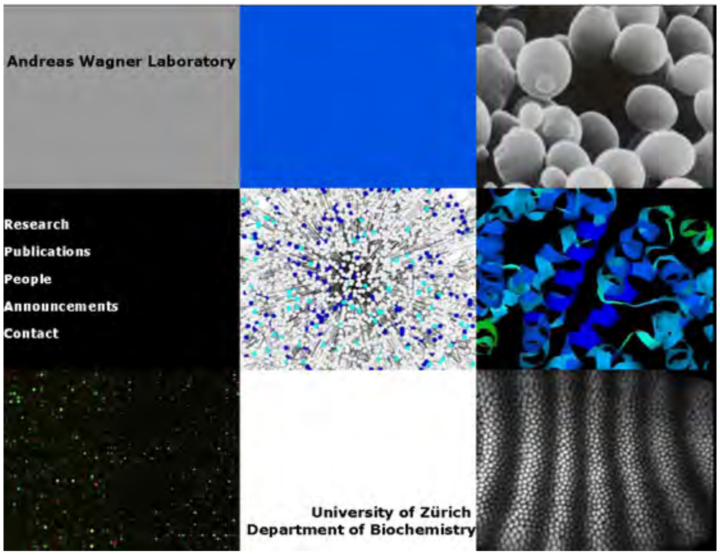


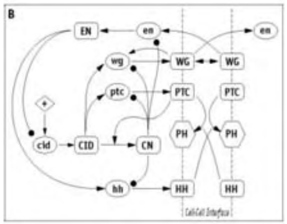
http://www.biochem.unizh.ch/wagner/
aw@bioc.unizh.ch

Genome-scale biological networks

Two kinds of biological networks

1. Small networks dedicated to a specific task
(up to dozens of gene products)

Chemotaxis
Cell-cycle regulation
Fruit fly segmentation
Flower development
...



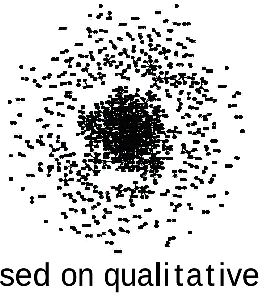
Von Dassow et al. 2000. *Nature* 406: 188-92

Mathematical characterization based on detailed, quantitative biochemical information

Two kinds of biological networks

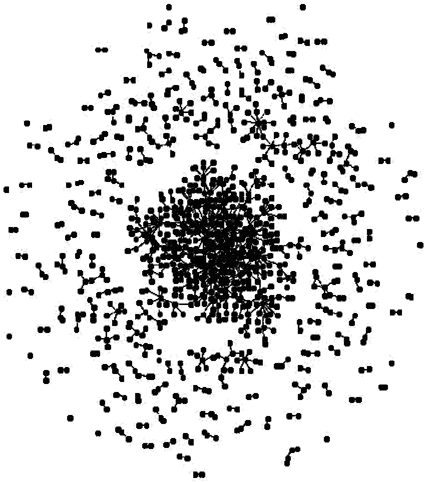
2. Genome-scale networks
(hundreds to thousands of genes products)

Protein interaction networks
Transcriptional regulation networks
Metabolic networks

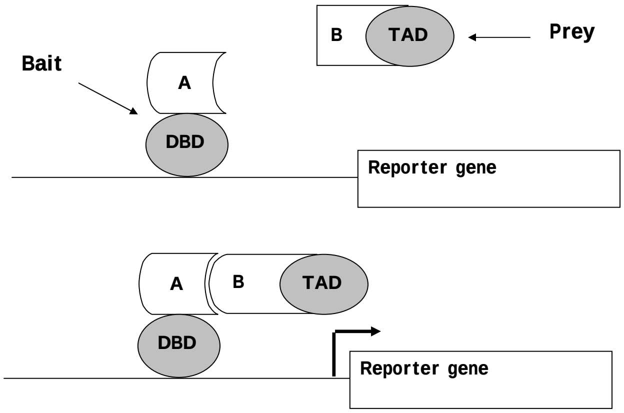


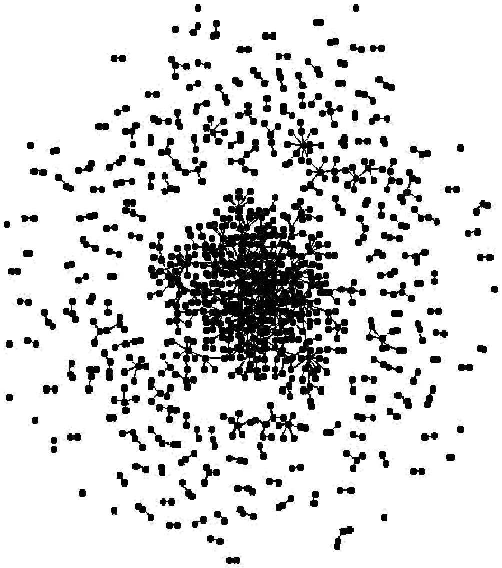
Mathematical characterization based on qualitative understanding of network topology

Protein interaction networks



The yeast two-hybrid assay can detect interactions between pairs of proteins





Some Limitations

Noisy data

- limited replicability

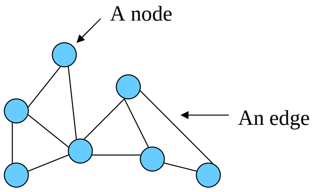
Interaction artefacts caused by chimaeric proteins

Membrane proteins

Interactions are functionally VERY heterogeneous

- structural, signaling, enzymatic ...

Graphs



A graph G=(V,E) comprises

a set V of nodes (vertices)

a set E of edges

$$V = \{V_1, \ldots, V_n\}$$

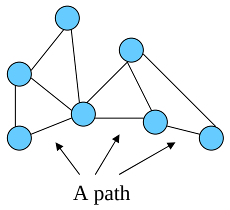
$$E = \{(V_i, V_j), \ldots, (V_k, V_l)\}$$

Protein interaction networks are undirected graphs
(Individual node pairs in *E* are unordered.)

Graphs are everywhere

Graph	*Nodes*	*Edges*
Computer networks	Computers	Data transmission lines
Friendship networks	People	Being acquainted
The world wide web	Web pages	Hyperlinks
Actor collaboration graph	Actors	Having acted in the same movie
Power grids	Transformers	Power lines
Citation network	Publication	Citation
Nematode CNS	Nerve cells	Axons

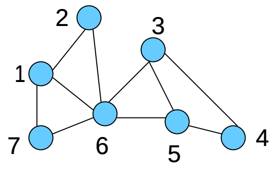
Graphs



A path is a sequence of alternating nodes and edges in which no node is visited more than once

A geodesic is the shortest path between two nodes.

Graphs can be represented by matrices

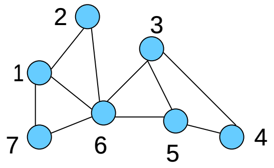


Adjacency matrix $A=(a_{ij})$

$a_{ij}=1$ $(V_i, V_j) \in E$

$a_{ij}=0$ otherwise

$$A=\begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$



The degree (connectivity) k_i of a node V_i is the number of edges incident with the node (e.g., k_1=3, k_6=5).

$$k_i = \sum_j a_{ij}$$

Graphs can be characterized according to their degree distribution P(k), the fraction of nodes having degree k.

