Bioinformatics: Network Analysis

Graph-theoretic Properties of Biological Networks

COMP 572 (BIOS 572 / BIOE 564) - Fall 2013

Luay Nakhleh, Rice University

Outline

- Architectural features
- Motifs, modules, and hierarchical networks
- Scale-free or geometric?

Architectural Features

Cellular Networks Are Scale-free

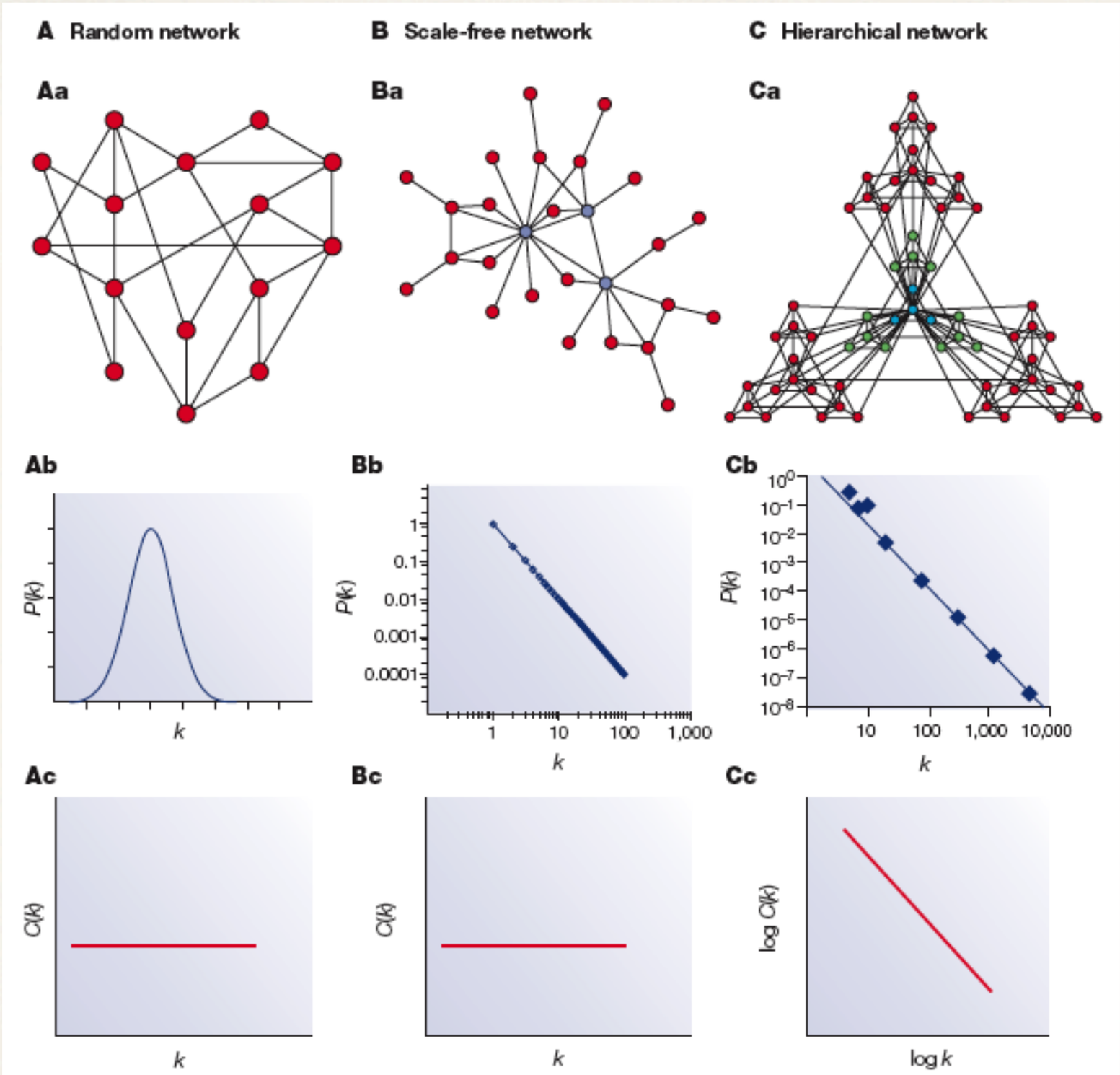
- Analysis of cellular networks of various types have indicated scale-free topologies

- The first evidence came from analyzing metabolism (vertices are metabolites, edges are enzyme-catalyzed biochemical reactions, and the edges are directed)

- P(k): degree distribution

- C(k): 2n/(k(k-1)) (n is the number of edges connecting neighbors of a node of degree k)

- It has been observed that $C(k) \sim k^{-1}$ reflects hierarchical structure of the network (in scale free networks)


A Random network
B Scale-free network
C Hierarchical network
Aa
Ba
Ca
Ab
Bb
Cb
P(k)
k
0.1
0.01
0.001
0.0001
1
10
100
1,000
10,000
Ac
Bc
Cc
C(k)
log C(k)
log k

