Bioinformatics I (BIO390)
Biological Networks

Gabriel Schweizer

Institute of Evolutionary Biology
and Environmental Studies, UZH
gabriel.schweizer@ieu.uzh.ch

November 10, 2020

A note on homework exercises for BIO390

The exercises are for you to solve on your own. You do not have to turn them in and they will not be graded. Even though solutions are provided at the end of this document, we highly recommend that you solve them and do so before looking at the solutions, because similar (not necessarily identical) problems will occur on the final exam.

Further reading

Complex networks in general

Newman, MEJ. 2003. The structure and function of complex networks. *SIAM Review* **45**, 167-256.

Fortunato, S., Hric, D. 2016. Community detection in networks: A user guide. *Physics Reports* **659**, 1-44.

Protein interaction networks

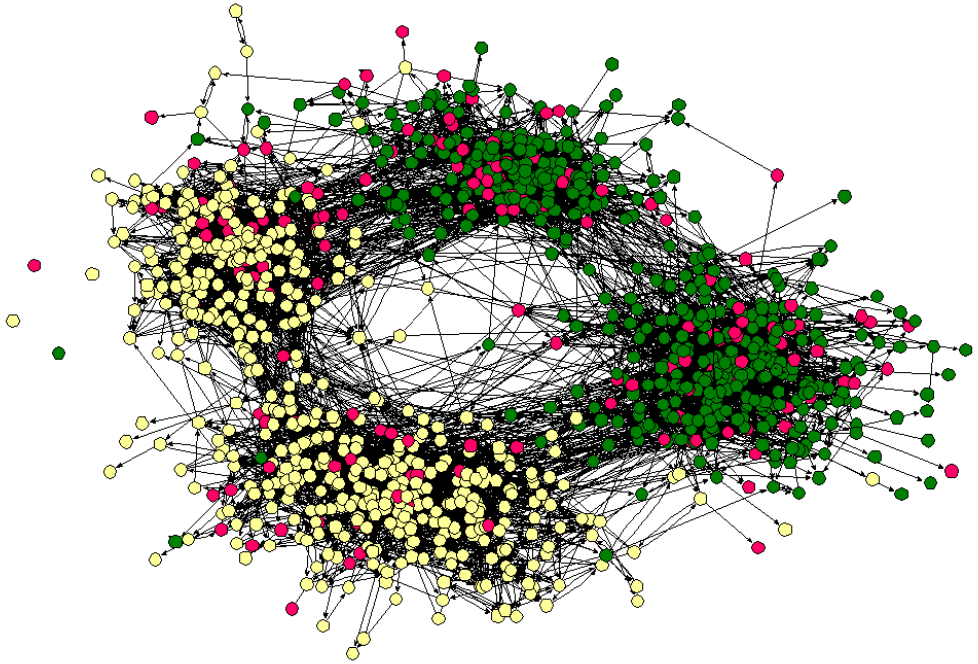
Xia et al. 2004. Analyzing cellular biochemistry in terms of molecular networks. *Annu Rev Biochem* **73**, 1051–1087.

Rajagopala et al. 2014. The binary protein-protein interaction landscape of *Escherichia coli. Nature Biotechnology* **32***,* 285-290.

Metabolic networks

Price et al. 2004. Nature Reviews Microbiology **2**, 886-897.

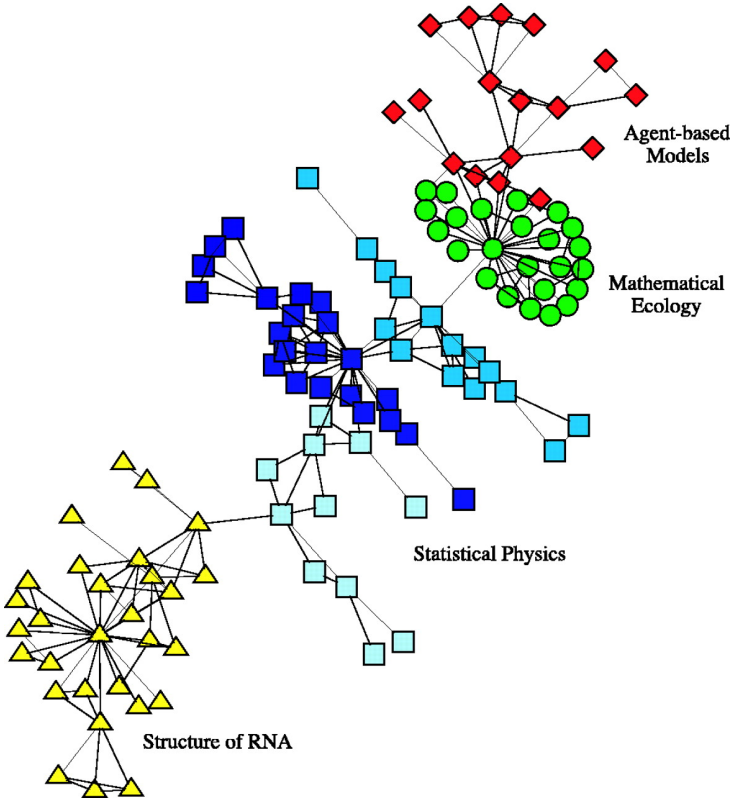
Networks everywhere



Middle and High school friendship network in a US school
Yellow - White Race; Green - Black Race; Pink - Other

(James M, 2001, *American Journal of Sociology* **107**, 679-716)

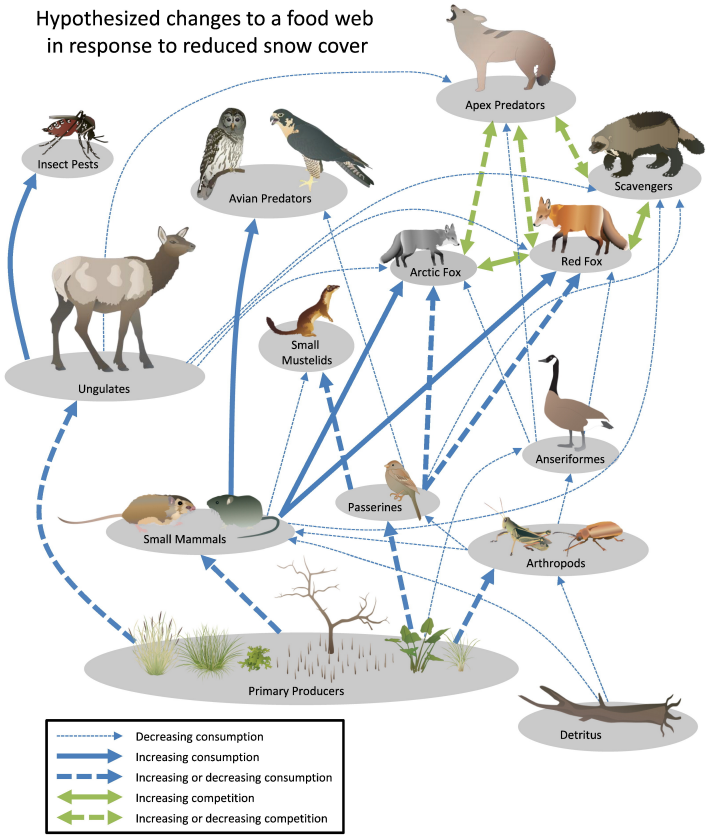
Networks everywhere



Collaborations in science
each node represents a scientist and node shapes indicate field of research

(Newman MEJ, 2004, *PNAS* **101**, 5200-5205)

Networks everywhere



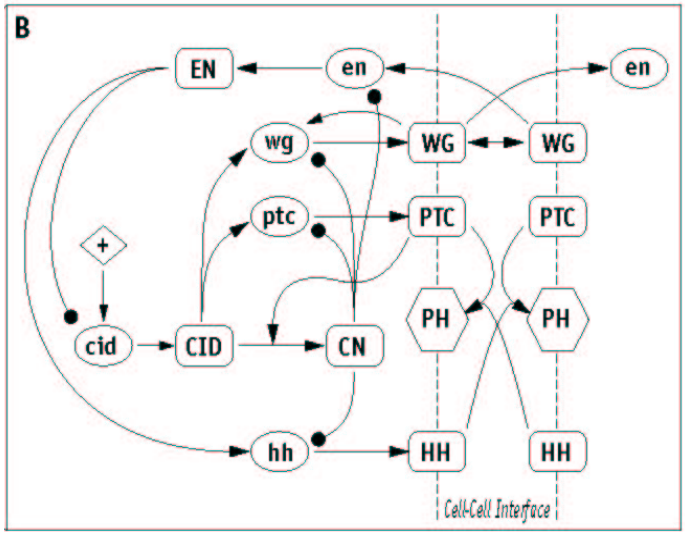
Hypothesized changes to an arctic food web in response to reduced snow cover

(Penczykowski RM et al., 2017, *Food Webs* **13**, 80-91)

Cell-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 G et al. 2000. *Nature* **406**,188-192)

mathematical characterization based on detailed, quantitative biochemical information (concentration, affinity, etc.)

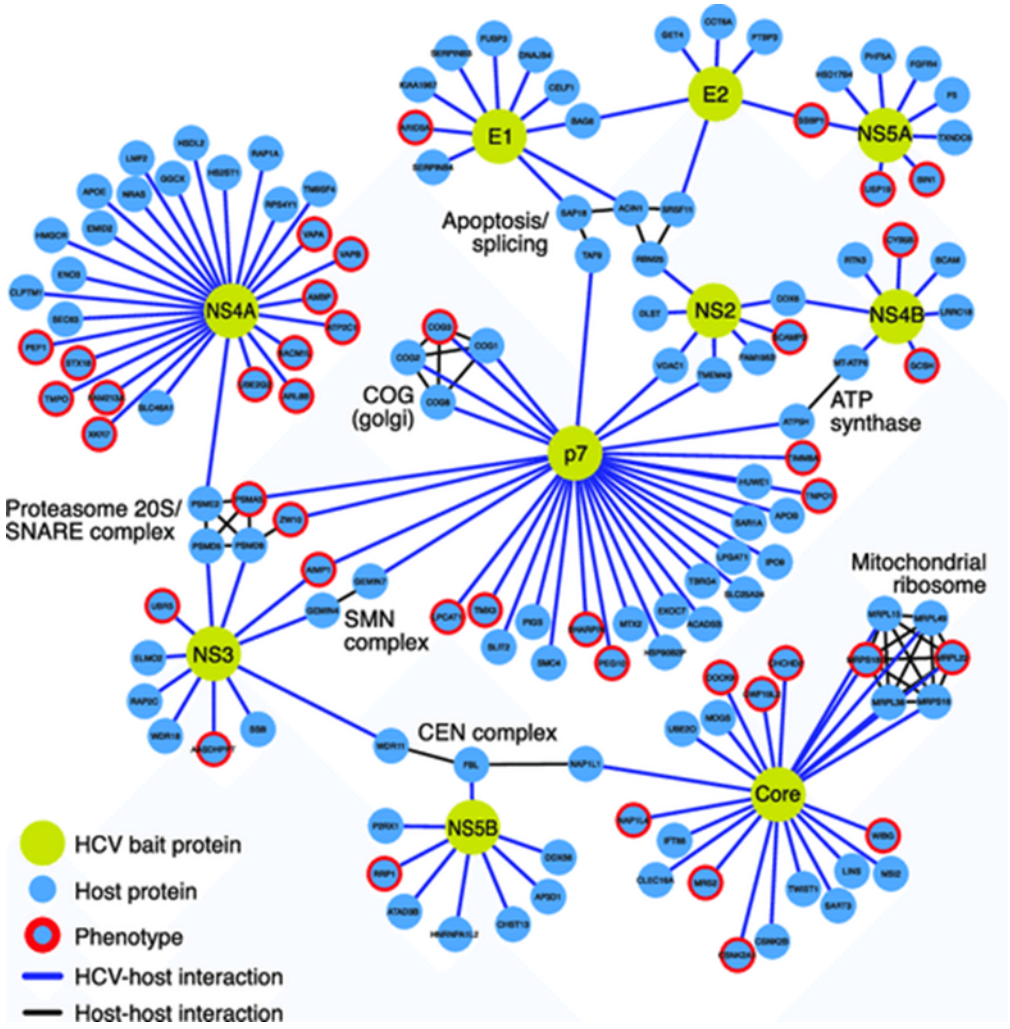
Cell-biological networks

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

Protein interaction networks
Metabolic networks
Transcriptional regulation networks
Genetic interaction networks
…

