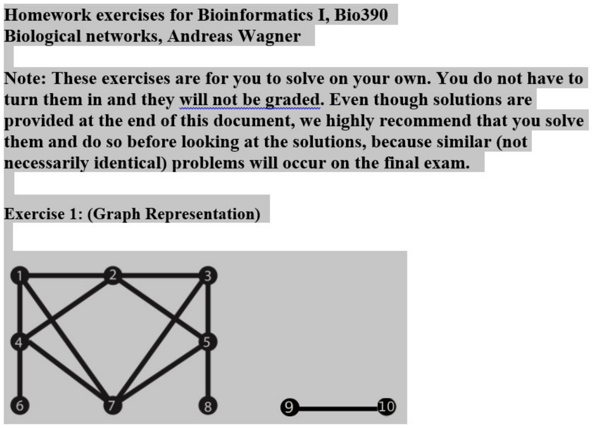
Bioinformatics I
Biological Networks

Andreas Wagner

Department of Evolutionary Biology
and Environmental Studies, UZH
andreas.wagner@ieu.uzh.ch

1

Available online



Homework exercises for Bioinformatics I, Bio390
Biological networks, Andreas Wagner

Note: These 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.

Exercise 1: (Graph Representation)

2

Further reading

Complex networks in general

Newman, MEJ. Networks (2nd edition). Oxford University Press. 2018

Newman, MEJ. Communities, modules and large-scale structure in networks. *Nature Physics* 8, 25–31 (2012)

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

Protein interaction networks

Wu, Z., Liao, Q., Liu, B. A comprehensive review and evaluation of computational methods for identifying protein complexes from protein–protein interaction networks. *Briefings in Bioinformatics*, 21(5), 2020, 1531–1548

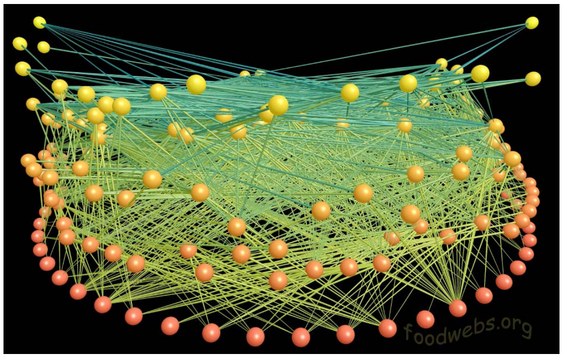
Rajagopala et a., The binary protein-protein interaction landscape of *Escherichia coli. Nature Biotechnology* **32**, *285-290, 2014*

Metabolic networks

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

3

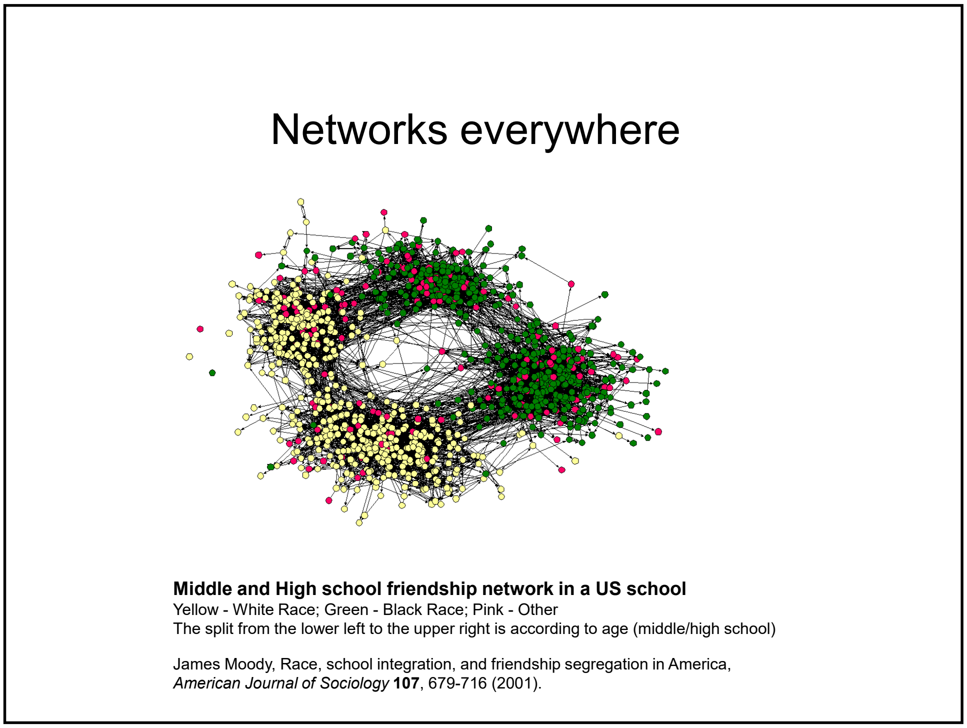
Networks everywhere



El Verde Rainforest trophic web, Puerto Rico

Dunne, J.A., R.J. Williams, and N.D. Martinez. 2002. *Food-web structure and network theory: The role of connectance and size.* PNAS, vol. 99, no. 20, pp. 12917-12922.