The degrees of nodes in a graph may be correlated

P(*k'*|*k*): conditional probability that a node with degree *k* is connected to a node with degree *k'*. In a graph with degree correlations, *P(k'|k) ≠P(k')*

Average nearest neighbor degree of a node

$$k_{nn,i} = \frac{1}{k_i} \sum_{\text{j, nearest neighbors of i}} k_j = \frac{1}{k_i} \sum_{j=1}^{N} a_{ij} k_j$$

Average nearest neighbor degree of all nodes with degree *k*

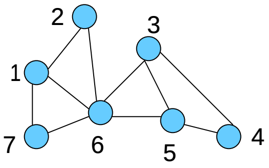
$$k_{nn}(k) = \sum_{k'} k' P(k'|k)$$

A graph is assortative if $k_{nn}(k)$ increases with k

nodes connect to nodes of similar connectivity

A graph is disassortive if $k_{nn}(k)$ decreases with k

Graphs can be represented by matrices



Matrix of shortest paths D=(*d*$_{ij}$)

$$D = \begin{pmatrix} 0 & 1 & 2 & 3 & 2 & 1 & 1 \\ 1 & 0 & 2 & 3 & 2 & 1 & 2 \\ 2 & 2 & 0 & 1 & 1 & 1 & 2 \\ 3 & 3 & 1 & 0 & 1 & 2 & 3 \\ 2 & 2 & 1 & 1 & 0 & 1 & 2 \\ 1 & 1 & 1 & 2 & 1 & 0 & 1 \\ 1 & 2 & 2 & 3 & 2 & 1 & 0 \end{pmatrix}$$

Connected graph: $d_{ij} < \infty$ **for all** i,j

Path length and diameter are measures of graph compactness

Diameter of a graph: $\max_{i,j} d_{ij}$

Mean (arithmetic) shortest path length
or characteristic path length

$$L = \frac{1}{N(N-1)} \sum_{i,j,i \neq j} d_{ij}$$

Mean (harmonic) shortest path length
or "efficiency" of a graph

$$L = \frac{1}{N(N-1)} \sum_{i,j,i \neq j} \frac{1}{d_{ij}}$$

(Better suited than characteristic path length for disconnected graphs)

A measure of node and edge centrality

Node betweenness or node load:
number of geodesics passing through a node

$$b_i = \sum_{j,k,\, j\neq k} \frac{n_{jk}(i)}{n_{jk}}$$

$n_{jk}(i)$ number of geodesics connecting *j* and *k* and passing through *i*
n_{jk} number of geodesics connecting *j* and *k*

Graph spectra

The spectrum of a graph is the set of eigenvalues of the adjacency matrix A.
It is intimately related to key graph properties

Examples:

1. An undirected graph is connected iff the largest eigenvalue μ_{max} of *A* has multiplicity one. Also, in a connected graph

$$k_{\min} < \langle k \rangle < \mu_{\max} < k_{\max}$$

2. Diam(*G*) is smaller than the number of distinct eigenvalues of *A*

3. For the Graph Laplacian *L*=*D*-*A*, where *D* is the diagonal matrix *D*=(d_{ii})=k_i, the multiplicity of the eigenvalue zero equals the number of components (maximal connected subgraphs)of the graph

High clustering and transitivity indicate locally dense neighborhoods in a graph

(How likely is it that my neighbors are also each other’s neighbors?)

Transitivity

fraction of connected node triplets that are triangles

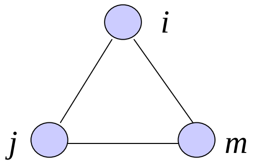
T= 3× (# of triangles) / (# of connected node triplets in G)

Clustering coefficient c_i of a node i

The fraction of a node’s neighbors that are neighbors of each other

$$c_i = \frac{\sum_{j,m} a_{ij} a_{jm} a_{mi}}{k_i (k_i - 1)}$$

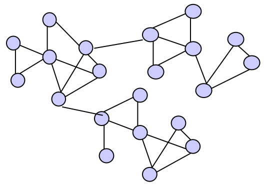
$$C = \frac{1}{k} \sum_i c_i$$



Modules, cohesive subgroups, or “communities”

subgraphs whose nodes are tightly connected or “cohesive”

Many measures of modularity are in use



1. Clique: a largest complete (=fully connected) subgraph

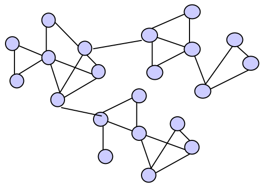
2. n-clique: a largest subgraph in which all geodesics have length ≤n
(A 1-clique is a clique)

3. k-plex: a largest subgraph of n nodes in which the degree k_i≥n-k for all n nodes
(a 1-plex is a clique)

Modules, cohesive subgroups, or "communities"

subgraphs whose nodes are tightly connected or "cohesive"

Many measures of modularity are in use



4. Modularity Q

$$Q = \sum_i (e_{ii} - a_i^2)$$

E=(e$_{\mathrm{ij}}$)
e_{ij}…fraction of edges that link vertices in module i to vertices in module j

$a_i = \sum_j e_{ij}$ fraction of edges connecting to nodes in module i

$e_{ij}=a_i a_j$ (thus $e_{\mathrm{ii}}=a_i^2$) if the probability that two nodes are connected is independent of their belonging to the same community

Q indicates the degree of correlation between edges joining two nodes and the nodes being in the same community.
Q ≈ 1 indicates strong community structure.

Some methods to identify modules

1. **Spectral graph partitioning**

Graph partitioning: Find a division of the vertex set V into two subsets with a minimum number of edges between subsets and a maximum number within each subset. (NP-complete)

Spectral bisection: Let λ_2 be the second-largest eigenvalue of the graph Laplacian (L=D-A) and v_2 the corresponding eigenvector. If λ_2 is close to zero, then the positive entries of v_2 correspond to vertices in one partition in one component, and the negative entries to the other.

To identify multiple communities, apply **repeated bisection.**

Limitations:

Difficult if modules are not well-defined
Repeated bisection is not guaranteed to give the best partition.
When to stop partitioning?

Some methods to identify modules

2. Girvan-Newman algorithm (Iterative Divisive Clustering)

Idea: Edges between modules would be those with the highest edge betweenness
Remove those edges and you get good module separation

Iterative procedure

1. Remove the edge with the highest betweenness score
2. Recalculate edge betweenness for the now-reduced graph
3. Determine modularity Q
3. Back to one until all nodes are isolated

The optimal partition is that with the highest Q

The best-studied mathematical models of graphs

k-regular graphs

N nodes, K=kN edges
every node has degree k

Erdős-Rényi random graphs

N nodes, K edges

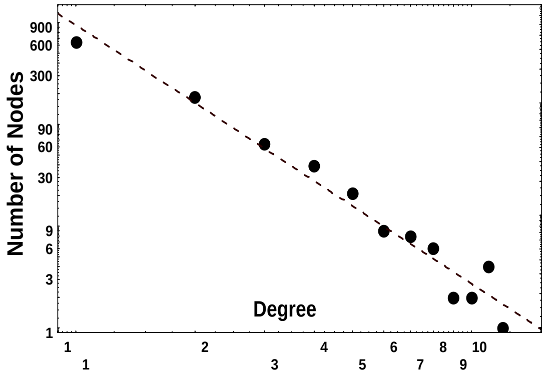
edges connect pairs of randomly chosen nodes (avoiding multiple edges)