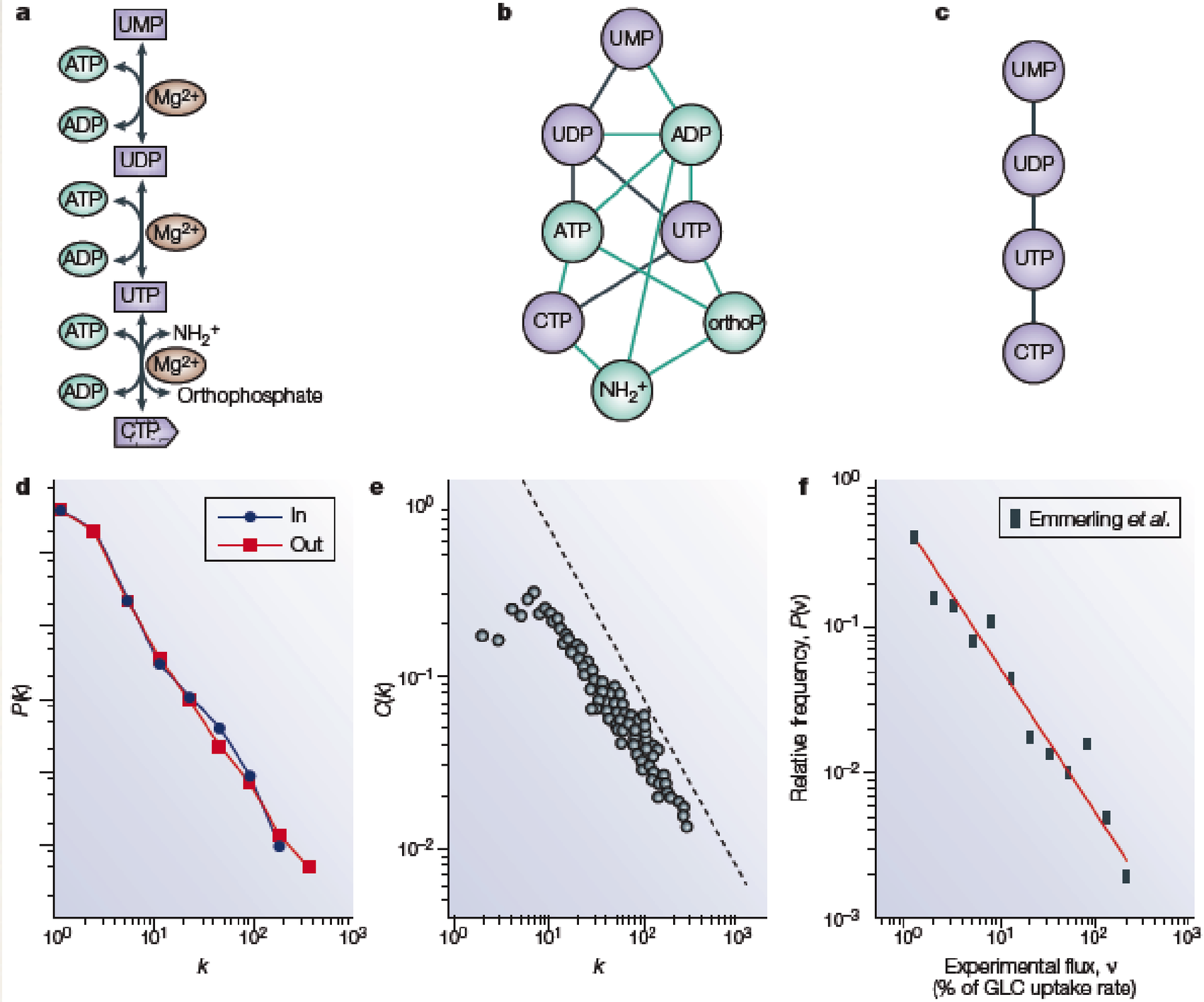


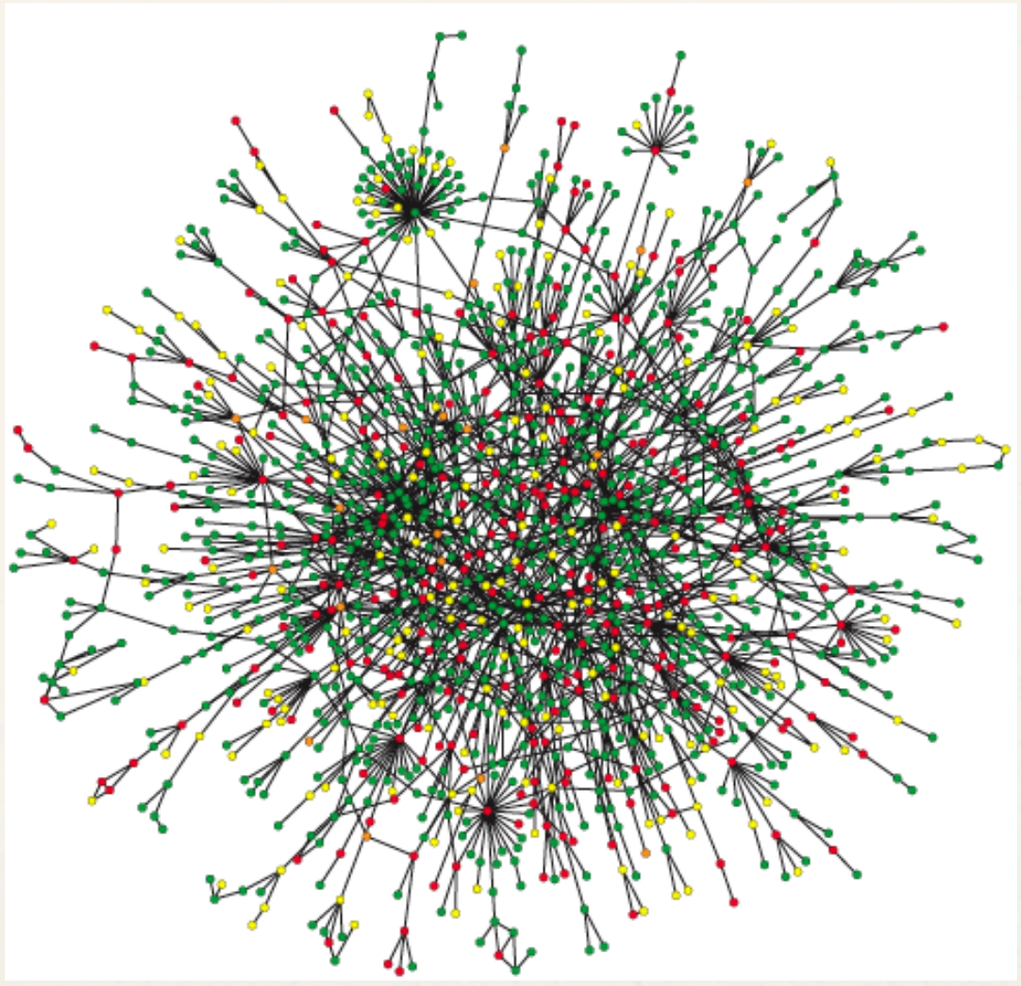
Figure 1 | **Characterizing metabolic networks.** To study the network characteristics of the metabolism a graph theoretic description needs to be established. Here, the graph theoretic description for a simple pathway (catalysed by Mg^{2+}-dependant enzymes) is illustrated (**a**). In the most abstract approach (**b**) all interacting metabolites are considered equally. The links between nodes represent reactions that interconvert one substrate into another. For many biological applications it is useful to ignore co-factors, such as the high-energy-phosphate donor ATP, which results in a second type of mapping (**c**) that connects only the main source metabolites to the main products. **d** | The degree distribution, $P(k)$ of the metabolic network illustrates its scale-free topology[16]. **e** | The scaling of the clustering coefficient $C(k)$ with the degree k illustrates the hierarchical architecture of metabolism[53] (The data shown in **d** and **e** represent an average over 43 organisms[16,53]). **f** | The flux distribution in the central metabolism of *Escherichia coli* follows a power law, which indicates that most reactions have small metabolic flux, whereas a few reactions, with high fluxes, carry most of the metabolic activity[91]. This plot is based on data that was collected by Emmerling *et al.*[106]. It should be noted that on all three plots the axis is logarithmic and a straight line on such log–log plots indicates a power-law scaling. CTP, cytidine triphosphate; GLC, aldo-hexose glucose; UDP, uridine diphosphate; UMP, uridine monophosphate; UTP, uridine triphosphate.

Scale-free Metabolic Networks

- The analysis of metabolic networks of 43 different organisms from all three domains of life (eukaryotes, prokaryotes, and archaea) indicates that the cellular metabolism has a scale-free topology, in which most metabolic substrates participate in only one or two reactions, but a few, such as pyruvate or coenzyme A, participate in dozens and function as metabolic hubs.