Mathematical characterization based on qualitative understanding of network topology

Protein interaction networks



(Ramage HR et al., 2015, *Molecular Cell* **57**, 329-340)

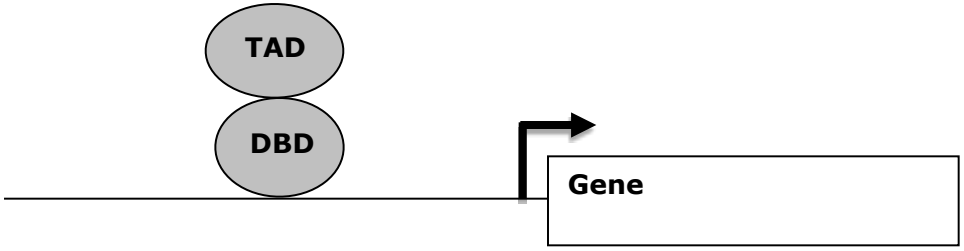
The yeast two–hybrid assay

A technique to identify interacting proteins

Relies on the modularity of eukaryotic transcriptional regulators

DBD: DNA binding domain

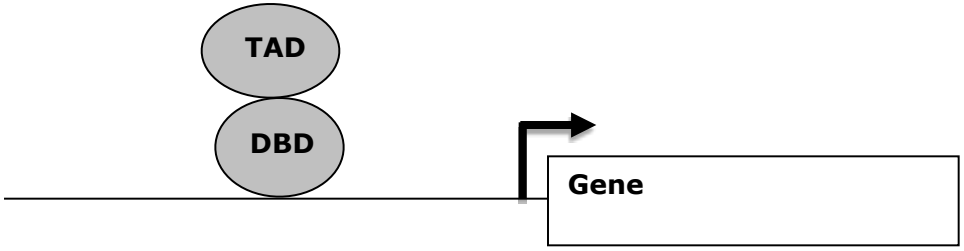
TAD: transcriptional activation domain



The yeast two –hybrid assay

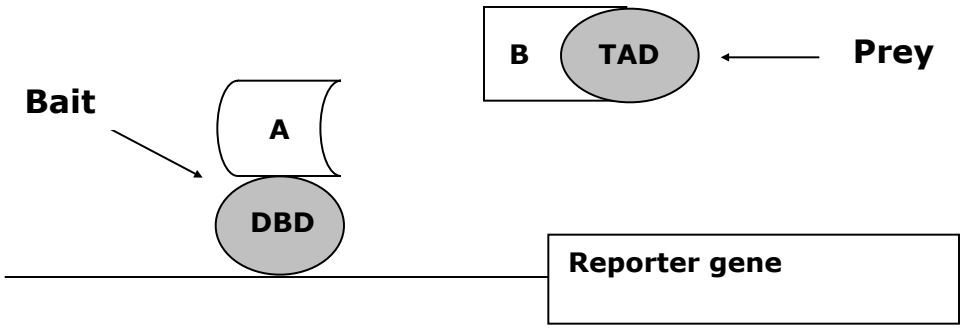
Carried out in cells of the yeast *Saccharomyces cerevisiae*

Can be applied to any two proteins (not just yeast proteins)

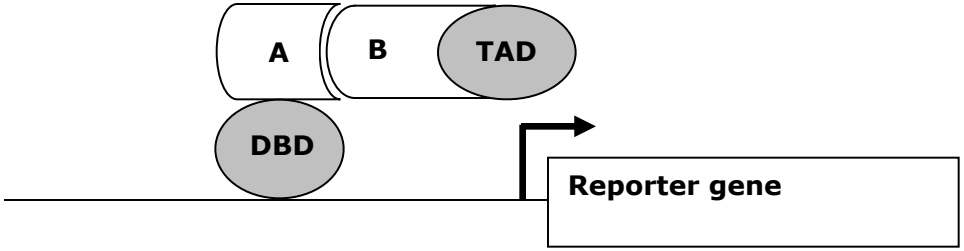


The yeast two –hybrid assay

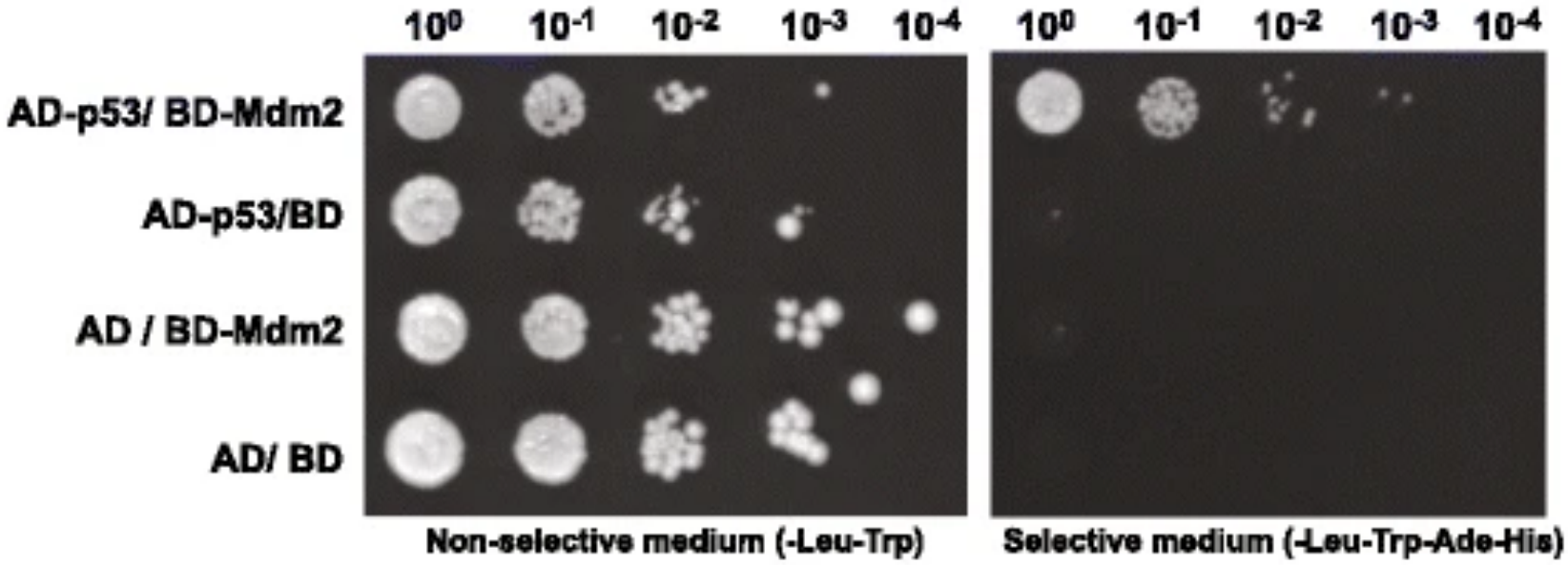
A,B: two proteins whose interaction is to be assayed (or part of a library)



Reporter gene: a gene whose activity is easily monitored

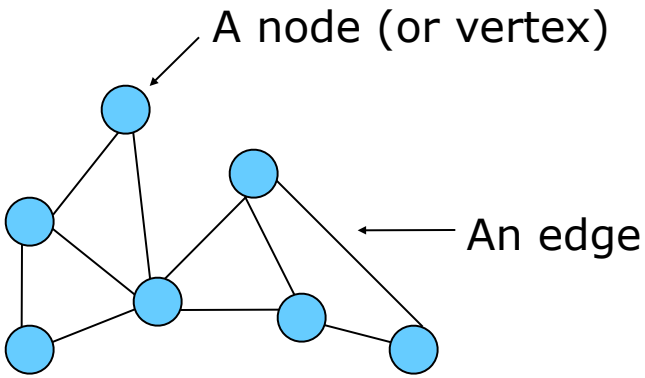


The yeast two –hybrid assay



(Wong JH et al., 2017, *BMC Biology* **15**, article 108)

Graphs

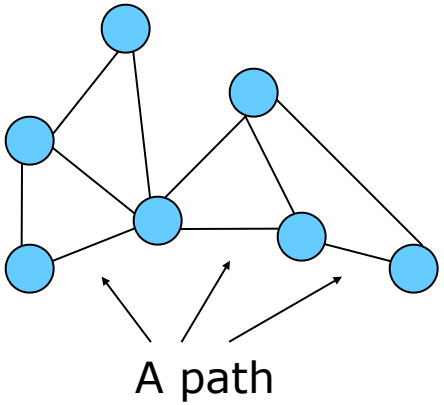


A graph G=(V,E) comprises
a set V of nodes (vertices)
a set E of edges

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

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

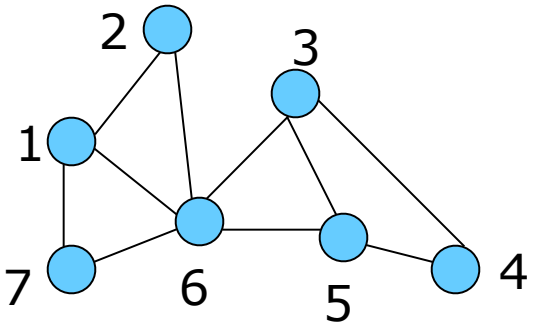
Graphs



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. (there may be several shortest paths)