Degree distribution is Poisson

$$P(k) = \exp(-\langle \bar{k} \rangle)\frac{\langle \bar{k} \rangle^{k}}{k!}$$

Biological networks are vastly more complex and heterogeneous than these models

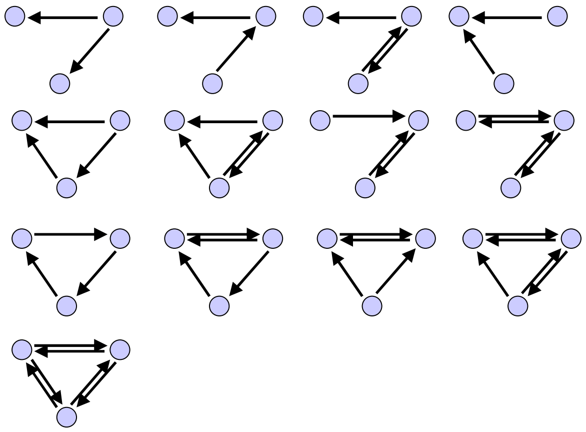
Protein interaction networks (and many other networks) have broad-tailed degree distributions.



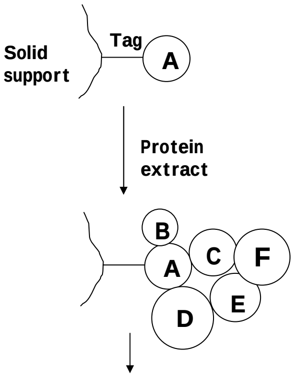
Wagner A, Proc. Roy. Soc. London 2003

A motif is a local pattern of connections in a graph

All possible 3-node motifs in a digraph



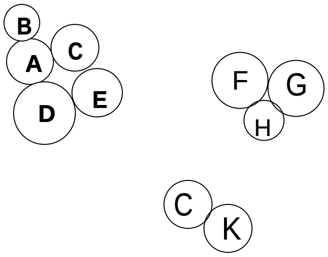
Affinity chromatography can identify protein complexes



Elution

Electrophoresis

Mass spectrometry



Different protein complexes may share proteins
(e.g., protein C)

Proteins within a complex need not interact directly
(B and E)

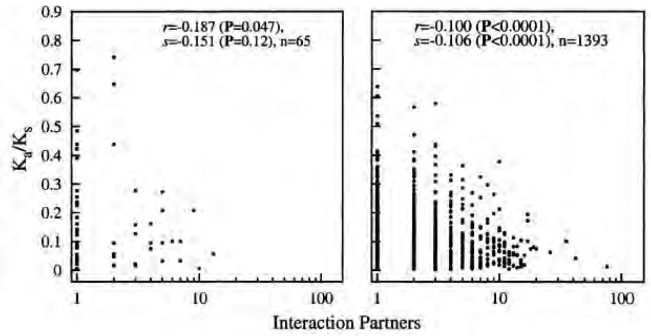
Hypergraphs are suited to study large assemblages of protein complexes

A hypergraph G=(V,E) comprises

a set V of nodes (vertices)
a set E of <u>hyperedges</u>

$$V = \{V_1, \ldots, V_n\}$$
$$E = \{(V_i, ..., V_j), \ldots, (V_k, ..., V_l)\}$$

Highly connected proteins tolerate fewer amino acid substitutions in their evolution



Hahn et al. Journal of Molecular Evolution 2004

Protein microarrays help take the heterogeneity out of molecular interactions

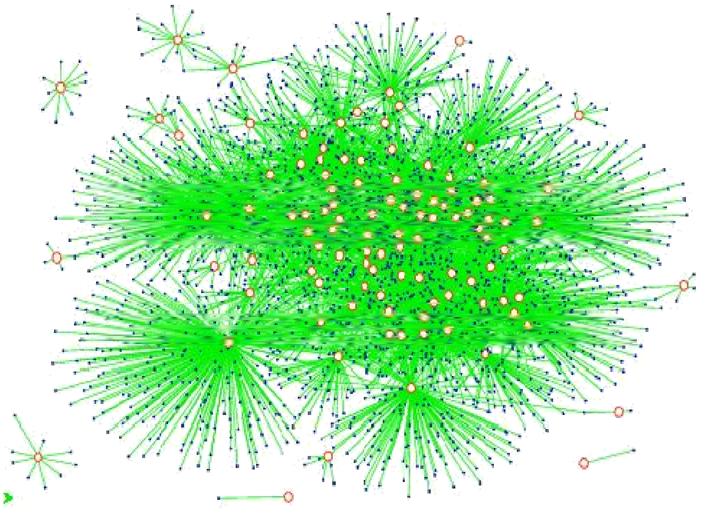
Collection of proteins is immobilized on a microarray

Array is exposed to ligand that is (directly or indirectly) labeled

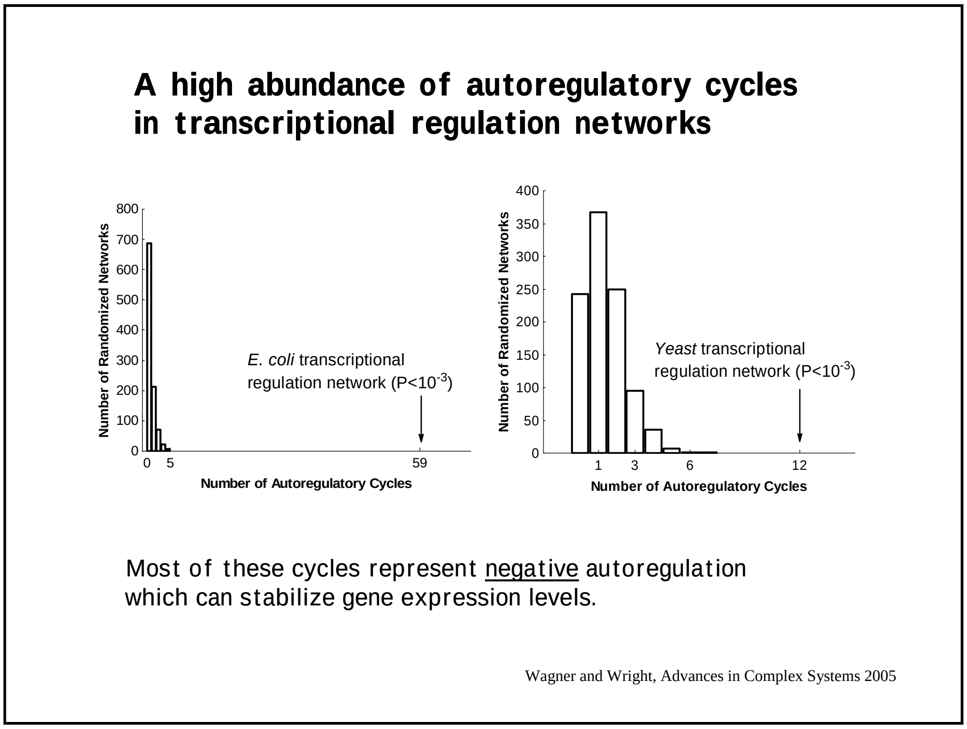
Calmodulin
Phosphoinositides
Protein kinase + phosphate

Proteins bound to ligand are detected through (flourescent or radioactive) signal

