Computational Systems Biology

Reconstruction and structural analysis of metabolic networks

Hongwu Ma

hma2@inf.ed.ac.uk

Computational Systems Biology Group

12 February 2010

Simple models

An enzyme reaction

$$A \xrightarrow{E} B$$

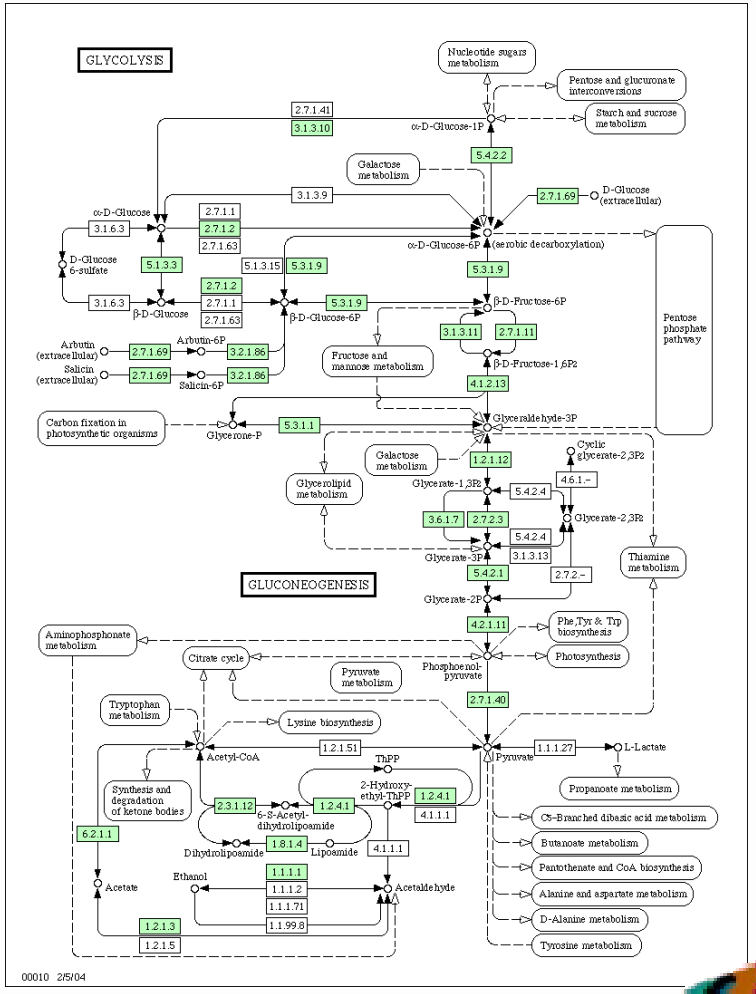
Enzyme kinetics

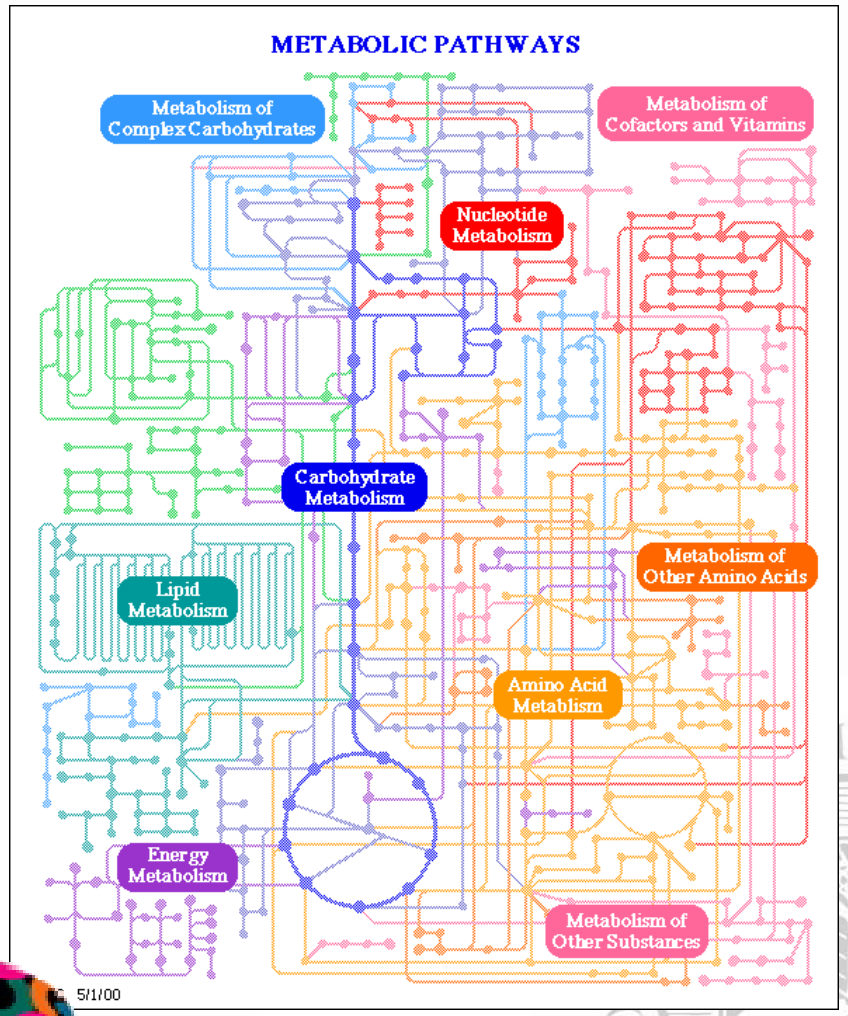
A linear pathway

$$X_0 \xrightarrow{E_1} S_1 \xrightarrow{E_2} S_2 \xrightarrow{E_3} X_1$$

Metabolic control analysis

The reality





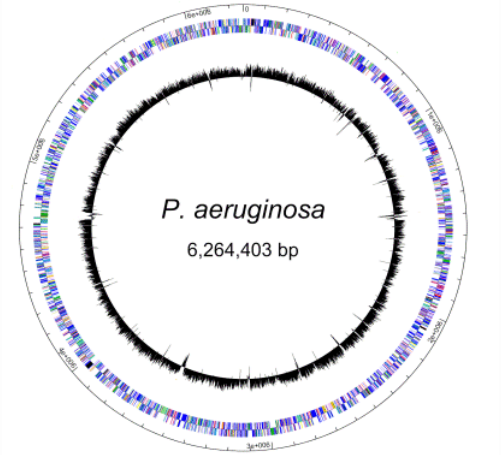
Glycolysis pathway

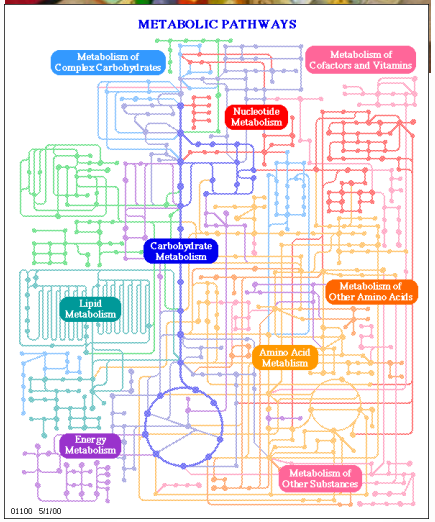
KEGG Kyoto Encyclopedia of Genes and Genomes

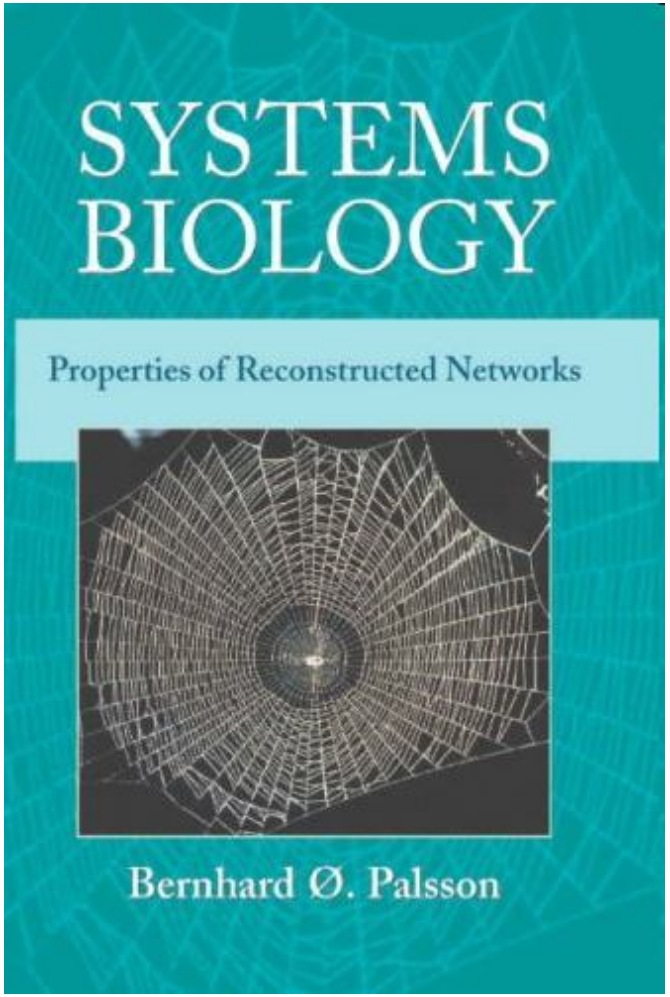
The whole network

Systems Biology: Network Reconstruction

- From part list to part organization
- Understanding the organization of the biosystem






