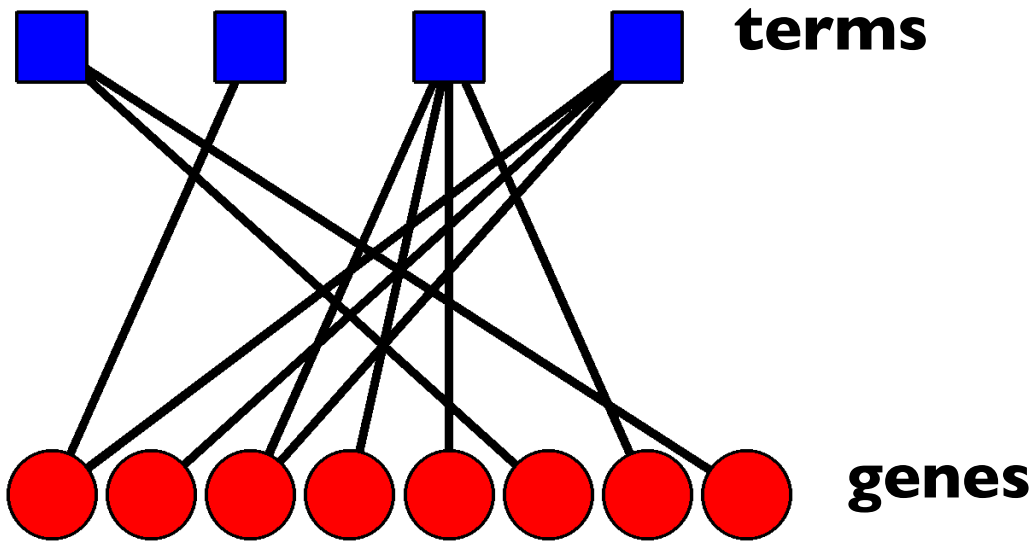
- The **Gene Ontology** is a hierarchical classification system for biological functions (terms).
- Hierarchy takes the form of a directed acyclic graph (DAG).



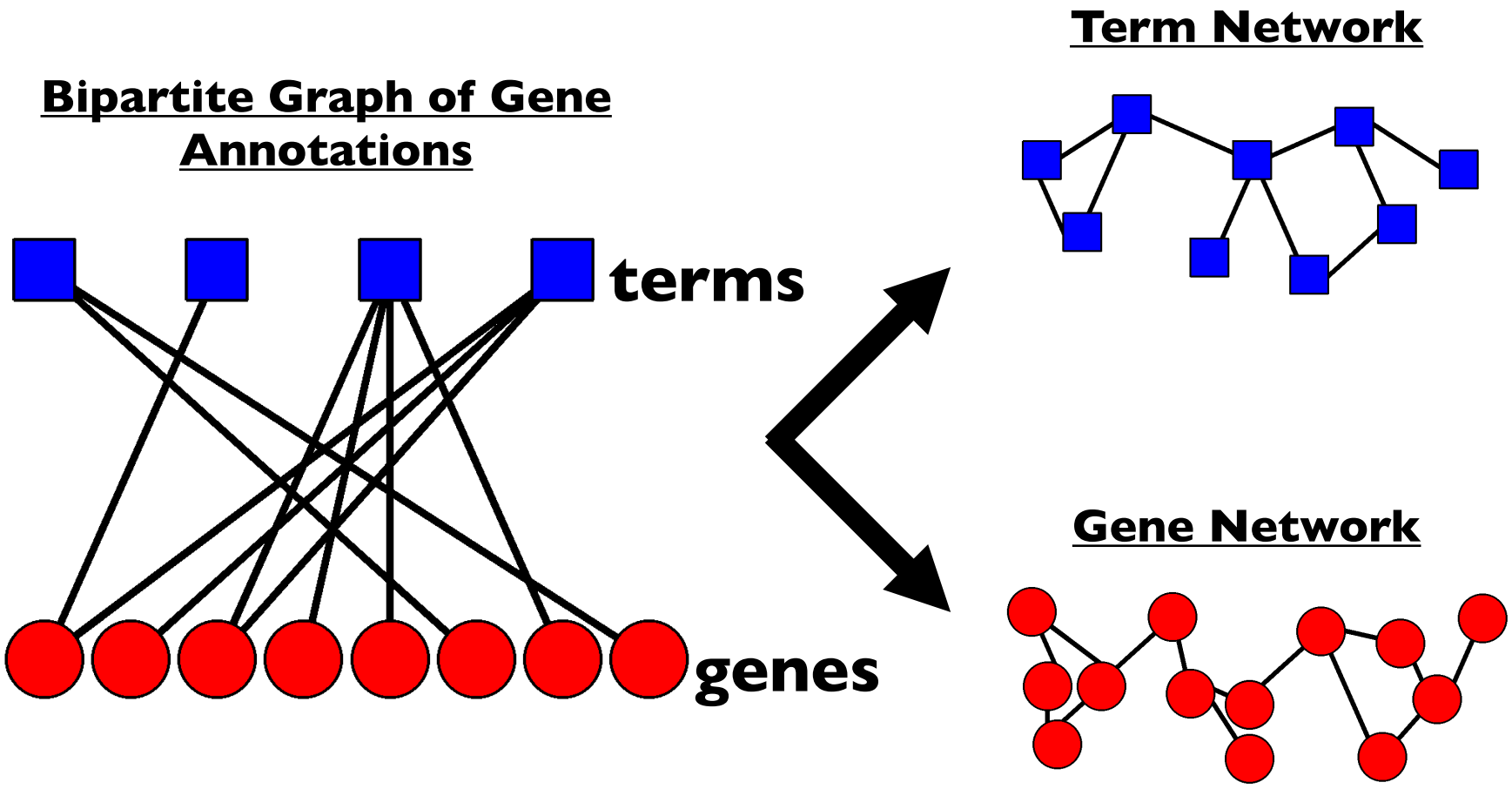
- Genes are assigned to terms. These assignments are transitive up the hierarchy.

Image from: *"Gene Ontology: Tool for the Unification of Biology"*

The graph structure of gene annotations



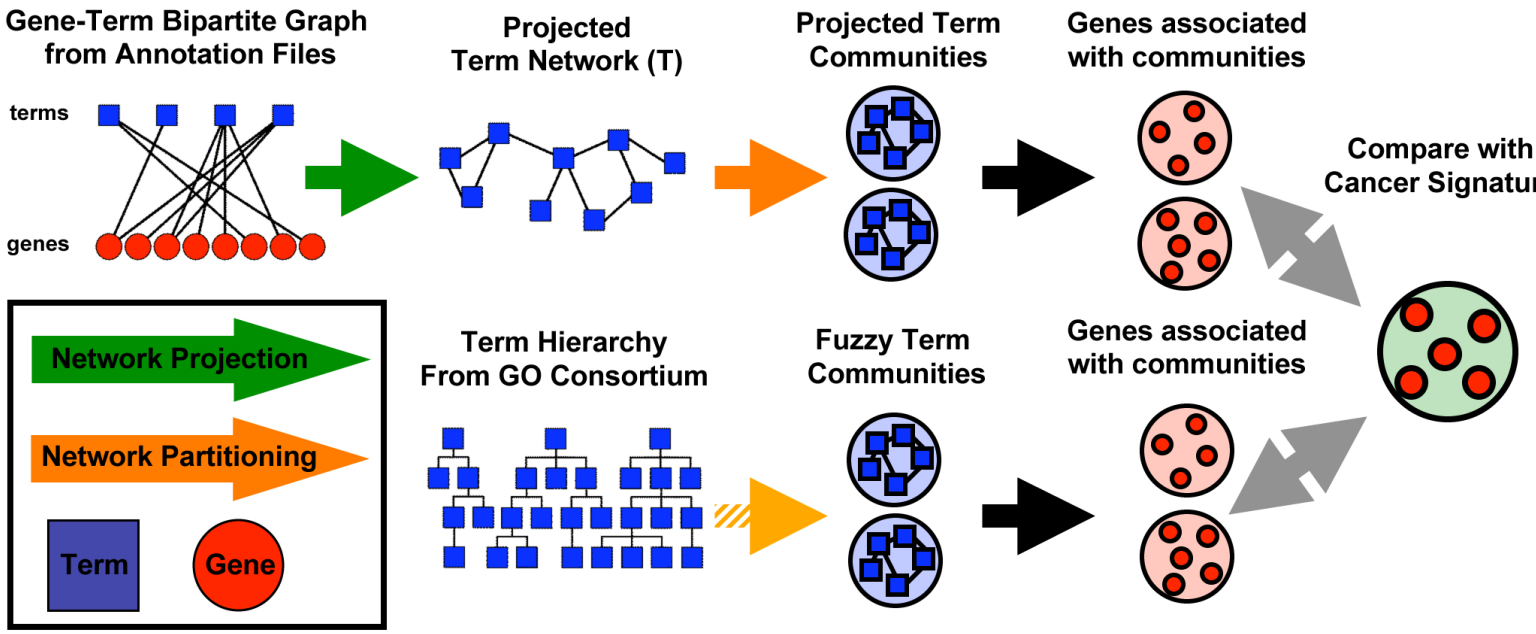
Creating Term and Gene Networks from the Bipartite Graph