Transcriptional regulation networks



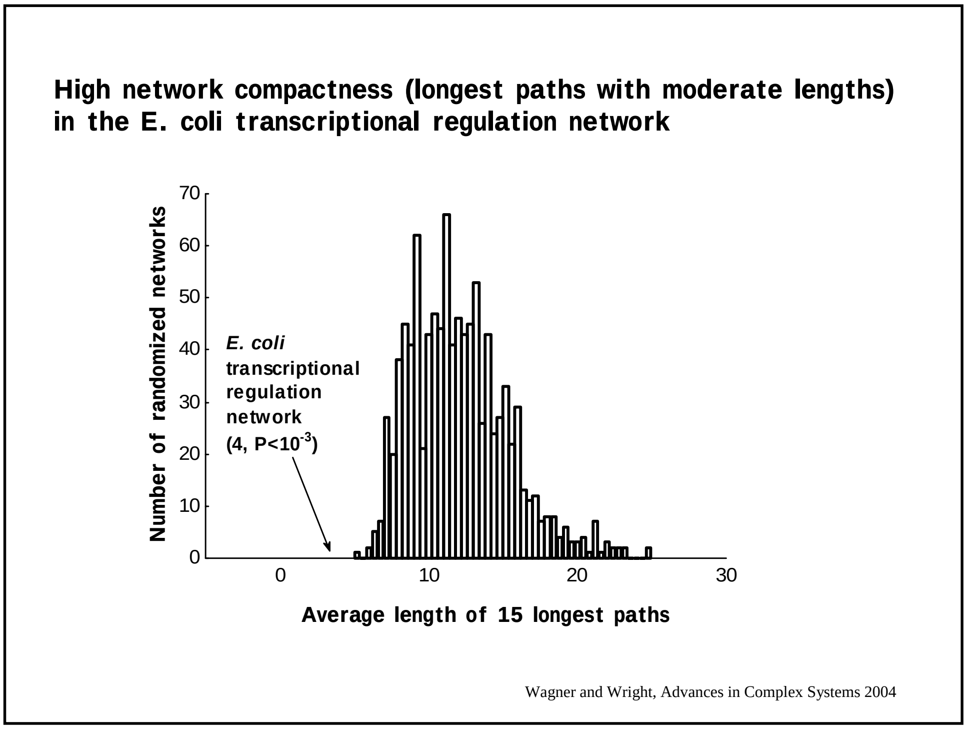
Evangelisti and Wagner 2004



A high abundance of autoregulatory cycles in transcriptional regulation networks

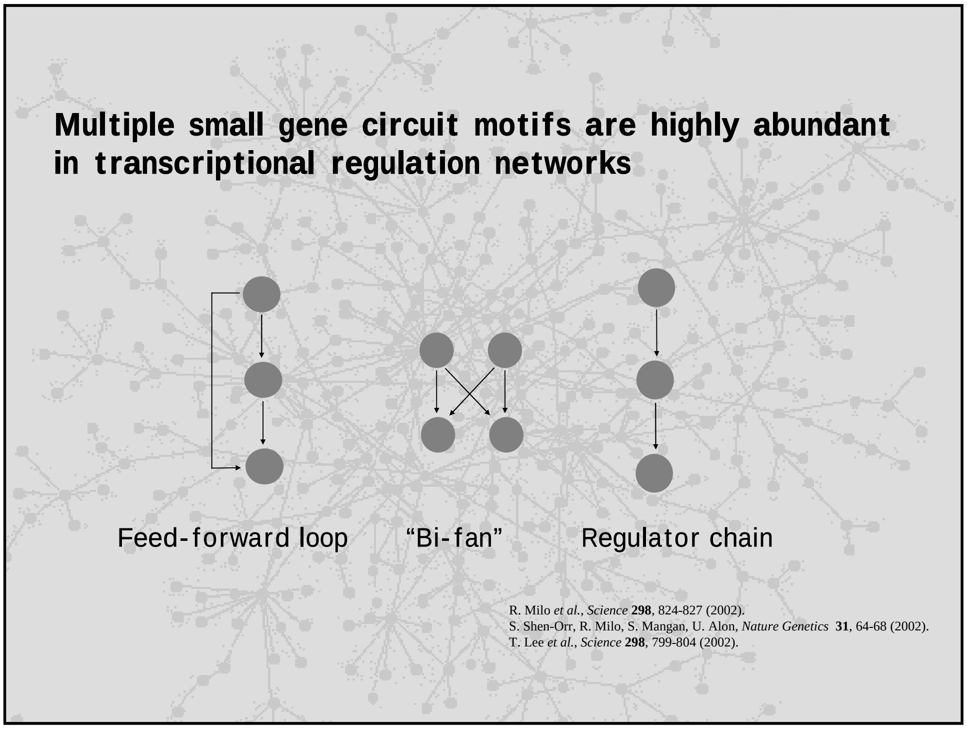
Most of these cycles represent negative autoregulation which can stabilize gene expression levels.

Wagner and Wright, Advances in Complex Systems 2005



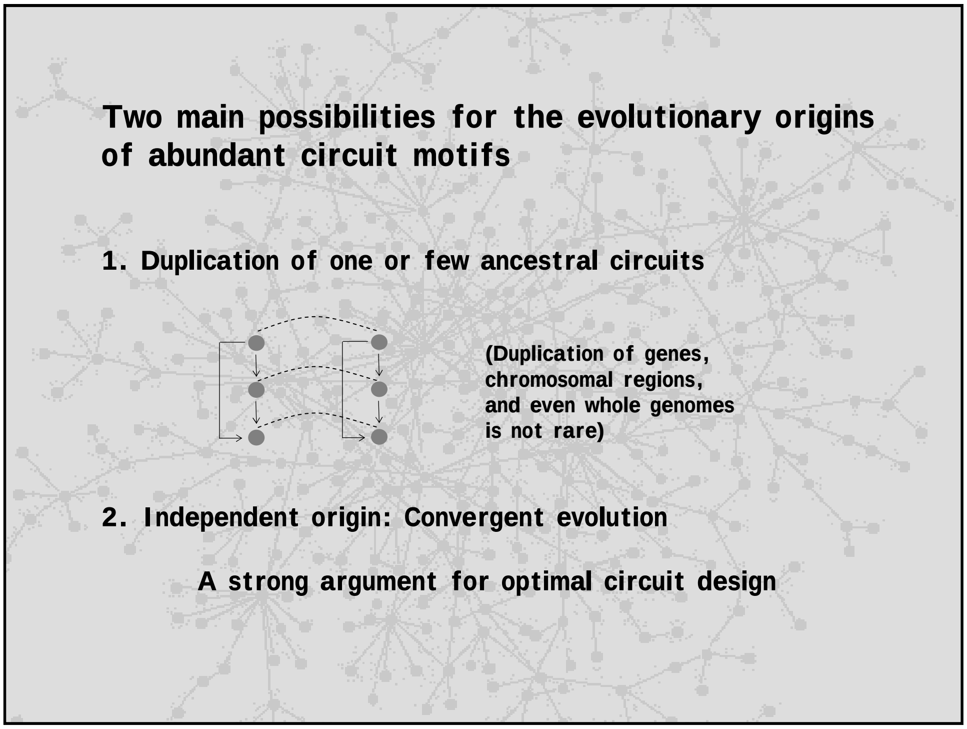
High network compactness (longest paths with moderate lengths) in the E. coli transcriptional regulation network

Wagner and Wright, Advances in Complex Systems 2004



Multiple small gene circuit motifs are highly abundant in transcriptional regulation networks

R. Milo *et al.*, *Science* **298**, 824-827 (2002).
S. Shen-Orr, R. Milo, S. Mangan, U. Alon, *Nature Genetics* **31**, 64-68 (2002).
T. Lee *et al.*, *Science* **298**, 799-804 (2002).