Graphs can be represented by matrices



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

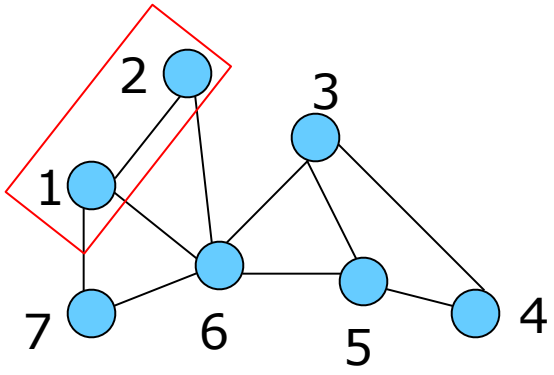
$a_{ij}=1 \quad (V_i, V_j) \in E$

$a_{ij}=0 \quad$ otherwise

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

adjacency matrix is symmetric for undirected graphs

Graphs can be represented by matrices



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

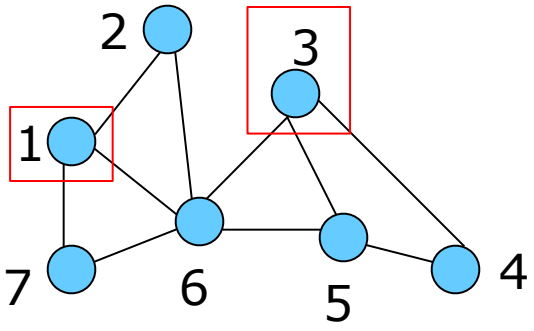
$a_{ij}=1 \quad (V_i, V_j) \in E$

$a_{ij}=0 \quad$ otherwise

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

adjacency matrix is symmetric for undirected graphs

Graphs can be represented by matrices



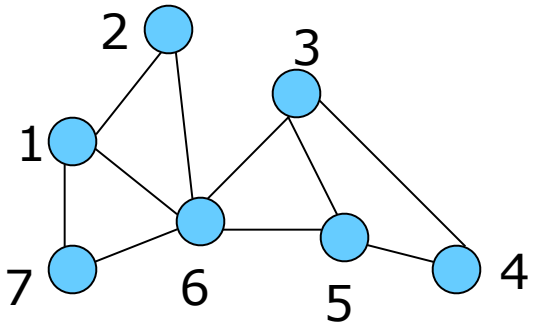
Adjacency matrix $A=(a_{ij})$

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

$a_{ij}=0 \quad$ otherwise

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

adjacency matrix is symmetric for undirected graphs



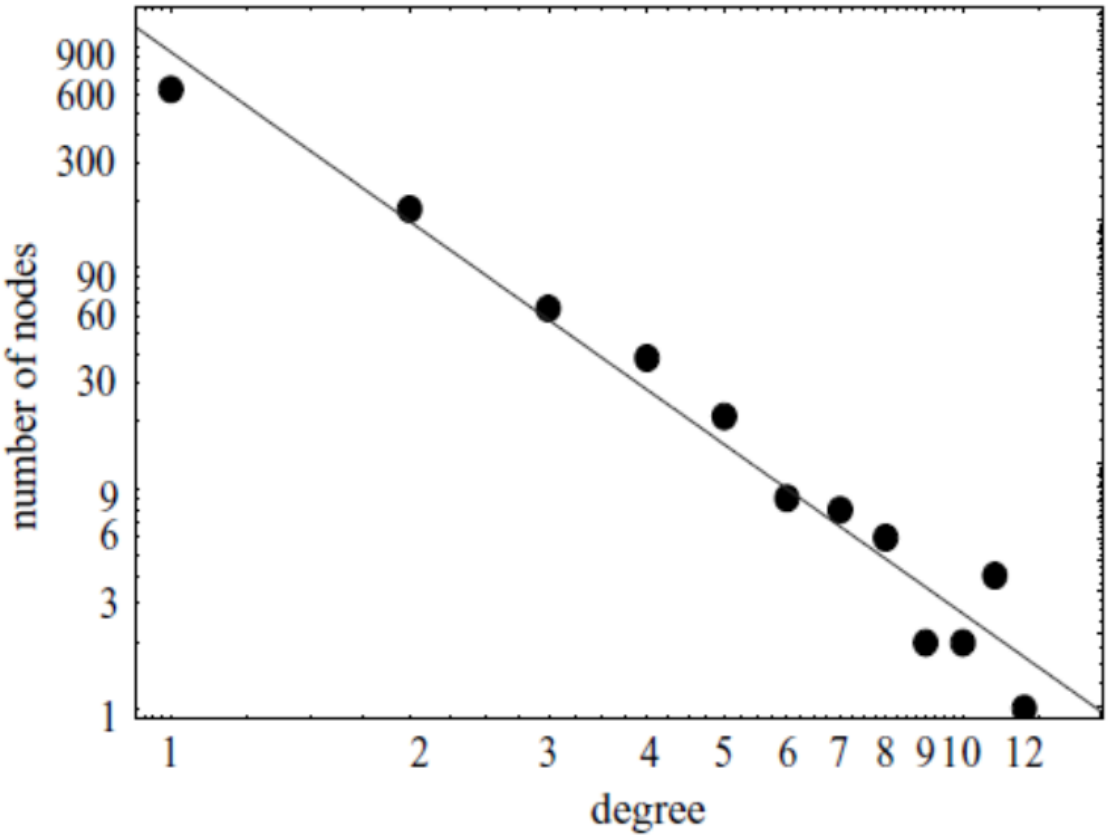
The degree (connectivity) k_i of a node V_i is the number of edges incident with the node (e.g., k_1=3, k_6=5).

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

corresponding to the row sums of an adjacency matrix

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

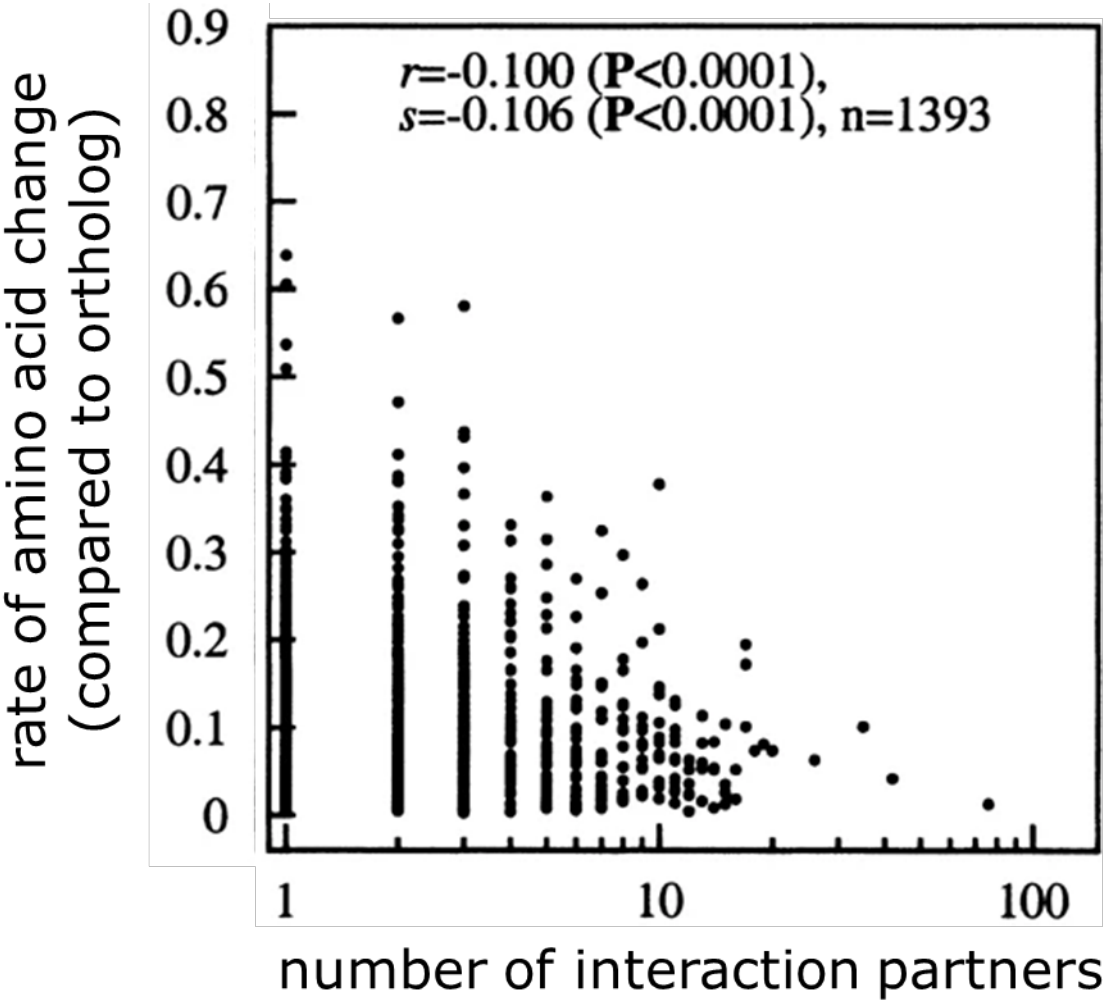
Protein interaction networks (and many other networks) have broad-tailed degree distributions.



more nodes with many interaction partners compared to random graph

(Wagner A, 2003, *Proc Roy Soc London B* **270**, 457-466)

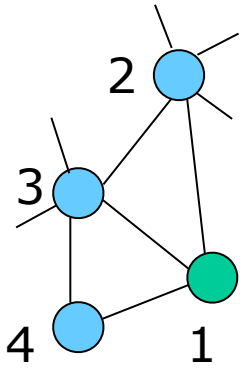
Highly connected yeast proteins tolerate fewer amino acid substitutions in their evolution



(Hahn M et al., 2004, *J Mol Evol* **58**, 203-2011)

The degrees of nodes in a graph may be correlated

Average nearest neighbor degree of a node



$k_1=3$

$k_2=5$

$k_3=5$

$k_4=2$

$k_{nn,1}=(1/3)(5+5+2)=4$

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

neighboring nodes in a graph may have similar degrees

The degrees of nodes in a graph may be correlated

Average nearest neighbor degree of all nodes with degree k

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

N_k ...number of nodes with degree k

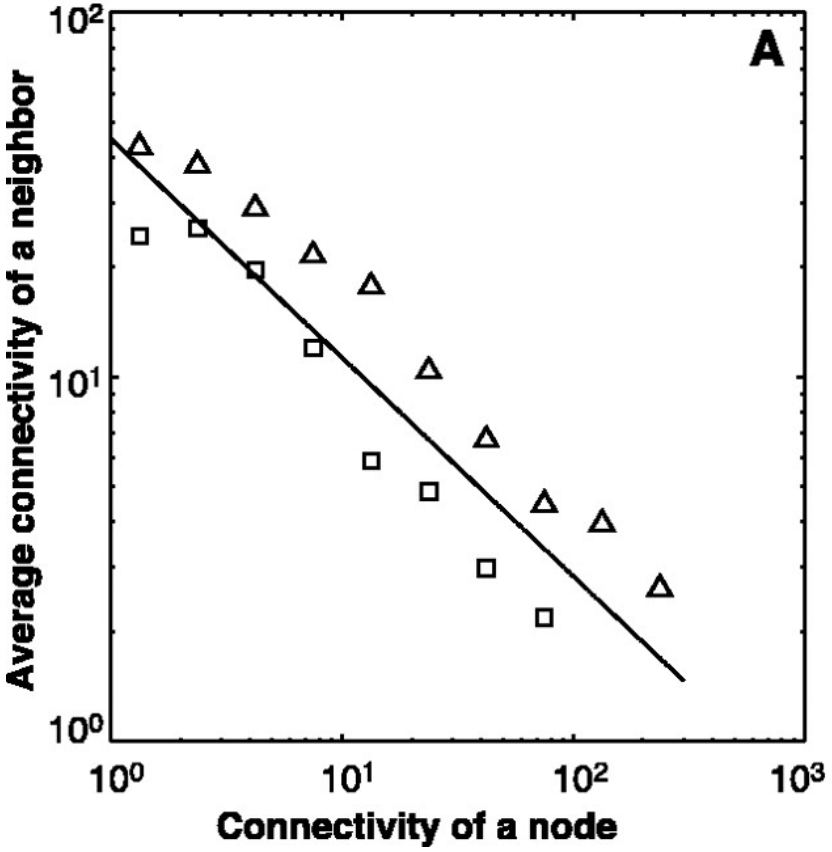
$$k_{nn}(k) = \frac{1}{N_k} \left(\sum_{\text{nodes with degree } k} k_{nn,k} \right)$$