Interpreting term and gene networks

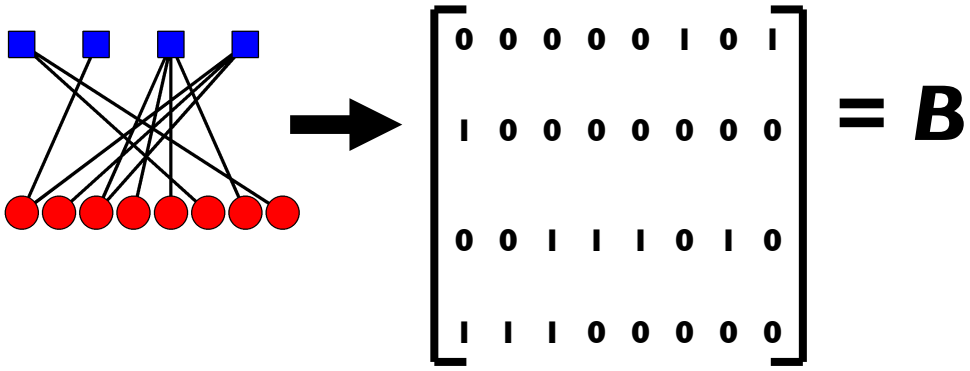
- Term networks can be used to group biological functions
- Gene networks can be used to understand/ predict interactions

Process for Analyzing the Structure of the Term Network



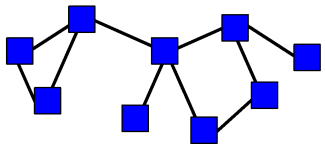
Term and Gene Networks

Gene Ontology
Bipartite Graph



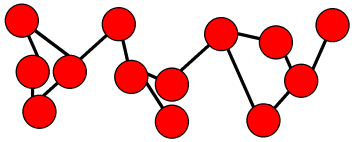
= *B*

Term Network



= *T* = *BB'*

Gene Network



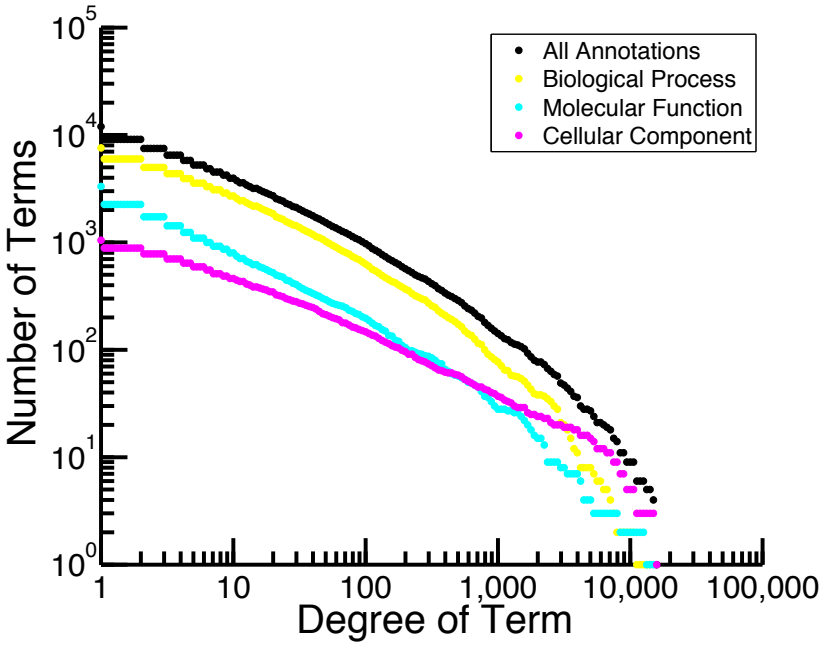
= *G* = *B'B*

$$T_{ij} = \sum_k B_{ik} B'_{kj} = \#\text{ of genes connecting term i to term j}$$

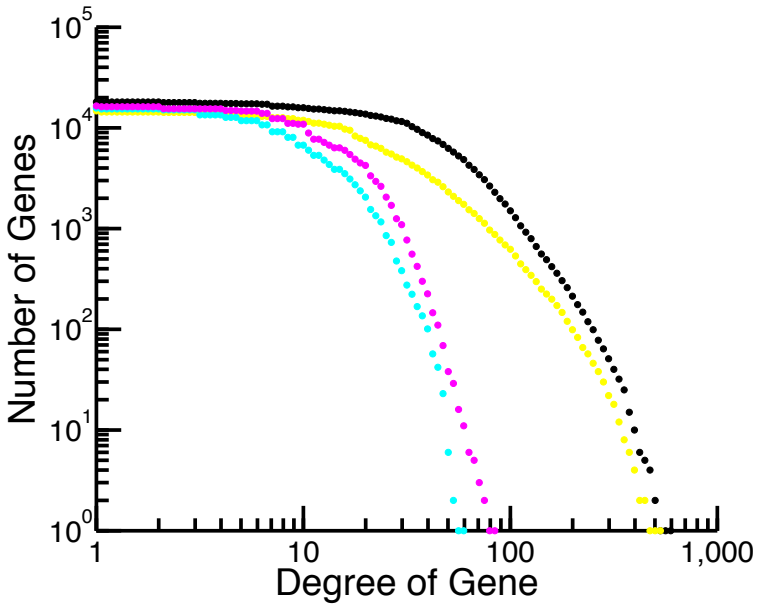
$$G_{ij} = \sum_p B'_{ip} B_{pj} = \#\text{ of terms connecting gene i to gene j}$$

Is it valid to weight term/gene connections by co-annotation?

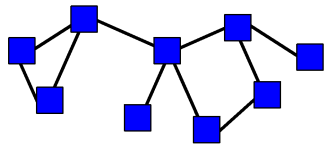
Degree distribution of GO Terms



Degree distribution of annotated genes

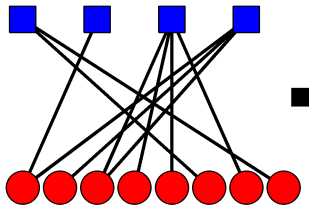


Weighting the Term Network



$$w_{ii} = \frac{1}{\sum_n B_{in}} = \frac{1}{\text{degree of term i}}$$

T = wBB'w'



$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} = B \qquad \begin{bmatrix} 1/2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1/4 & 0 \\ 0 & 0 & 0 & 1/3 \end{bmatrix} = w$$

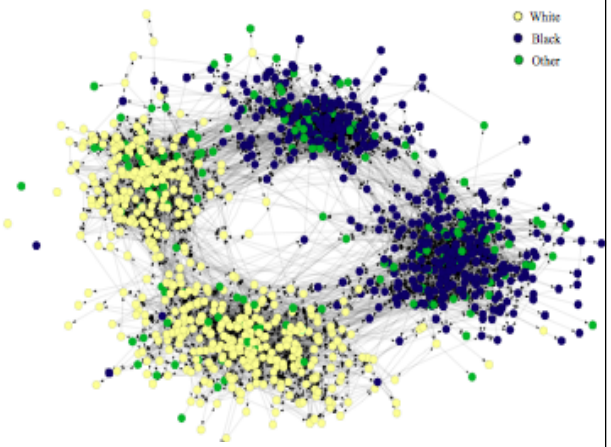
$$T_{ij} = \frac{\sum_k B_{ik} B'_{kj}}{\sum_n B_{in} \sum_m B'_{mj}} = \frac{\#\text{ of genes connecting term i and term j}}{\text{degree of term i} \times \text{degree of term j}}$$

Consequences of weighting T

$$T_{ij} = \frac{\#\text{ of genes connecting term i and term j}}{\text{degree of term i} \times \text{degree of term j}}$$

- T_{ij} takes on a maximal value of 1 when term i and term j share each only have the same single gene annotation.
- T_{ij} takes on a minimal value of 0 when term i and term j share no common annotations.
- T_{ij} gets small when term i and term j are both high degree and share few common annotations.

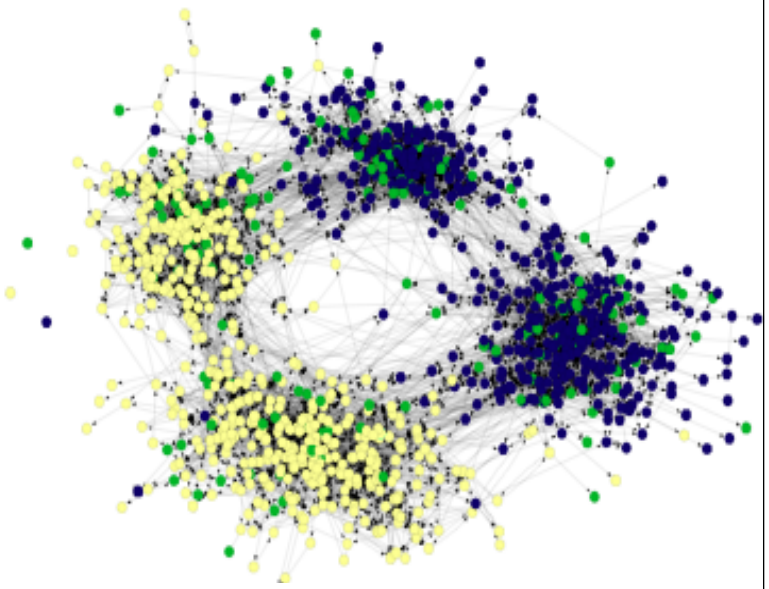
Community Structure in the Term Network



- Having constructed the term network, we want to identify groups of strongly connected terms.
- To do this, we can use any one of a variety of network community finding techniques.