Two main possibilities for the evolutionary origins of abundant circuit motifs

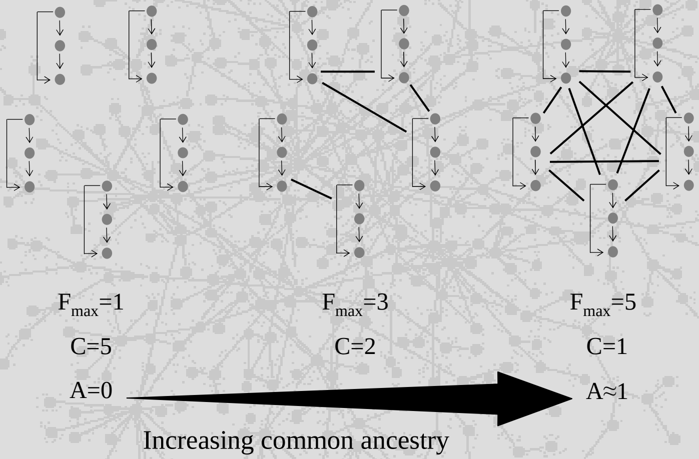
1. Duplication of one or few ancestral circuits

(Duplication of genes, chromosomal regions, and even whole genomes is not rare)

2. Independent origin: Convergent evolution

A strong argument for optimal circuit design

A circuit duplication graph to detect common ancestry among gene circuits



$F_{max}=1$ | $F_{max}=3$ | $F_{max}=5$

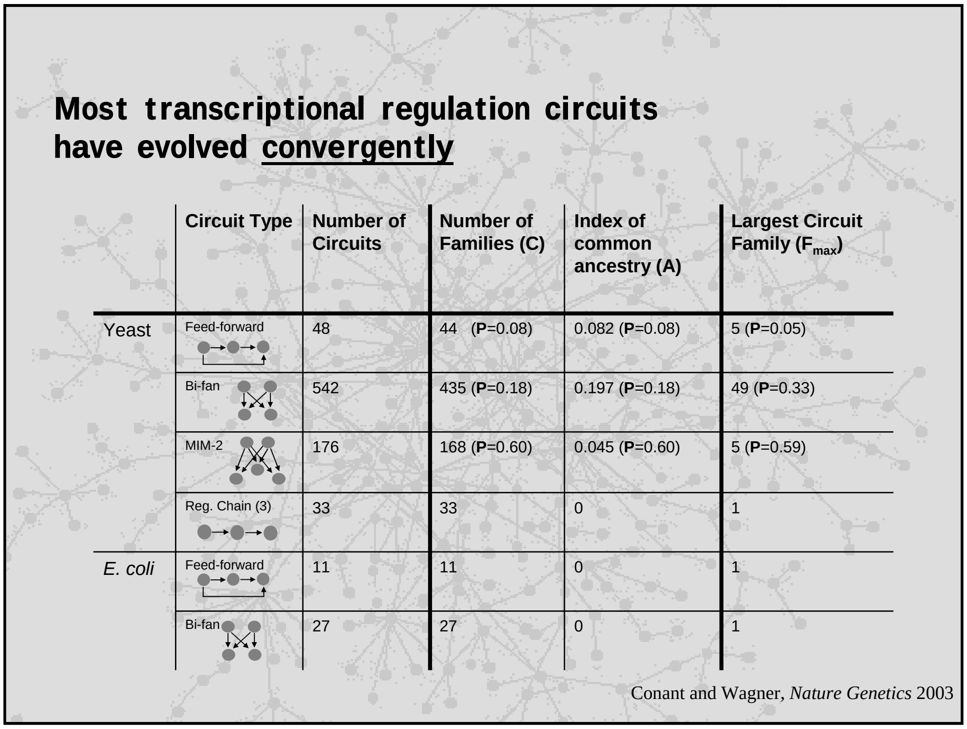
$C=5$ | $C=2$ | $C=1$

$A=0$ | $A\approx1$

Increasing common ancestry

Circuit Type	Number of Circuits	Number of Families (C)	Index of common ancestry (A)	Largest Circuit Family (F_{max})
Bi-fan	542	435 (**P**=0.18)	0.197 (**P**=0.18)	49 (**P**=0.33)

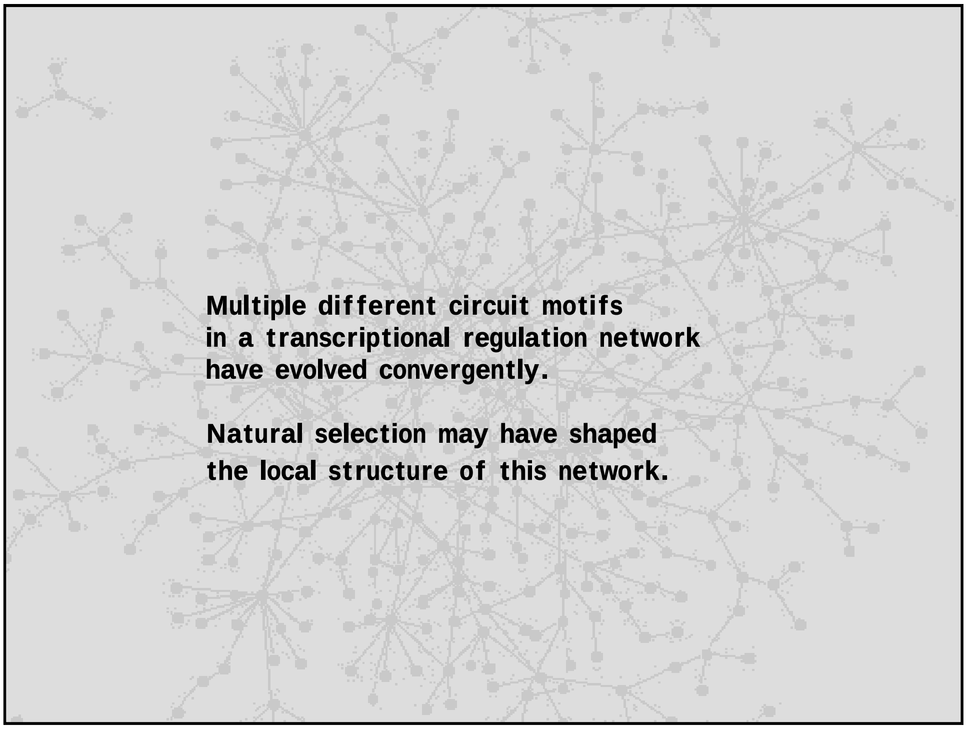
Conant and Wagner, *Nature Genetics* 2003



