Network Analysis

BCH 5101: Analysis of -Omics Data

Network Analysis

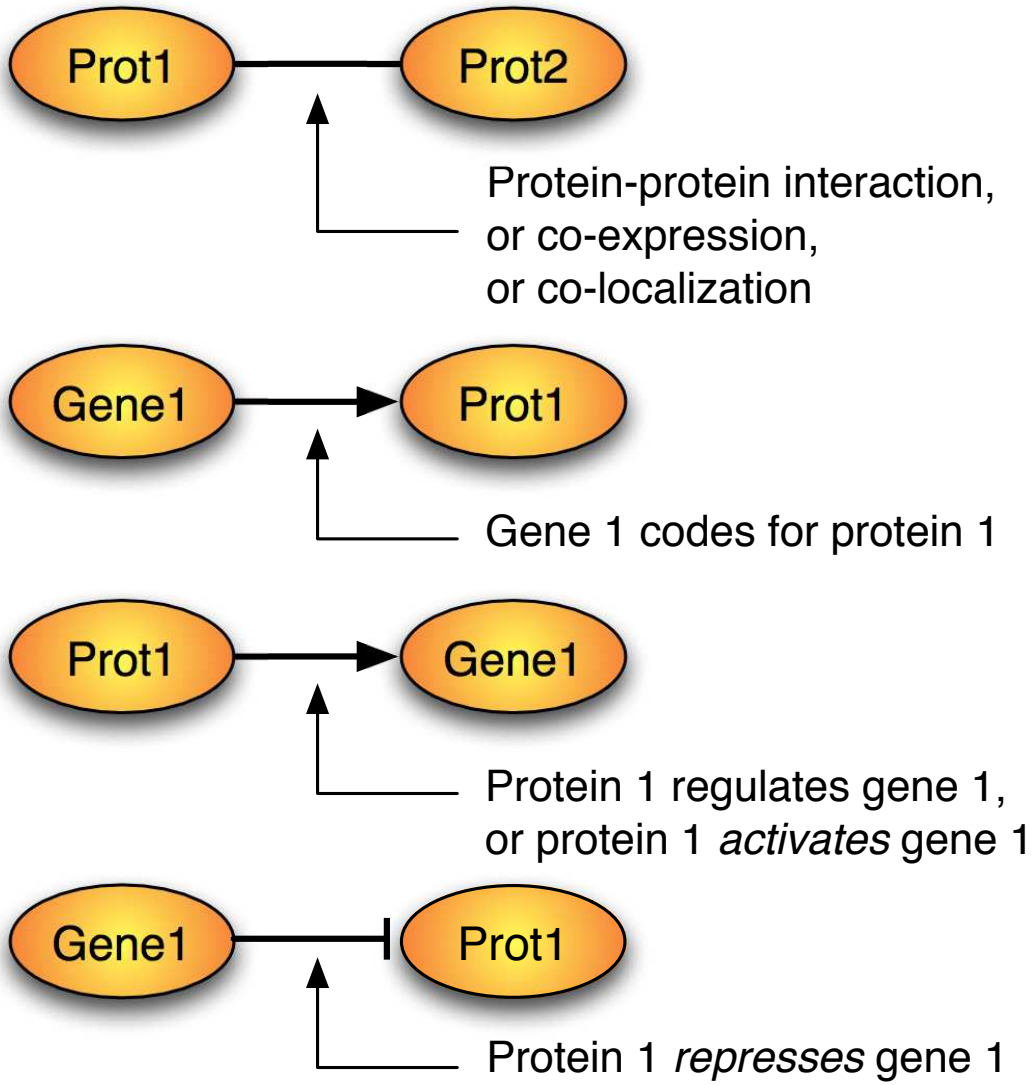
- Graphs as a representation of networks
- Examples of genome-scale graphs
- Statistical properties of genome-scale graphs
- The search for meso-scale subgraphs/modules
- Network motifs

Networks

Mathematically, a network is often represented as a *graph*, having:

- A set of *vertices*, or *nodes* — typically corresponding to genes, mRNAs, proteins, complexes, etc.
- A set of *edges* — representing interactions, regulation, co-expression, co-location, etc.
- Edges may be *directed* ($A \rightarrow B$) or *undirected* ($A - B$)

Common edge types and meanings



Networks (again)

Mathematically, a network is often represented as a *graph*, having:

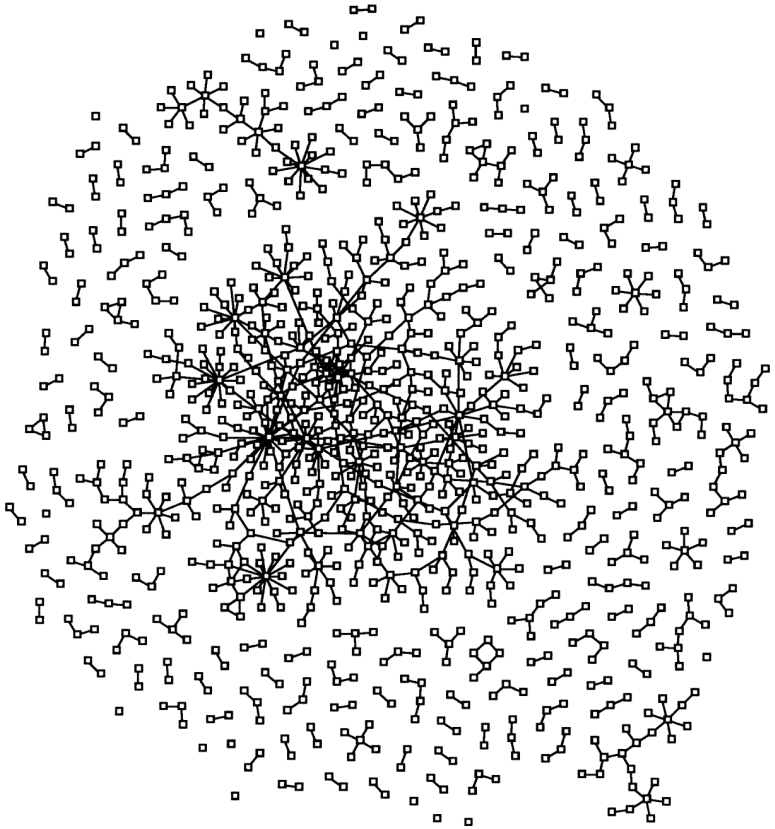
- A set of *vertices*, or *nodes* — typically corresponding to genes, mRNAs, proteins, complexes, etc.
- A set of *edges* — representing interactions, regulation, co-expression, co-location, etc.
- Edges may be *directed* ($A \rightarrow B$) or *undirected* ($A - B$)

Sometimes, we want to extend this traditional notion of graphs:

- Edges may have types (e.g., activation or repression)
- Edges or vertices may be *weighted* (assigned a real number, indicating importance, evidence, relevance, etc.)
- Edges may point to other edges