The problem of identifying community structure in networks

- The goal: Given an arbitrary network, develop a method to divide the network into groups, or communities, such that within-group edges are relatively dense.
- Important caveat: We do not want to specify the number of groups a priori. Rather, we would like to find a "natural" division of the network into communities.



Adolescent friendship network, from Jim Moody

Quantifying the community structure

- The strength of a given partition of a network into *k* communities can be quantified by the modularity function:

$$Q = \sum_{i=1}^{k} \left[\frac{e_i}{m} - \left(\frac{d_i}{2m} \right)^2 \right]$$

- where e_i is the number of edges that connect vertices in community *i,* d_i is the number of edge ends that connect to vertices in community *i*, and *m* is the total number of edges.
- The modularity measures observed within-community density vs. expected within community density.

Newman and Girvan, PRE 2004

Modularity Maximization

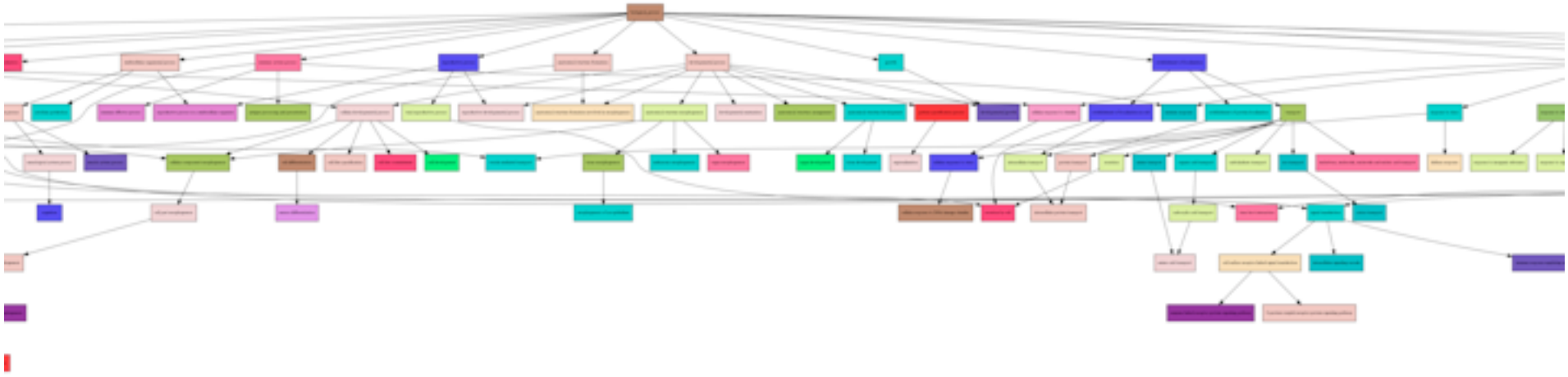
$$Q = \sum_{i=1}^{k} \left[\frac{e_i}{m} - \left(\frac{d_i}{2m} \right)^2 \right]$$

- The problem: find the partition that maximizes the modularity function.
- NP hard, but many heuristics work well in practice:
  - Greedy agglomeration
  - Spectral methods
  - Simulated annealing

Brandes et al. 2007, Clauset et al. 2004, Newman 2006, Massen and Doye 2006

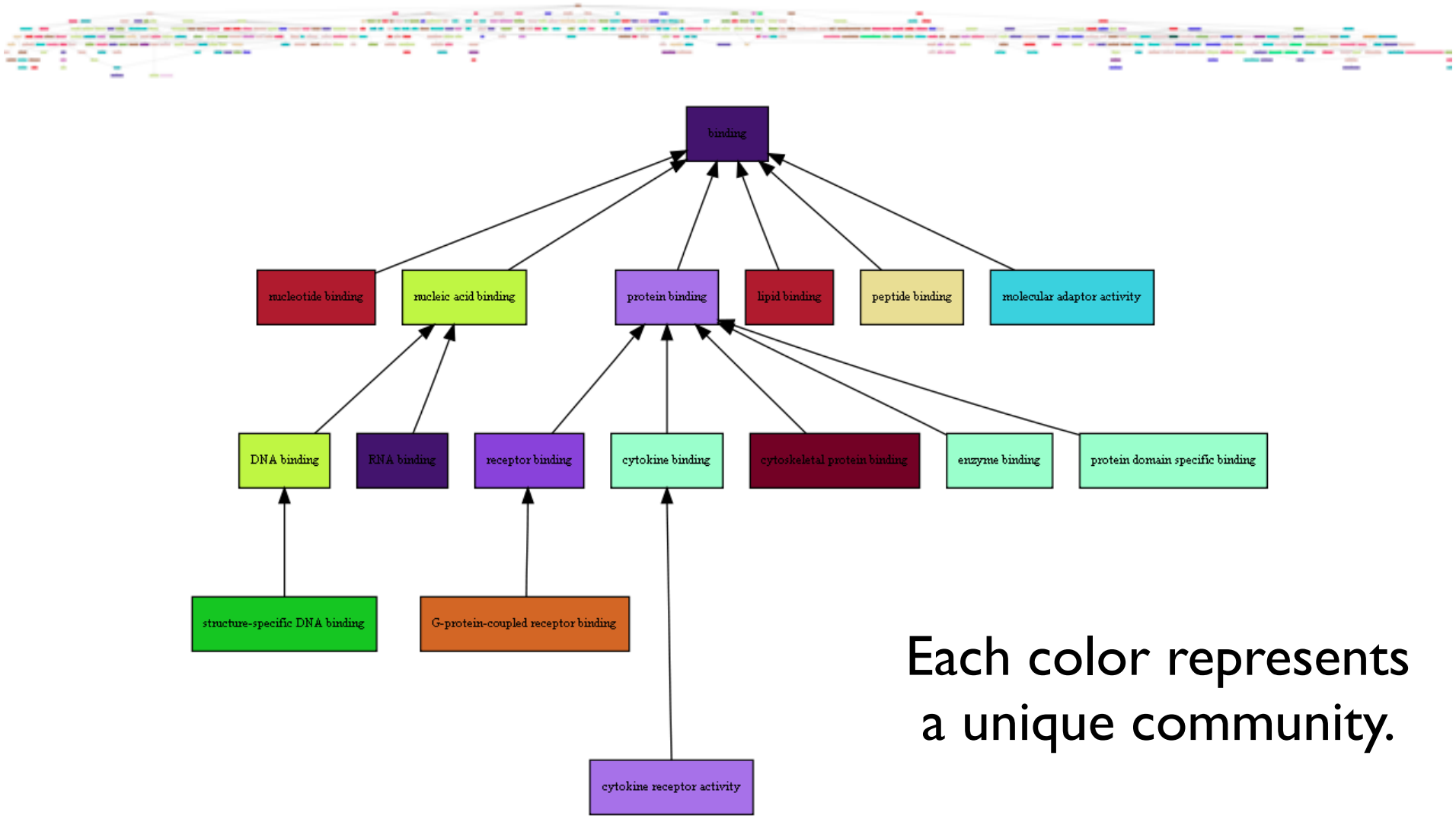
Community Structure in the Term Network

Communities of Terms are largely independent of the Hierarchical structure.



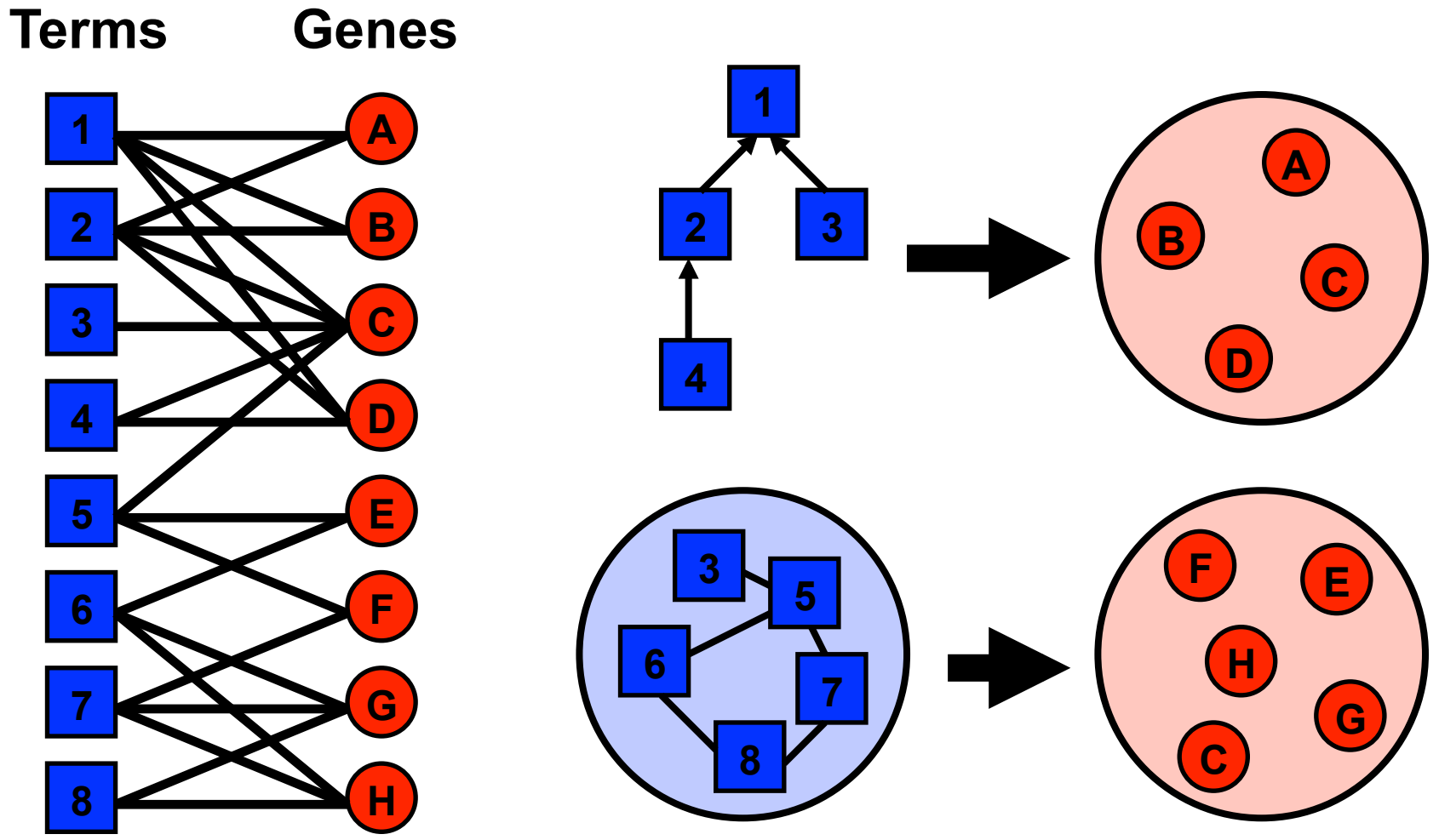
Each color represents a unique community.