SYSTEMS
BIOLOGY
Properties of Reconstructed Networks
Bernhard Ø. Palsson

Network reconstruction: the background

- Fully sequenced genomes for several hundreds organisms
- Bioinformatics methods for gene function annotation (2D annotation)
- Public available internet database for storage and management of the annotation data

Biological networks

- Metabolic network, well studied, based on enzyme annotation (easy to reconstruct)
- Gene regulatory network: based on literature, computational prediction (motif) or ChIP on chip)
- Signal transduction network, based on literature
- Protein-protein interaction network: yeast 2 hybrid experiments, low reliability

Genome based MN reconstruction

Genome sequence

CTGAGGTCGACTCTAGAGGATCCCCCTTCCAGATGTGTAAGTTA
TTTGAGTGAAATAGTGCCAGTTTAACCATAGTTCTAGTAAGCT
TTTGAGTGAAATAGTGCCAGTTTAACCATAGTTCTAGTAAGCTG

ORF

Gene recognition

GeneMark, Glimmer

Sequence similarity search, Blast

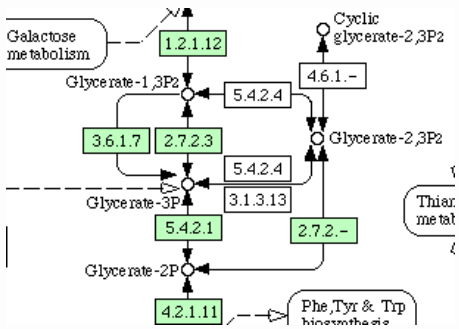
Database

Knowledge transfer

Gene annotation

ID	Name	start	end	function
EG11277	thrL	190	255	Regulatory leader peptide for thrABC oper
EG10998	thrA	337	2799	Aspartokinase I-homoserine dehydrogenas
EG10999	thrB	2801	3733	Homoserine kinase [EC 2.7.1.39]
EG11000	thrC	3734	5020	Threonine synthase [EC 4.2.3.1]

Network



Gene & Protein databases

- NCBI GenBank: www.ncbi.nlm.nih.gov
- EBI EMBL: http://www.ebi.ac.uk/embl/
- EBI UniProt: http://www.uniprot.org
- Organism specific databases such as EcoGene for E. Coli, SubtiList for B. Subtilis, et

Many of them do not contain reaction and even EC number information, make it difficult for MN reconstruction

More information on database: Nucleic Acid Research, Database issue (published in January every year)

More specific: metabolic and enzyme databases

- KEGG: http://www.genome.ad.jp/kegg/
- Biocyc: http://biocyc.org/
- Expasy: http://www.expasy.ch
- Brenda: http://www.brenda.uni-koeln.de/

Function: Link the enzyme with metabolic reactions and metabolites. KEGG is especially useful as it contains standardized reaction and compound information.

Ability required: writing program to extract information from websites (web service) or downloadable text files.

Example: KEGG Enzyme information

http://www.genome.ad.jp/dbget-bin/www_bget?enzyme+1.1.1.81

ENTRY	**EC 1.1.1.81**
NAME	**Hydroxypyruvate reductase**
CLASS	**Oxidoreductases**
	Acting on the CH-OH group of donors
	With NAD+ or NADP+ as acceptor
SYSNAME	**D-Glycerate:NADP+ 2-oxidoreductase**
REACTION	**D-Glycerate + NAD+ or NADP+ = Hydroxypyruvate + NADH or NADPH**
PATHWAY	**PATH: MAP00260 Glycine, serine and threonine metabolism**
	PATH: MAP00630 Glyoxylate and dicarboxylate metabolism
GENES	**YPE: YPO2536**
	PAE: PA1499
	RSO: RS03094(ttuD1) RS05749(ttuD2)
	MLO: mlr5146
	SME: SMa1406(ttuD3) SMb20678(ttuD2) SMc04389(ttuD1)
	ATU: Atu3232 Atu5334

KEGG Reaction database

From downloaded text file to database

More than 8000 reactions

.....

ENTRY R02739

NAME alpha-D-Glucose 6-phosphate ketol-isomerase

DEFINITION alpha-D-Glucose 6-phosphate <=> beta-D-Glucose 6-phosphate

EQUATION C00668 <=> C01172

ENZYME 5.3.1.9 5.1.3.15

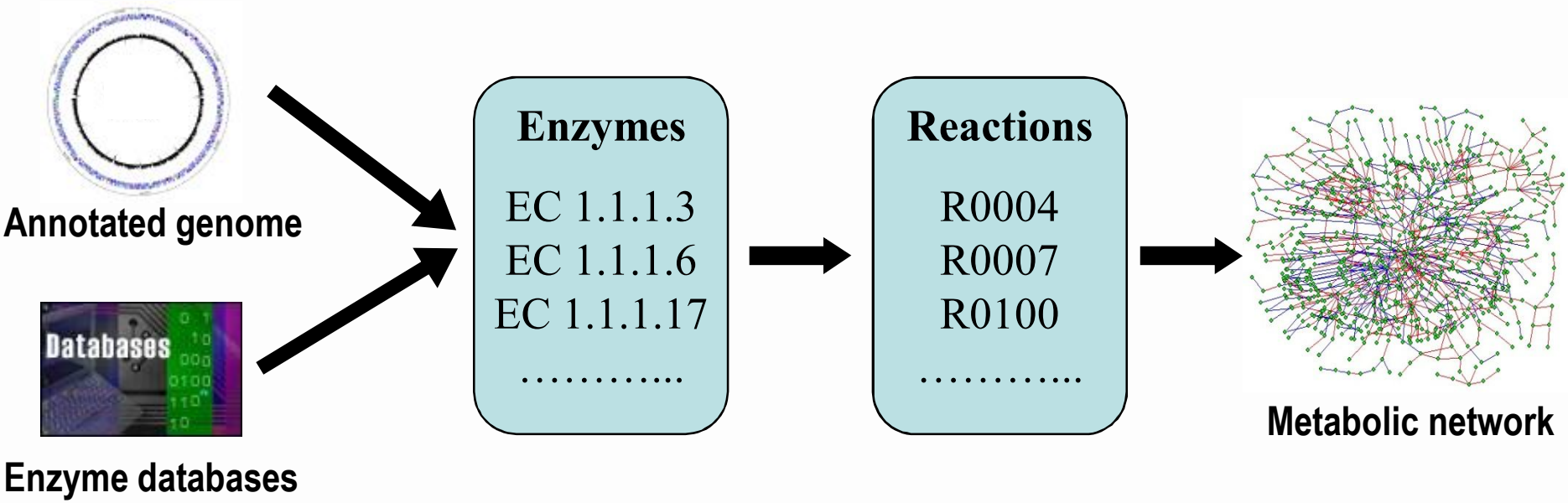
.....

J19 =

	A	B	C	D	E	F	G
1	R00001		L-Methionine -	C00073 + C0034	1	3.6.1.10	
2	R00002		16 ATP + 16 H	16 C00002 + 16	1	1.18.6.1	
3	R00004	1	Pyrophosphat	C00013 + C0000	1	3.6.1.1	
4	R00005	1	Urea-1-carbox	C01010 + C0000	1	3.5.1.54	
5	R00006	1	2 Pyruvate = 2	2 C00022 = C00	1	4.1.3.18	
6	R00007		4-Hydroxy-4-m	C04184 = 2 C00	1	4.1.3.17	
7	R00008	1	Parapyruvate =	C06033 = 2 C00	1	4.1.3.17	
8	R00009		2 H2O2 = Oxy	2 C00027 = C00	1	1.11.1.6	
9	R00010	1	alpha,alpha-Tr	C01083 + C0000	1	3.2.1.28	
10	R00011		2 Manganese	2 C00034 + 2 C0	1	1.11.1.13	
11	R00012		2 GTP = Pyro	2 C00044 = C00	1	2.7.7.45	
12	R00013	1	2 Glyoxylate =	2 C00048 = C01	1	4.1.1.47	
13	R00014	1	2-(alpha-Hydro	C00068 + C0002	2	4.1.3.18	1.2.4.1
14	R00015		2 Sucrose = D	2 C00089 = C00	1	2.4.1.99	
15	R00016		2 D-Glucose 1	2 C00103 = C00	1	2.7.1.41	
16	R00017		H2O2 + 2 Ferr	C00027 + 2 C00	1	1.11.1.5	
17	R00018	1	2 Putrescine =	2 C00134 = C06	1	2.5.1.44	
18	R00019	1	2 Reduced fer	2 C00138 + 2 C0	1	1.18.99.1	