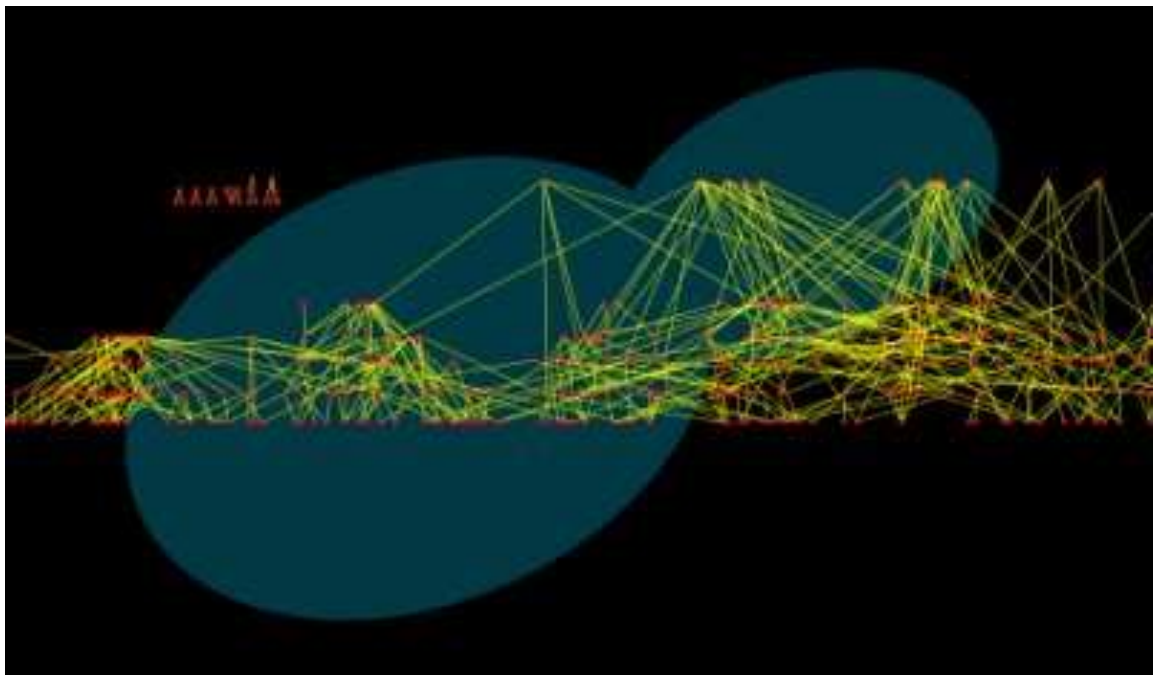
Example: Uetz et al. yeast PPI network

- Uetz *et al.* (Nature,2000) performed two variants of yeast two-hybrid screening to test for interactions between proteins in *S. cerevisiae*
- They found 957 interactions (far fewer than there really are!) involving 1004 proteins