Community Structure in the Term Network

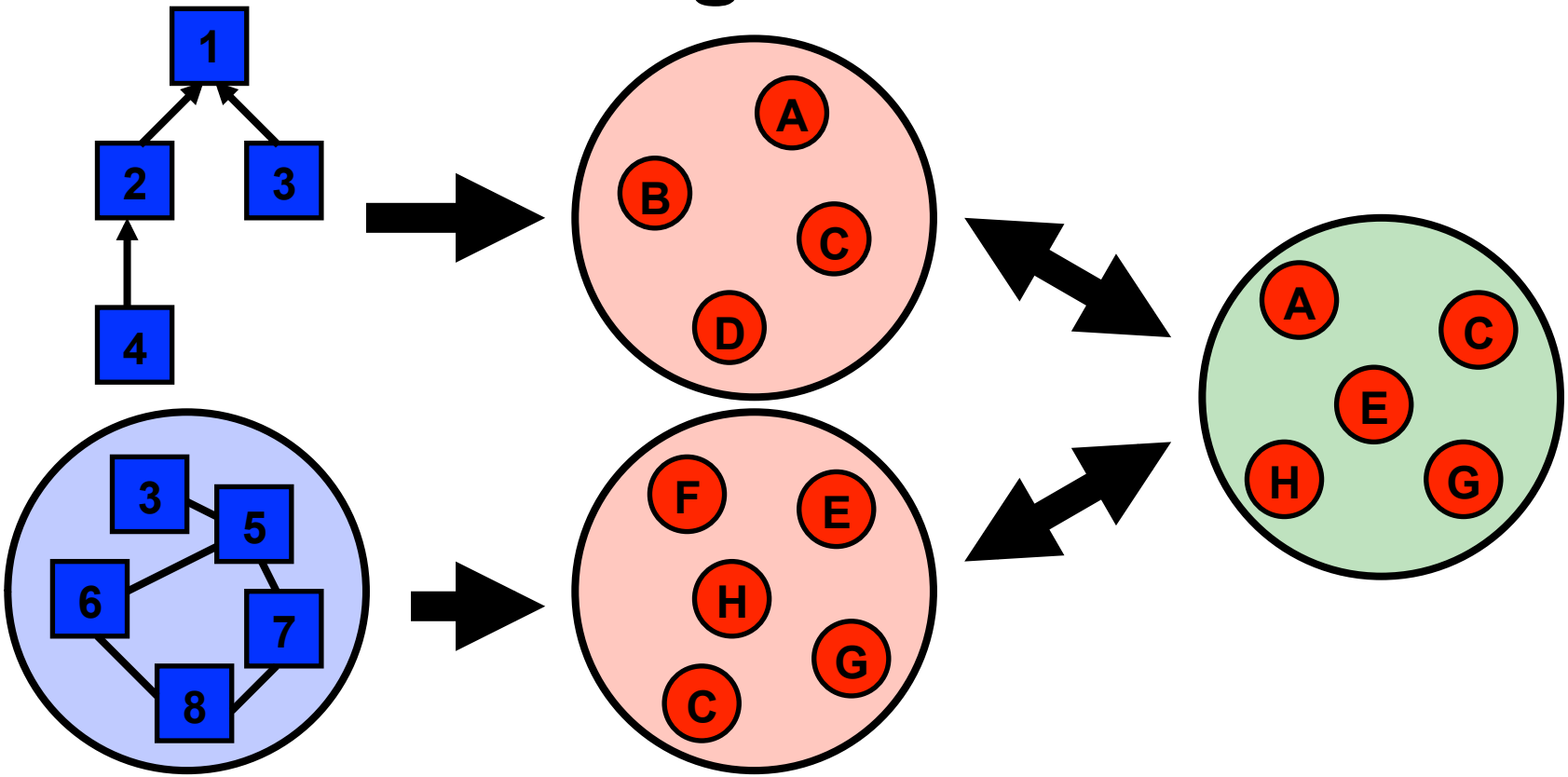


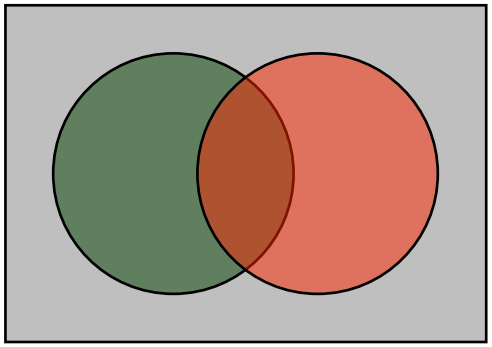
Each color represents a unique community.

Comparing the biological significance of communities and branches



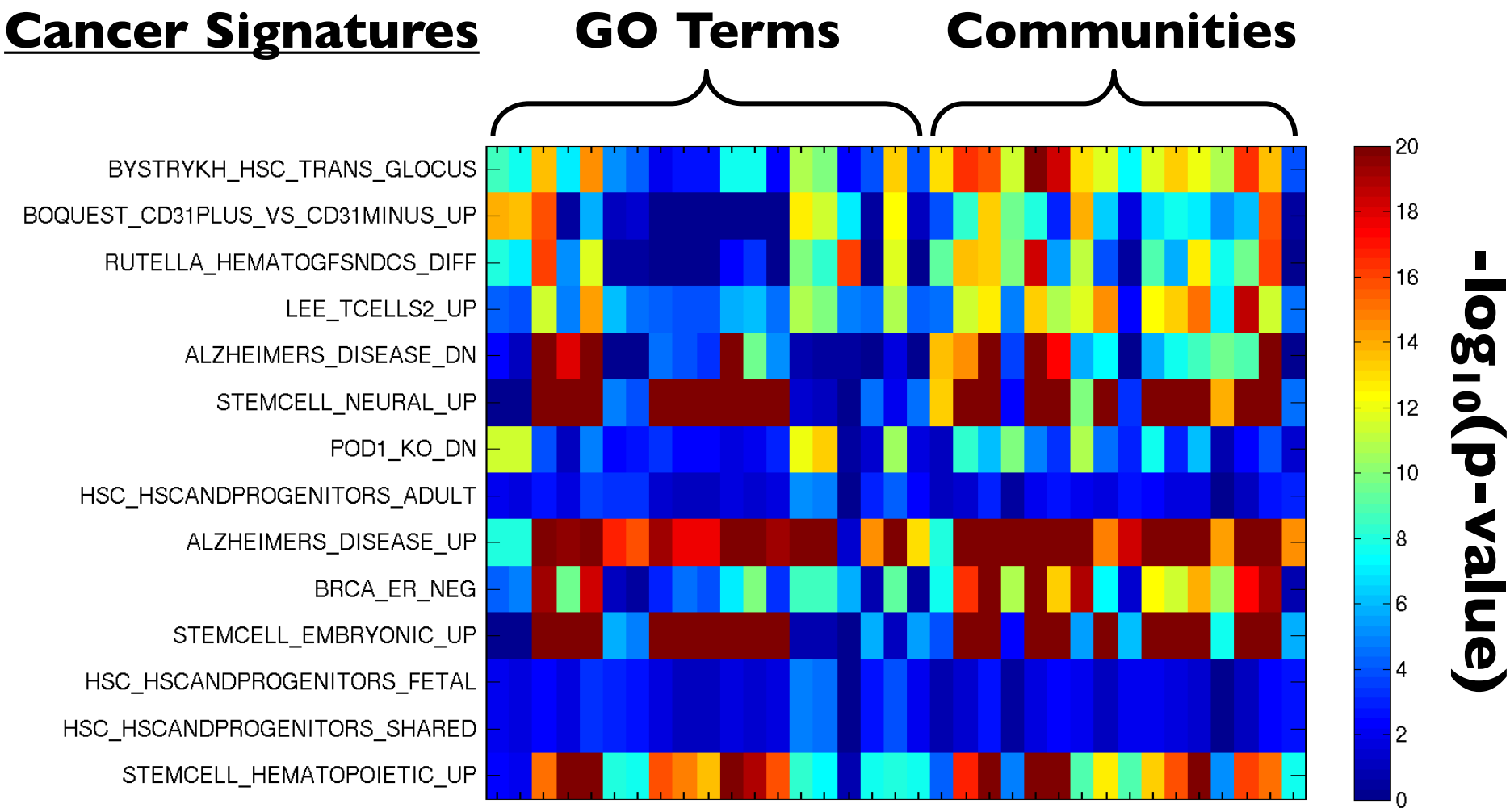
Community Enrichment in Cancer Signatures