irreversibility

The reconstruction process



KEGG
Brenda

KNEVA: web tool for MN reconstruction

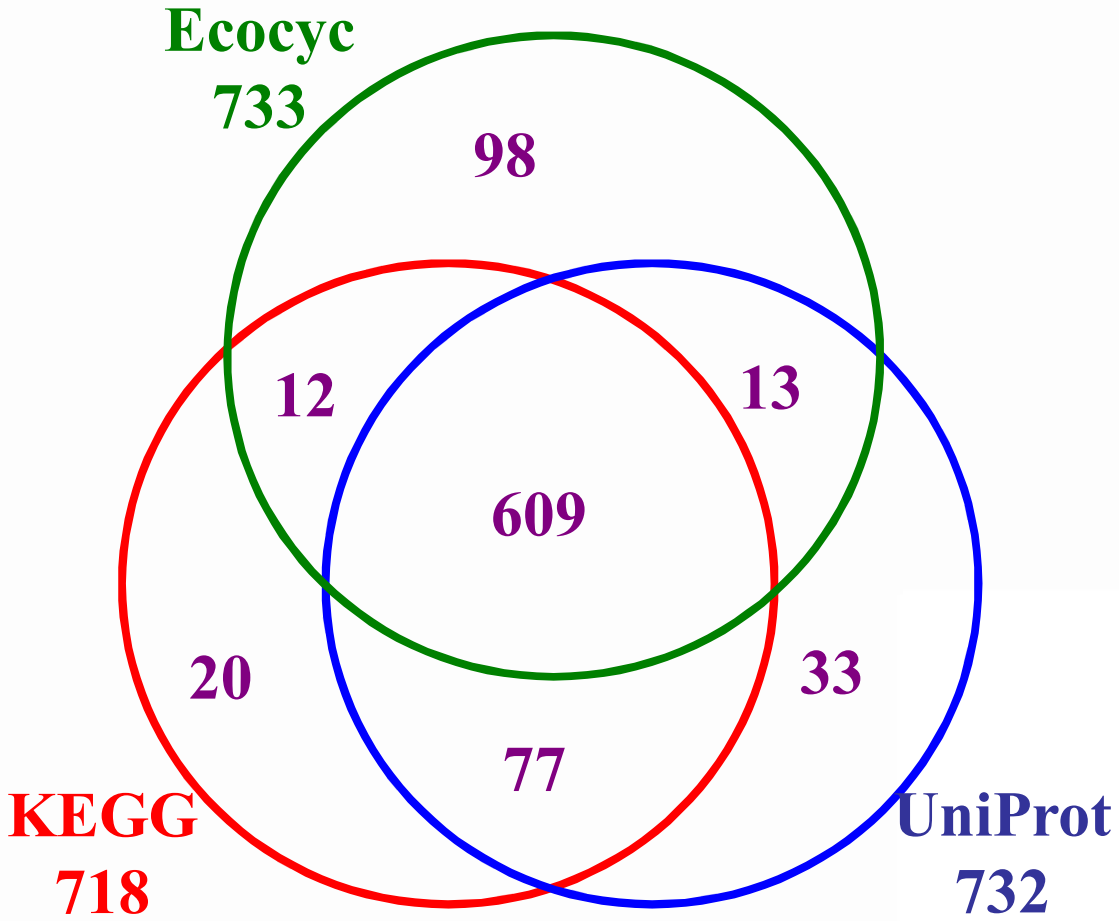
Csb.inf.ed.ac.uk/kneva

- Get a network for an organism or multiple organisms by typing the KEGG organism abbreviation
- Reconstruction based on EC and/or KO
- Network reconstruction from submitted ECs, KOs and reactions
- Subnetwork reconstruction for selected pathways.

Problems in the high throughput methods

- Incomplete and inconsistent annotation and EC assignment
- Nonenzyme catalyzed reactions link
- Unclear enzymes like 1.-.-.-
- Unsequensed enzymes (gene for a known enzyme not identified, more than 40 in *E. coli*) link
- Non organism specific enzyme-reaction relationships for unsepcific enzymes (ex, unspecific monooxygenase 1.14.14.1 and alcohol dehydrogenase 1.1.1.1)

E. Coli enzyme annotation in different databases



High quality network reconstruction

Annotated genome

Enzyme databases

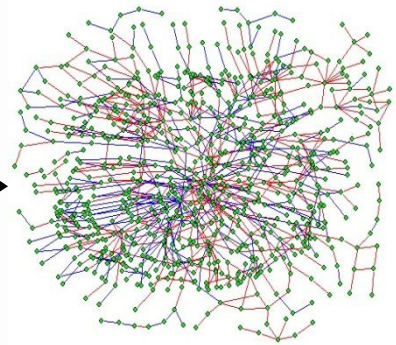
KEGG
Expasy
Brenda

Enzymes

EC 1.1.1.3
EC 1.1.1.6
EC 1.1.1.17
……..

Reactions

R0004
R0007
R0100
……..



Metabolic network

Human curation and examination

Available high quality networks

E. coli: Ecocyc, Palsson's group (at least 3 different versions)

Yeast: Palsson and Nielsen's group (YSBN)

Helicobacter pylori, Haemophilus influenzae

Bacillus subtilis, Mycobacterium tuberculosis

Human: EHMN and Human Recon 1 (several man-years)

www.ehmn.bioinformatics.ed.ac.uk

Comparing networks reconstructed from different groups is often difficult due to compound synonyms

What you actually get after reconstruction

A list of reactions which make up of the network (FBA analysis)

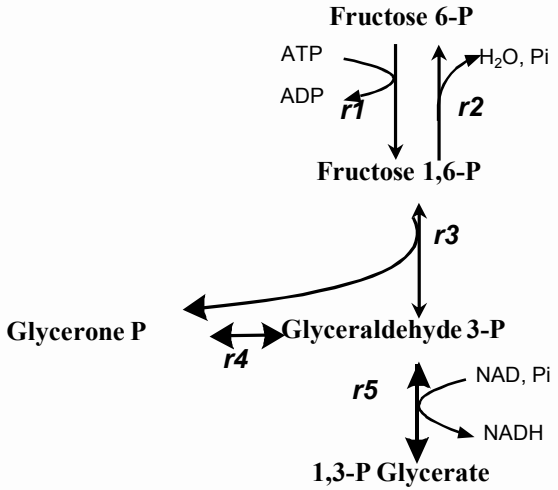
A list of metabolites in the network

Reaction-gene relationships (may link through enzymes)

You get the data, but not the picture.

Visual and mathematical representation of metabolic networks

Mathematical representation of MN



$$\begin{array}{c} \\ F6P \\ FDP \\ T3P1 \\ T3P2 \\ 13PG \end{array} \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

$$\begin{array}{c} \\ r_1 \\ r_2 \\ r_3 \\ r_4 \\ r_5 \end{array} \begin{array}{c} \begin{array}{cccccccccccc} F6P & FDP & T3P1 & T3P2 & ATP & 13PG & ADP & NADH & NAD & Pi & H_2O \end{array} \\ \begin{pmatrix} -1 & 1 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & -1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 1 & 0 & 0 & 1 & -1 & -1 & 0 \end{pmatrix} \end{array}$$

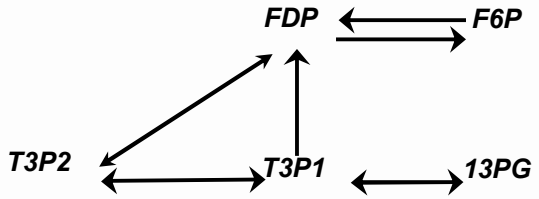
Stoichiometric matrix

$$\begin{array}{c} \\ r_1 \\ r_2 \\ r_3 \\ r_4 \\ r_5 \end{array} \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{pmatrix}$$

Connectivity matrix

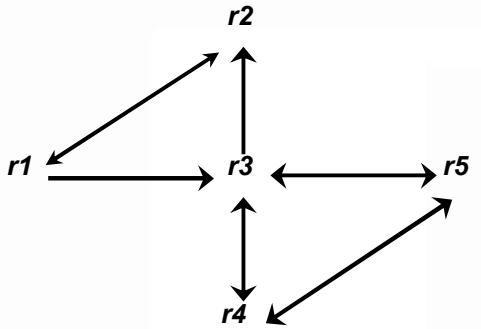
Connectivity matrix to graph

$$\begin{array}{c} \\ F6P \\ FDP \\ T3P1 \\ T3P2 \\ 13PG \end{array} \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$



Metabolite graph

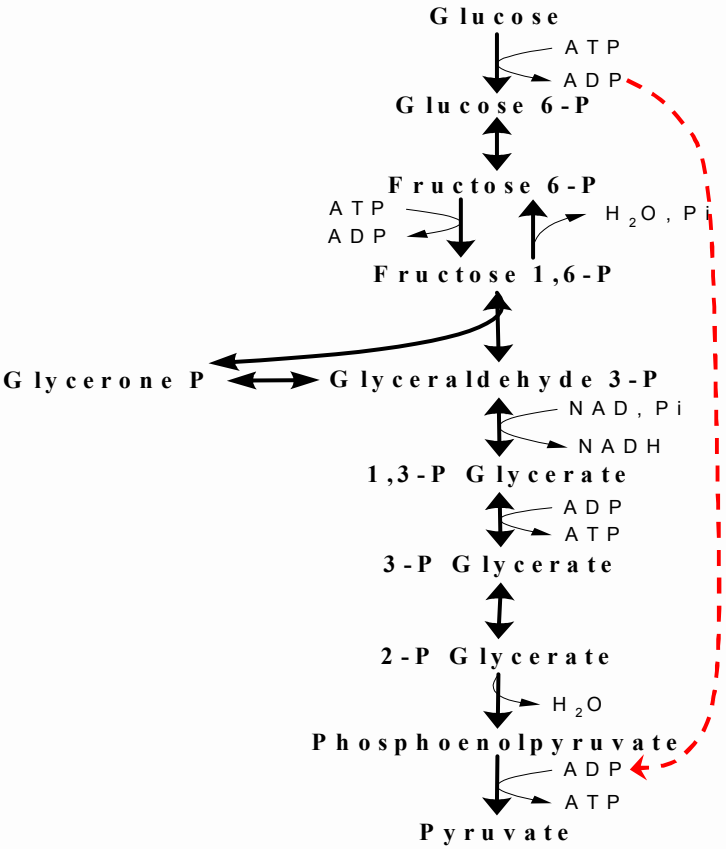
$$\begin{array}{c} \\ r_1 \\ r_2 \\ r_3 \\ r_4 \\ r_5 \end{array} \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{pmatrix}$$