Yeast transcriptome

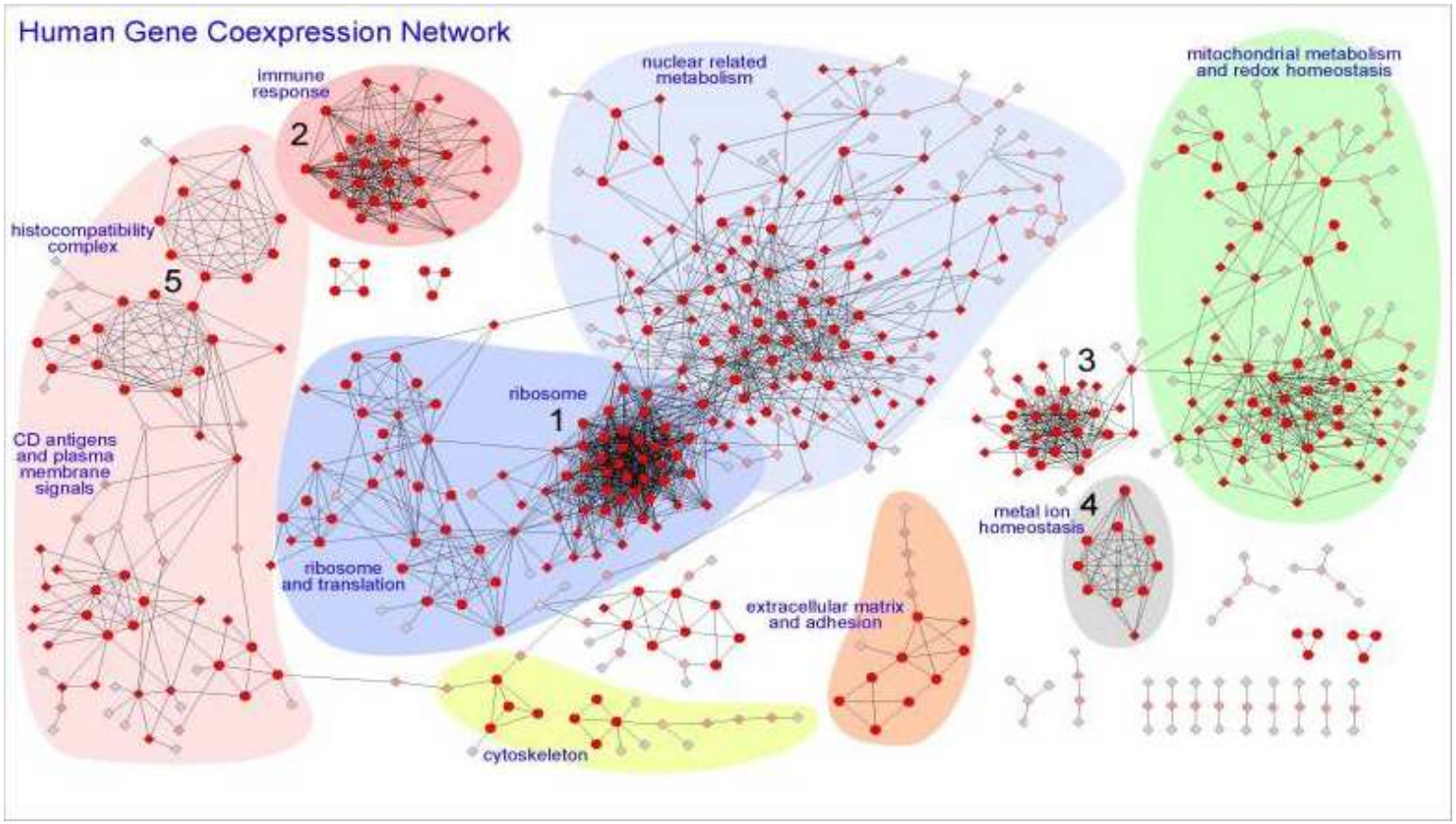
- Lee et al. (Nature, 2002) used ChIP-chip to determine transcription factor $\rightarrow$ gene promoter binding relationships
- 106 of 141 known transcription factors were successfully tested, resulting in $\approx$ 4000 relationships at a stringent p-value threshold