4

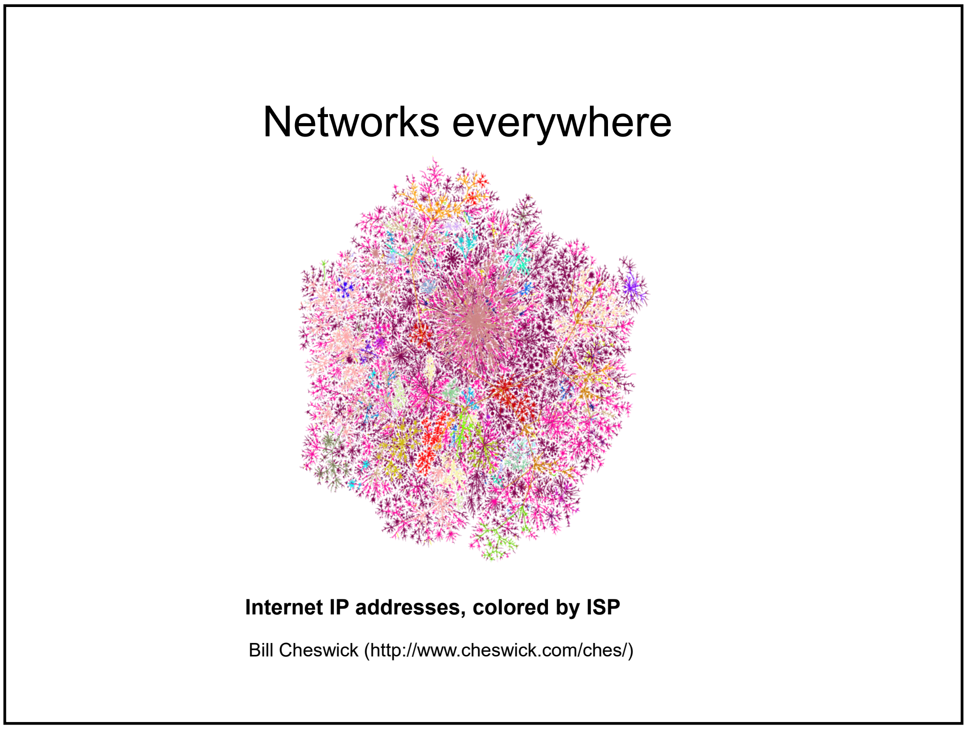


Networks everywhere

Middle and High school friendship network in a US school
Yellow - White Race; Green - Black Race; Pink - Other
The split from the lower left to the upper right is according to age (middle/high school)

James Moody, Race, school integration, and friendship segregation in America,
American Journal of Sociology **107**, 679-716 (2001).

5



Networks everywhere

Internet IP addresses, colored by ISP

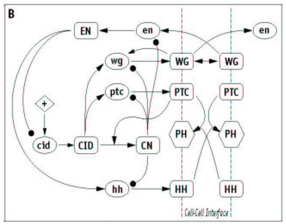
Bill Cheswick (http://www.cheswick.com/ches/)

6

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

Mathematical characterization based on detailed,
quantitative biochemical information

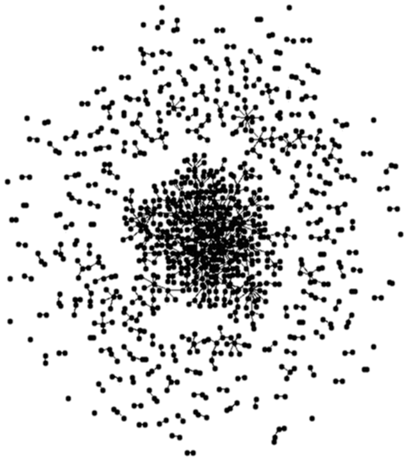
Cell-biological networks

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



experimental data: yeast two-hybrid assay

9

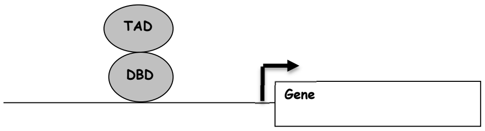
The yeast two–hybrid assay (review)

A technique to identify interacting proteins

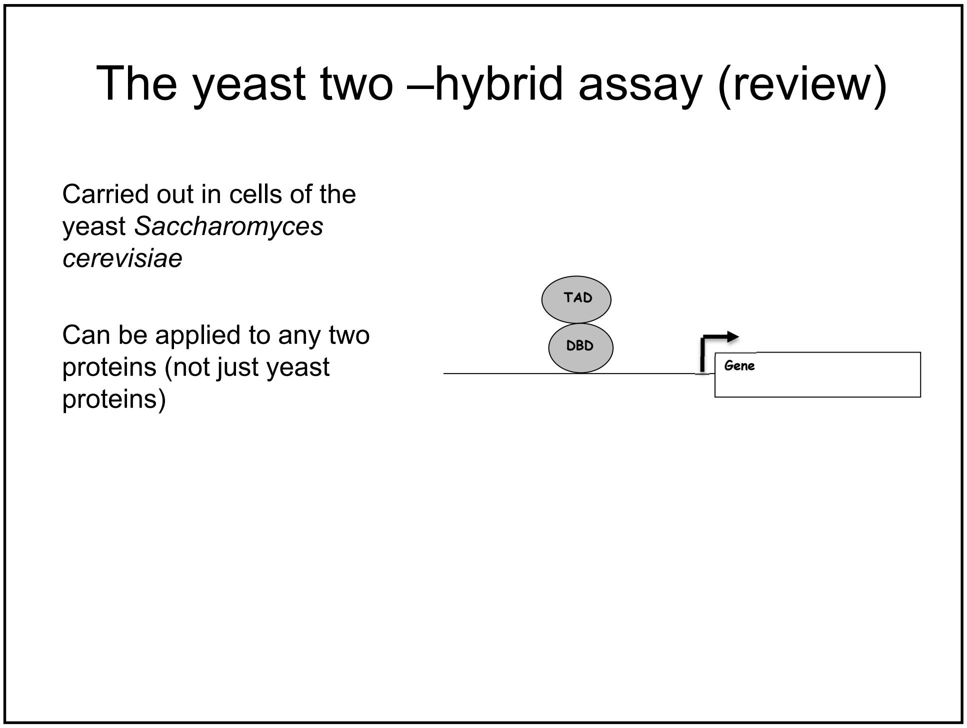
Relies on the modularity of eukaryotic transcriptional regulators

DBD: DNA binding domain

TAD: transcriptional activation domain



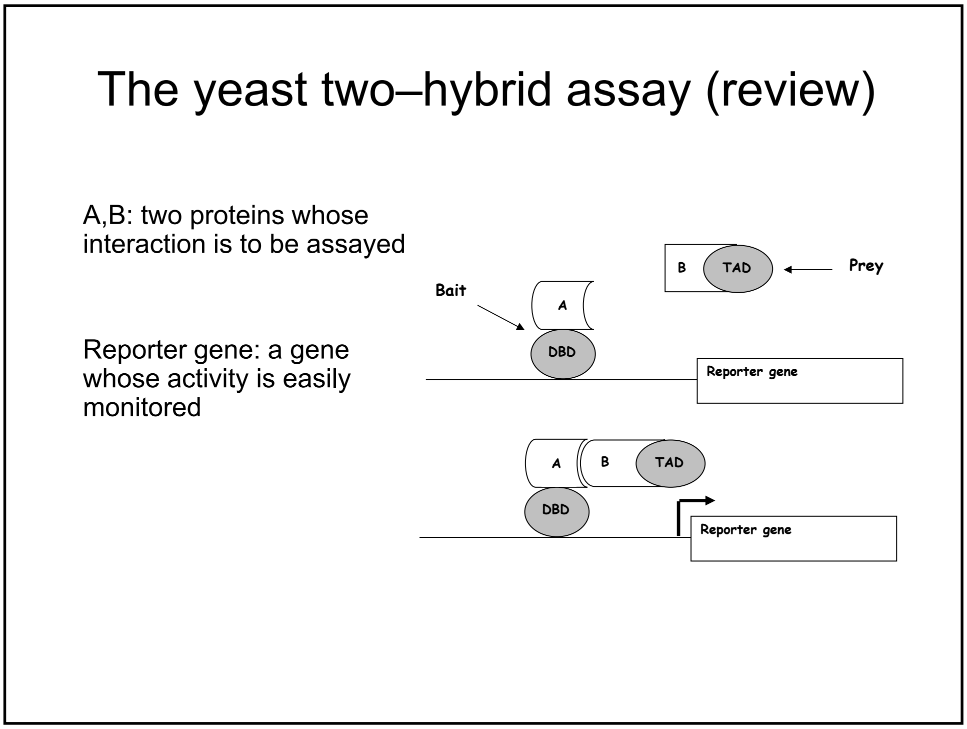
10



The yeast two –hybrid assay (review)