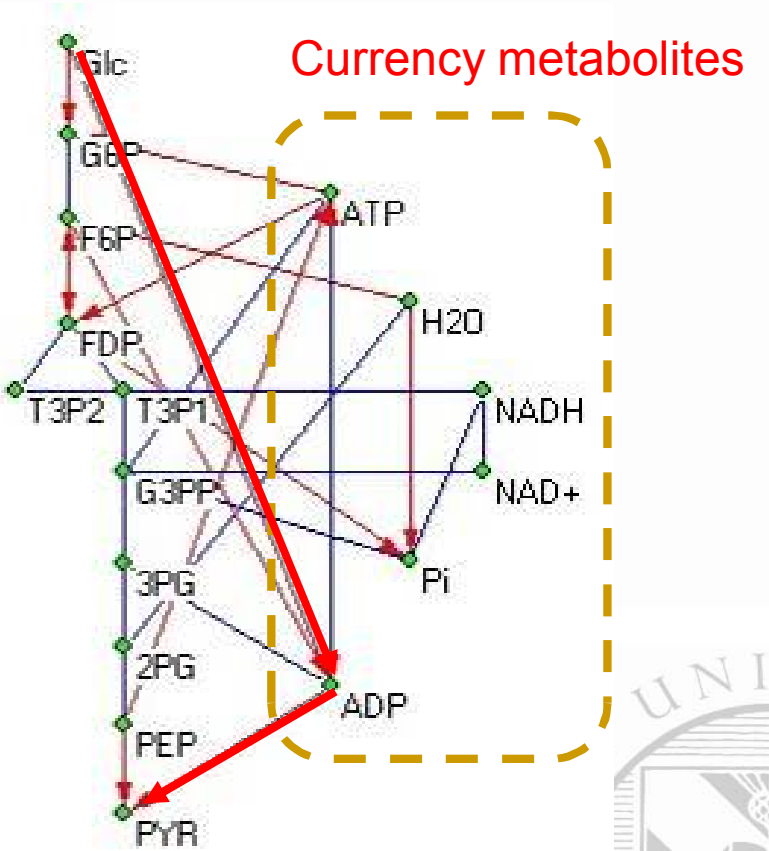


Connectivity (Adjacency) matrix

Reaction graph

Currency metabolites in graph analysis

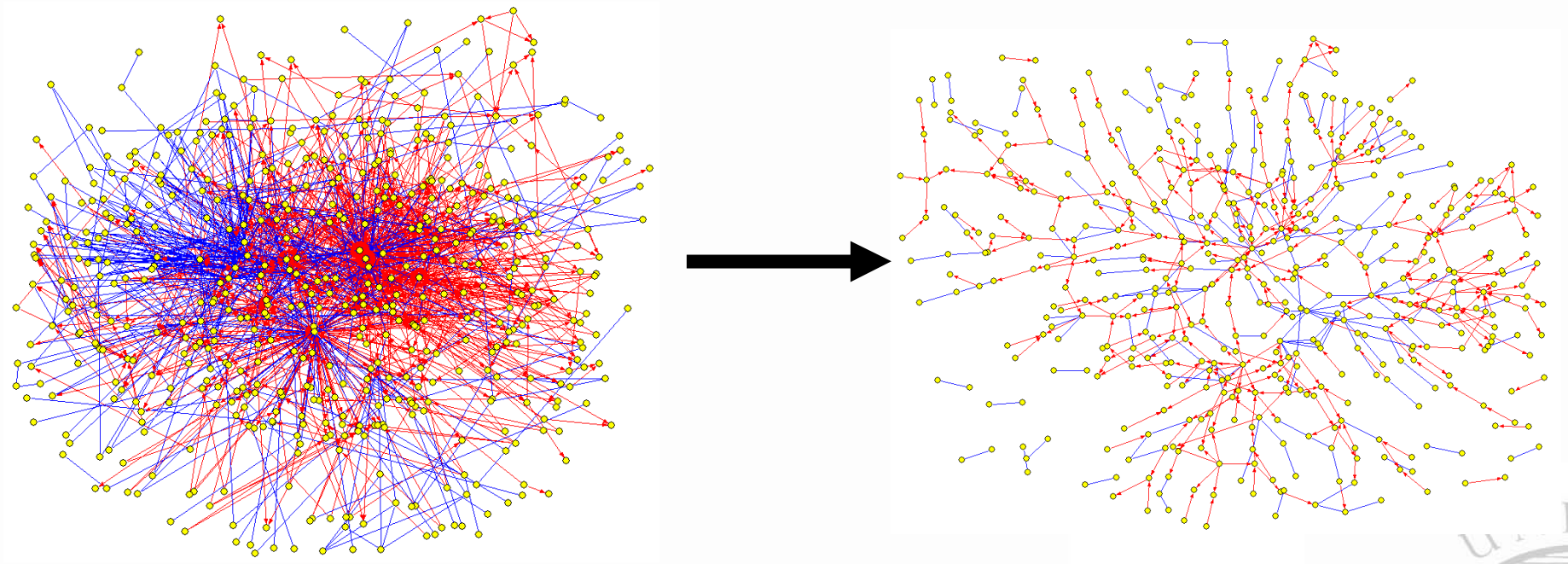




From glucose to pyruvate, ADP can not be used as a link.

Otherwise path length will be 2 instead of 9
(Jeong et al. 2000 Nature 407:651)

With or without currency metabolites



Metabolic network of *S. pneumonia* (616 reactions)

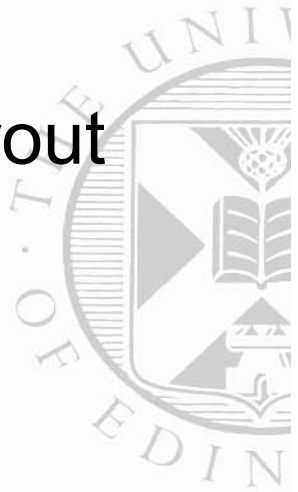
Objective: find biologically meaningful pathways

Other bionetworks represented in a similar way

- Protein-protein interaction network (proteins as nodes)
- Signal transduction network (proteins or effectors as nodes)
- Gene regulatory network (genes or proteins as nodes)
- Other relational networks: co-expression network, disease-gene network, drug target network, etc
- All complex relationships represented as networks

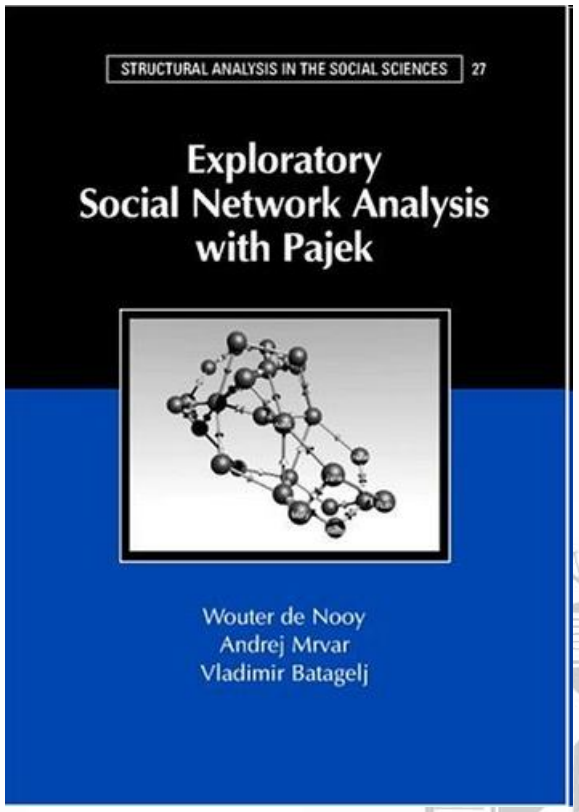
Tools for network analysis

- Pajek: http://pajek.imfm.si/doku.php
- Cytoscape http://www.cytoscape.org/ (for Biological networks, mapping data)
- Ucinet http://www.analytictech.com/ucinet.htm
- Bioconductor and R (SNA)
- Java and Python packages (NetworkX)
- A main issue in visualization: automatic layout (Graphviz)



Pajek, the software used for network visualization and analysis

- Written in C, very fast, many functions, 1M
- Free and updated with new function frequently
- good manual, theory introduction with how-to-do in the software



Structural analysis of large scale network

- Connection degree distribution
- Path length (efficiency)
- Node centrality (the most important nodes)
- Global connectivity of the network
- Network decomposition and hierarchical Modular organization

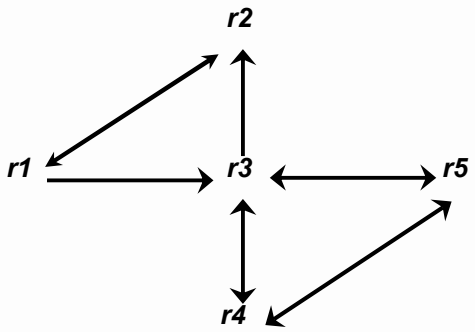
Neighbours and degree

Neighbours: directly linked nodes

K-neighbours: nodes linked with a node in k steps.