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

Protein interaction networks are disassortative



Few interactions between hubs

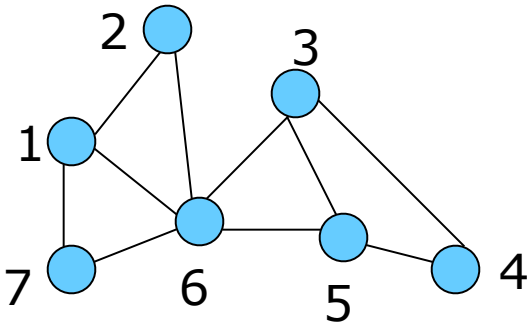
Many interactions between hubs and neighbors with low degree

So far no good explanation was found

Plot of $P_{nn}(k)$ against k for the
yeast protein interaction network (triangles)
and the transcriptional regulation network (squares)

(Maslov and Sneppen, 2002, *Science* **296**, 910-913)

Path length and diameter are measures of graph compactness



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: $\max_{i,j} d_{ij}$ ('longest shortest path')

Mean (arithmetic) shortest path length or characteristic path length (for undirected graphs)

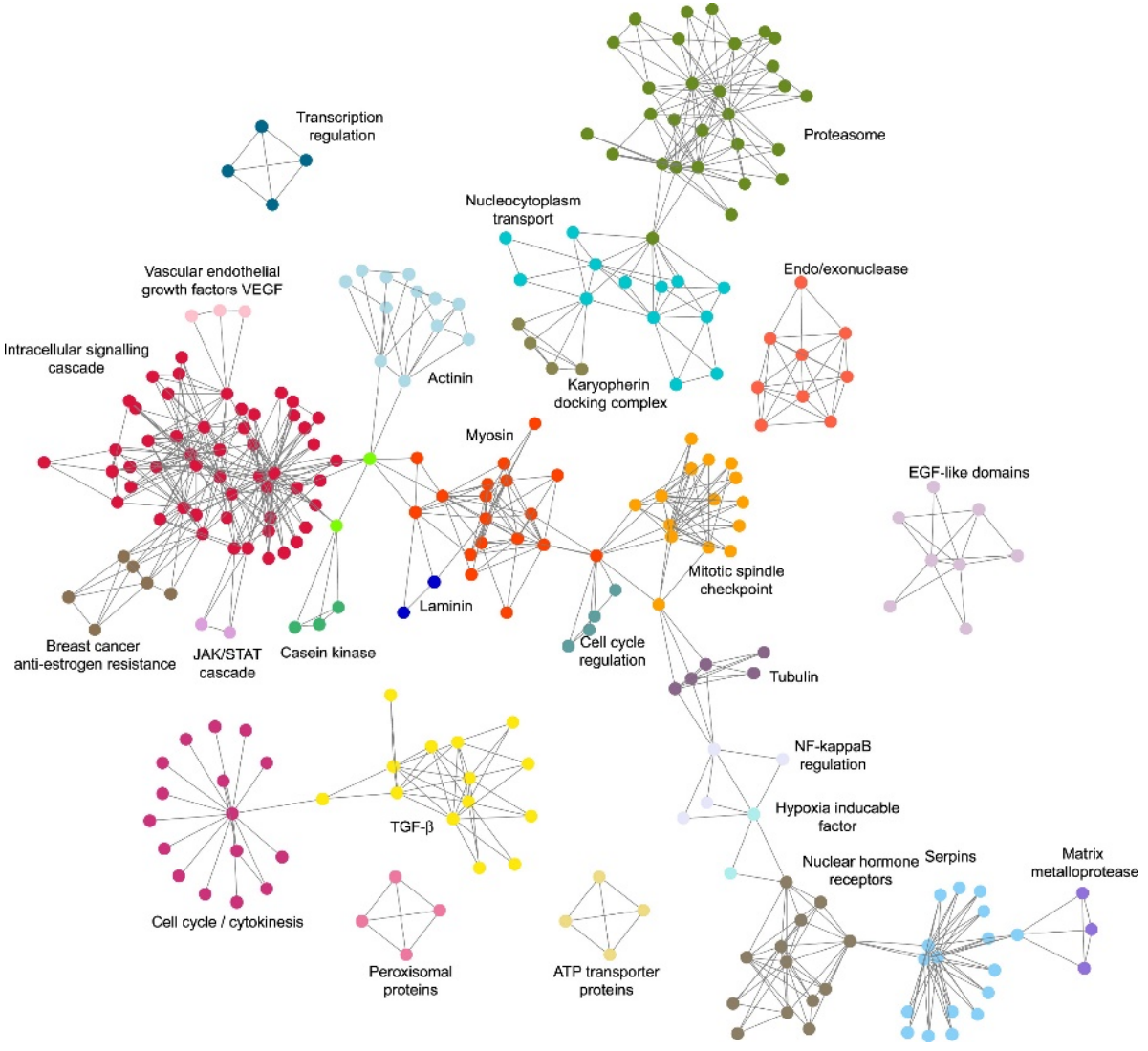
$$L = \frac{1}{N(N-1)} \sum_{i,j,i \neq j} d_{ij}$$

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)

Many graphs can be subdivided into "communities"

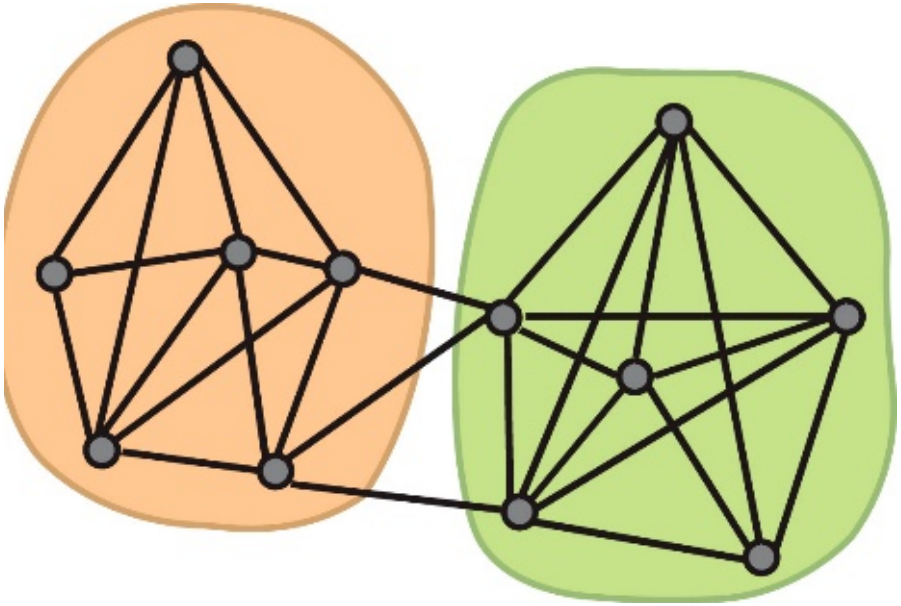


Community structure of a rat protein interaction network

(Fortunato and Hric, 2016, *Physics Reports* **659**, 1-44)

Many graphs can be subdivided into "communities"

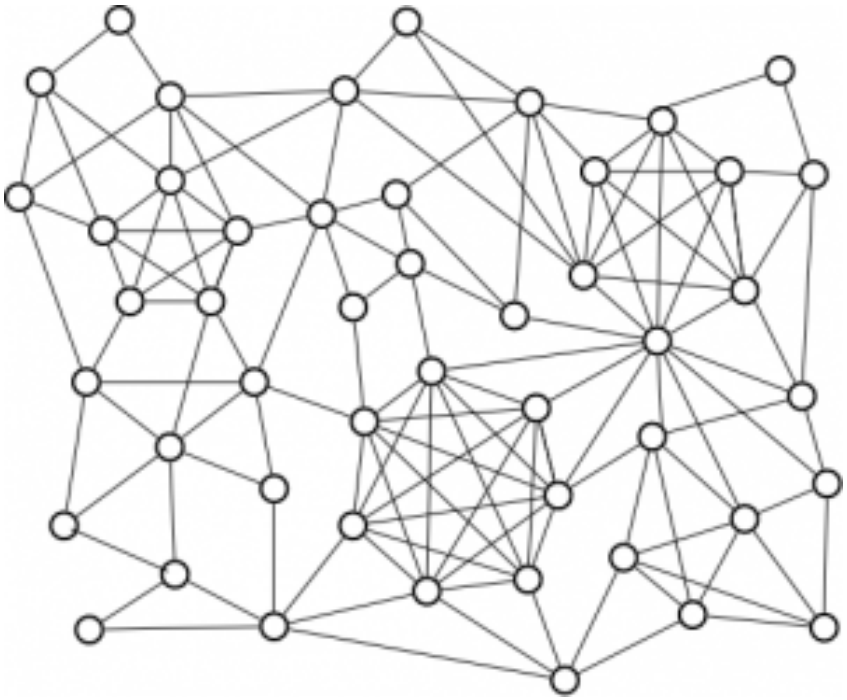
In a graph that can be subdivided into communities (clusters, modules) nodes fall into groups that share more edges with each other than with nodes outside the community



(Fortunato and Hric, 2016, *Physics Reports* **659**, 1-44)

The most densely connected communities are cliques

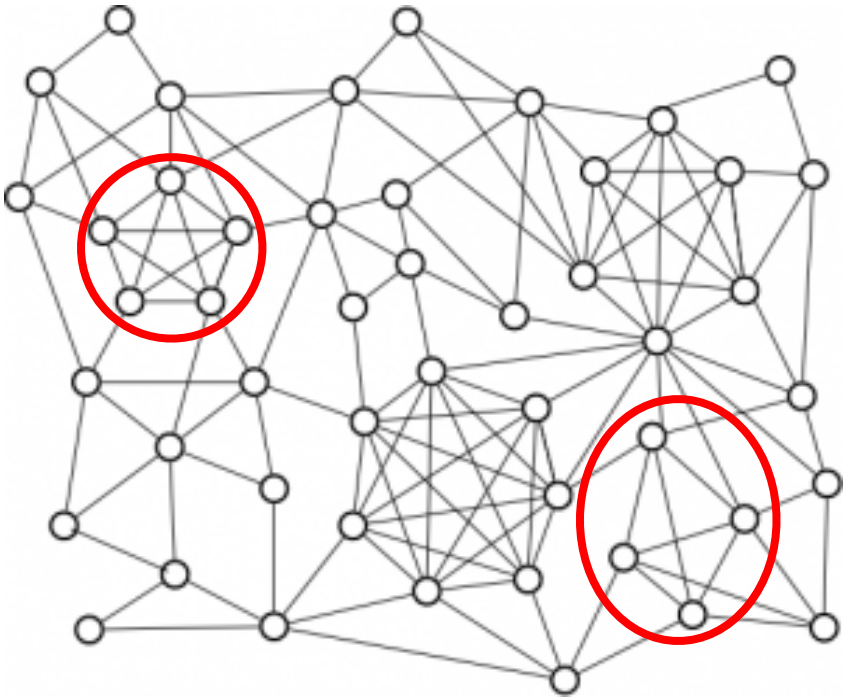
clique: a largest complete (=fully connected) subgraph



http://skipperkongen.dk/2010/11/1

The most densely connected communities are cliques

clique: a largest complete (=fully connected) subgraph



Cliques are usually not fully connected in biological data

http://skipperkongen.dk/2010/11/1

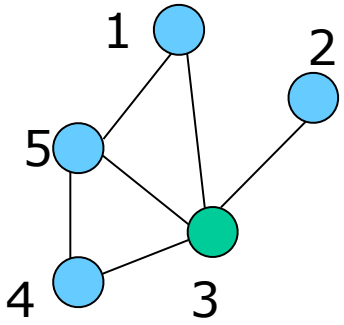
The clustering coefficient is a measure of edge density

Clustering coefficient c_i of a node *i*.