Carried out in cells of the yeast *Saccharomyces cerevisiae*

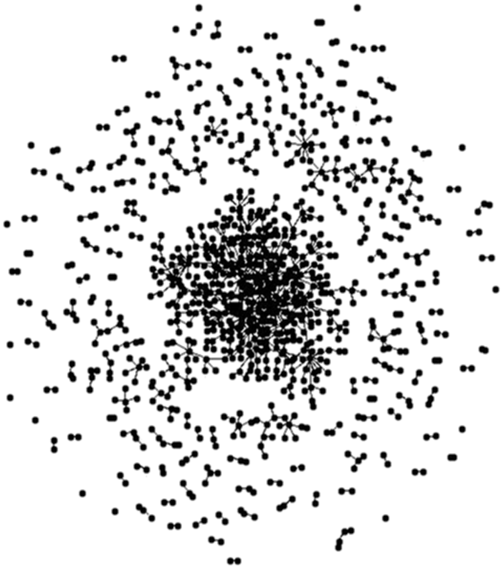
Can be applied to any two proteins (not just yeast proteins)



The yeast two–hybrid assay (review)

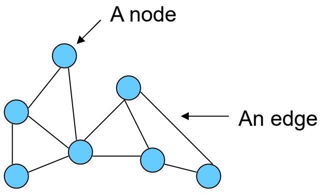
A,B: two proteins whose interaction is to be assayed

Reporter gene: a gene whose activity is easily monitored



13

Graphs

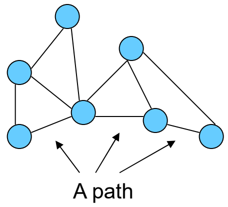


A graph G=(V,E) comprises
a set V of nodes (vertices)
a set E of edges (pairs of nodes)

Protein interaction networks are undirected graphs
Edges are straight lines
Other graphs are directed.
Edges are “arrows”

14

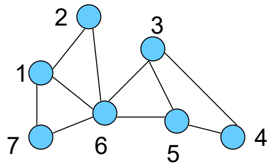
Graphs



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



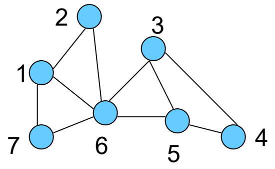
Adjacency matrix $A=(a_{ij})$

a_{ij}=1 $V_{\text{i,}}\ V_{\text{j}}$ connected by an edge

a_{ij}=0 otherwise

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

Undirected graph: A is symmetric
Directed graph: A is asymmetric

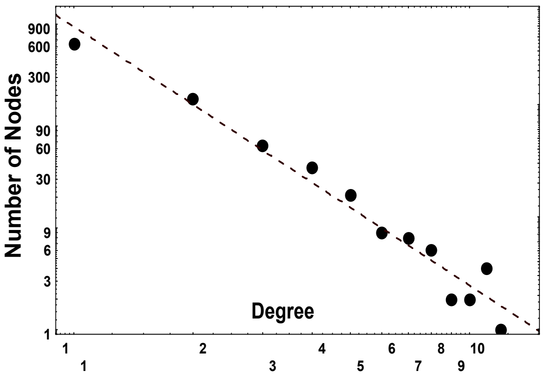


The degree (connectivity) k_i of a node V_i is the number of edges incident with the node (e.g., k_1=3, k_6=5).

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

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

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



Wagner A, Proc. Roy. Soc. London 2003

The best-studied mathematical models of graphs

k-regular graphs

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

Erdős-Rényi random graphs

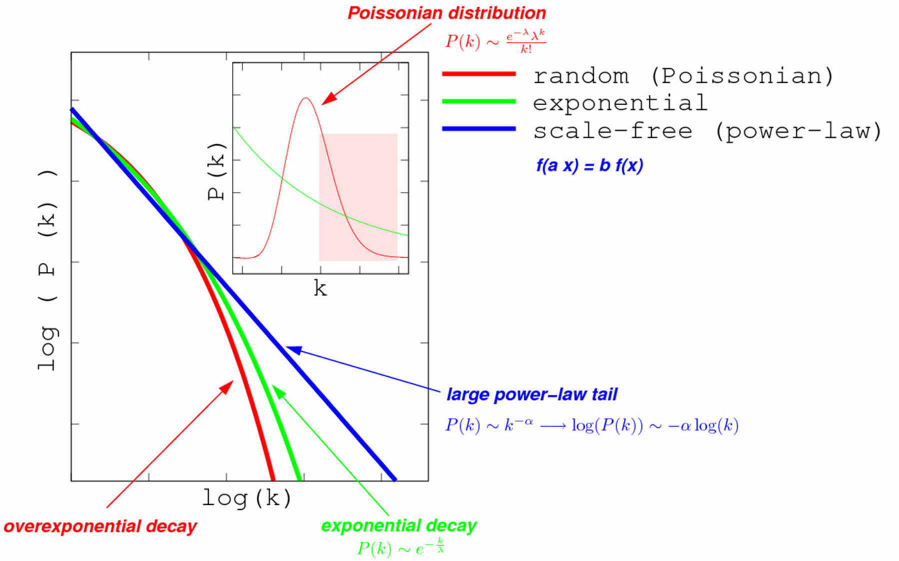
N nodes, K edges

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

Degree distribution is Poisson $P(k) = \exp(-\langle \bar{k} \rangle)\dfrac{\langle \bar{k} \rangle^k}{k!}$

Biological networks are more complex and heterogeneous than predicted by these models