Degree: number of links to its neighbours for a node (may not equal to the number of neighbours).

For directed network: input and output degree.

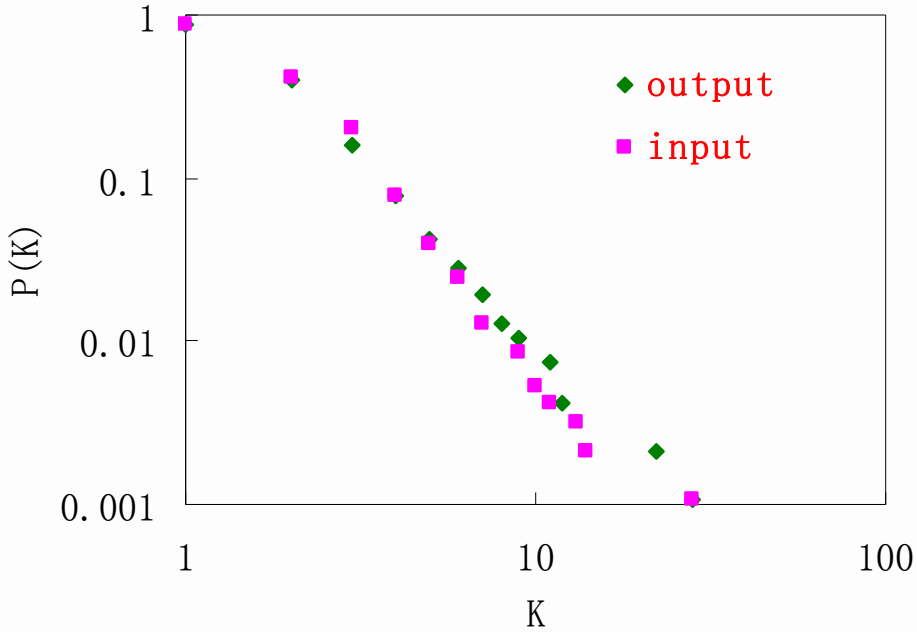


For r2, neighbours are 2, 2-neighbours are 4

Degree is 2, input degree is 2 and output degree is 1.

Connection degree distribution

How node degrees distributed in a network.

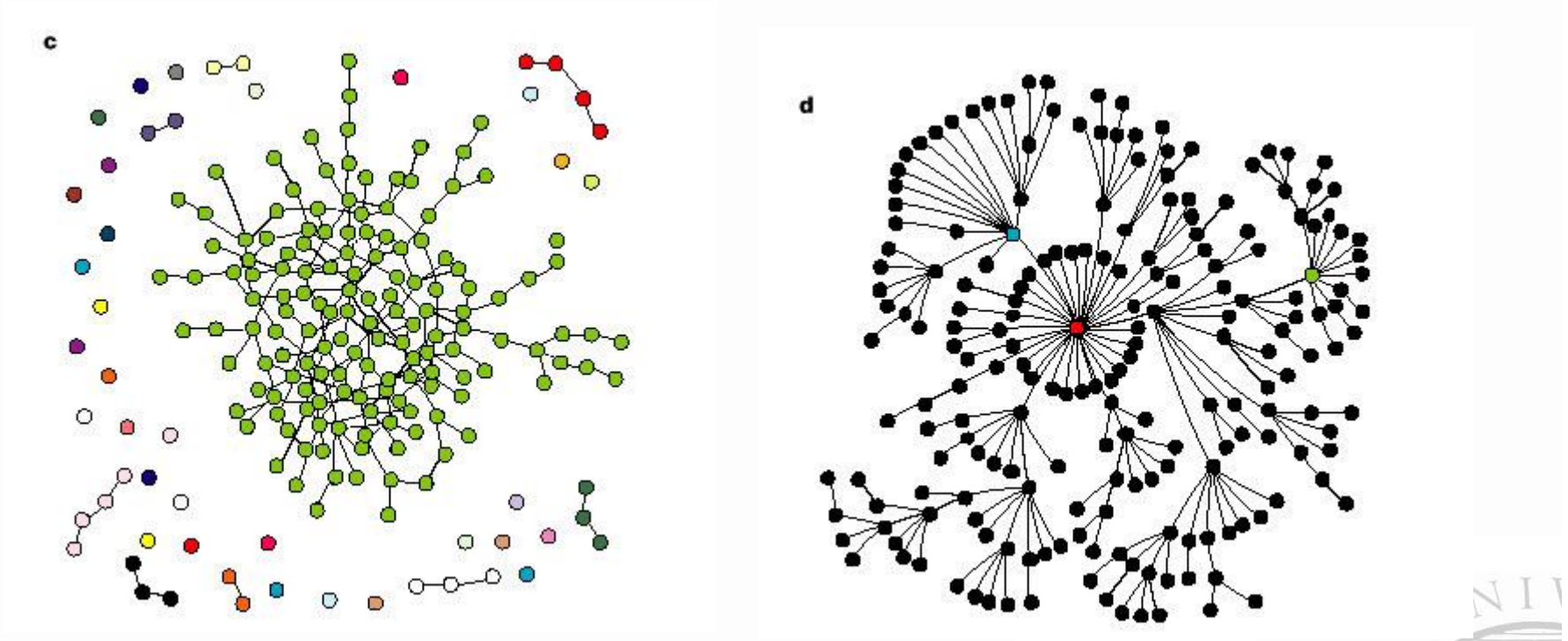


$$P(k) = ak^{-\gamma}$$

P(k): Percentage of nodes with a degree k or not less than k (Cumulative distribution).

Power law degree distribution indicates a scale free network: A few nodes (hubs) have very high degree while most nodes have very low degree.

Random network and scale free network



Many real networks are scale free networks.

Robust on random failure but vulnerable under aimed attack

Hub metabolites

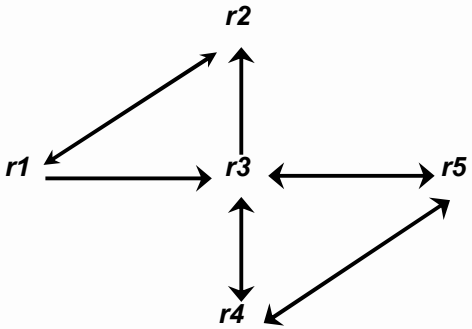
E. Coli metabolic network

Glycerate-3-phosphate, D-Ribose-5-phosphate, Acetyl-CoA, Pyruvate, D-Xylulose 5-phosphate
D-Fructose 6-phosphate, 5-Phospho-D-ribose 1-diphosphate, L-Glutamate, D-Glyceraldehyde 3-phosphate, L-Aspartate, Propanoyl-CoA, Malonyl-ACP, Succinate, Acetate, Isocitrate, Fumarate

Most hubs are in central pathways. However, if currency metabolites are included in the network, Most hubs would be currency metabolites

Clustering coefficient

- Number of edges (k) between neighbors of vertex divided by total possible edges (n) between neighbors of vertex (often not consider direction)



For r3, k=4, n=4*(4-1)/2=6

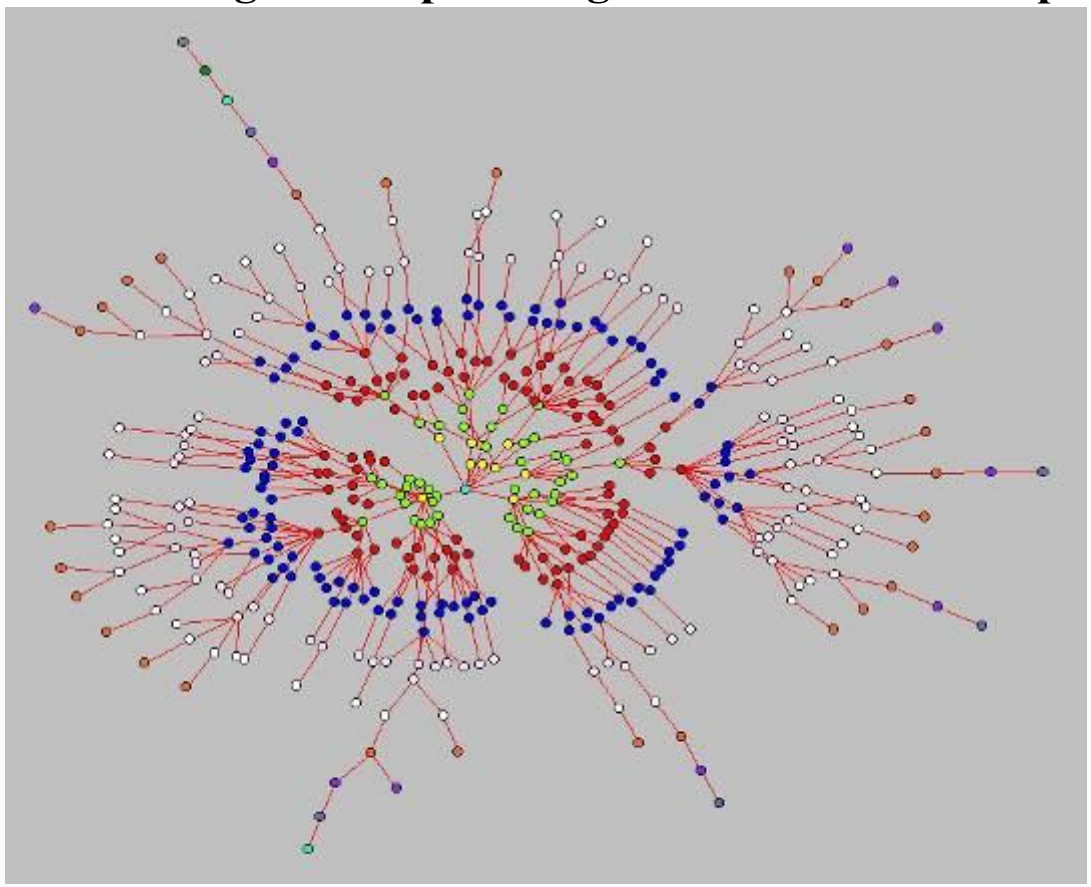
Then c=k/n=0.67

Average clustering coefficient: average for all nodes in a network.