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

$$c_i = \frac{E_i}{\frac{k_i(k_i-1)}{2}}$$

E_i … number of edges among neighbors of i
k_i … degree of i



$$c_3 = \frac{2}{\frac{4(3)}{2}} = \frac{1}{3}$$

Clustering coefficient *c* of a graph

The average of the clustering coefficients of all nodes

(In a clique, all nodes have c_i=1, so c=1 for a graph that is a clique.)

The clustering coefficient is a measure of edge density

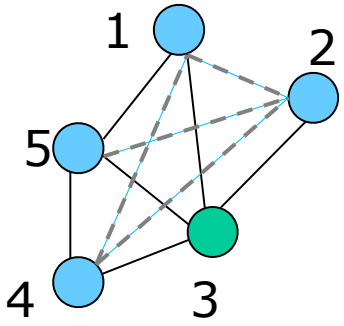
Clustering coefficient c_i of a node *i.*

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

$$c_i = \frac{E_i}{\frac{k_i(k_i-1)}{2}}$$

E_i … number of edges among neighbors of i

k_i … degree of i



number of edges between neighbors that are found (2) / number of all possible edges between neighbors (6)

Clustering coefficient *c* of a graph

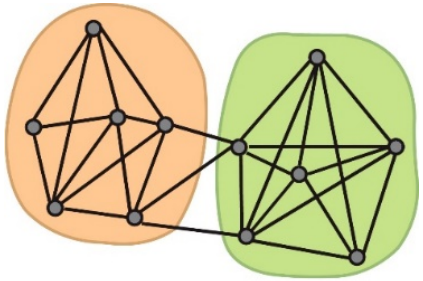
The average of the clustering coefficients of all nodes

(In a clique, all nodes have c_i=1, so c=1 for a graph that is a clique.)

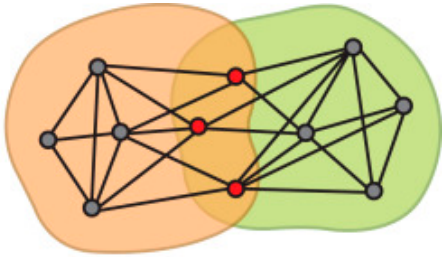
Many computational methods aim to detect communities in networks

Some require information about the total number of communities (easier), others don't (more difficult).

Hard-clustering methods generate non-overlapping communities (easier)



Soft-clustering methods allow overlapping communities (more difficult)



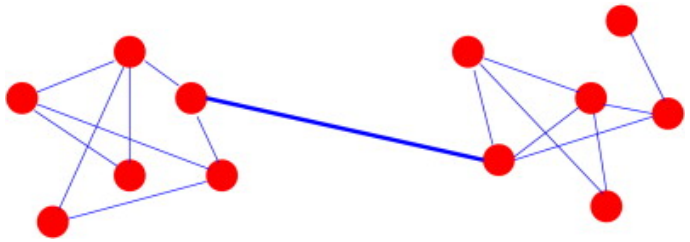
The Girvan-Newman algorithm is a popular heuristic to cluster large graphs

It does not guarantee to find the best possible clustering

It relies on the concept of edge betweenness

Edge <u>betweenness</u> (centrality, load):
the number of shortest paths passing through an edge i

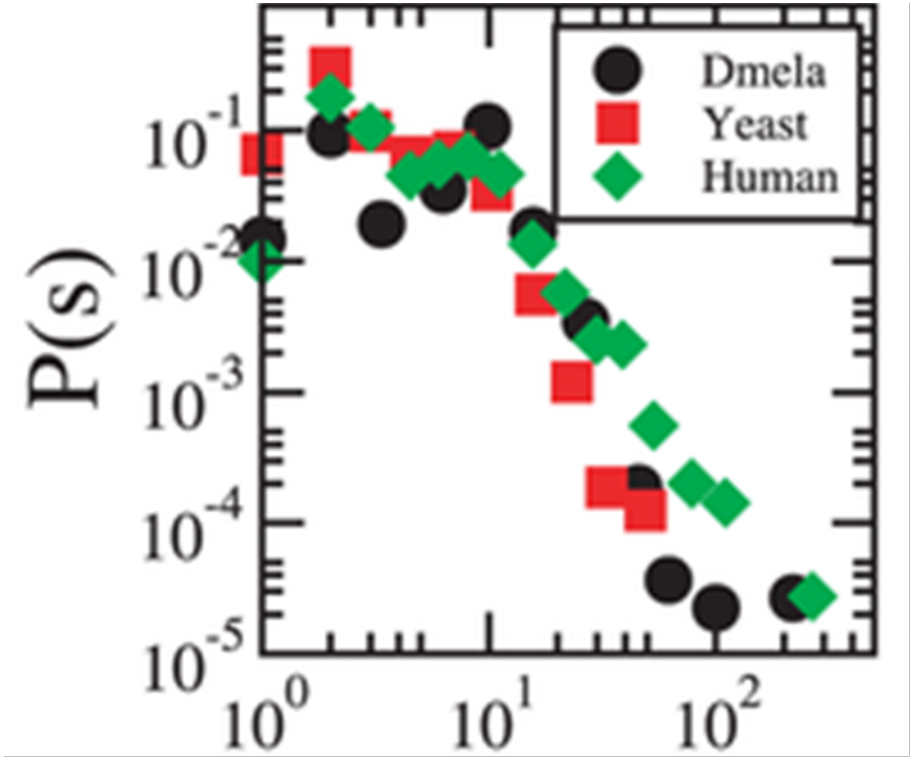
$$b_i = \sum_{j,k,j \neq k} \frac{n_{jk}(i)}{n_{jk}}$$



$n_{jk}(i)$ number of shortest paths connecting node j and k and passing through edge i
n_{jk} number of shortest paths connecting node j and k

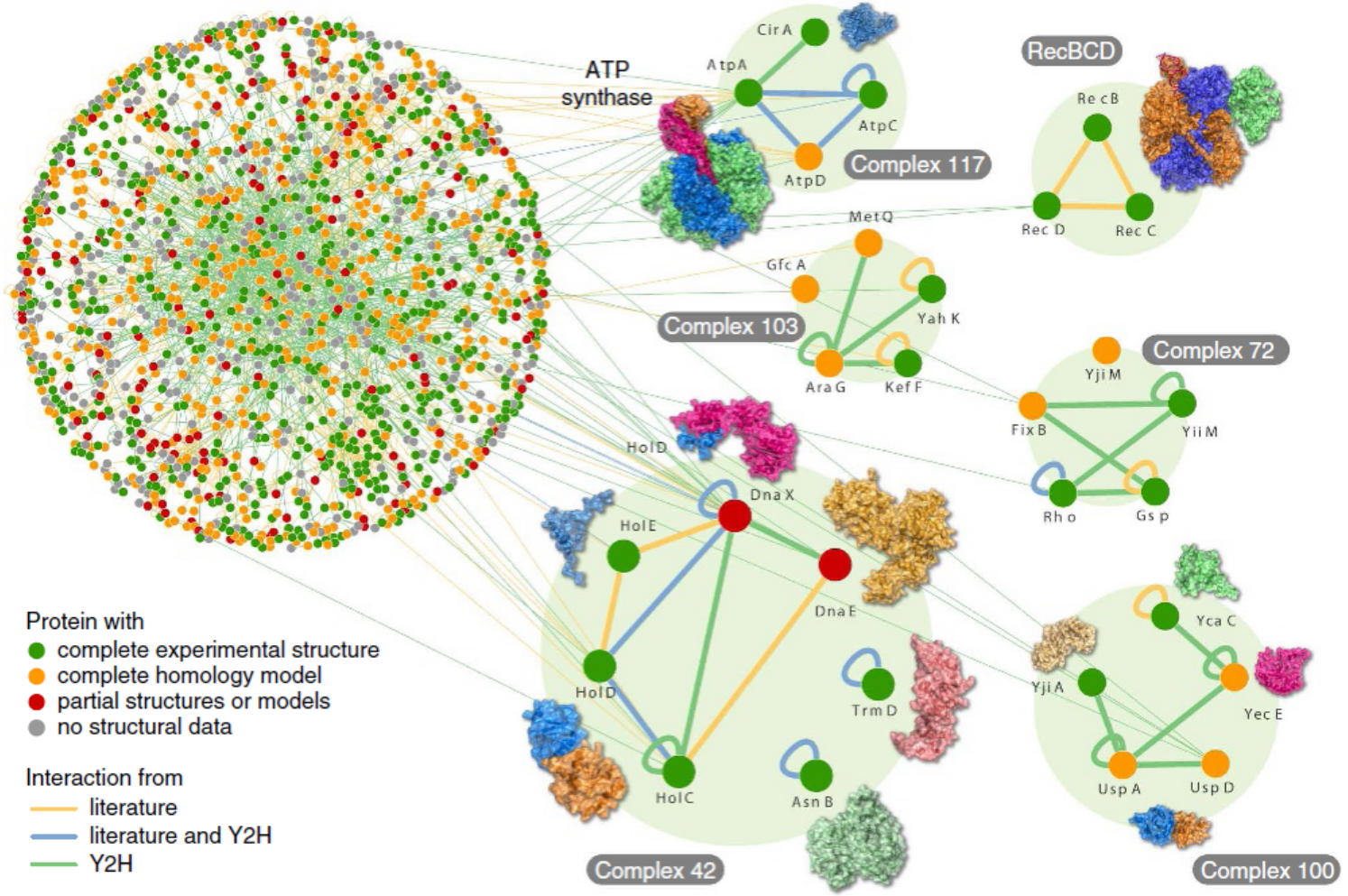
edges with high betweenness separate clusters in a graph