Scale-free PPI Networks

- Several recent publications indicate that PPI networks have a scale-free topology



PPI network of the yeast Saccharomyces cerevisiae

Other Types of Networks

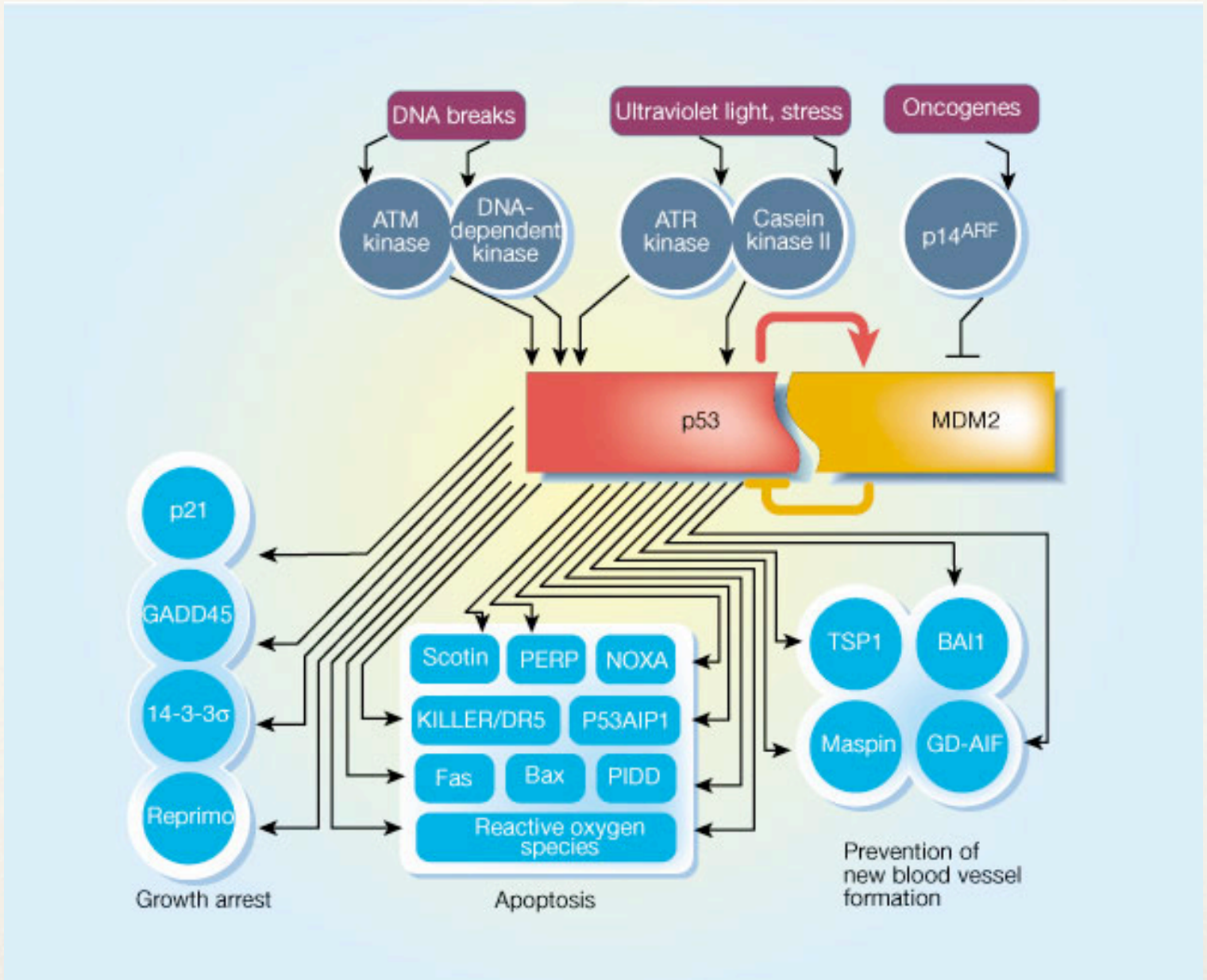
- Genetic regulatory networks (vertices are genes and edges are expression correlations) also exhibit scale-free topologies

- Transcription regulatory networks (vertices are genes and transcription factors, and edges are interactions) exhibit mixed scale-free and exponential distributions:

  - The distribution of the number of genes that a transcription factor interacts with follows a power-law (scale-free). Most TFs regulate only a few genes, but a few TF's regulate many genes.

  - The distribution of the number of transcription factors that interact with a given gene follows an exponential distribution. Most genes are regulated by 1-3 TFs.

- While establishing scale-free properties is hard when information is available on only a few nodes, a salient feature of cellular networks is the presence of hubs, from regulatory webs to the p53 module



Source: "Surfing the p53 network", Vogelstein et al., Nature 408: 307-310, 2000.

Small-world Effect

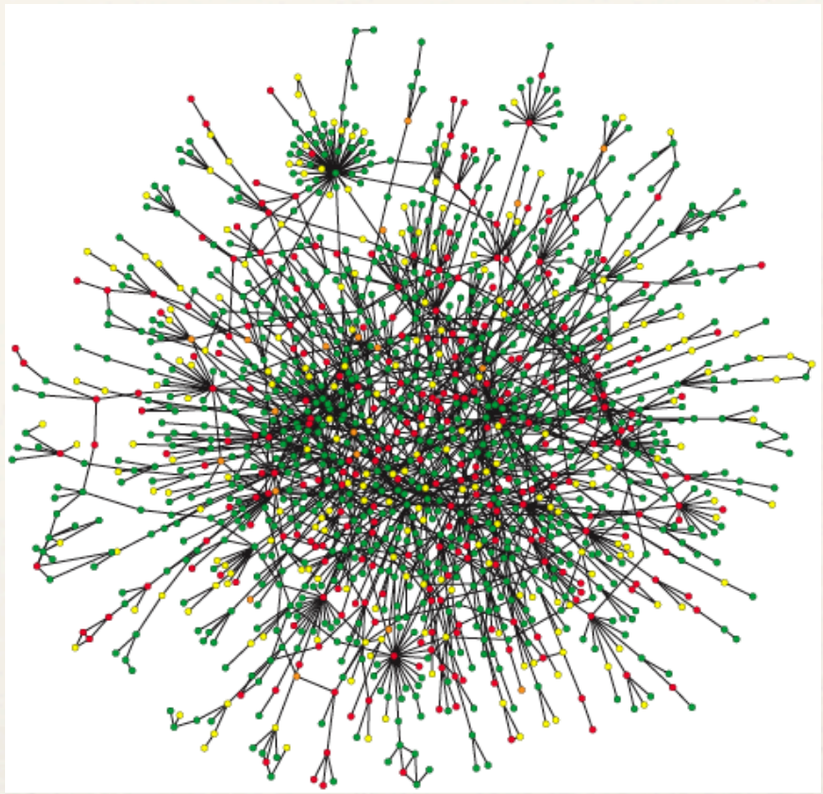
- Although the small-world effect is a property of random networks, scale-free networks are ultra small

- In metabolism, paths of only 3-4 reactions can link most pairs of metabolites (implication: local perturbations in metabolite concentrations could reach the whole network very quickly)

- Interestingly, the metabolic network of a parasitic bacterium has the same mean path length as the much larger and more developed network of a large multicellular organism (implication: certain evolutionary mechanisms have maintained the average path length during evolution?)