Most transcriptional regulation circuits have evolved <u>convergently</u>

	Circuit Type	Number of Circuits	Number of Families (C)	Index of common ancestry (A)	Largest Circuit Family (F_{max})
Yeast	Feed-forward	48	44 (**P**=0.08)	0.082 (**P**=0.08)	5 (**P**=0.05)
	Bi-fan	542	435 (**P**=0.18)	0.197 (**P**=0.18)	49 (**P**=0.33)
	MIM-2	176	168 (**P**=0.60)	0.045 (**P**=0.60)	5 (**P**=0.59)
	Reg. Chain (3)	33	33	0	1
E. coli	Feed-forward	11	11	0	1
	Bi-fan	27	27	0	1

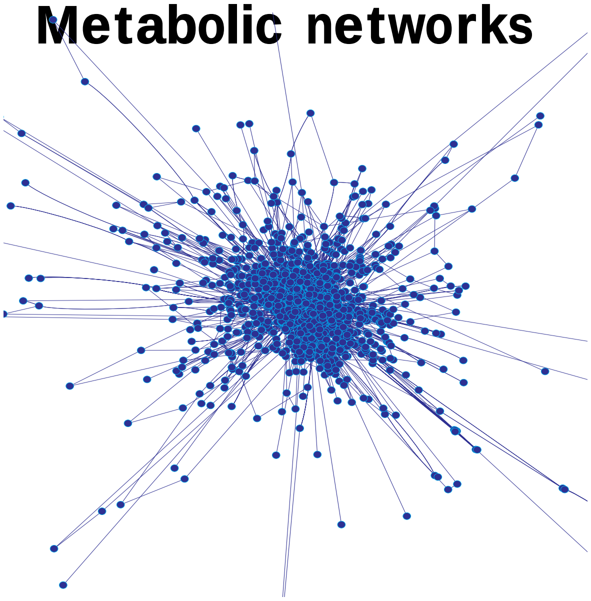
Conant and Wagner, *Nature Genetics* 2003



Multiple different circuit motifs in a transcriptional regulation network have evolved convergently.

Natural selection may have shaped the local structure of this network.

Metabolic networks



A metabolic network is a set of chemical reactions that produces

energy
(for maintenance of cell functions and for biosyntheses)

molecular building blocks for biosyntheses

These reactions are catalyzed by enzymes that are encoded by genes.

In free-living heterotrophic organisms, several hundred such enzymatic reactions are necessary to fulfill these functions.

Graphs can (crudely) represent large chemical reaction networks

Stoichiometric Equations

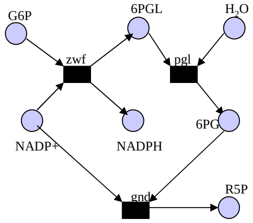
1 Glucose 6-phosphate (G6P) + 1 NADP$^+$ $\overset{zwf}{\Rightarrow}$ 1 6-Phosphoglucono δ-lactone (6PGL) + 1 NADPH

1 6-Phosphoglucono δ-lactone + 1 H_2O $\overset{pgl}{\Rightarrow}$ 1 6-Phosphogluconate (6PG)

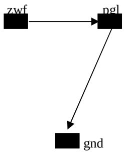
1 6-Phosphogluconate + 1 NADP$^+$ $\overset{gnd}{\Rightarrow}$ 1 Ribulose 5-phosphate (R5P) + 1 NADPH

1 Ribulose 5-phosphate $\overset{rpe}{\Leftrightarrow}$ 1 Xylulose 5-phosphate (X5P)

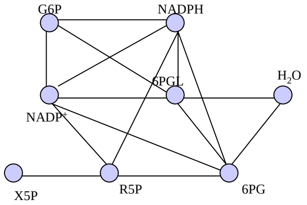
Bipartite graph



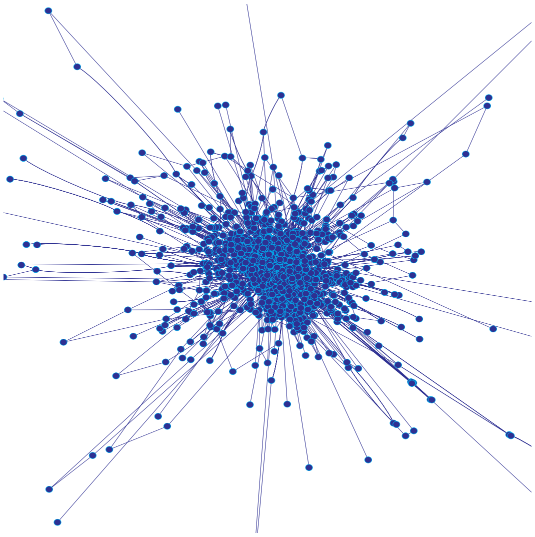
Enzyme graph



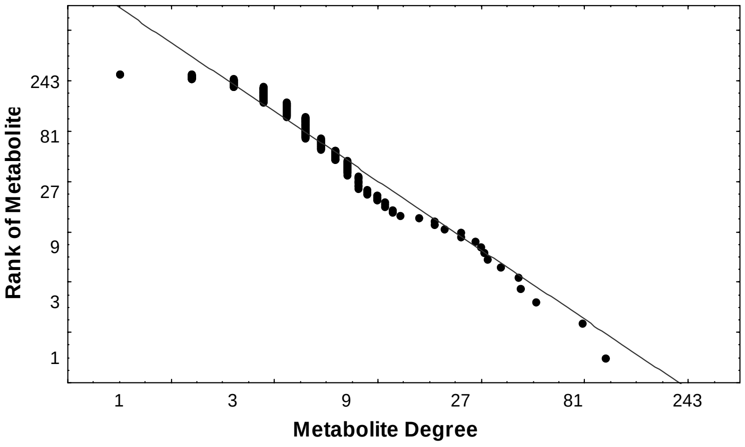
Substrate graph



An enzyme graph representation of the metabolic network of the yeast Saccharomyces cerevisiae



Metabolic networks have a broad-tailed degree distribution



Substrate network of E. coli

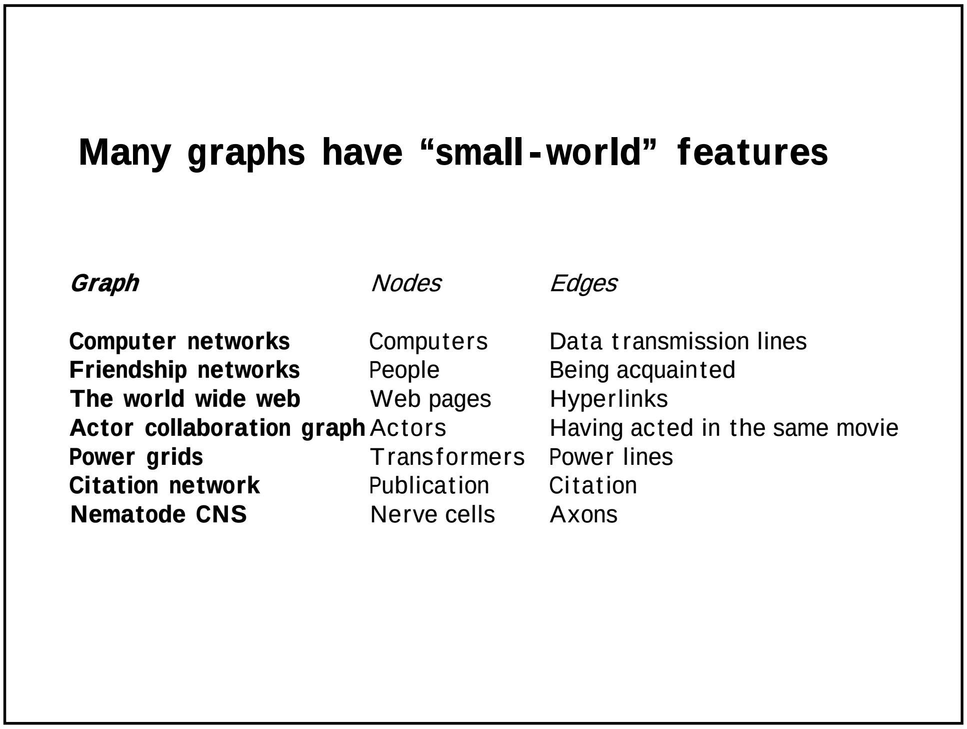
Wagner and Fell, Proc. Roy. Soc. London B 2001