19

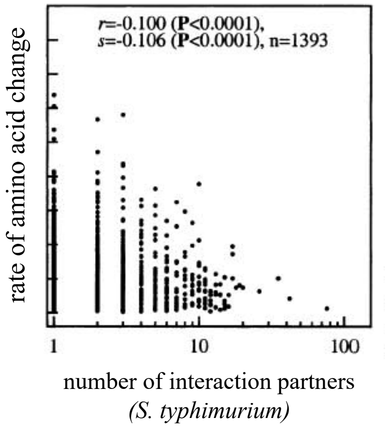


P(k): proportion of nodes with degree k

Figure courtesy of A. Caflisch

20

Highly connected proteins tolerate fewer amino acid substitutions in their evolution

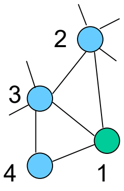


Hahn et al. Journal of Molecular Evolution 2004

21

The degrees of nodes in a graph may be correlated

Average nearest neighbor degree of a node



$k_1=3$

$k_2=5$
$k_3=5$
$k_4=2$

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

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

22

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

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

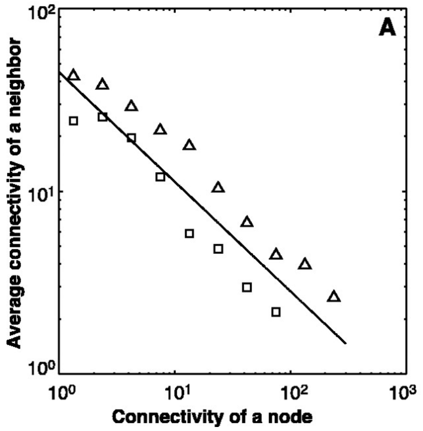
A graph is assortative if $k_{nn}(k)$ increases with k

nodes connect to nodes of similar connectivity

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

23

Protein interaction networks are disassortative



Few interactions between hubs

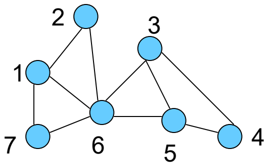
Many interactions between hubs and neighbors with low degree

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

Maslov and Sneppen, Science 2002

24

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: $\max_{i,j} d_{ij}$

Mean (arithmetic) shortest path length
or characteristic path length

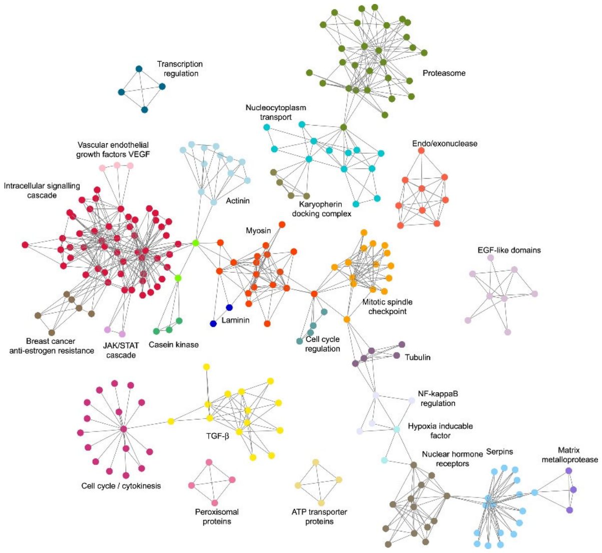
$$L=\frac{1}{N(N-1)}\sum_{i,j,i\neq j} d_{ij}$$

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)

Many graphs can be subdivided into “communities”

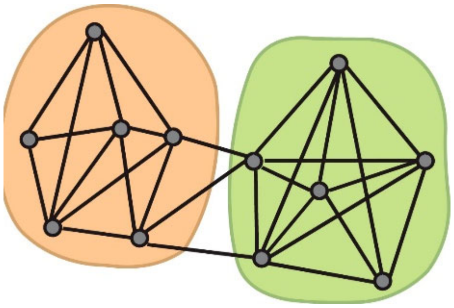


Community structure of a rat protein interaction network

Fortunato and Hric, Physics Reports 659, 1-44, 2016

Many graphs can be subdivided into “communities”

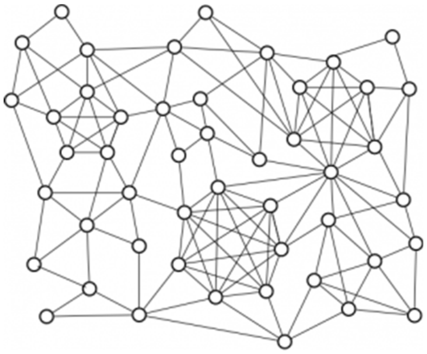
In a graph that can be subdivided into communities (clusters, modules) nodes fall into groups that share more edges with each other than with nodes outside the community



Fortunato and Hric, Physics Reports 659, 1-44, 2016

The most densely connected communities are cliques

clique: a largest complete (=fully connected) subgraph



A graph with multiple cliques

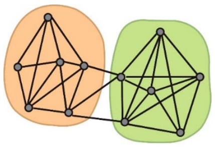
http://skipperkongen.dk/2010/11/

29

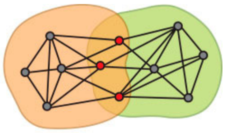
Many computational methods aim to detect communities in networks

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)



30

Optimization methods for community detection aim to maximize a quantity that indicates to what extent a network clusters into different communities

A very popular such quantity

Modularity Q for a network that is subdivided into *n* modules

$$Q = \sum_{i=1\ldots n} (e_{ii} - a_i^2)$$

e_{ij}…fraction of edges that connect nodes in module *i* and module *j*

e_{ii}…fraction of edges that connect nodes within module *i.*

$a_i = \sum_{j=1\ldots n} e_{ij}$ fraction of edges that begin or end in module *i*

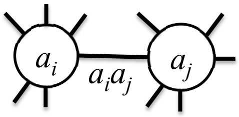
31

Optimization methods for community detection aim to maximize a quantity that indicates to what extent a network clusters into different communities

A very popular such measure

Modularity Q $$Q = \sum_{i=1\ldots n} (e_{ii} - a_i^2)$$

If you have subdivided a graph into *n* putative modules, but these modules do not reflect the graph's actual structure, then the fraction of edges that connect two such "spurious" modules *i* and *j* is given by the product rule of probabilities as $e_{ij}=a_i a_j$.



A special case is $e_{ii}=a_i^2$

Thus, if a graph does not have a modular structure, then Q ≈ 0.