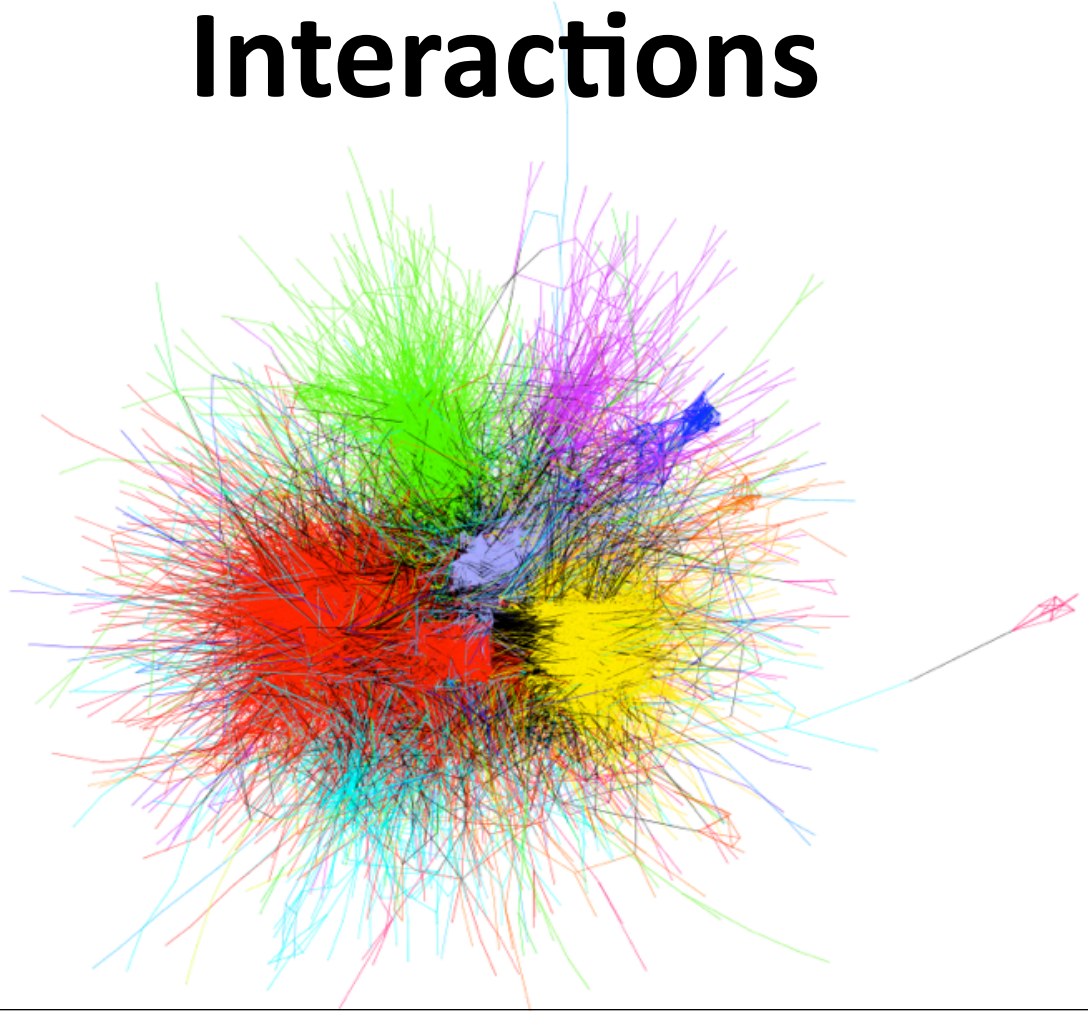
Hypergeometric probability returns a p-value for the similarity of the cancer signature to the genes annotated to terms in the branch of the hierarchy and for the similarity of the signature to genes annotated to terms in a community.

Community Enrichment in Cancer Signatures



Signatures defined in *"Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles"*

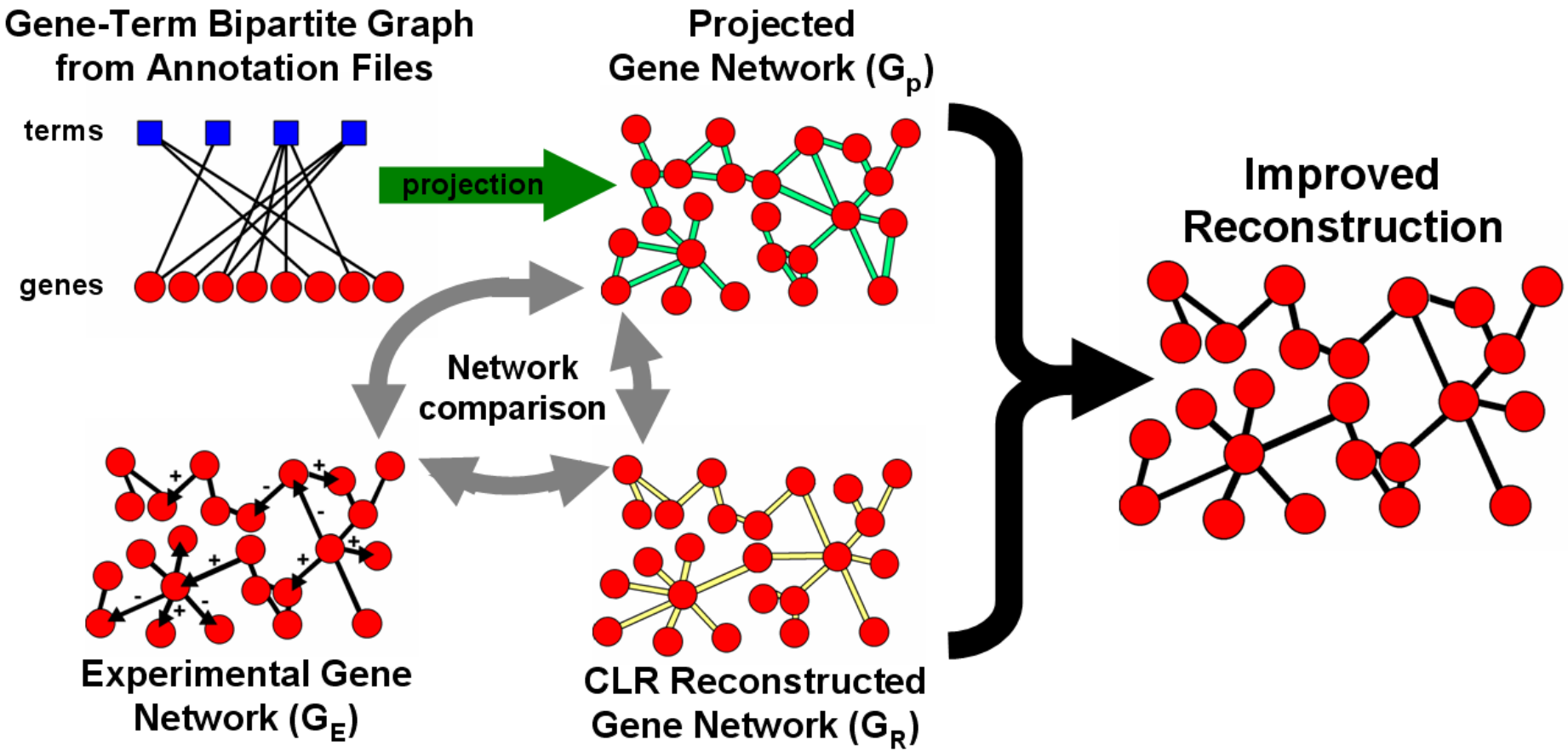
Implications of Functional Similarity for Gene Regulatory Interactions



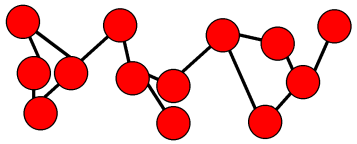
Why make a gene network from gene annotations?

- Is a cheap, easy way to generate a gene network for species for which there is no or limited experimental gene networks.
- Can be used to interpret known gene regulatory networks.
- Can be used to evaluate and/or improve existing network reconstruction algorithms.