The *E. coli* core metabolism is a small-world network

It is sparse

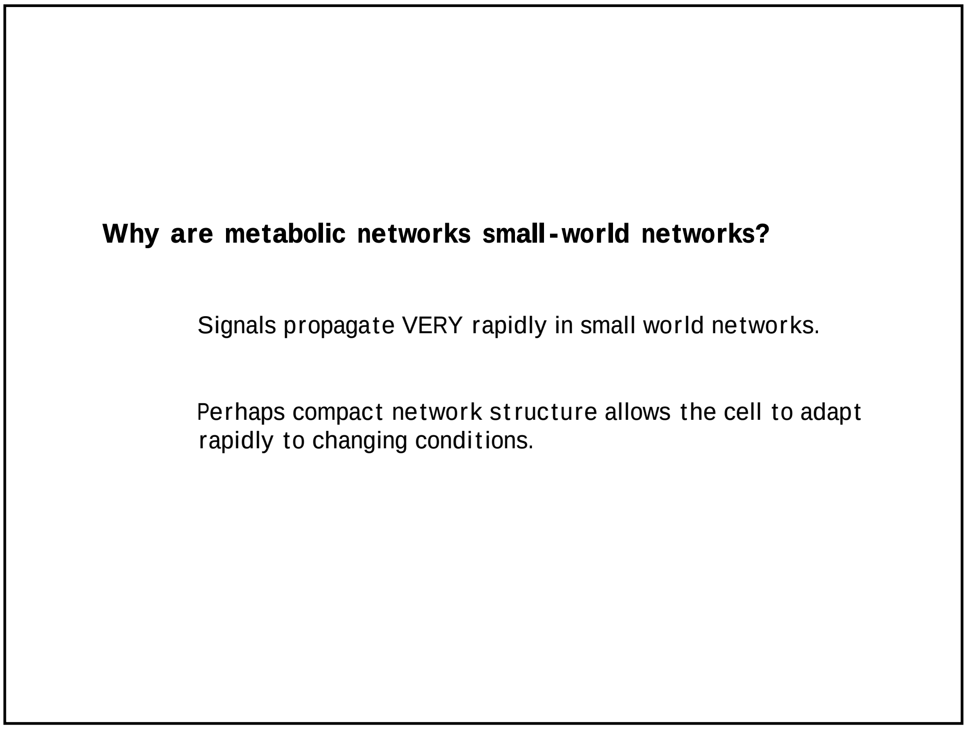
It is highly clustered

It has short characteristic path length



Many graphs have “small-world” features

Graph	*Nodes*	*Edges*
Computer networks	Computers	Data transmission lines
Friendship networks	People	Being acquainted
The world wide web	Web pages	Hyperlinks
Actor collaboration graph	Actors	Having acted in the same movie
Power grids	Transformers	Power lines
Citation network	Publication	Citation
Nematode CNS	Nerve cells	Axons



Why are metabolic networks small-world networks?

Signals propagate VERY rapidly in small world networks.

Perhaps compact network structure allows the cell to adapt rapidly to changing conditions.

Studying only the structure of metabolic networks neglects their function

One needs to analyze the flow (flux) of matter through these networks

For optimal cell growth, metabolic networks need to produce biochemical precursors in well-balanced amounts.

This necessitates a specific distribution of metabolic fluxes through enzymatic reactions in the network.

(Metabolic flux: the rate at which an enzyme converts substrate into product per unit time.)

Flux balance analysis requires a list of chemical reactions known to be catalyzed by enzymes in a given organism.

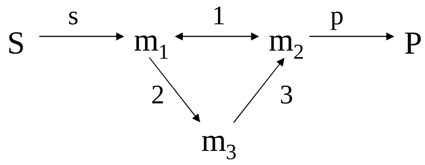
(For example, in yeast
>1100 reactions,
>500 metabolites,
>100 nutrients or waste products.)

Flux balance analysis has two tasks

Identify allowable metabolic fluxes through a metabolic network (fluxes that do not violate the law of mass conservation)

Within the set of allowable fluxes, identify fluxes that are associated with desirable properties (e.g., maximal rate of biomass production, maximal biomass yield per unit of carbon source.)

A simple chemical reaction network



Metabolite concentrations change according to the equations

$$\frac{dm_1}{dt} = v_s - v_1 - v_2$$

$$\frac{dm_2}{dt} = v_1 + v_3 - v_p$$

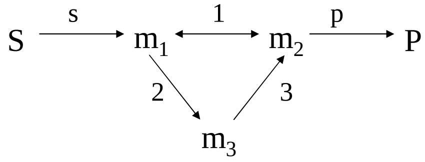
$$\frac{dm_3}{dt} = v_2 - v$$

$$\frac{d\vec{m}}{dt} = \mathbf{S}\vec{v}$$

$$\mathbf{S} = \begin{pmatrix} 1 & -1 & -1 & 0 & 0 \\ 0 & 1 & 0 & 1 & -1 \\ 0 & 0 & 1 & -1 & 0 \end{pmatrix}$$

Stoichiometry matrix

A simple chemical reaction network



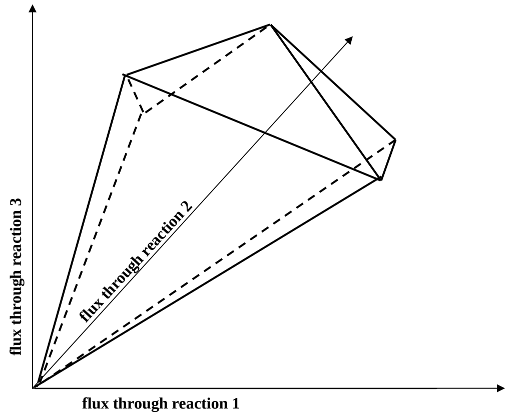
In steady state

$$\frac{d\vec{m}}{dt} = 0$$

$$\mathbf{S}\vec{v} = 0$$

The solutions of these equations form the <u>null space of S</u>

The null space of a metabolic network forms a high-dimensional “flux cone” (a convex polytope)



Several important properties of a metabolic network can be expressed as weighted sums of fluxes

$$Z(\vec{v}) = \sum_{i=1}^{m} c_i v_i$$

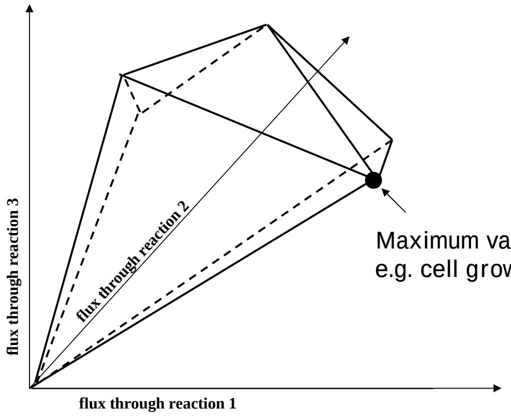
Example:

In the biomass growth flux ,

v_i is the rate at which essential biochemical precursor *i* is produced by a metabolic network.

c_i is a constant that reflects the relative contribution of precursor *i* to biomass (can be estimated from the biomass composition of a cell.)