Module sizes in protein interaction networks have a broad-tailed distribution



(Lancichinetti et al, 2010, *PLOS ONE* **5**, e11976)

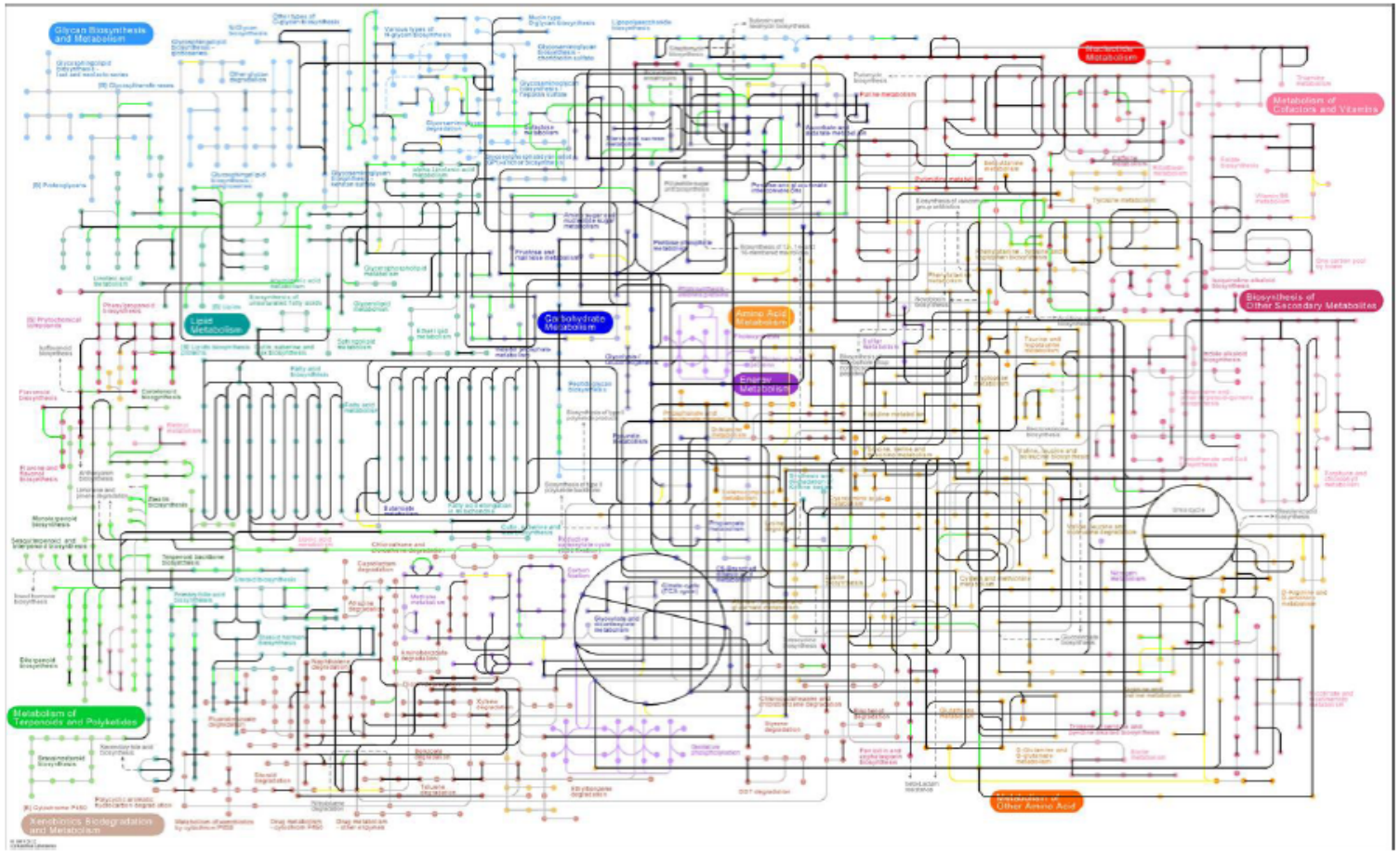
The best maps of protein interaction networks integrate different kinds of information



An *E.coli* protein interaction network

(Rajagopala et al., 2014, *Nature Biotechnology* **32**, 285-290)

Metabolic networks



(Duan et al., 2013, *Metabolites* **3**, 1-23)

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.

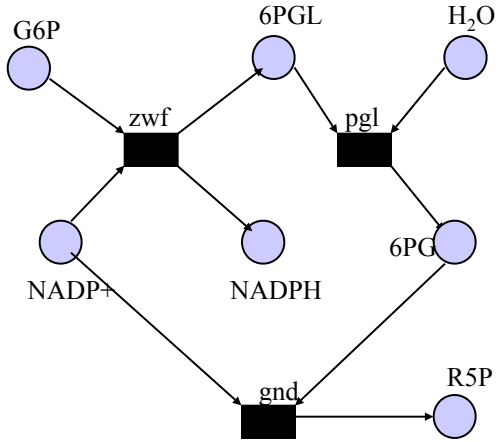
In free-living heterotrophic organisms, several hundred such enzymatic reactions are necessary to fulfill these functions.

Graphs can (crudely) represent large chemical reaction networks

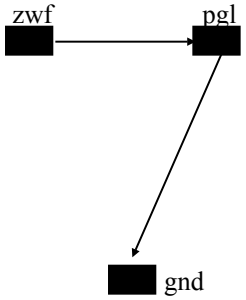
Stoichiometric Equations

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)

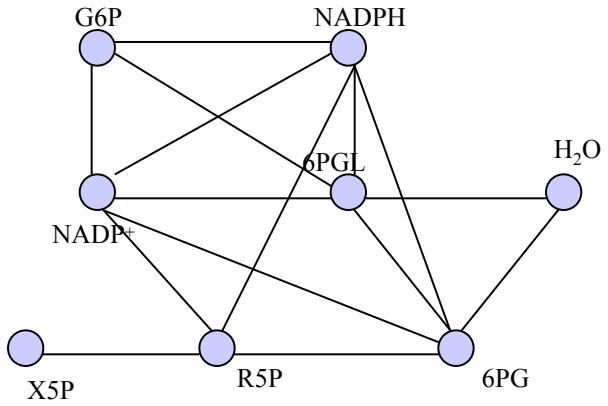
bipartite graph



enzyme graph

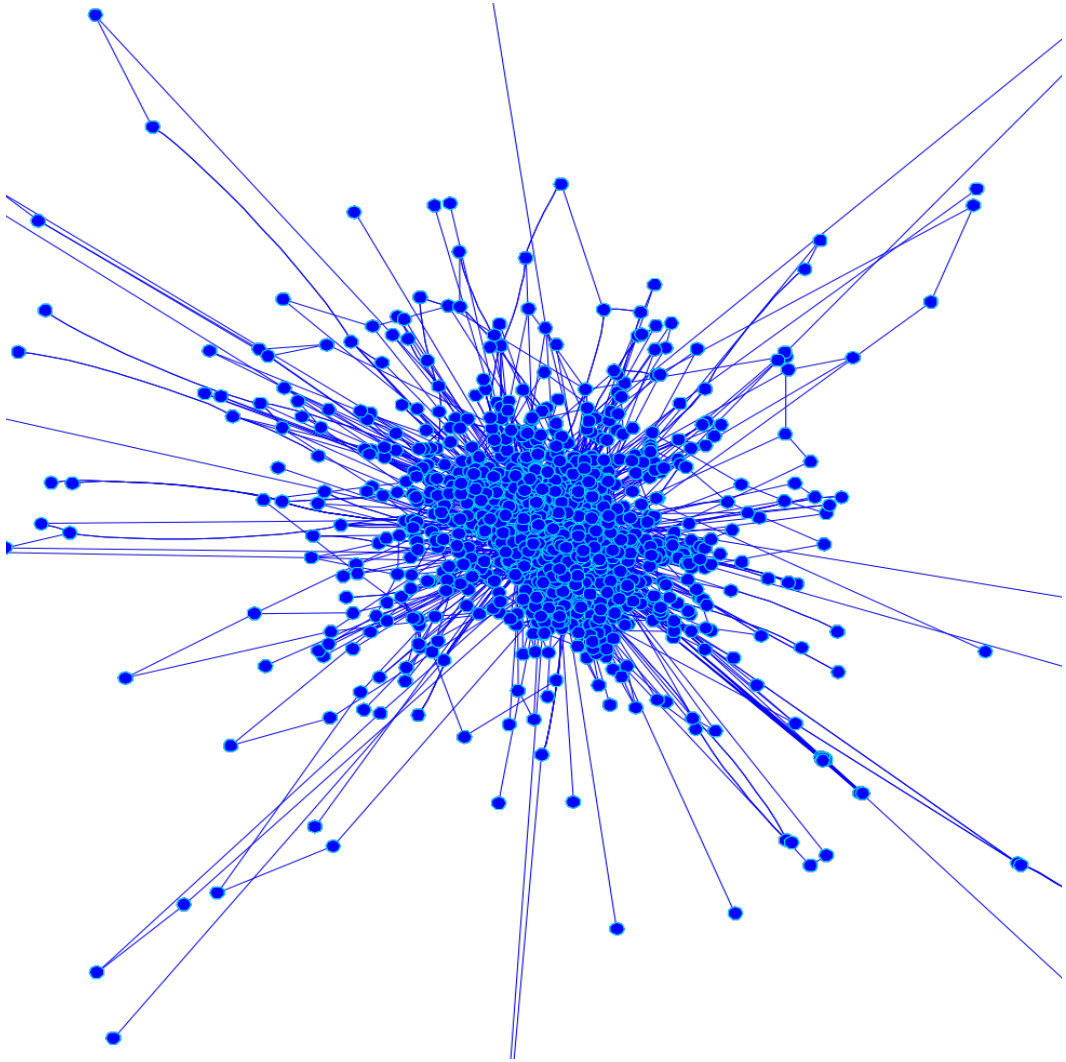


substrate graph (small molecules)