Understanding and Improving Gene Network Reconstruction using Functional Relationships

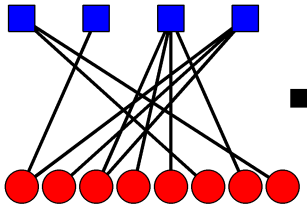


Weighting the Gene Network



$$w_{ii} = \frac{1}{\left(\sum_j B_{ij}\right)^{\alpha}} = \frac{1}{(\text{degree of term i})^{\alpha}}$$

G = B'wB



$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} = B$$

$$\begin{bmatrix} 1/2^{\alpha} & 0 & 0 & 0 \\ 0 & 1^{\alpha} & 0 & 0 \\ 0 & 0 & 1/4^{\alpha} & 0 \\ 0 & 0 & 0 & 1/3^{\alpha} \end{bmatrix} = w$$

In the limit of large α, edges in G to take a particular ordering such that those genes connected through many low degree terms have the highest weight.

Consequences of weighting G with large α

$$G_{ij} = \sum_{p} \frac{B'_{ip} B_{pj}}{\sum_{l} B_{pl}} = \sum_{\substack{\text{terms annotated to} \\ \text{both genes i and j}}} \frac{1}{(\text{degree of term})^{\alpha}}$$

- G_{ij} is largest when gene i and gene j are connected through many low degree terms.
- G_{ij} takes on a minimal value of 0 when gene i and gene j share no common annotations.
- G_{ij} is small when gene i and gene j are only connected through a single high degree term.

Comparing the Gene Network to Experimental Data

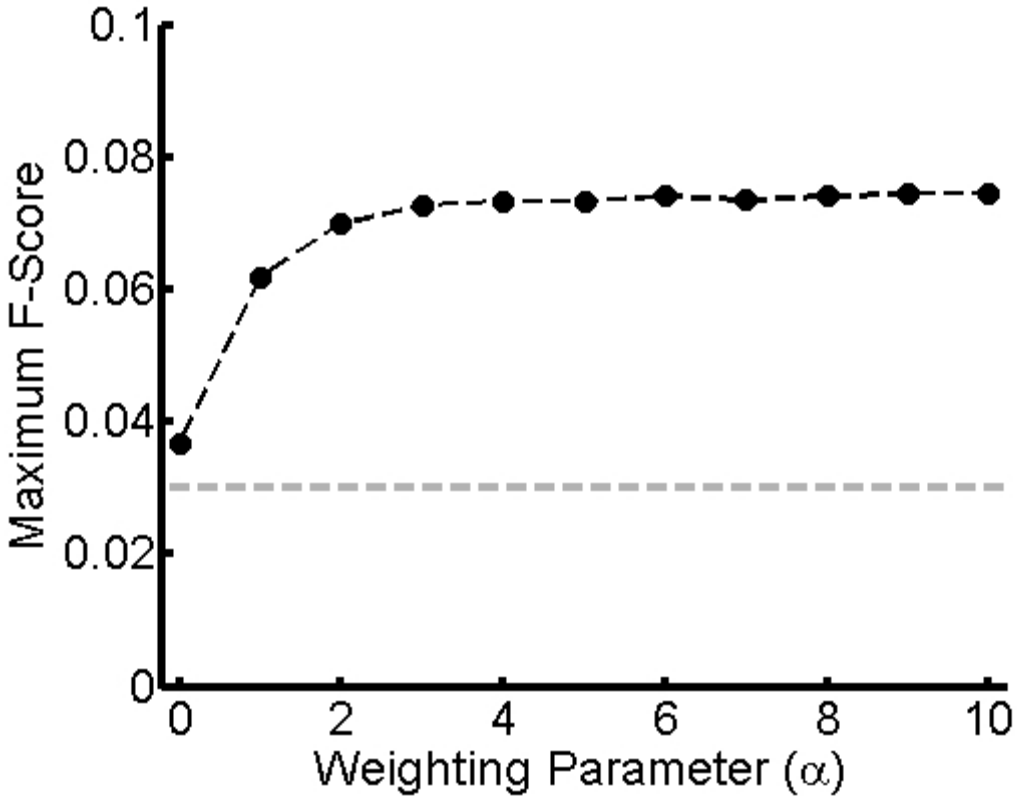
- We apply a threshold to the gene-gene network we create from annotation data such that every gene pair whose G_{ij} is above the threshold is considered connected.
- We compare this network to an experimentally derived regulatory network.
- For each threshold, we calculate the f-score to measure the utility of our gene-gene network for capturing true regulatory interactions.

$$F = 2\,\frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

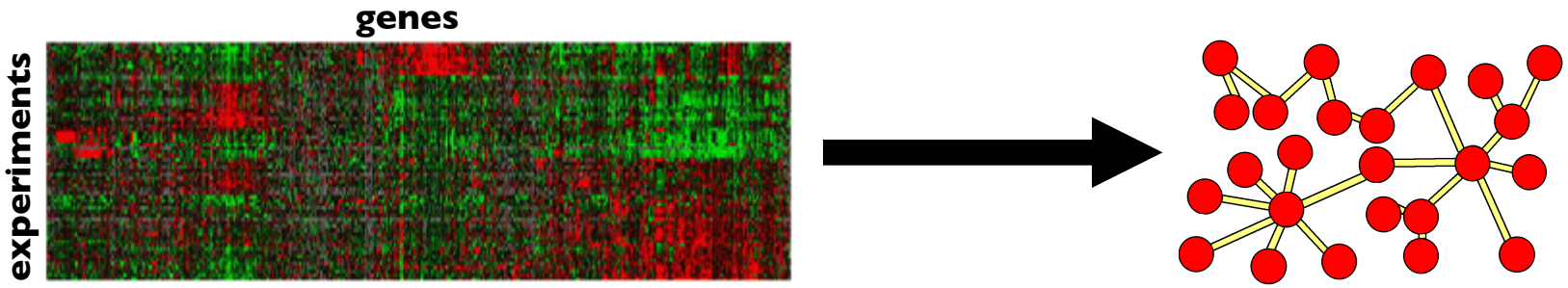
$$\text{Precsion} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

Inference power as a function of α



A gene network reconstructed from high-throughput data (G_R)



Context-Likelihood-of-Relatedness

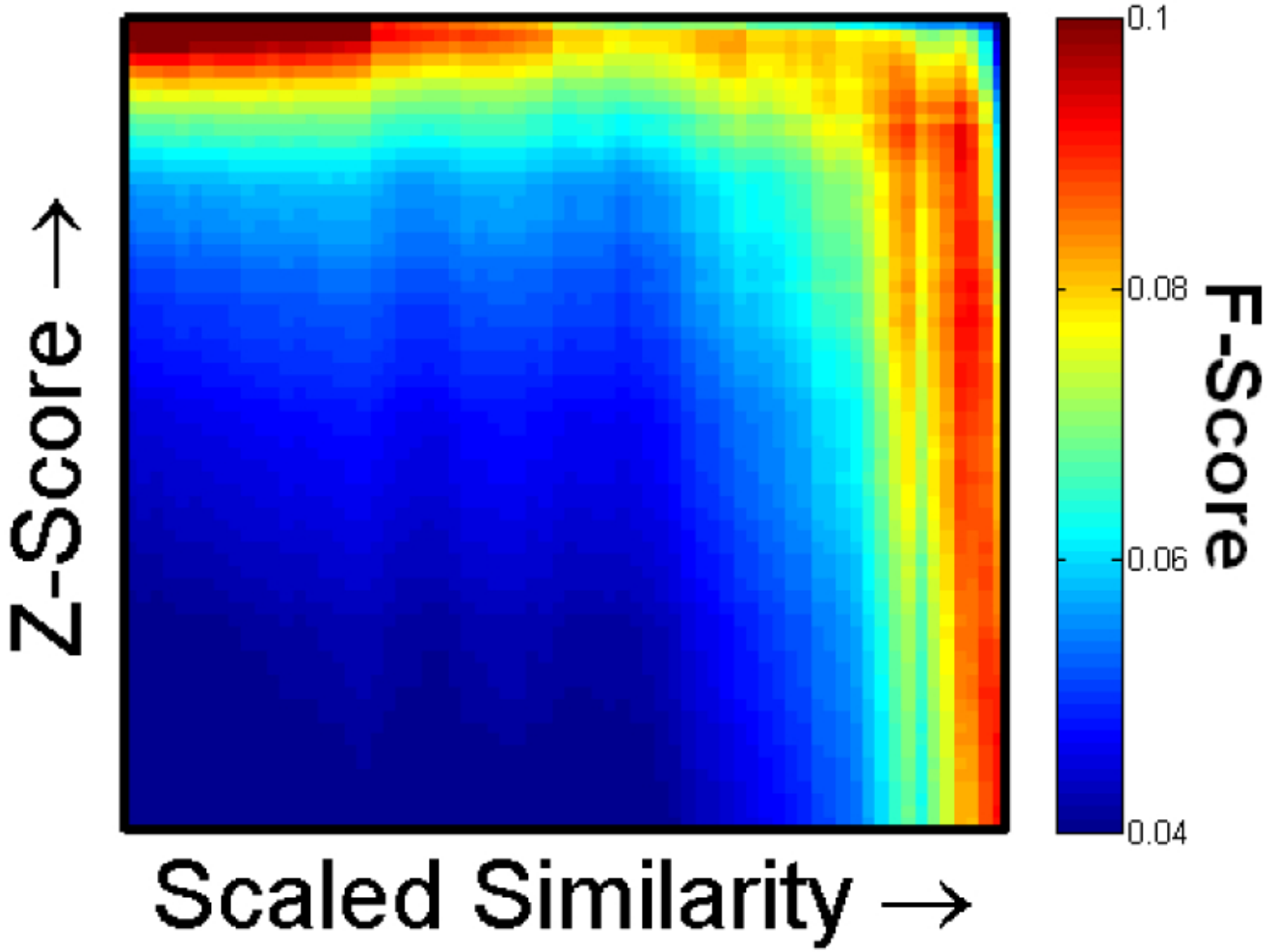
- Calculates the mutual information between pairs of genes using expression data.
- Uses that mutual information profile to calculate a Z-Score for these pairs of genes.