32

Optimization methods aim to maximize a quantity that indicates to what extent the network clusters into different communities

A very popular such measure

Modularity Q

$$Q = \sum_{i=1\ldots n} (e_{ii} - a_i^2)$$

Q is larger for graphs and communities in which pairs of connected nodes tend to reside in the same module

Q≈1 for graphs with the most pronounced modular structure

This occurs if all values of e_{ii} are large, i.e., almost all edges connect nodes within the same module, while a_i^2 is small, i.e., by chance alone one would expect that very few edges connect nodes within modules

33

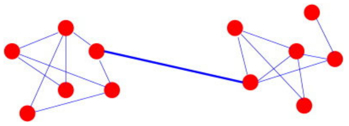
The Girvan-Newman algorithm is a popular heuristic to cluster large graphs

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*

34

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

It is an iterative divisive clustering algorithm

Idea: Edges between modules are those with the highest edge betweenness

Remove those edges and you get good module separation

Procedure

1. Remove the edge with the highest betweenness
2. Recalculate edge betweenness for the now-reduced graph

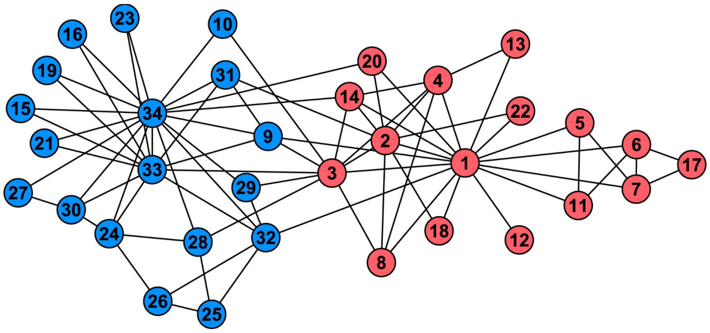
(3. Determine modularity Q)

4. Back to one until all nodes are isolated

The optimal partition is that with the highest Q

35

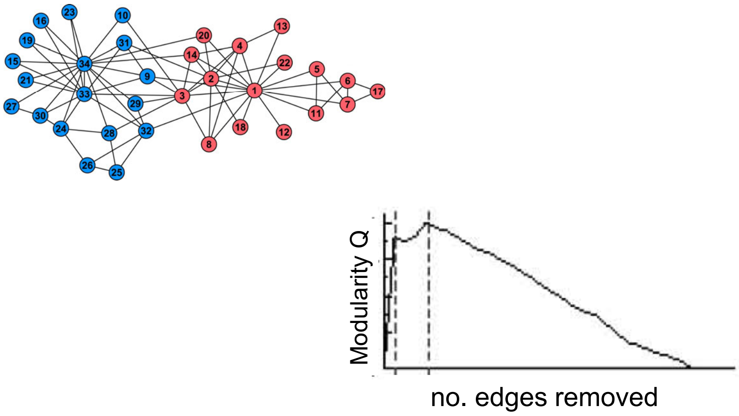
The “Karate club” network of Zachary has served as a benchmark for many community detection algorithms



Fortunato and Hric, Physics Reports 659, 1-44, 2016

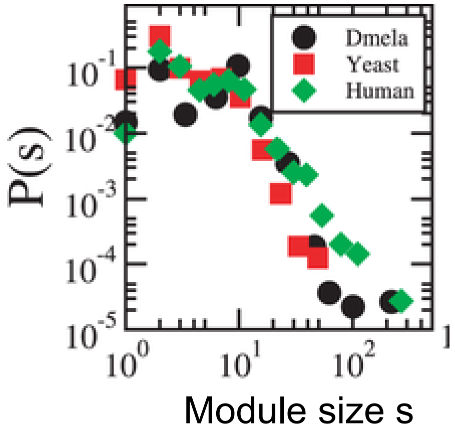
36

The “Karate club” network of Zachary has served as a benchmark for many community detection algorithms

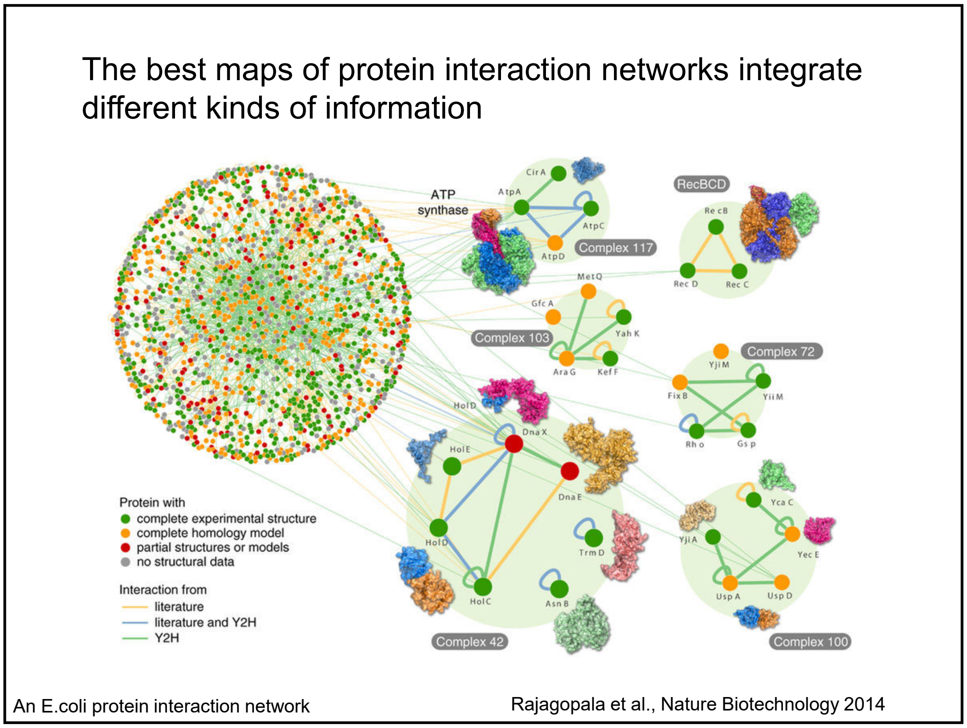


Boccaletti et al. 2006

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



Lancichinetti et al. (2010) Characterizing the Community Structure of Complex Networks. PLOS ONE 5(8): e11976.