Assortativity

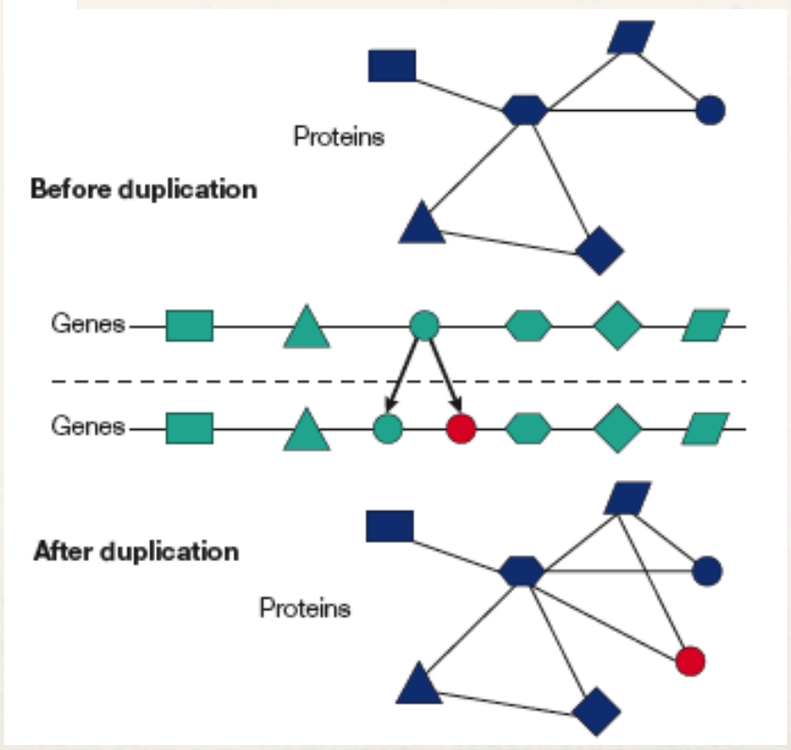
- Cellular networks seem to be disassortative: hubs avoid linking directly to each other and instead connect to vertices with only a few connections
- The origin of disassortativity in cellular networks remains unexplained



Evolutionary Origin of Scale-free Networks?

- Recall the two processes underlying the development of real networks: growth (new nodes joining the network) and preferential attachment (nodes prefer to connect to nodes that have many edges)

- In protein networks, growth and preferential attachment have a possible evolutionary explanation that is rooted in gene duplication



Evolutionary Origin of Scale-free Networks?

- Duplicated genes produce identical proteins that interact with the same protein partners

- Highly connected nodes get more new links: not that they have a higher probability of duplicating, but a higher probability to have a link to a duplicated gene

- The role of gene duplication has been shown only for PPI networks, but not for regulatory or metabolic networks

Evolutionary Origin of Scale-free Networks?

- An inspection of the metabolic hubs indicates that the remnants of the RNA world, such as coenzyme A, NAD and GTP, are among the most connected substrates of the metabolic network, as are elements of some of the most ancients metabolic pathways, such as glycolysis and TCA cycle.

- Recall the correlation between the age and degree of a node in the scale-free model

Motifs, Modules, and Hierarchical Networks

A Brief Overview

[More on this topic later]

- Cellular functions are likely to be carried out in a highly modular manner: a group of physically or functionally linked molecules (nodes) work together to achieve a distinct function

- Biology is full of examples of modularity

- Questions of interest: Is a given network modular? What are the modules in a network? What are their relationships in a given network?

High Clustering in Cellular Networks

- In a network representation, a module appears as a highly interconnected group of nodes

- Each module can be reduced to a set of triangles, and the clustering coefficient can be computed to quantify modularity

- In the absence of modularity, the clustering coefficient of the real and random networks are comparable

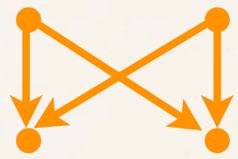
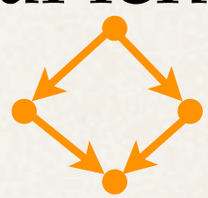
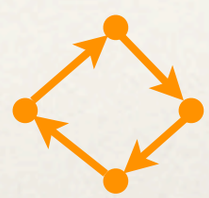
- Metabolic, PPI, and protein domain networks have all exhibited high clustering

Motifs in Cellular Networks

- Not all subgraphs occur with equal frequency
- Motifs are subgraphs that are over-represented compared to a randomized version of the same network

- To identify motifs:
  - Identify all subgraphs of n nodes in the network
  - Randomize the network, while keeping the number of nodes, edges, and degree distribution unchanged
  - Identify all subgraphs of n nodes in the randomized version
  - Subgraphs that occur significantly more frequently in the real network, as compared to the randomized one, are designated to be the motifs

Recall:
Special directed subgraphs

- Bi-fan: 
- Feed-forward loop: two non-intersecting directed paths from a start to an endpoint 
- Bi-parallel: two non-intersecting paths of identical length from a start to an endpoint 
- Feed-back loop: a directed cycle 