$$MI(a,b) = \sum_{i,j} p(a_i,b_j) \log \frac{p(a_i,b_j)}{p(a_i)p(b_j)}$$

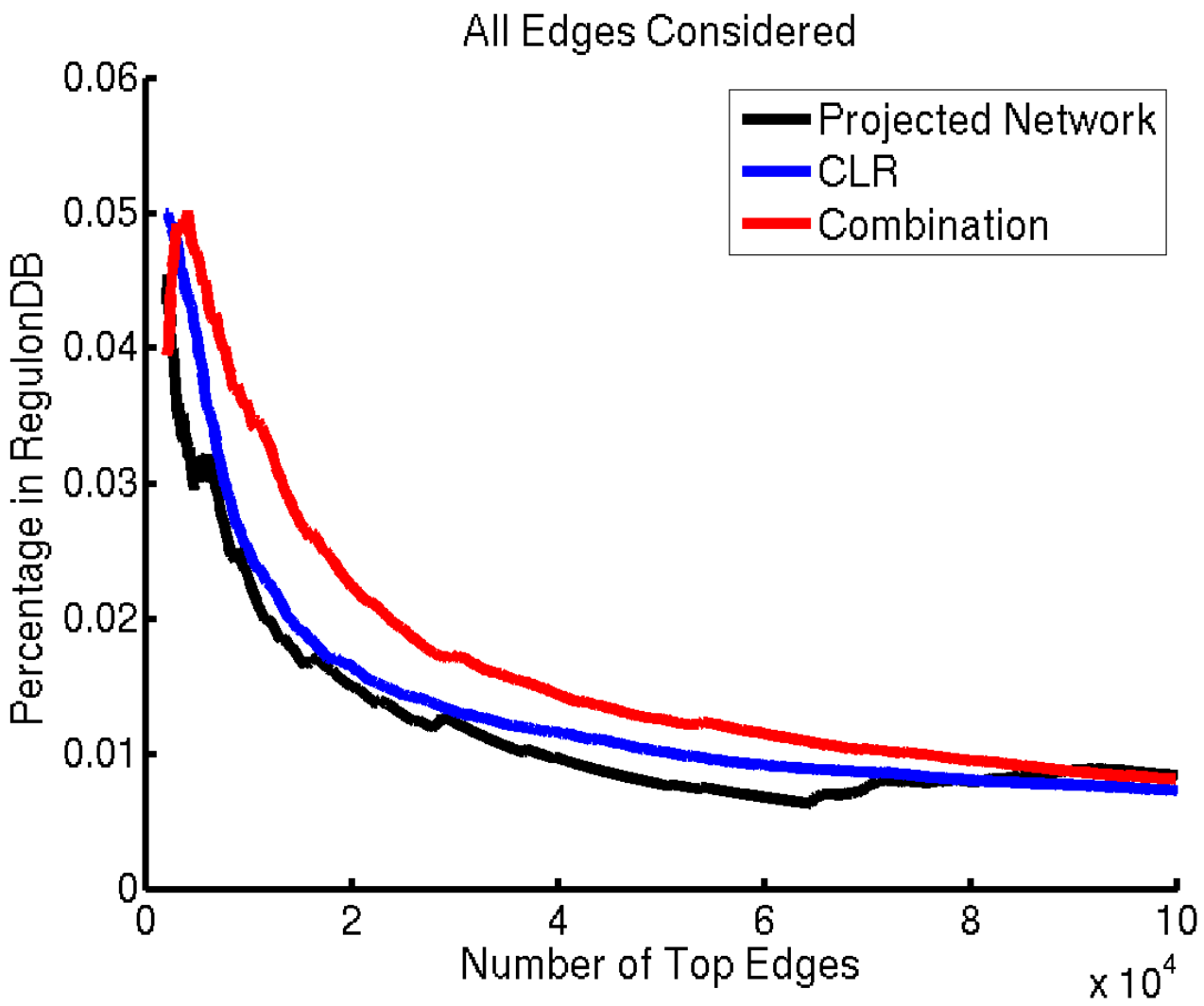
- Z-Score value meant to predict true regulatory interactions.

reference for CLR algorithm: Faith, *PLoS Biology*, 2007.