(Picture from Beyer lab website.)

Human co-expression network

- Prieto *et al.* (PLoS One, 2008) used microarrays on a set of human tissue samples to determine co-expression.
- They found 15841 high-confidence relationships between 3327 genes.

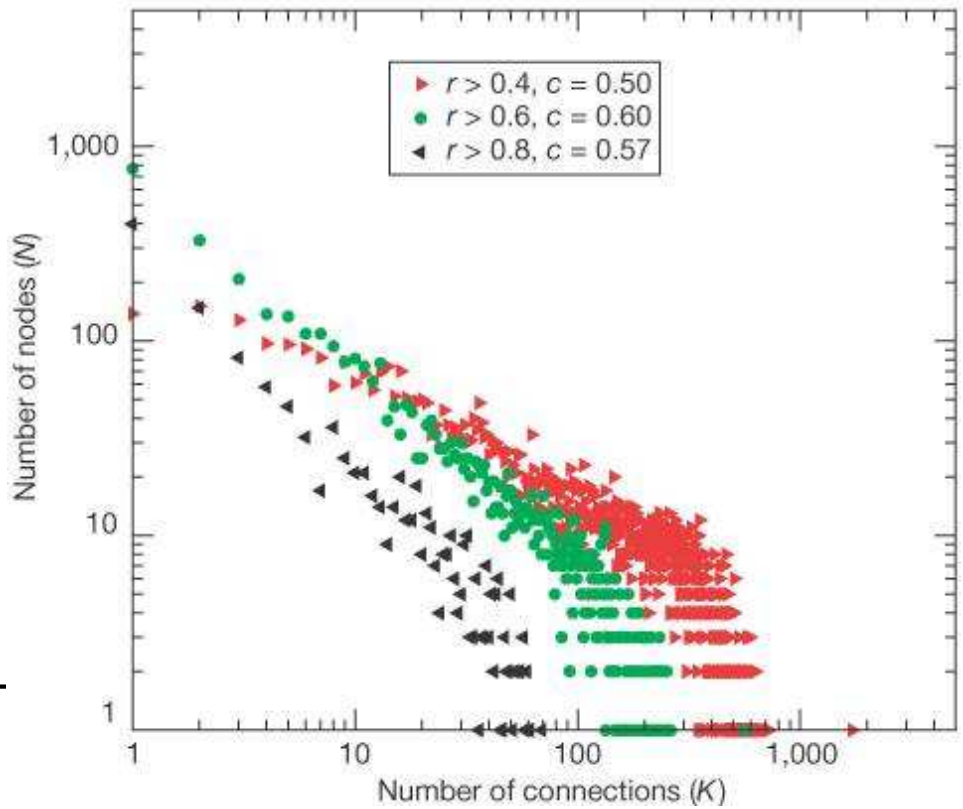


Typical structure of large-scale newtorks

- Such networks are often found to be approximately "scale-free":
  - A few vertices ("hubs") have many edges
  - Most vertices ("leaves") have just a few
  - Formally, the number vertices with k edges is proportional to $1/k^{\alpha}$ for some real number $\alpha > 1$
- (See previous graphs, especially Uetz *et al.*)

Yeast co-expression network scale-free

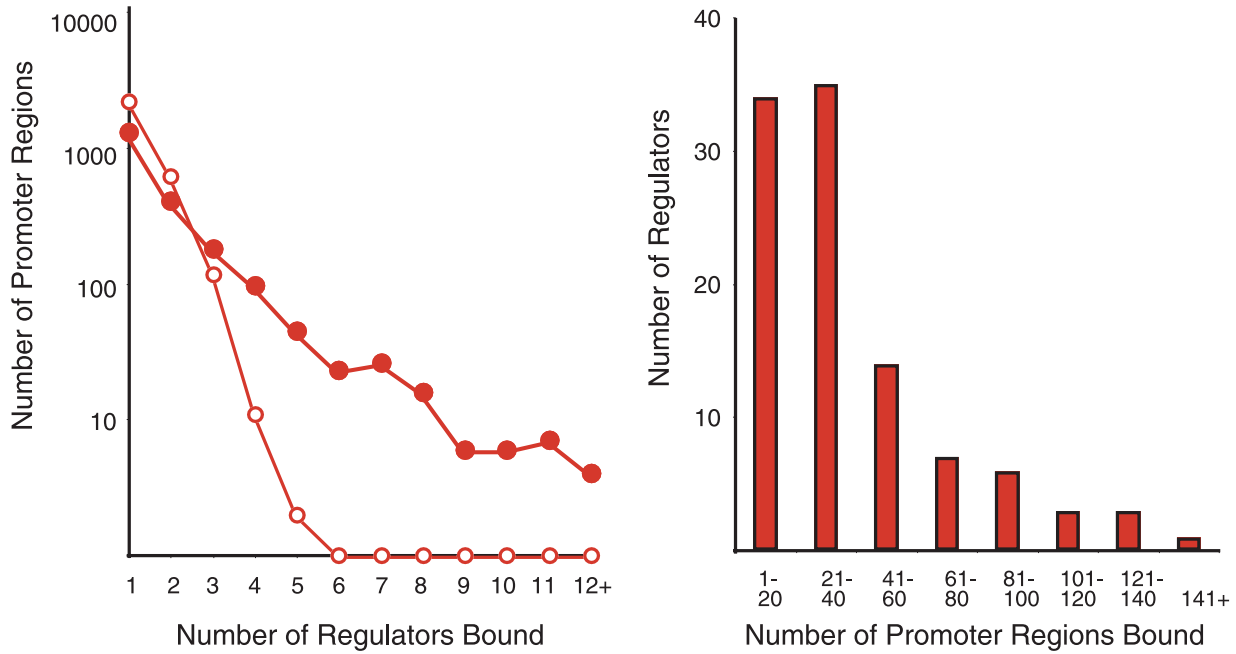
- van Noort *et al.* (EMBO Rep, 2004) analyzed co-expression in the data of Hughes *et al.* (Cell, 2000)
- 4077 genes (nodes) are connected by 65,430 undirected edges using a correlation threshold of 0.6
- At that level of correlation, and at other levels, the degree distribution is approximately scale-free



"Scale-free" topology of Yeast transcriptional network

(Lee *et al.*, Science 2002)