Higher average clustering coefficient indicates small world network

Average path length

Path length: number of the steps in the shortest paths from one node to another
APL: average of the path lengths for all connected pairs of nodes



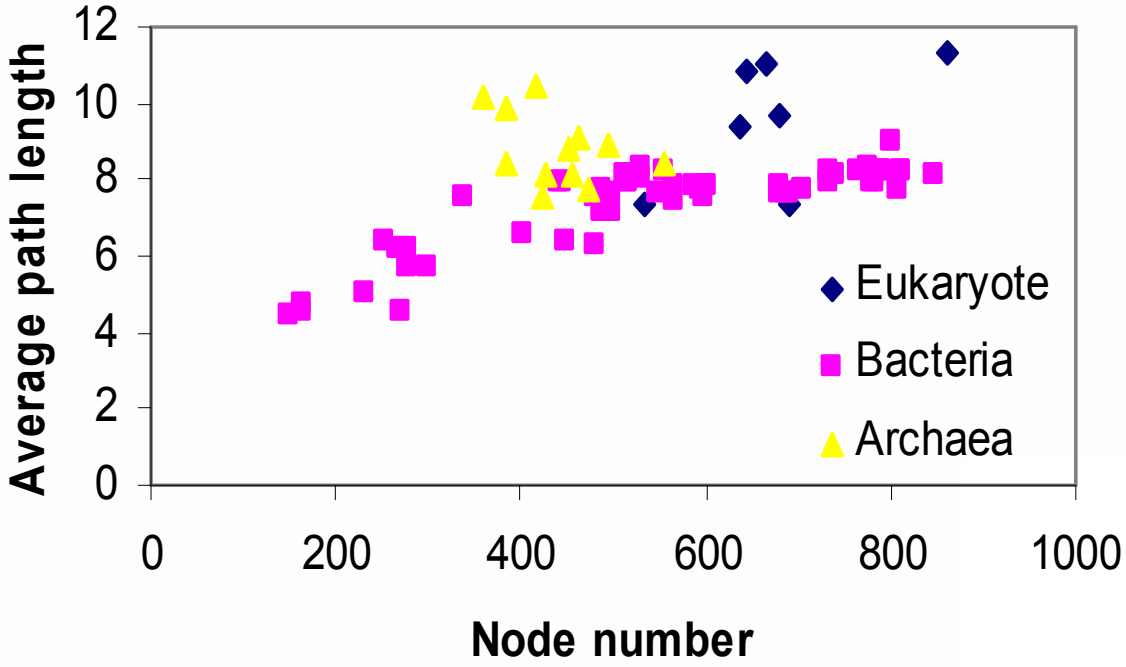
Input (output) domain

The set of nodes that can reach (or can be reached) by a node.

Breadth first searching method to find the shortest paths from one node to all other nodes

Average path length in MN

Jeong *et al*. (2000) Nature: constant AL (about 3) for all the 43 organisms studied.



Structural differences among MNs of the three domains of organism

Ma & Zeng (2003) Bioinformatics. 19, 270

Node Centrality

Closeness centrality of node *x:*

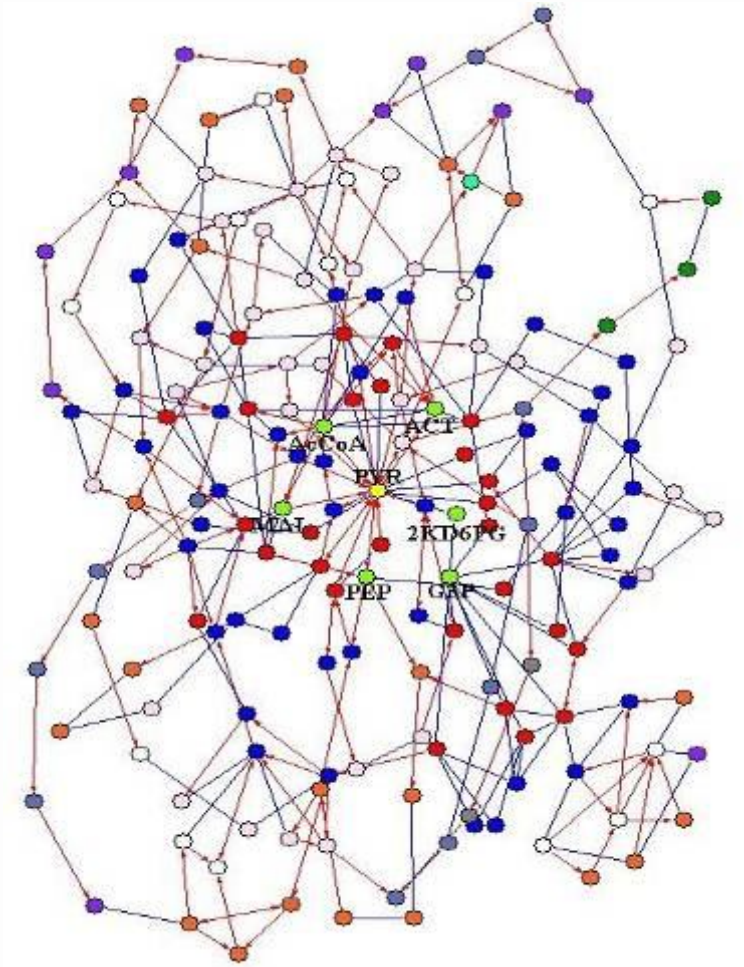
$$C(x) = \frac{n-1}{\sum_{y \in U,\, y \neq x} d(x,y)} = \frac{1}{\overline{d}}$$

d(x,y)	**the path length between node *x* and node *y***
U	**the set of all nodes**
$\overline{d}$	**average path length between *x* and the other nodes**

The central nodes have short path lengths to other nodes in the network

The most central metabolites in the metabolic network of *E. coli*

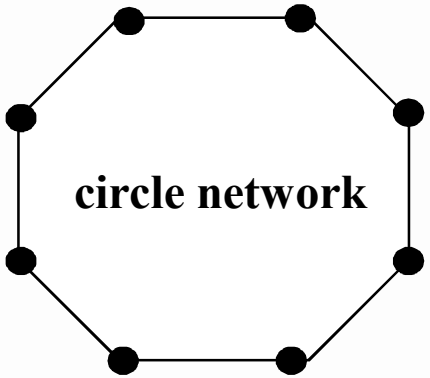
Metabolite	Mean distance
Pyruvate	4.44
Actyl-CoA	4.76
Malate	4.89
2KD6PG	4.93
Acetate	4.98
Acetaldehyde	5.03
G3P	5.06