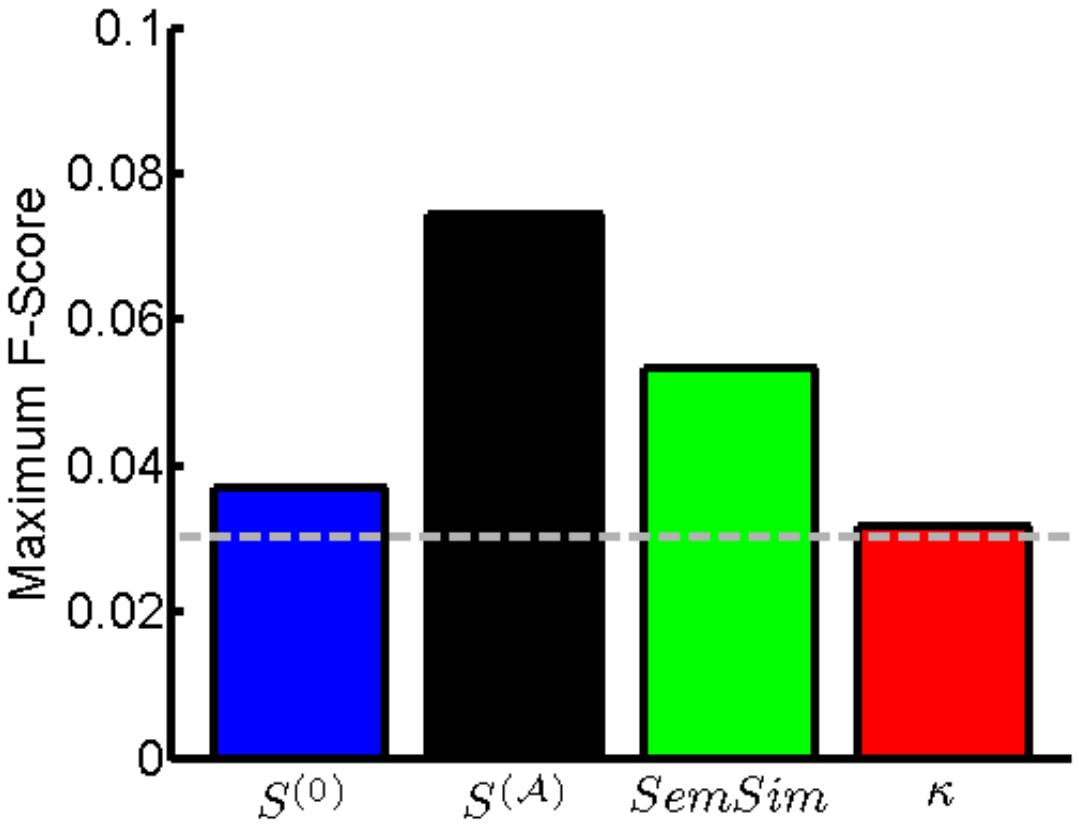
Comparison to CLR Reconstruction



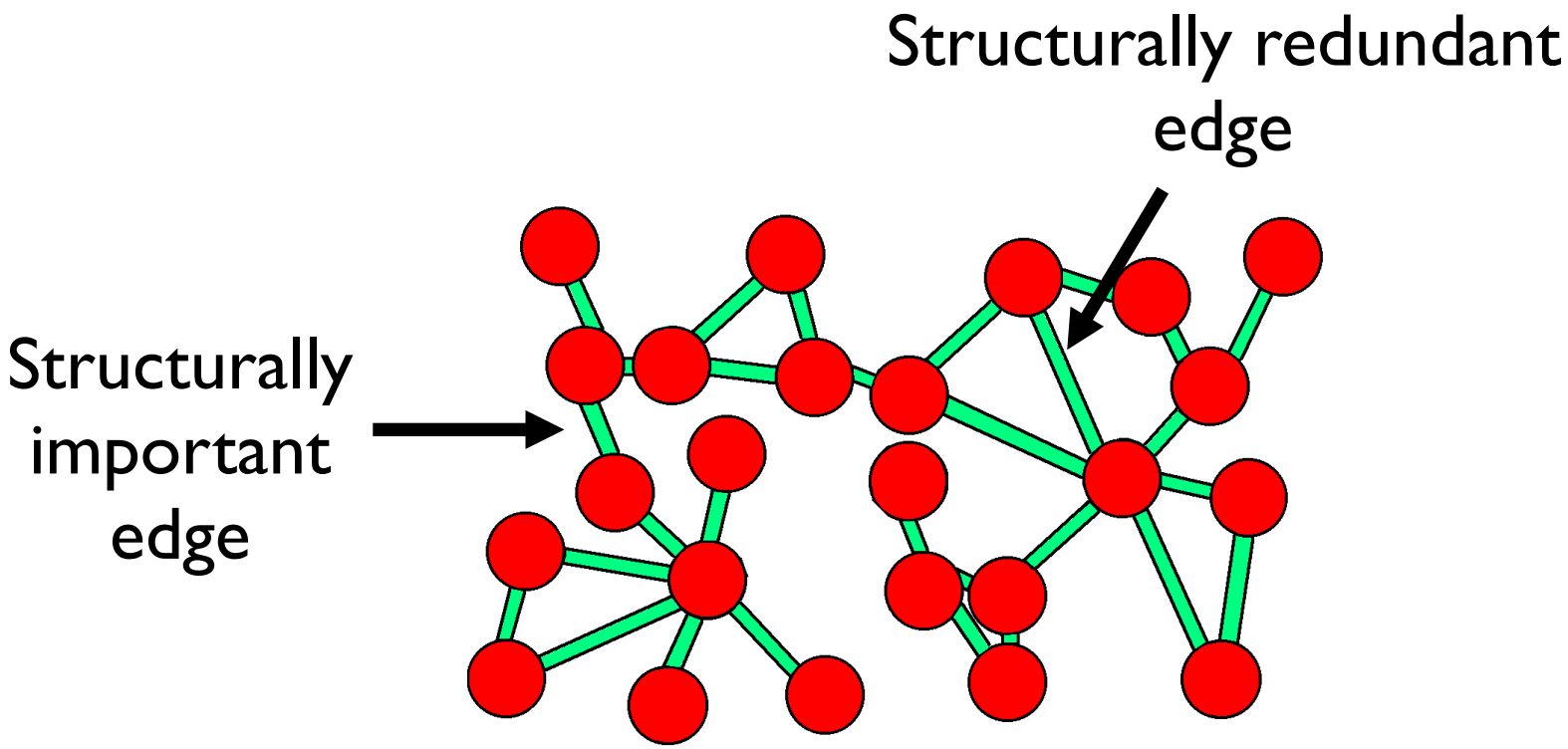
Improving Network Reconstruction



Comparison with other measures of functional similarity

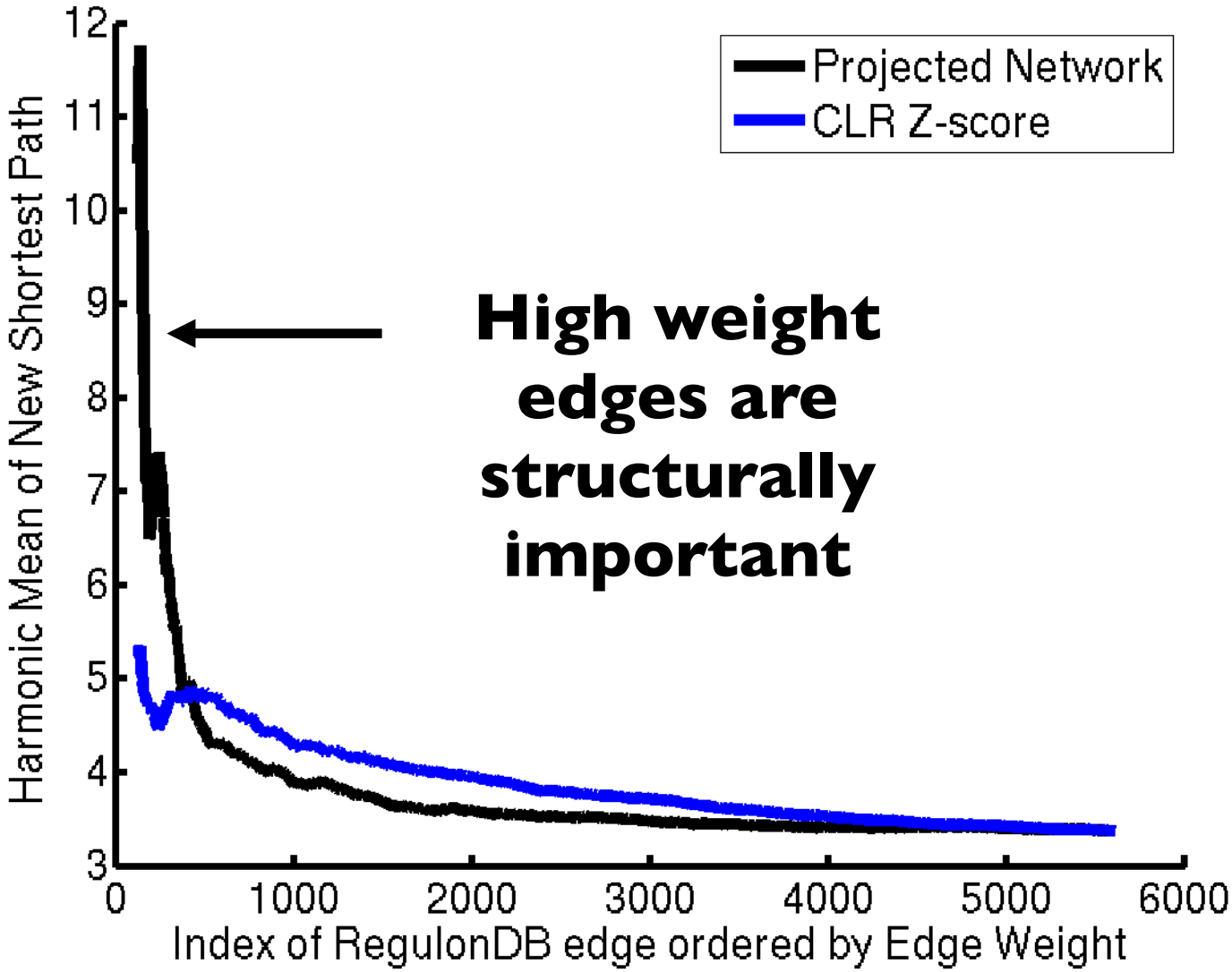


What does it mean to have functional similarity?



To measure how structurally important or redundant an edge is in G_E, we calculated the new shortest path between nodes upon the removal of that edge.

A biological interpretation of functional similarity



Conclusions

- There is an alternate natural way to group GO terms, unique from the hierarchy, which provides an independent framework with which to describe and predict the functions of experimentally identified groups of genes.
- GO can be used to create a gene-network entirely based on functional annotations. Properties of this network are correlated with known regulatory interactions.
- This gene network identifies a different subset of regulatory interactions than those predicted by the CLR algorithm and can be combined with CLR further to improve predictive power.

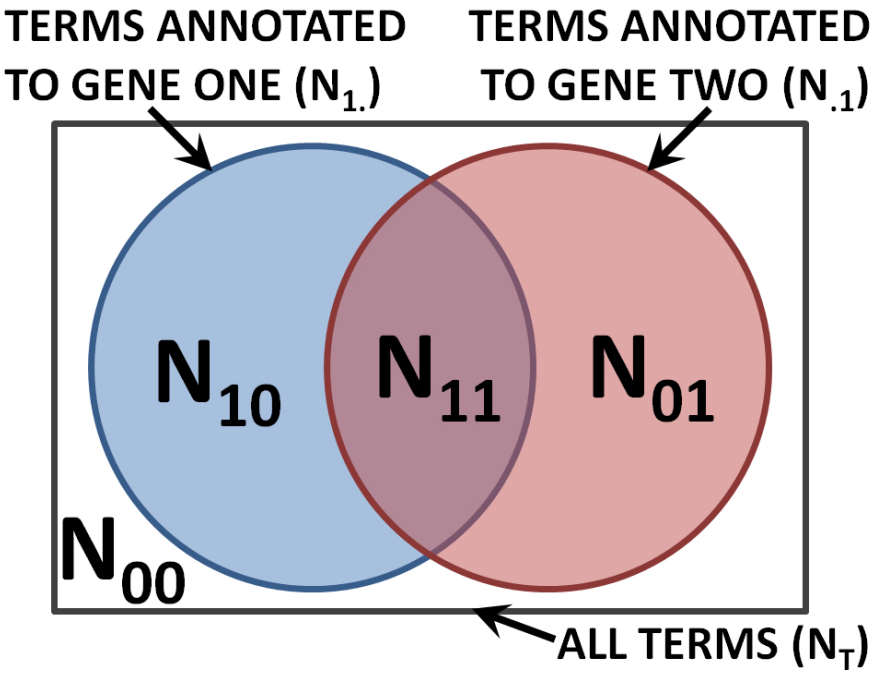
Semantic Similarity

- Define the probability, p(t), of observing a term t as the number of gene annotations made to that term, divided by the number of gene annotations made to the parent node of the branch to which the term belongs.
- The semantic similarity between two terms is then defined as
- where $T(t_1,t_2)$ is the set of parent terms shared by the two terms.

$$SemSim(t_1,t_2) = -\log \min_{t \in T(t_1,t_2)} p(t)$$

- In order to find the semantic similarity between two genes, G1 and G2, one constructs an $n_{G1} x n_{G2}$ where n_{G1} (n_{G2}) is the number of terms annotated to G1 (G2), and populates it with the semantic similarity values between all the pairs of terms. The semantic similarity between the two genes is then determined by taking the average of all values in the matrix.

Kappa statistics



$$X = \frac{N_{11} - N_{00}}{N_T}$$

$$\kappa = \frac{X - \langle X \rangle}{1 - \langle X \rangle}$$