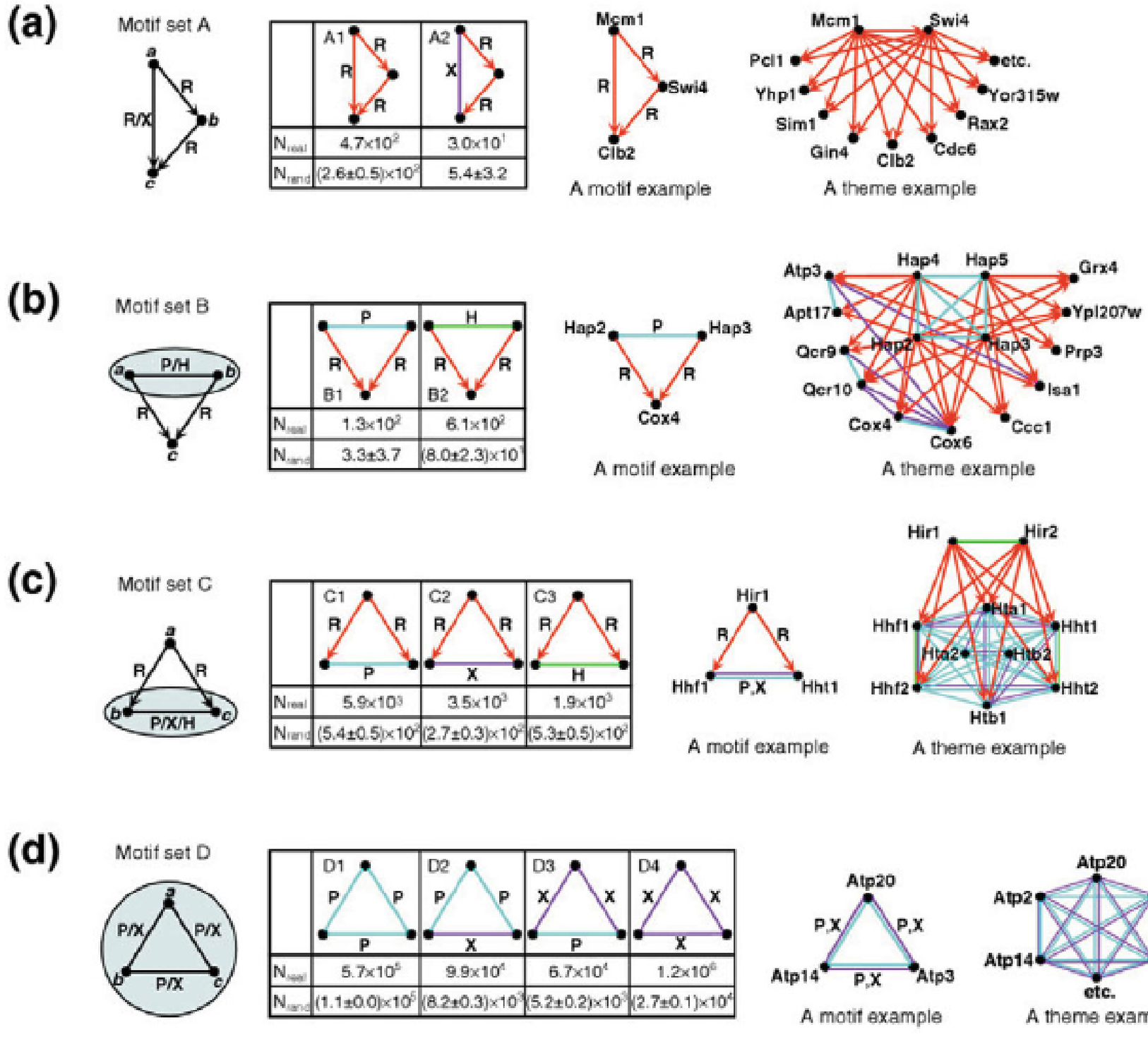
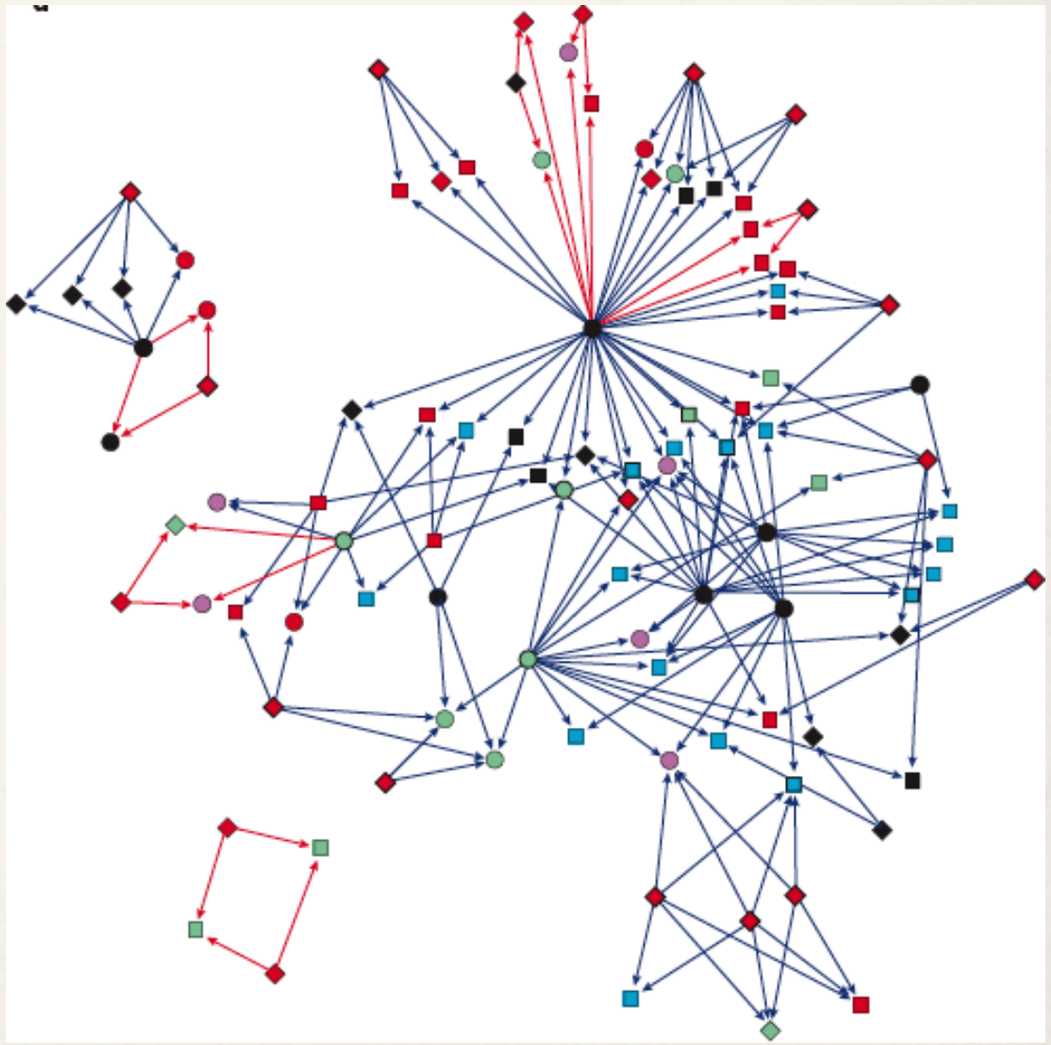


Fig. 10. Network motifs and themes in the integrated *S. cerevisiae* network. Edges denote transcriptional regulation (R), protein interaction (P), sequence homology (H), correlated expression (X) or synthetic lethal interactions (S). (a) Motifs corresponding to the 'feed-forward' theme are based on transcriptional feed-forward loops; (b) motifs in the 'co-pointing' theme consist of interacting transcription factors that regulate the same target gene; (c) motifs corresponding to the 'regulonic complex' theme include co-regulation of members of a protein complex; (d) motifs in the 'protein complex' theme represent interacting and coexpressed protein cliques. For a given motif, N_{real} is the number of corresponding subgraphs in the real network, and N_{rand} is the number of corresponding subgraphs in a randomized network. Figure reproduced with permission from BioMed Central (Zhang et al., 2005).

Motif Clusters

- The motifs in a network are not independent
- The figure shows the 209 bi-fan motifs in the E. coli transcription regulatory network
- 208 of the 209 motifs form two extended motif clusters and only one motif remains isolated
- Motif clusters seem to be a general property of all real networks

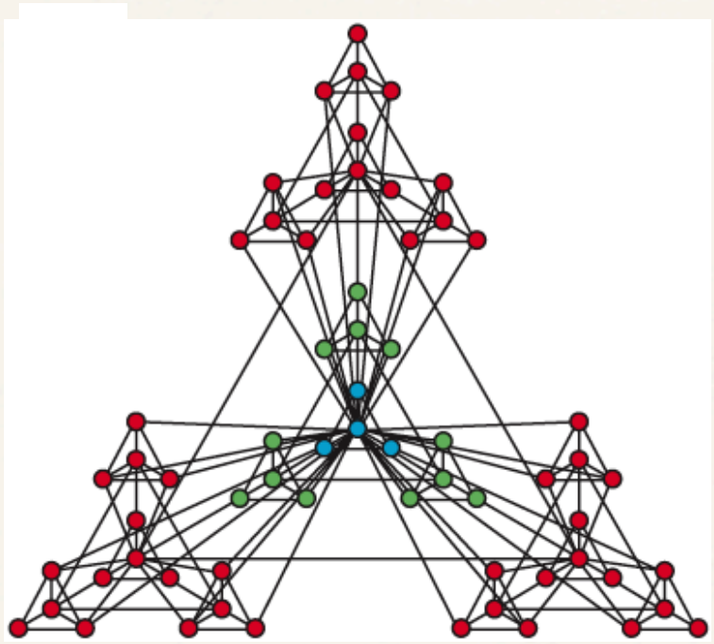


Hierarchical Organization of Topological Modules

- At face value, the scale-free property and modularity seem to be contradictory: the former implies the existence of nodes that are connected to a high fraction of nodes which makes the existence of relatively isolated modules unlikely, and the latter implies the existence of groups of nodes that are relatively isolated from the rest of the system