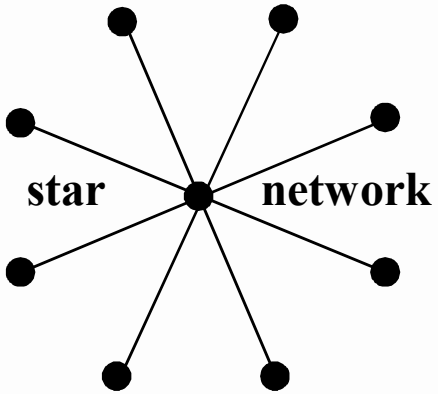


most central nodes ≠ highly connected nodes

Network Centrality

Distribution of the node centrality in the network





C = 0

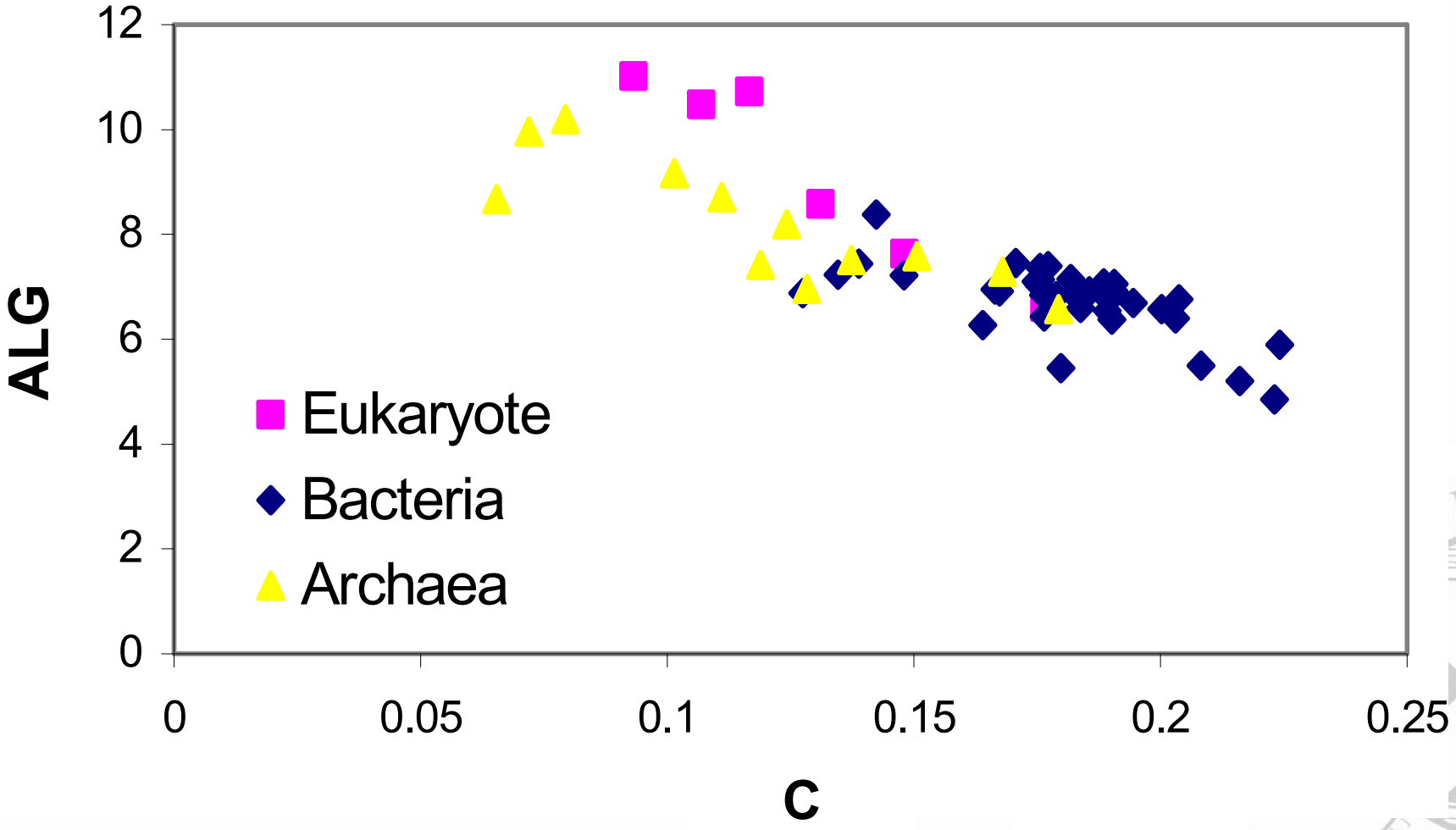
C = 1

Network closeness centralization index

$$C = \frac{(2n-3)\sum_{x \in U}(C^* - C(x))}{(n-1)(n-2)}$$

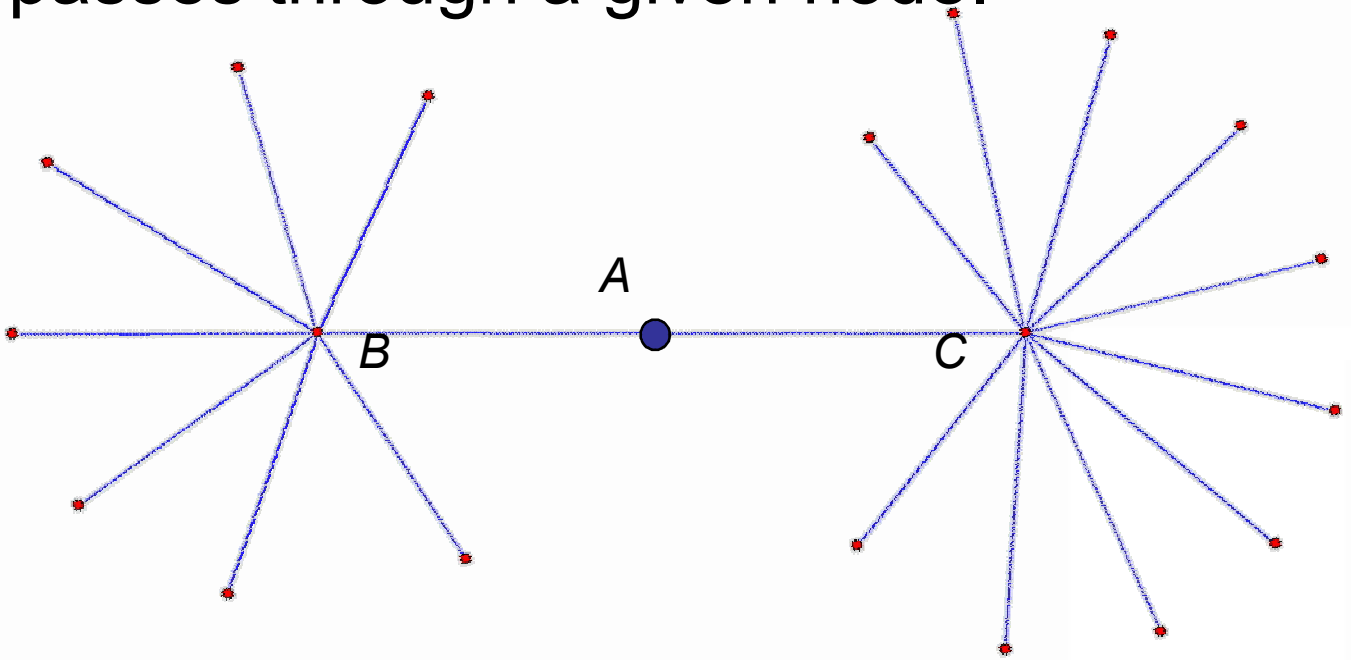
n	**node number in the network**
C(x)	**the closeness centrality for node *x***
C*	**the highest value of node closeness centrality**

Average path length vs. network closeness centralization index



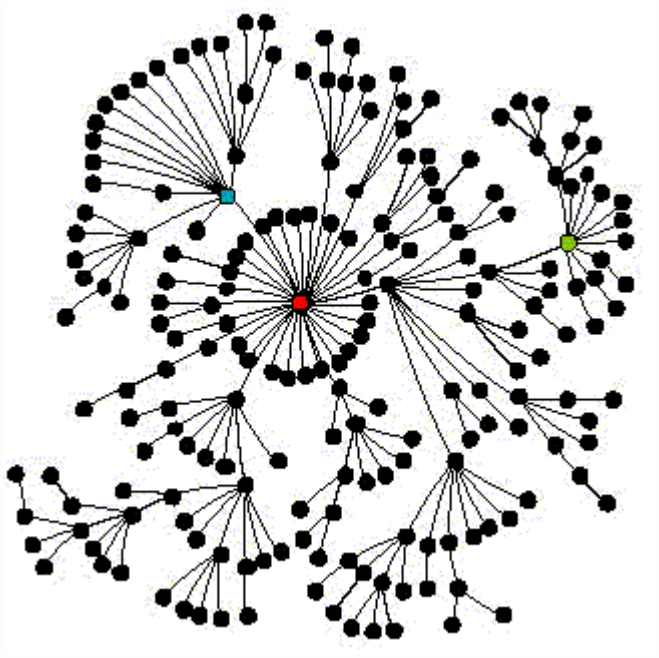
Other centrality measures

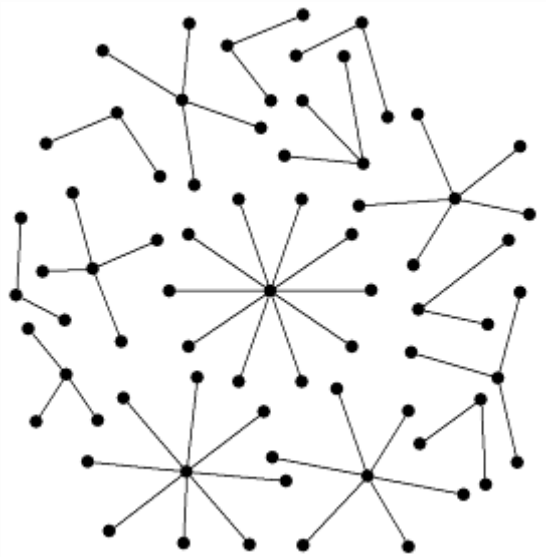
- Betweenness centrality: the fraction of shortest paths between pairs of nodes that passes through a given node.