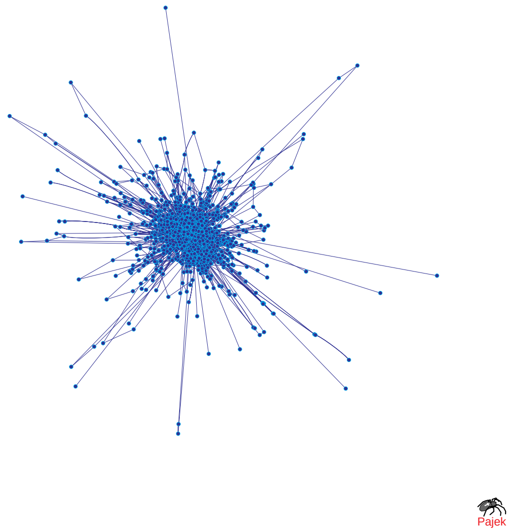
The best maps of protein interaction networks integrate different kinds of information

An E.coli protein interaction network

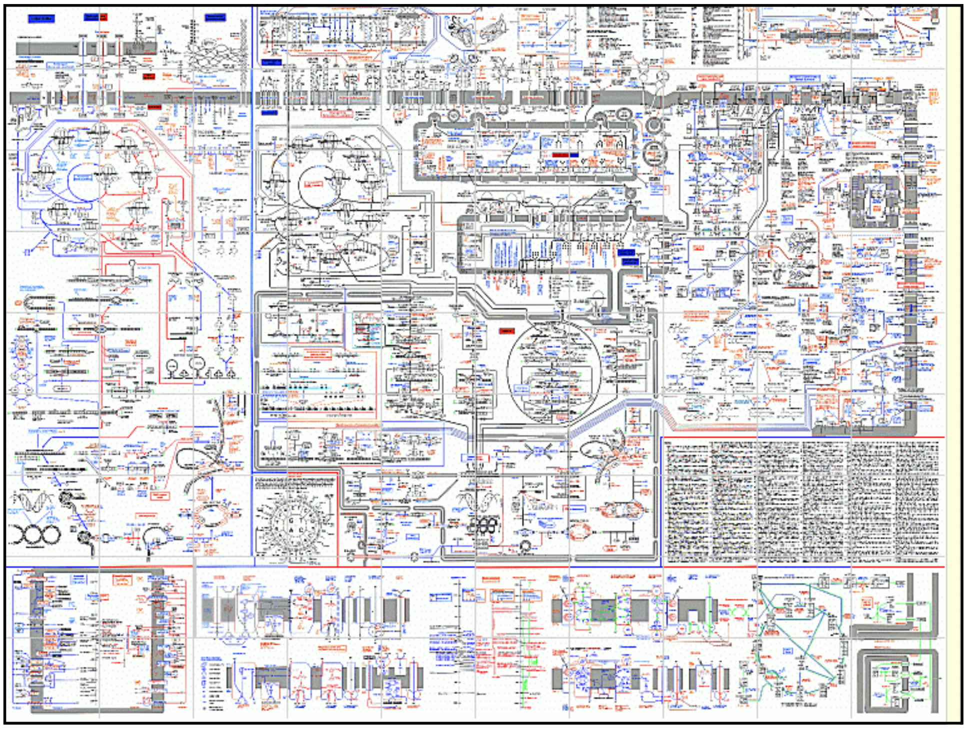
Rajagopala et al., Nature Biotechnology 2014

39

Metabolic networks



40



41

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.

42

Graphs can (crudely) represent large chemical reaction networks

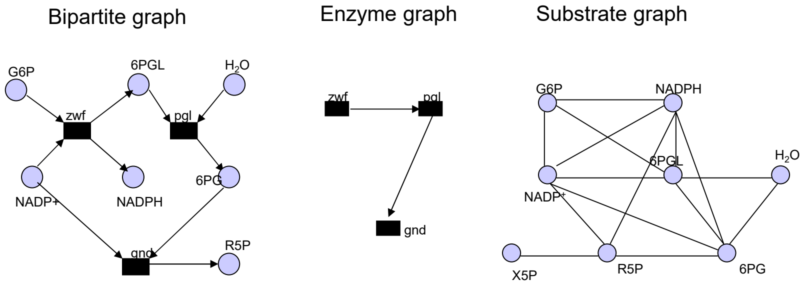
Stoichiometric Equations

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

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

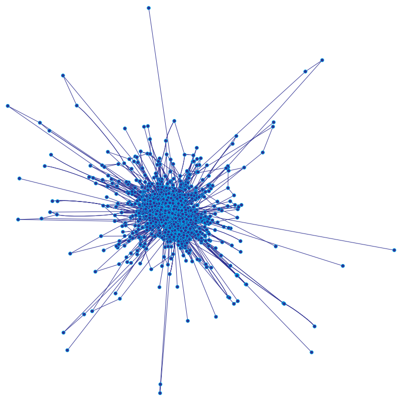
1 6-Phosphogluconate + 1 NADP$^+$ $\overset{gnd}{\Rightarrow}$ 1 Ribulose 5-phosphate (R5P) + 1 NADPH