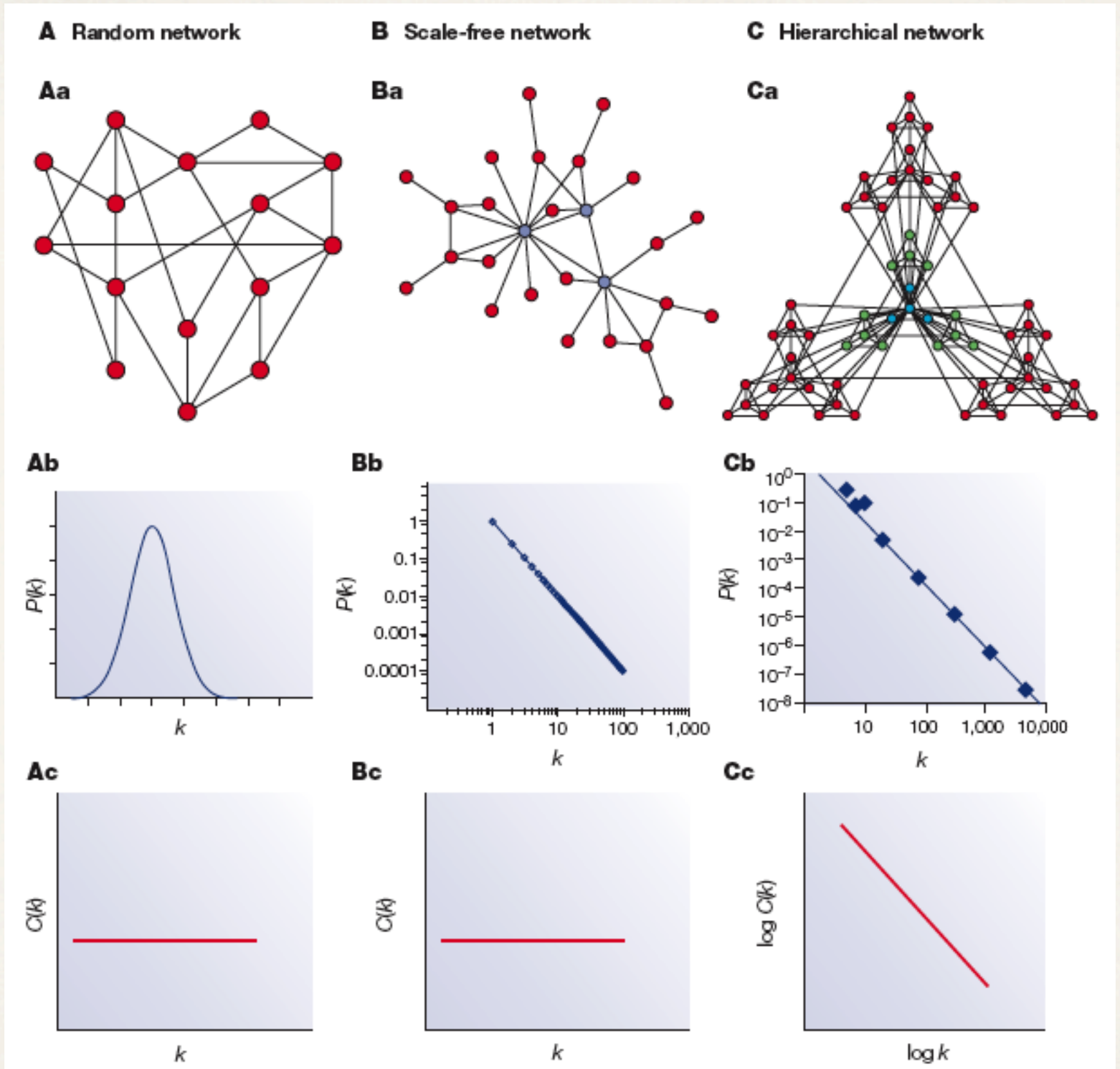
- However, clustering and hubs naturally coexist, which indicates that topological modules are not independent, but combine to form a hierarchical network

Hierarchical Networks



1. a cluster of four densely connected nodes is first created (blue)
2. three replicas of this module are generated and the three external nodes of the replicated clusters are connected to the central node of the old cluster (green)
3. three replicas of the 16-node module are generated and the 16 peripheral nodes are connected to the central node of the old module (red)

The model combines scale-free and modularity properties


A Random network
B Scale-free network
C Hierarchical network
Aa
Ba
Ca
Ab
Bb
Cb
P(k)
k
0.1
0.01
0.001
0.0001
1
10
100
1,000
10,000
Ac
Bc
Cc
C(k)
log C(k)
log k

Identifying Modules

- Various clustering techniques have been developed or adapted to identifying modules in networks

- Different methods return different decompositions of the networks

- At present there are no objective mathematical criteria for deciding that one decomposition is better than another

Scale-free or Geometric?

- Pržulj et al. studied the fit of four different network models to PPI networks of Saccharomyces cerevisiae (yeast) and Drosophila melanogaster (fruitfly)

- Findings:

  - The scale-free model fails to fit the data, and a random geometric model provides a much more accurate model

Geometric Random Graphs

- A geometric graph $G(V,r)$ with radius r is a graph with node set V of points in a metric space and edge set $E=\{(u,v): u,v\in V, 0\leq d(u,v)\leq r\}$, where $d(.,.)$ is an arbitrary distance norm in this space.

- In other words, imagine a set of points in a metric space, with an edge between two points if the distance between them is at most r

- Usually, two-dimensional space is considered, containing points in the unit square $[0,1]^2$ or unit disc, and $0<r<1$

- Typical distance norms between two points (x_1,y_1) and (x_2,y_2): L_1 norm $[|x_1-x_2|+|y_1-y_2|]$, L_2 norm $[((x_1-x_2)^2+(y_1-y_2)^2)^{1/2}]$, L_∞ norm $[\max(|x_1-x_2|,|y_1-y_2|)]$

- A random geometric graph $G(n,r)$ is a geometric graph with n nodes which correspond to n independently and uniformly distributed points in a metric space

Graphlet Analysis of PPI Networks

- Pržulj et al. considered graphlets (connected network with a small number of nodes), and used an approach similar to that of identifying motifs to assess the fit