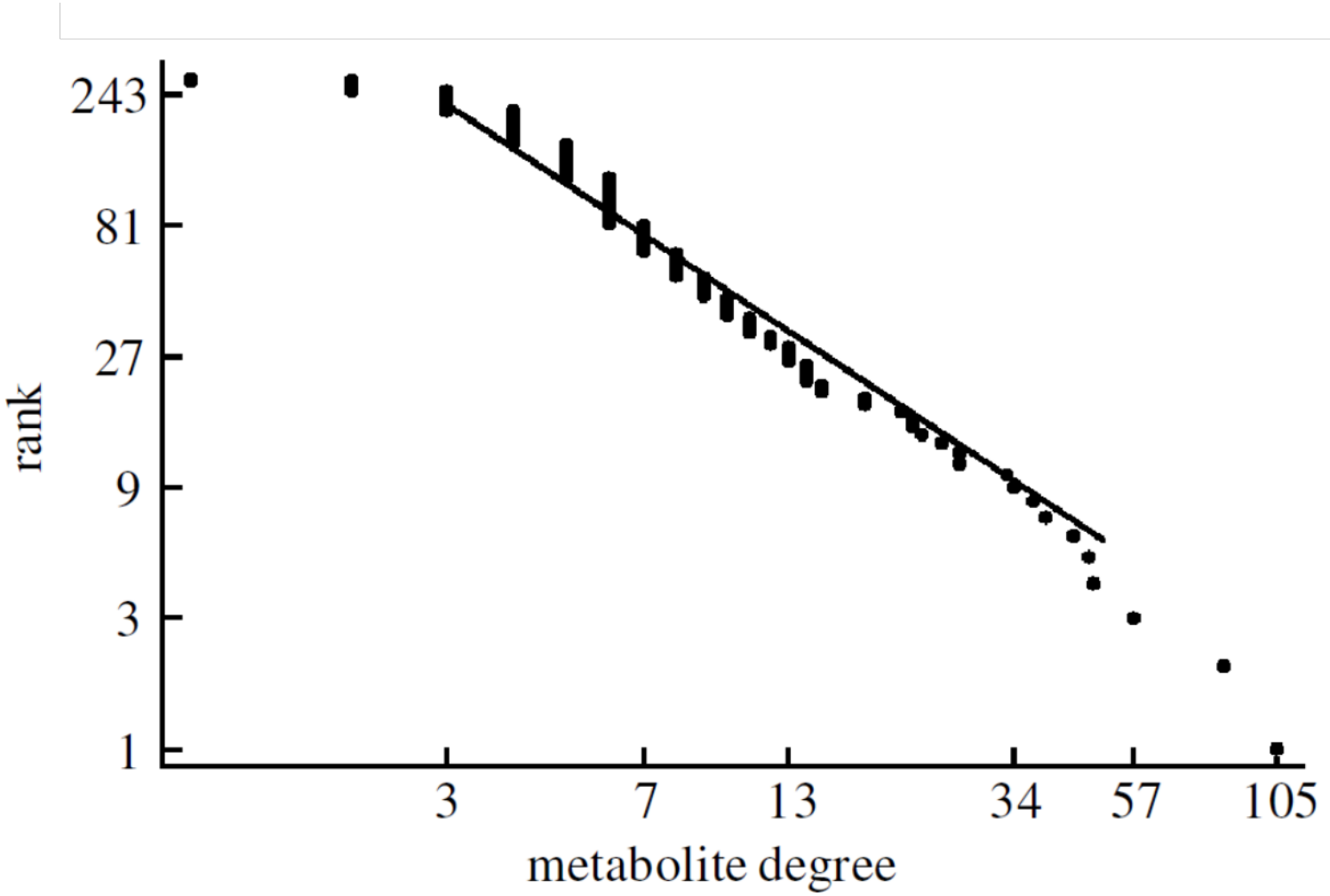


two kinds of nodes (enzymes and substrates)

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



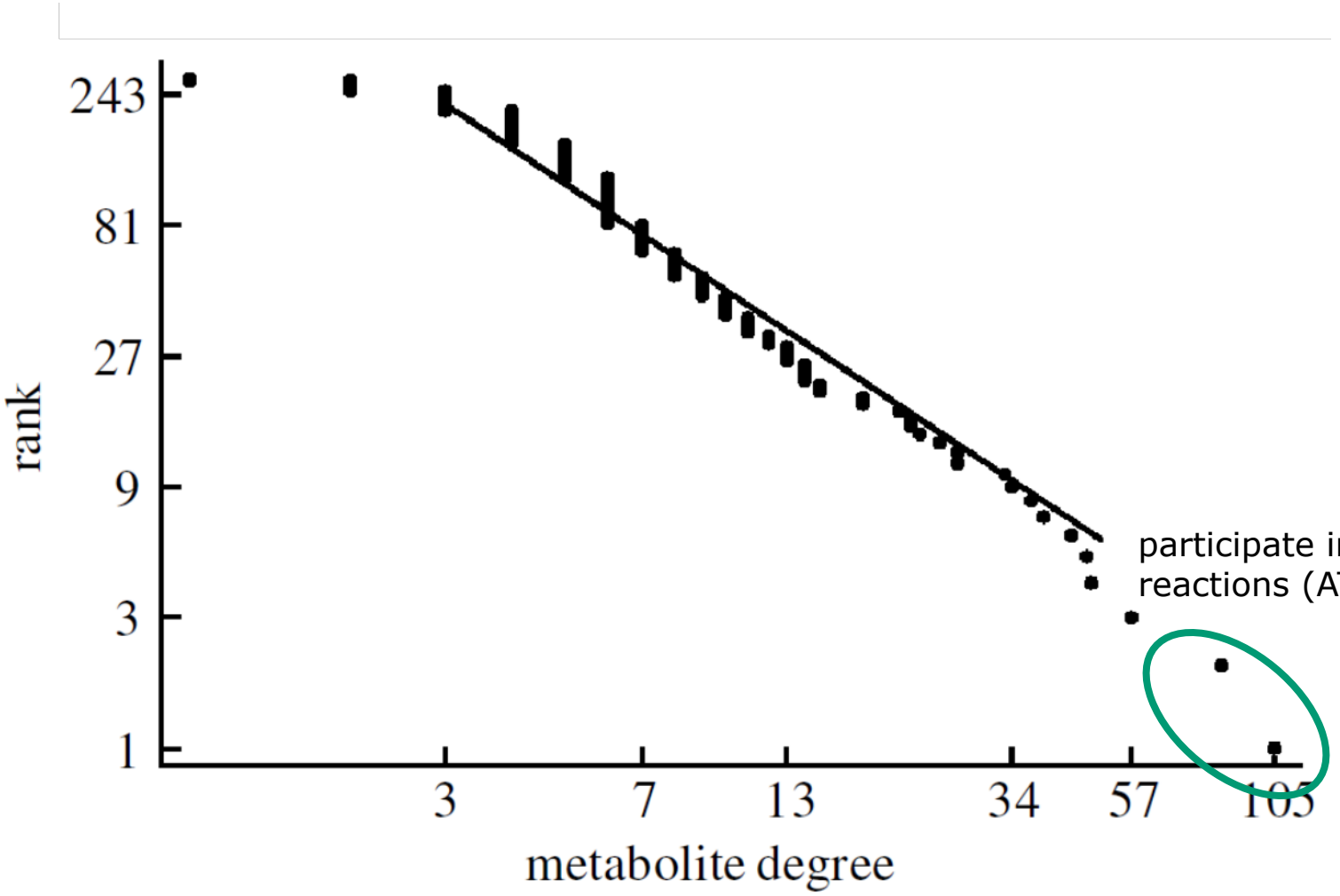
Metabolic networks have a broad-tailed degree distribution



Substrate network of *E. coli*

(Wagner and Fell, 2001, *Proc Roy Soc London B* **268**, 1803-1810)

Metabolic networks have a broad-tailed degree distribution



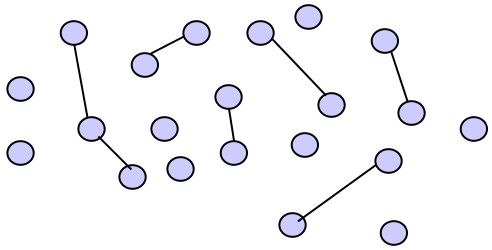
Substrate network of *E. coli*

(Wagner and Fell, 2001, *Proc Roy Soc London B* **268**, 1803-1810)

Key features of small-world graphs

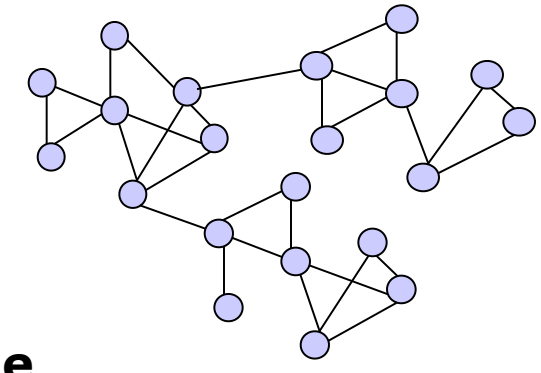
1. They are sparse

relatively small number of edges



2. They are "cliquish"

as measured by a high clustering coefficient



3. Despite 1 and 2, paths from any one node to any other node are VERY short

(short mean path length, "small-worldness")

(Watts and Strogatz, 1998, *Nature* **393**, 440-442)

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
world wide web	web pages	hyperlinks
actor collaborations	actors	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?

Not really clear, some ideas are:

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

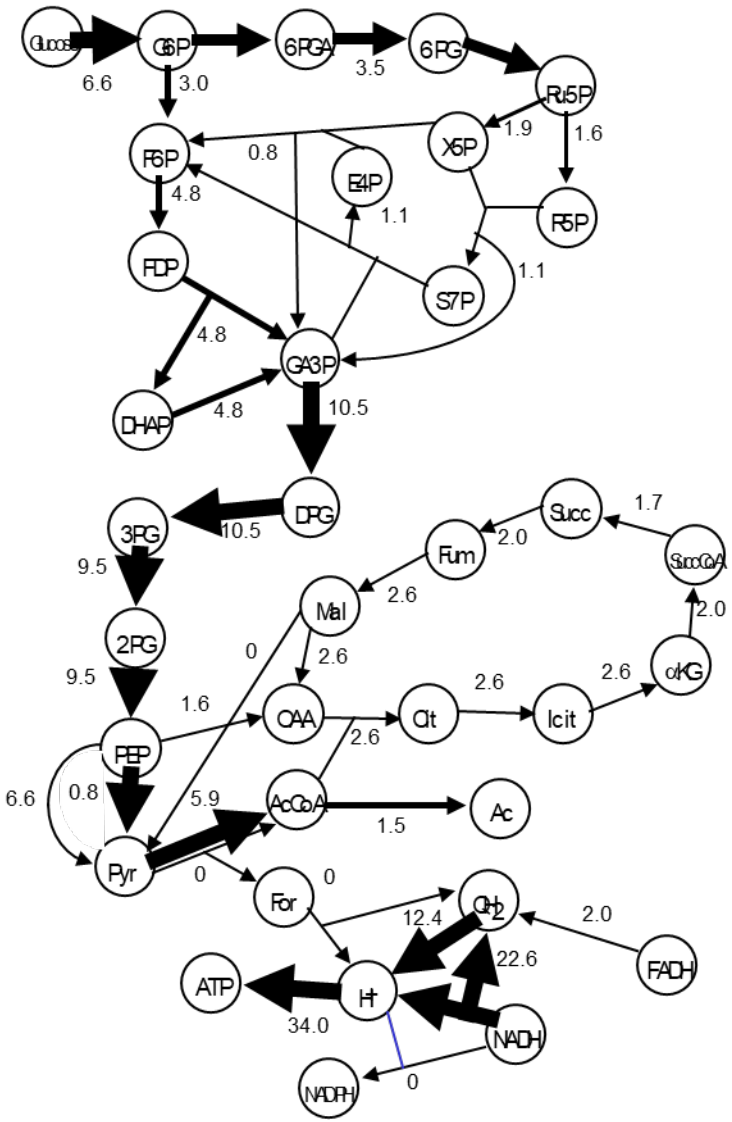
One needs to analyze the **flow (flux) of matter**
through these networks

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.)