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



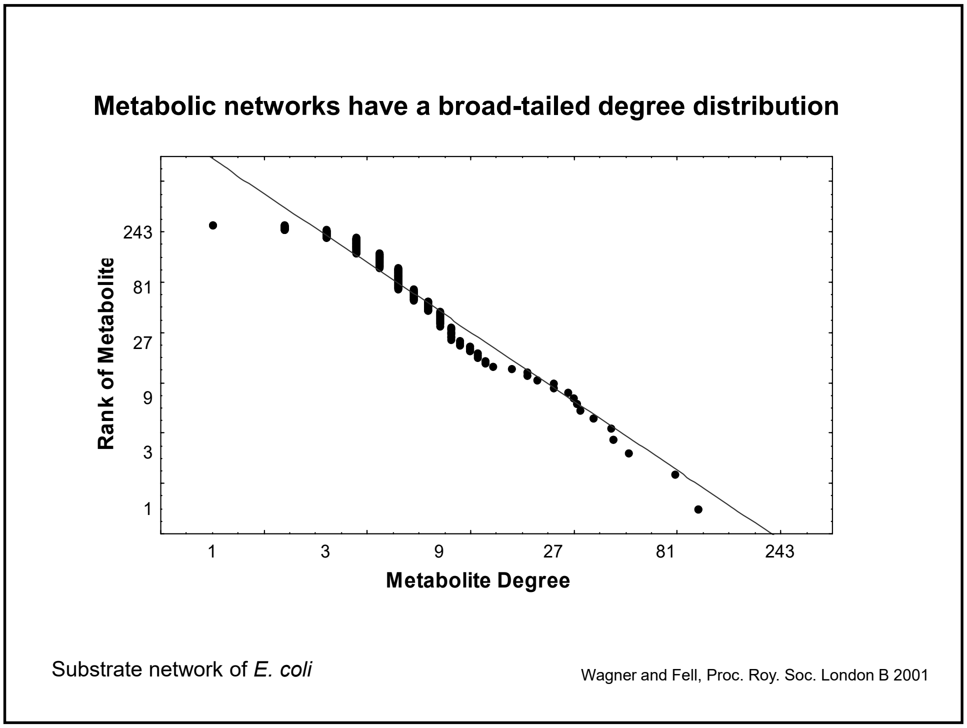
43

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



Pajek

44

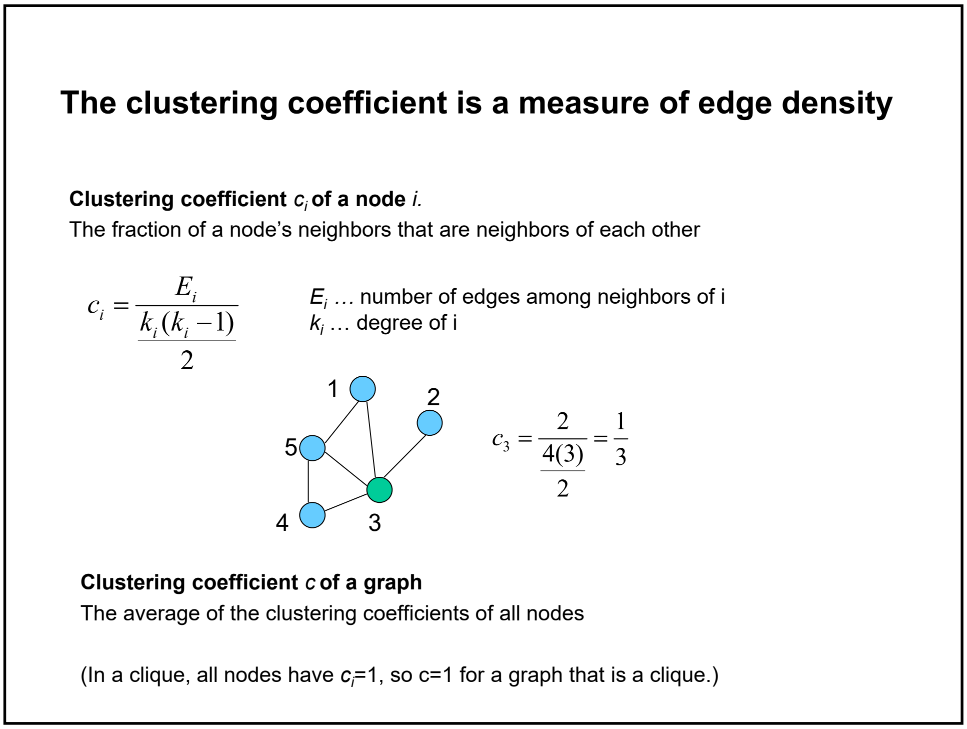


Metabolic networks have a broad-tailed degree distribution

Substrate network of *E. coli*

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

45



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.)

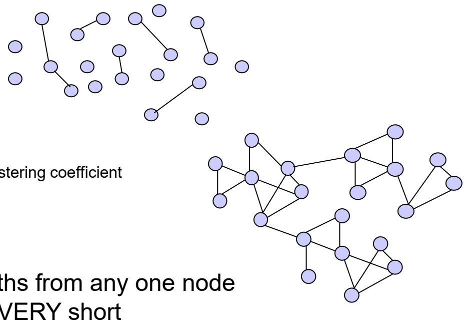
46

Key features of small-world graphs

1. They are sparse

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, Nature 1998

47

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

It is sparse

It is highly clustered

It has short characteristic path length

48

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

49

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.

50

Studying only the structure of metabolic networks neglects their function

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

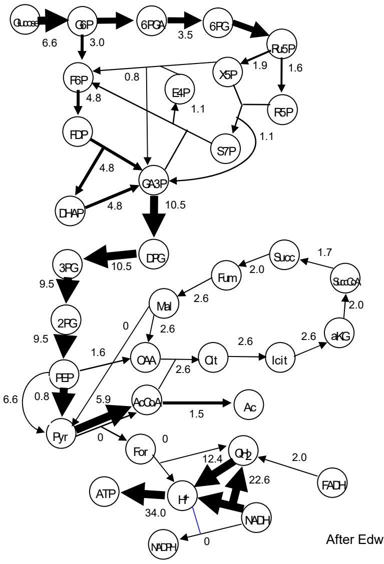
For optimal cell growth, metabolic networks need to produce biochemical precursors in well-balanced amounts.

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.)

51

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



After Edwards JS, Palsson BO. 2000. *PNAS* 97: 5528-33

52

Flux balance analysis (FBA)

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

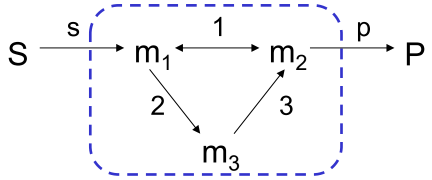
(For example, in yeast
>1100 reactions,
>500 metabolites,
>100 nutrients or waste products.)

FBA 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 m_i change according to the equations

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

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

$$\frac{dm_3}{dt} = v_2 - v_3$$

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

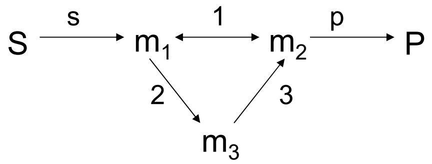
$$\mathbf{S} = \begin{pmatrix} 1 & -1 & -1 & 0 & 0 \\ 0 & 1 & 0 & 1 & -1 \\ 0 & 0 & 1 & -1 & 0 \end{pmatrix}$$

Stoichiometry matrix

Rows: metabolites
Columns: reactions

v_i metabolic flux through reaction i

$$\vec{v} = (v_S, v_1, v_2, v_3, v_P)^{\mathsf{T}}$$



FBA assumes that metabolism is in a steady state where the concentrations of metabolites no longer change