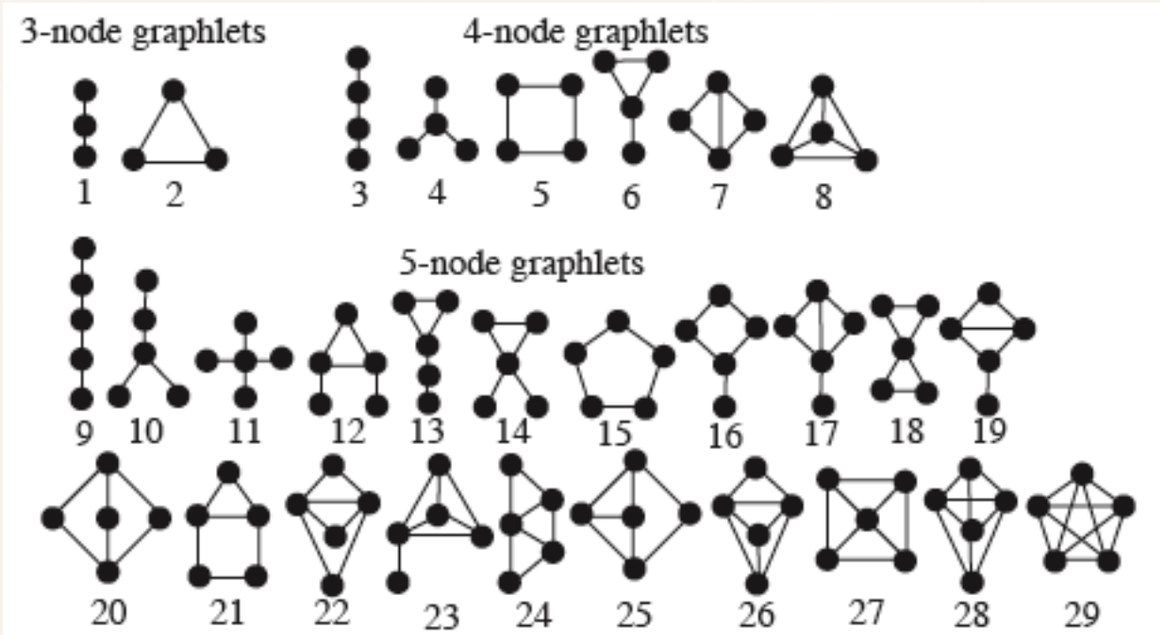


Fig. 1. All 3-node, 4-node and 5-node connected networks (graphlets), ordered within groups from the least to the most dense with respect to the number of edges when compared to the maximum possible number of edges in the graphlet; they are numbered from 1 to 29.

Graphlet Analysis of PPI Networks

- Compared the frequency of the appearance of these graphlets in PPI networks with the frequency of their appearance in four different types of random networks:
  - ER: Erdös-Rényi randm networks with same number of nodes and edges as the corresponding PPI networks
  - ER-DD: Erdös-Rényi randm networks with same number of nodes, edges, and degree distribution as the corresponding PPI networks
  - SF: Scale-free random networks with the same number of nodes, and the number of edges within 1% of those of the corresponding PPI networks
  - GEO: several types of geometric random graphs with the same number of nodes, and the number of edges within 1% of those of the corresponding PPI networks (three versions: GEO-2D, GEO-3D, and GEO-4D, with Euclidean distance)

Graphlet Analysis of PPI Networks

- They analyzed four PPI networks of the yeast C. cerevisiae and fruitfly D. melanogaster

- They quantified the fit using the measure $D(G,H) = \sum_{i=1}^{29} |F_i(G) - F_i(H)|$ where

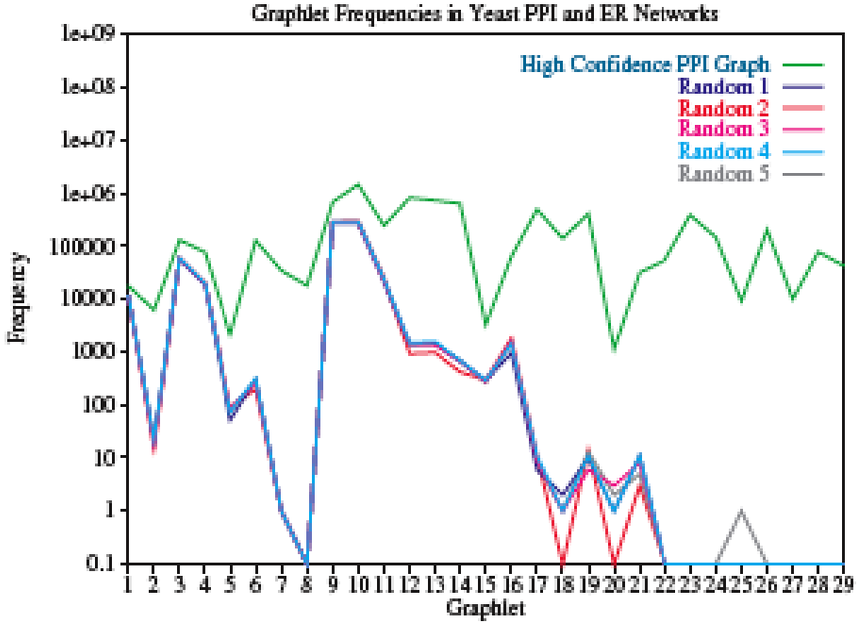
$F_i(G) = -\log(N_i(G)/T(G))$

$N_i(G)$ is the number of graphlets of type i

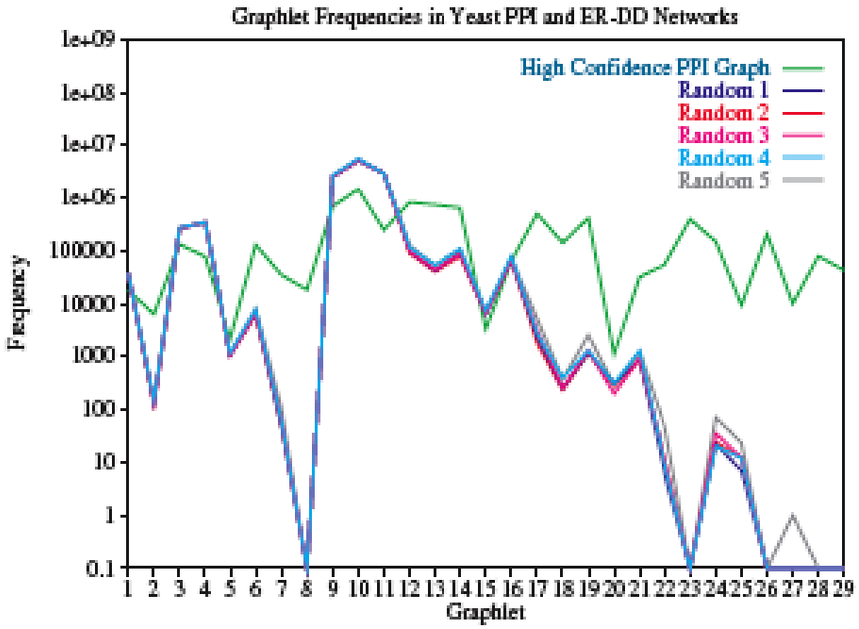
$T(G) = \sum_{i=1}^{29} N_i(G)$

(G and H are the PPI network and its randomized counterpart)

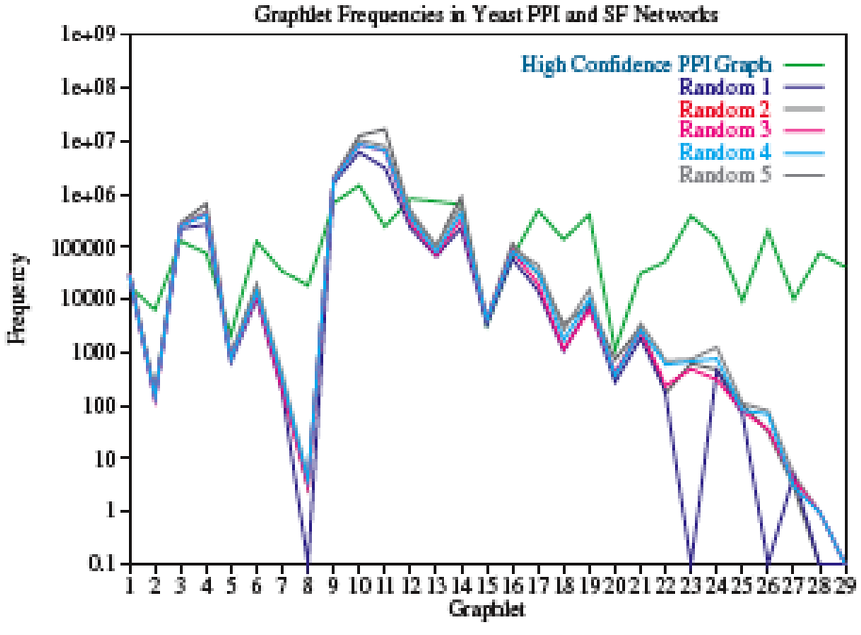
(A)