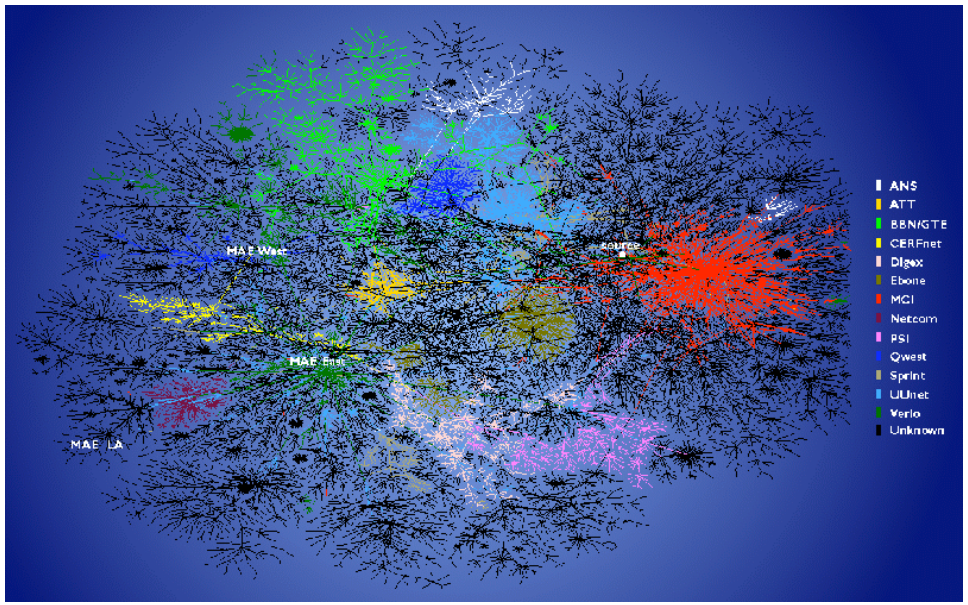
- In-degree and out-degree of resulting graph shows hubs and leaves.



(Ignore the open red circles; they're a null model.)

Barabasi

- Modern excitement over scale-free networks began with work of Barabasi (Nature, 1999; Science, 1999)
- Many other natural and artificial networks show scale-free degree distributions, e.g., internet, www, friendship, roads, metabolism, etc.



- Barabasi proposed *preferential attachment*, or "rich grow richer" model of network growth to explain scale-free nature — new nodes more likely to make connections to existing high degree nodes

Preferential attachment

The Barabasi-Albert model works as follows:

- We begin with a "core" connected network that is "small"
- We repeatedly add a new vertex to the graph, along with m edges
- The probability that an edge is added between the new vertex and existing vertex i (P_i) is proportional to the degree of i (d_i)

$$P_i = \frac{d_i}{\sum_j d_j}$$

Asymptotically, this produces graphs with a power-law degree distribution; the number of vertices with degree d is $N(d) \propto d^{-3}$. (Independent of m.)

Other variants can produce different exponents, different average degrees, etc.

Are high-degree nodes in genome networks important?

- There have been attempts to relate graph structure / high-degree nodes to biological properties:
  - Robustness: does the network survival deletion or mutation of a gene?
  - Conservation: are higher-degree nodes more constrained?
  - Expression: are higher-degree nodes expression more often?
- Results are mixed...
- High-degree nodes may also be "promiscuous", "sticky", "nonspecific" — or experimental artifacts.
- ...still, high-degree nodes are a natural place to start an investigation.

Explaining scaling in biological networks

- Could preferential attachment explain powerlaw scaling in biological networks?

Explaining scaling in biological networks

- Another answer: evolutionary drift
  - Assume genes duplicated at random (singly, or in blocks) – carrying their regulatory or interacting links with them
  - Assume genes deleted at random
  - Assume interactions have random chance of creation or deletion between existing genes
- Then one can show a power-law degree distribution will result (See, e.g., Wagner (Proc. R. Soc. Lond. B, 2003) for protein interaction networks.)

$\Rightarrow$ Note: no explicit evolutionary selection for individual genes or for network structure as a whole!