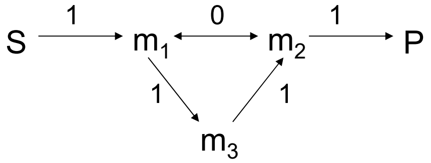
$$\frac{d\vec{m}}{dt} = 0$$

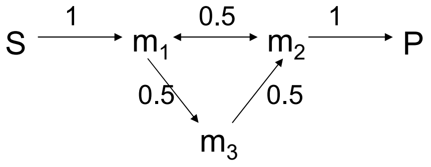
$$\mathbf{S}\vec{v} = 0$$

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

55

Two allowable flux distributions for our example network

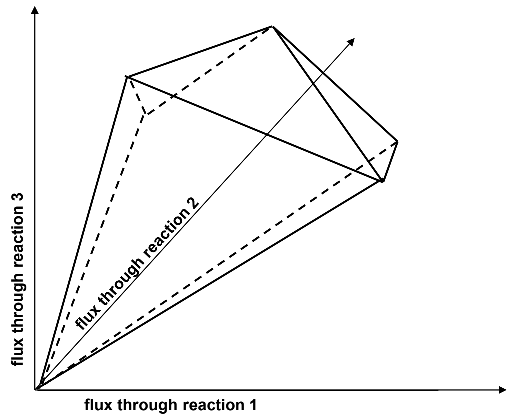




All fluxes of the form (1,x,1-x,1-x,1), 0≤x≤1 are allowable

56

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



57

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:

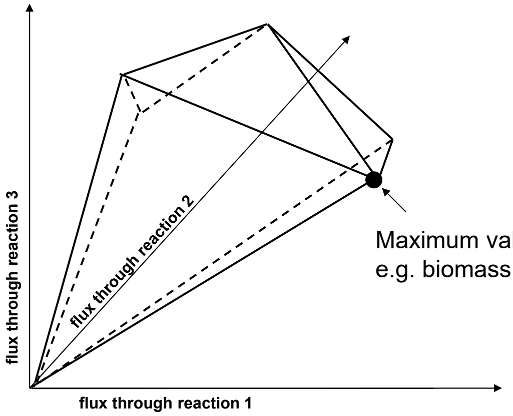
In the biomass growth flux ,

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.)

58

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

59

Example questions for flux balance analysis

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

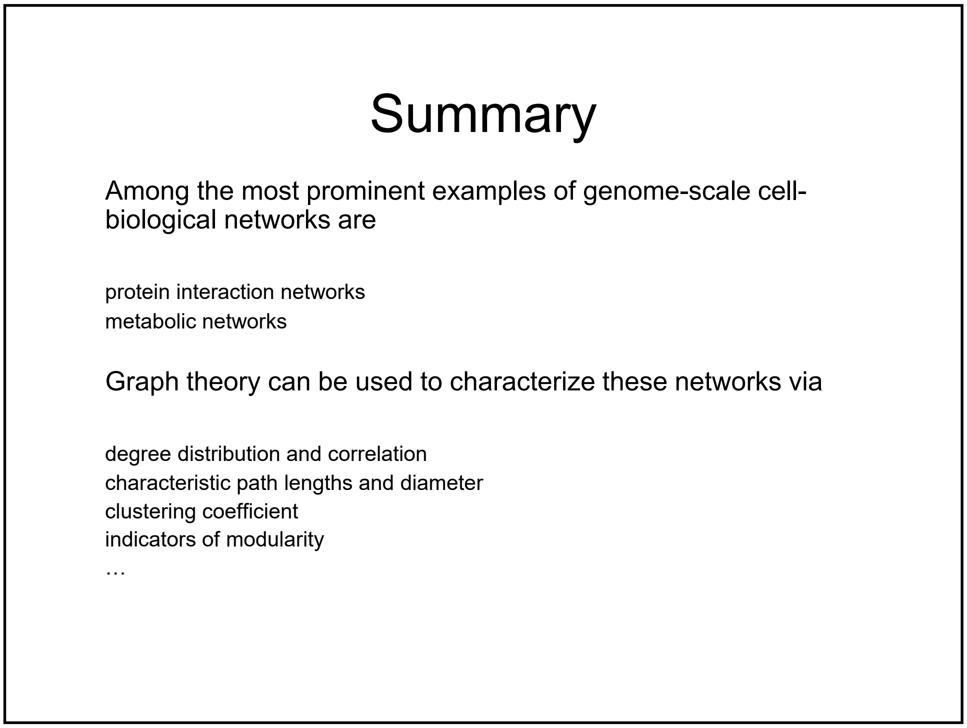
How fast could it grow in this environment?

Why are many enzymatic reactions dispensable in any one environment?

Does network function and flux influence network evolution?

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

60



Summary

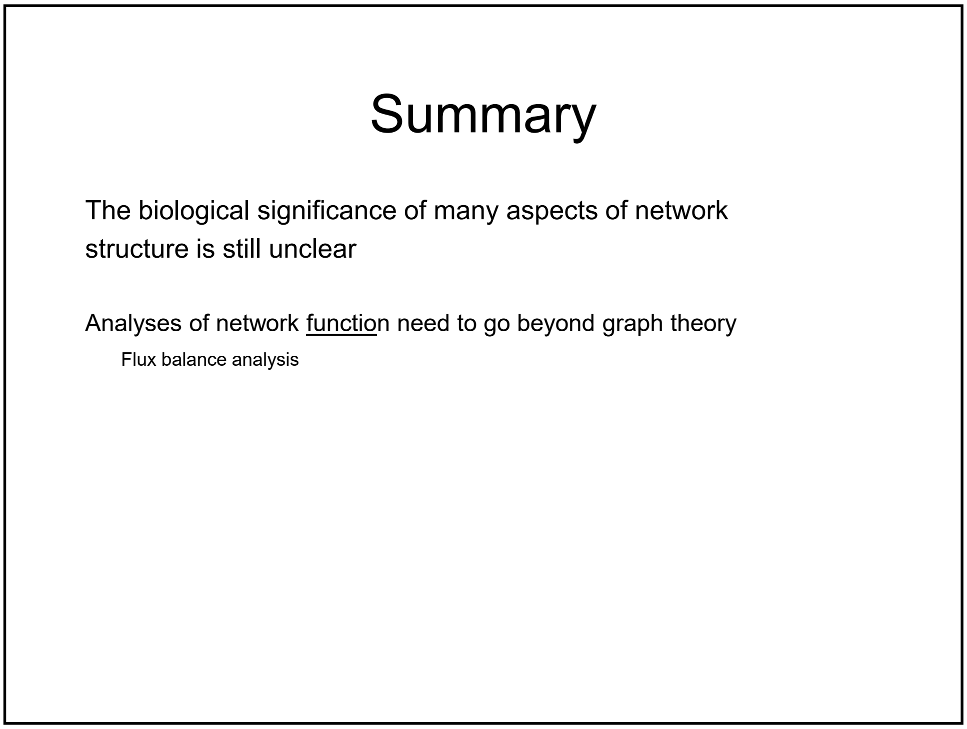
Among the most prominent examples of genome-scale cell-biological networks are

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
…

61



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

62