Linear programming can be used to determine regions within the flux cone where some linear function Z of the fluxes will be maximized.



Maximum value of Z,
e.g. cell growth flux

Example questions for flux balance analysis

FBA shows what is possible for a metabolism.
Is this metabolic potential realized in an organism?

Often not.

Can an organism evolve towards its full metabolic potential?

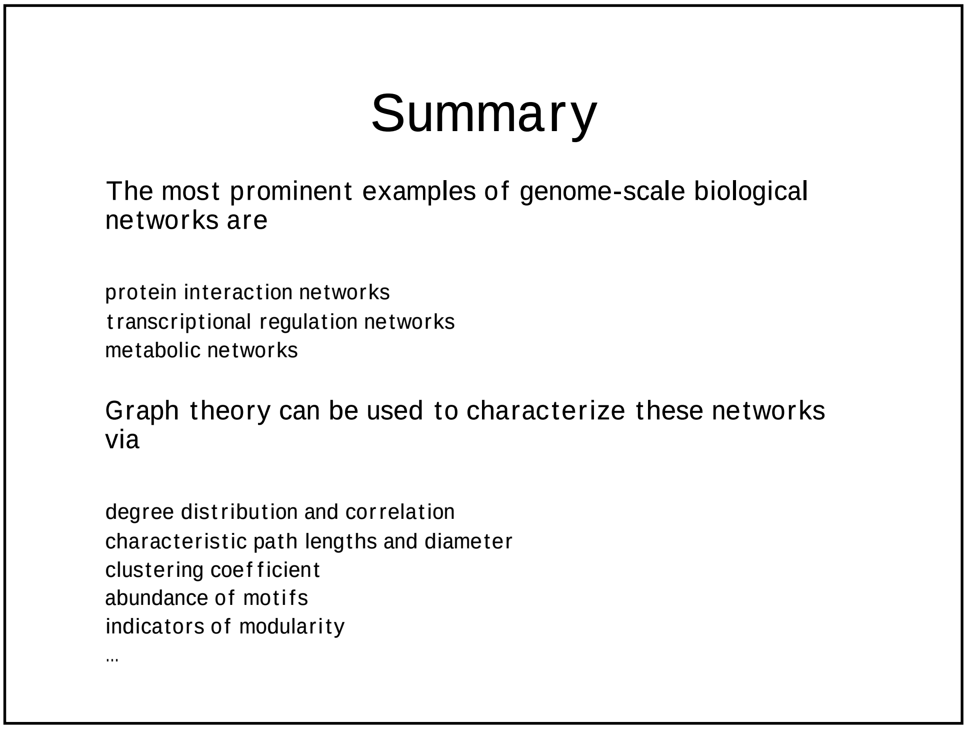
Yes, and quickly

Many enzymatic reactions (and thus the genes encoding them) are dispensable in any one environment? Why?

non-use (reaction is silent)
redundancy (multiple genes for same enzymatic function)
flux rerouting around blocked reactions

Does network function and flux influence network evolution

Yes. High-flux enzymes accumulate fewer amino acid substitutions.



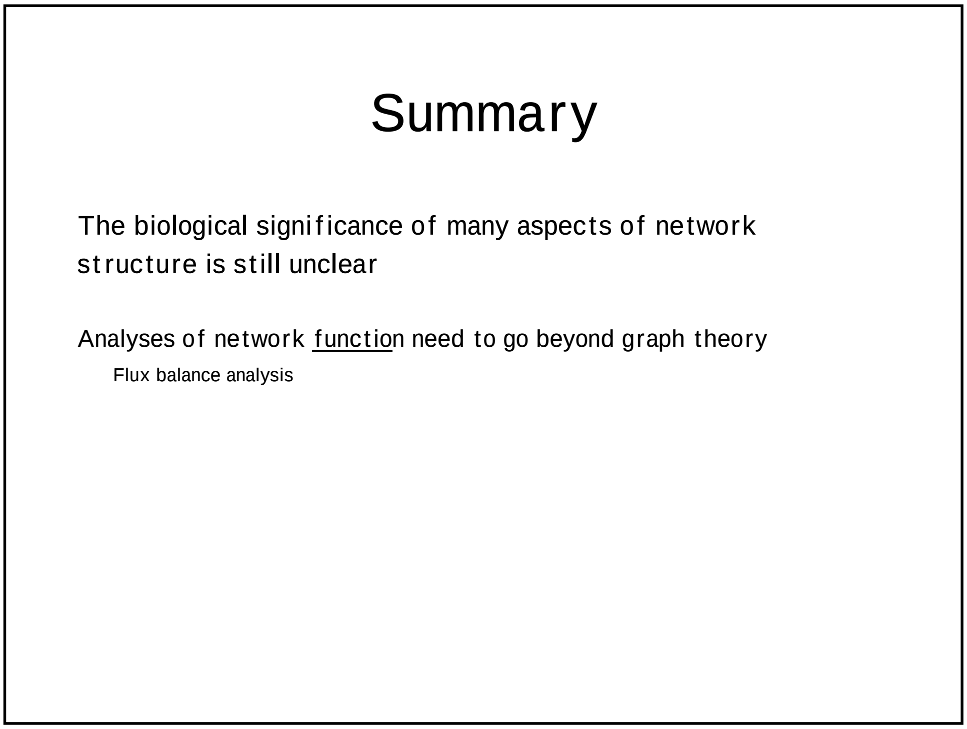
Summary

The most prominent examples of genome-scale biological networks are

- protein interaction networks
- transcriptional regulation networks
- metabolic networks

Graph theory can be used to characterize these networks via

- degree distribution and correlation
- characteristic path lengths and diameter
- clustering coefficient
- abundance of motifs
- indicators of modularity
- …

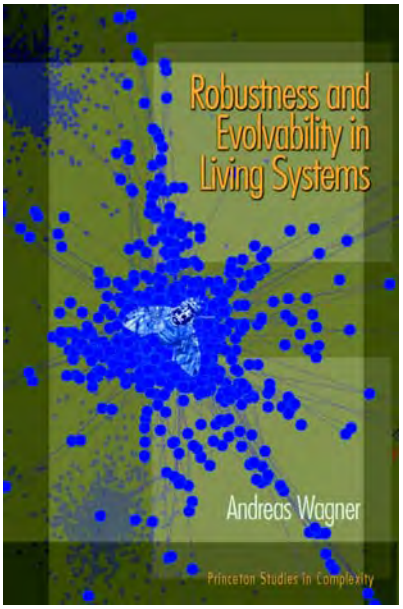


Summary

The biological significance of many aspects of network structure is still unclear

Analyses of network function need to go beyond graph theory

- Flux balance analysis



Robustness and Evolvability in Living Systems

Andreas Wagner

Princeton Studies in Complexity