Metabolic flux through central carbon metabolism of *E.coli* growing at a maximally possible rate in a glucose-minimal medium



glucose is the only C-source in the medium

arrow thickness is the reaction rate of enzymes needed to achieve maximum growth rate

(modified from Edwards and Palsson, 2000, *PNAS* **97**, 5528-5533)

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**

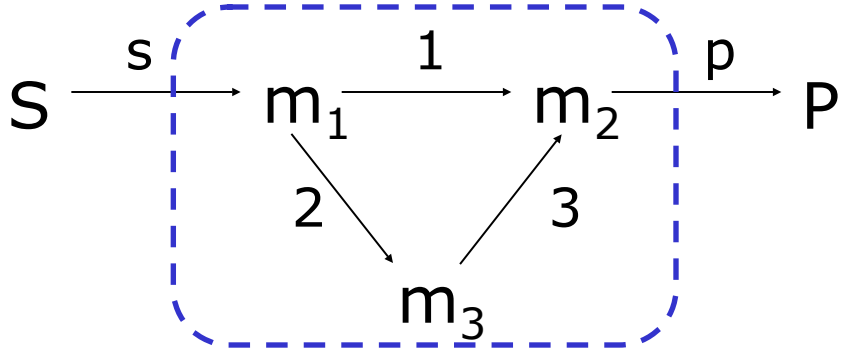
chemical aspect

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

biological aspect

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 m_i change according to the equations

$$\frac{dm_1}{dt} = v_s - v_1 - v_2$$
$$\frac{dm_2}{dt} = v_1 + v_3 - v_p$$
$$\frac{dm_3}{dt} = v_2 - v_3$$

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

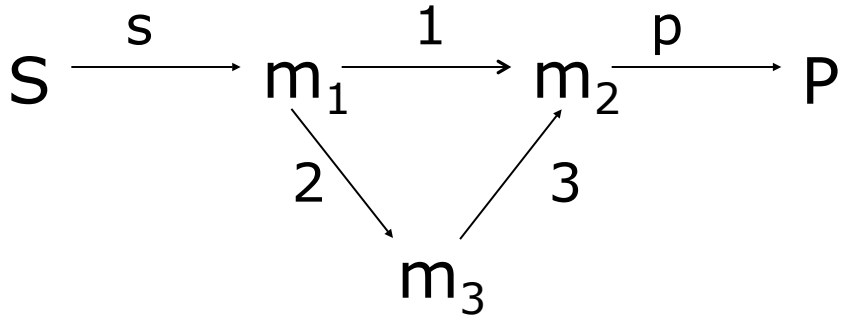
$$\mathbf{S} = \begin{pmatrix} 1 & -1 & -1 & 0 & 0 \\ 0 & 1 & 0 & 1 & -1 \\ 0 & 0 & 1 & -1 & 0 \end{pmatrix}$$

Stoichiometry matrix

v_i metabolic flux through reaction i

$$\vec{v} = (v_s, v_1, v_2, v_3, v_P)^\top$$

Rows: metabolites
Columns: reactions



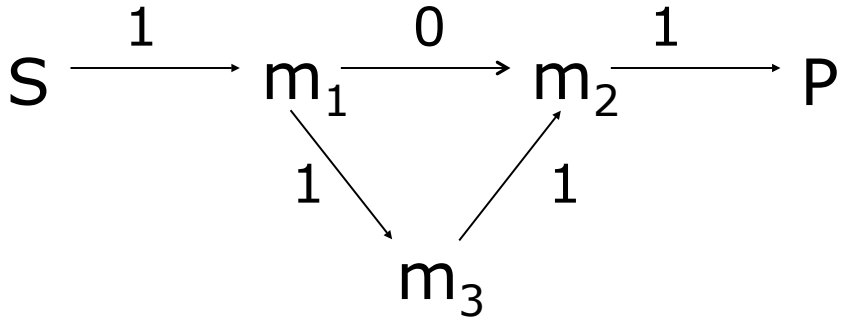
FBA assumes that metabolism is in a steady state where the concentrations of metabolites no longer change (the environment does not change)

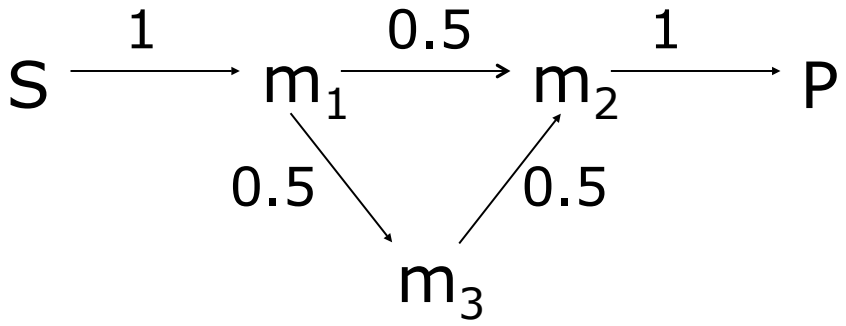
$$\frac{d\vec{m}}{dt} = 0$$

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

The solutions of these equations are the allowable metabolic fluxes. They form the so-called <u>null space of S</u>

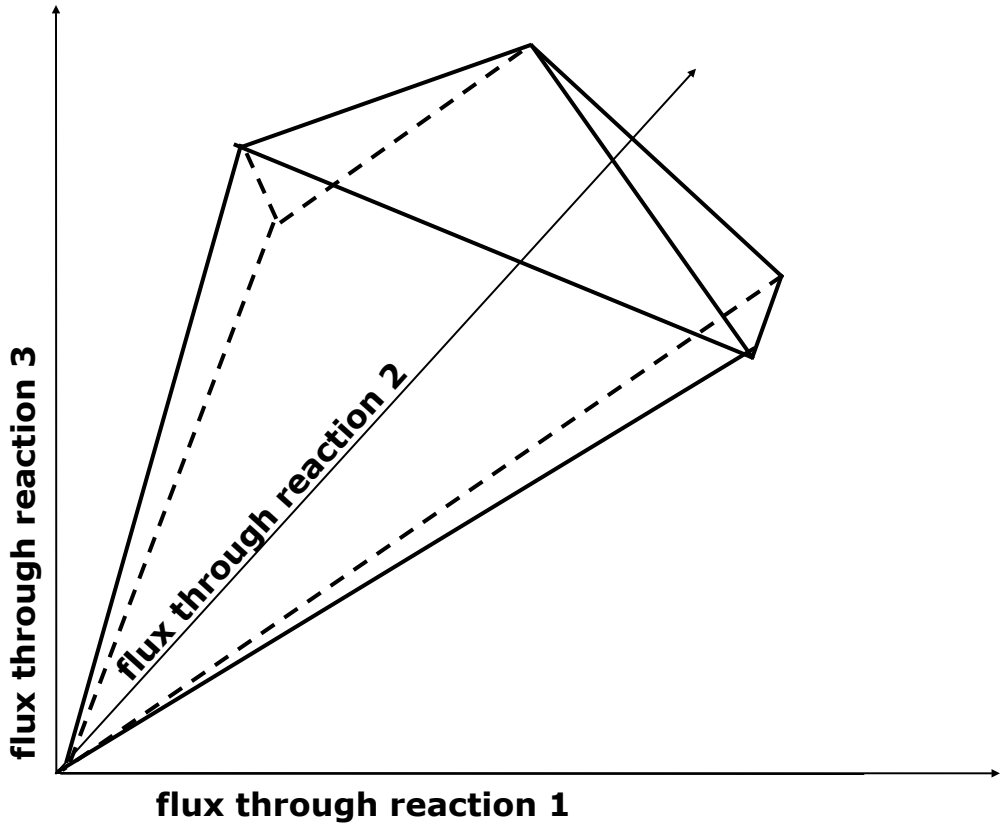
Two allowable flux distributions for our example network





All fluxes of the form (1,x,1-x,1-x,1), $0 \leq x \leq 1$ are allowable (remember to not violate the law of mass conservation)

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

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

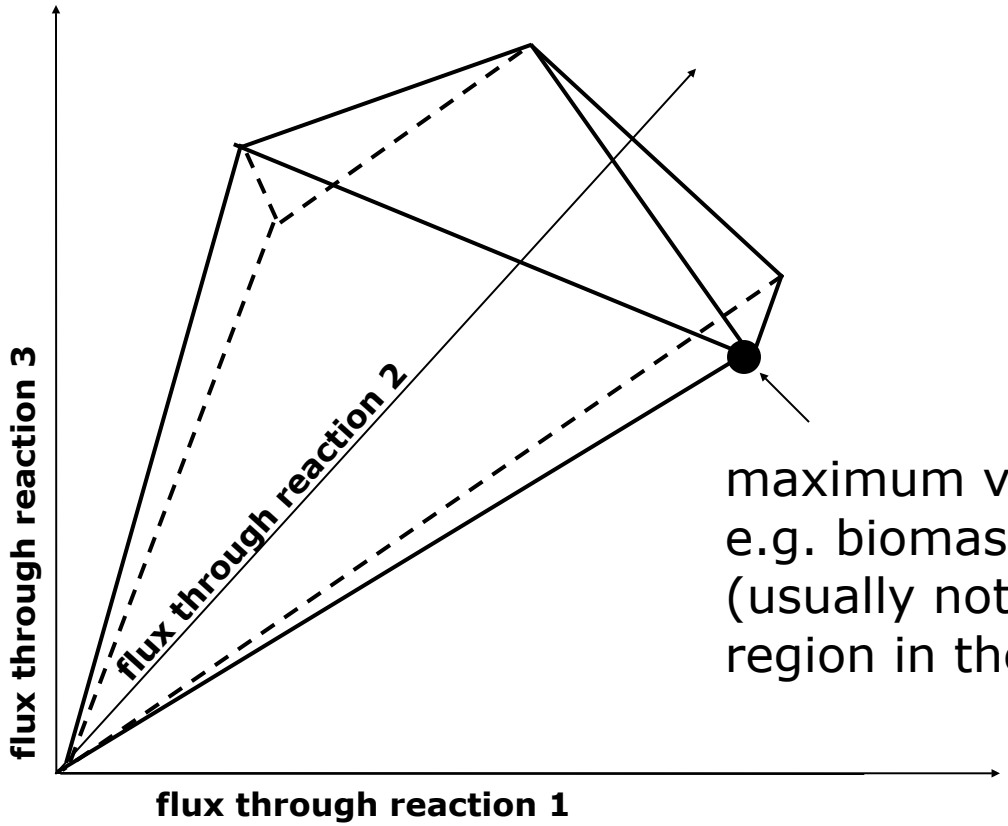
Example:

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. biomass growth flux
(usually not a point, but a region in the null space)

Example questions for flux balance analysis

Can a given organism (metabolism) survive in environment X?

How fast could it grow in this environment?

Why are many enzymatic reactions dispensable in any one environment?

Why do some metabolisms have many reactions, while others have few?

Does network function and flux influence network evolution

Is it possible to design “resistance-proof” antimetabolic drugs?

Summary

Among the most prominent examples of genome-scale cell-biological networks are

protein interaction networks
metabolic networks

Graph theory can be used to characterize these networks via

degree distribution and correlation
characteristic path lengths and diameter
clustering coefficient
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