Graphlet Frequencies in Yeast PPI and ER Networks
High Confidence PPI Graph
Random 1
Random 2
Random 3
Random 4
Random 5
Frequency
Graphlet

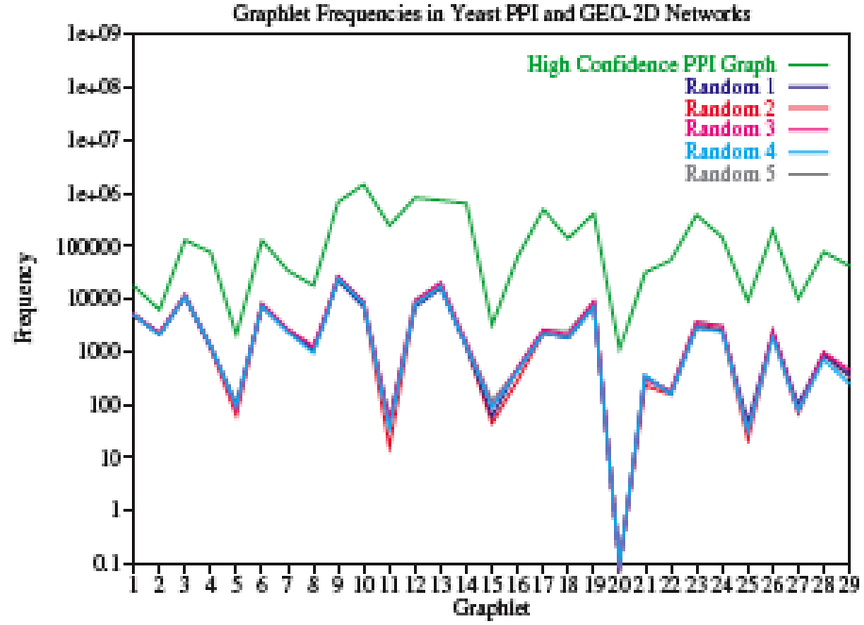
(B)


Graphlet Frequencies in Yeast PPI and ER-DD Networks
High Confidence PPI Graph
Random 1
Random 2
Random 3
Random 4
Random 5
Frequency
Graphlet

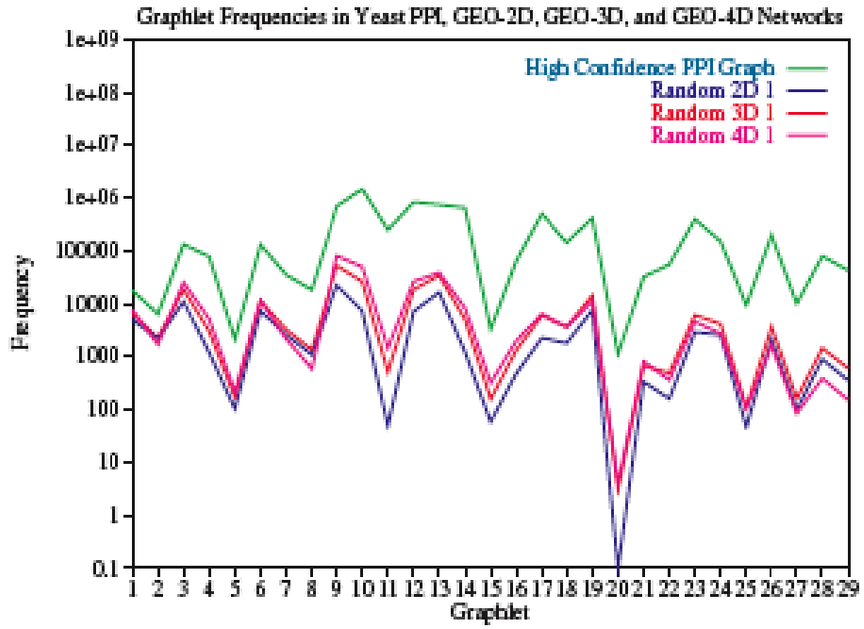
(C)


Graphlet Frequencies in Yeast PPI and SF Networks
High Confidence PPI Graph
Random 1
Random 2
Random 3
Random 4
Random 5
Frequency
Graphlet

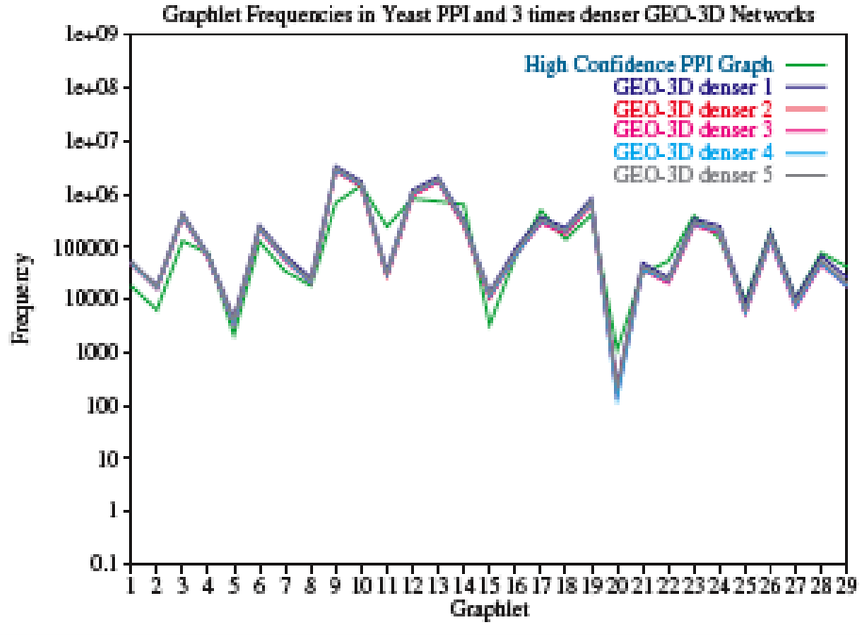
(D)


Graphlet Frequencies in Yeast PPI and GEO-2D Networks
High Confidence PPI Graph
Random 1
Random 2
Random 3
Random 4
Random 5
Frequency
Graphlet

(E)


Graphlet Frequencies in Yeast PPI, GEO-2D, GEO-3D, and GEO-4D Networks
High Confidence PPI Graph
Random 2D 1
Random 3D 1
Random 4D 1
Frequency
Graphlet

(F)


Graphlet Frequencies in Yeast PPI and 3 times denser GEO-3D Networks
High Confidence PPI Graph
GEO-3D denser 1
GEO-3D denser 2
GEO-3D denser 3
GEO-3D denser 4
GEO-3D denser 5
Frequency
Graphlet

Acknowledgments

- Materials in this lecture are mostly based on:
  - "Network Biology: Understanding the Cell's Functional Organization", by Barabáśi and Oltvai.
  - "Scale-free Networks in Cell Biology", by R. Albert.
  - "Modeling interactome: Scale-free or Geometric?", by Pržulj, Corneil, and Jurisica