The most effective target to break down the network

Network Global Connectivity



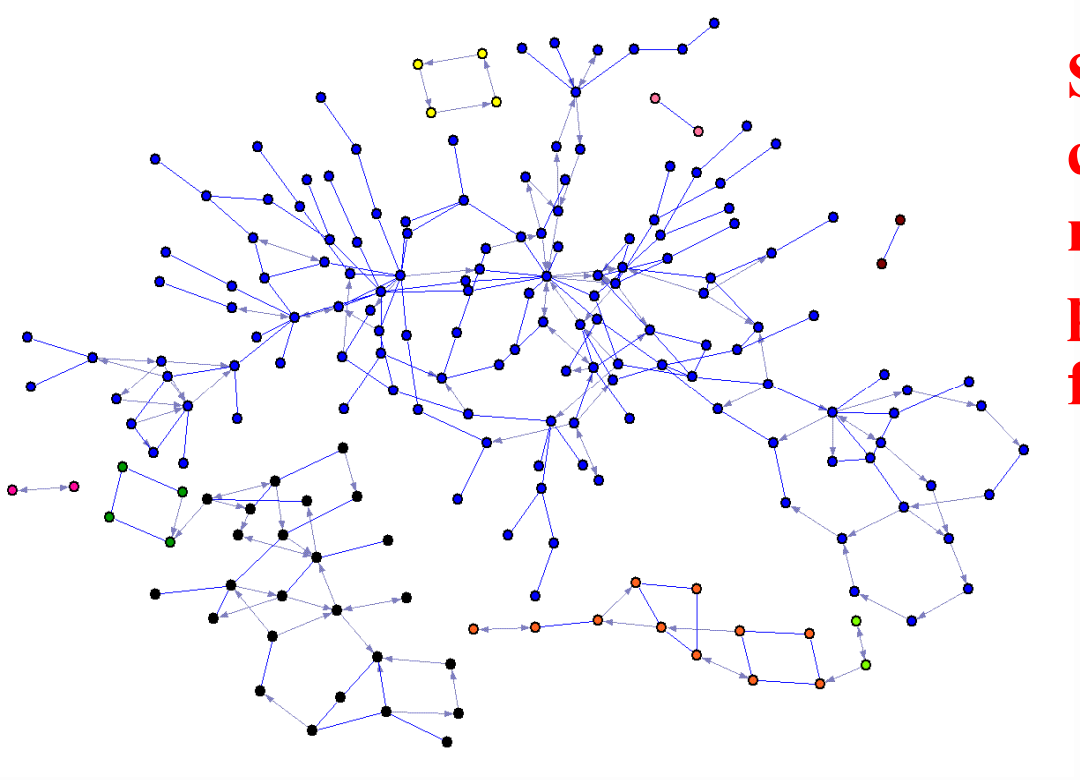


Degree distribution tells nothing about global connectivity

The right network can have short average path length though not connected at all

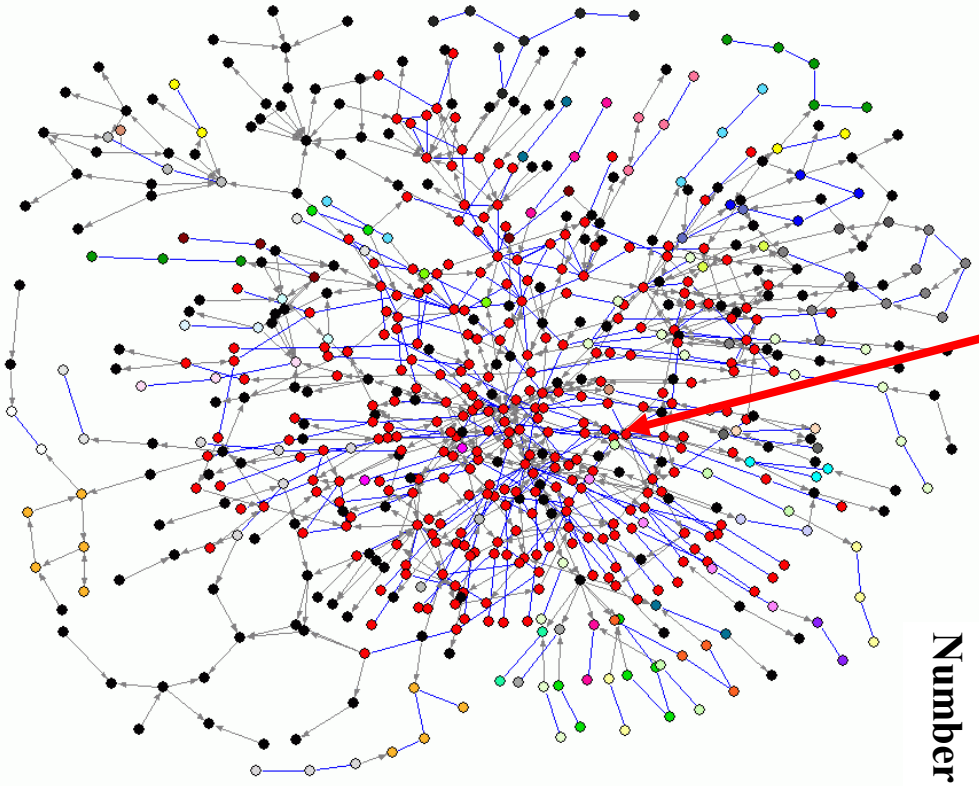
Strongly or weakly connected components

a **connected component** is a maximal connected subgraph. Two nodes are defined to be in the same connected component if there exists a path between them.

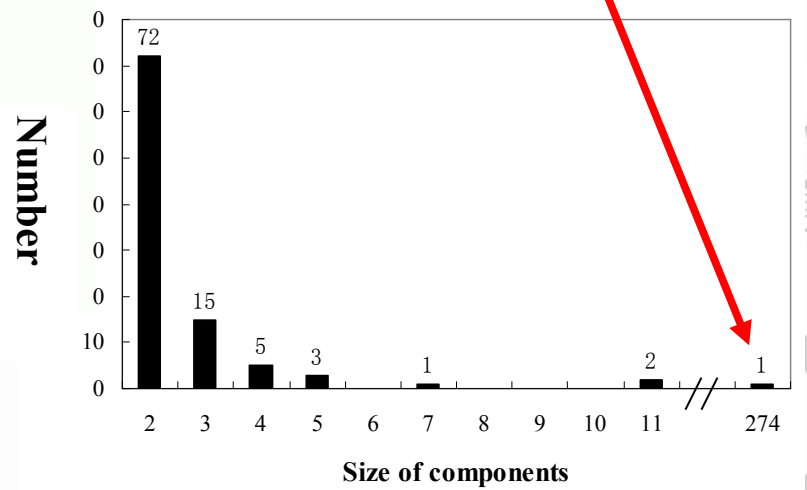


Strongly connected component: For any two nodes a, b in it, there is a path from a to b and a path from b to a

SC distribution in a metabolic network



GSC: Giant strong component

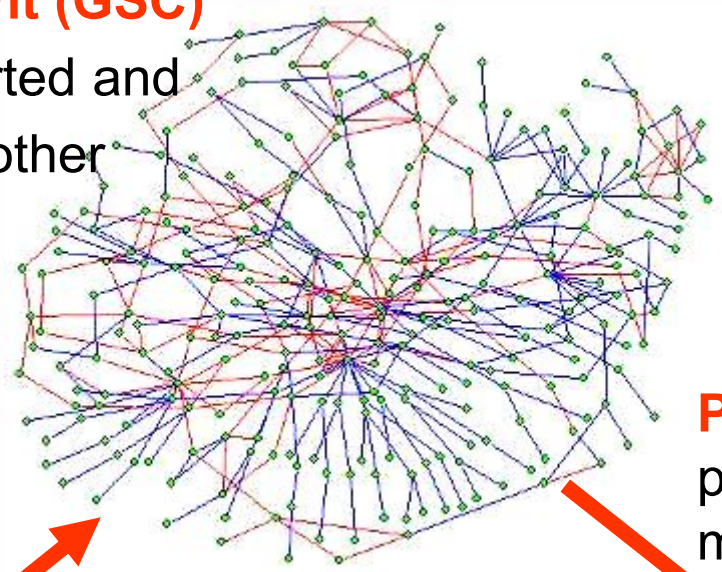


One big SC and many small SCs

Connectivity structure of MN

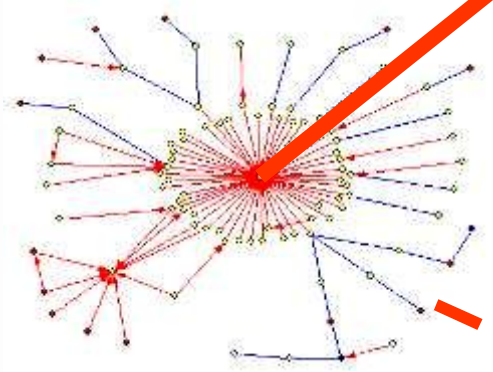
Giant strong component (GSC)

metabolites fully converted and convertible to each other

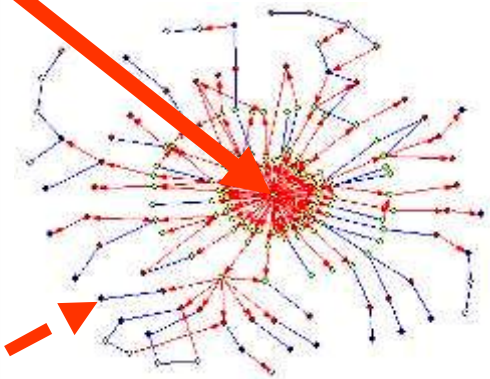


274 out of total 811 metabolites

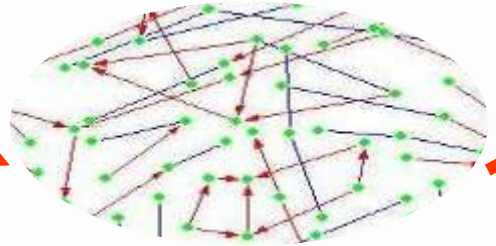
Substrate subset (93) converted to metabolite in GSC



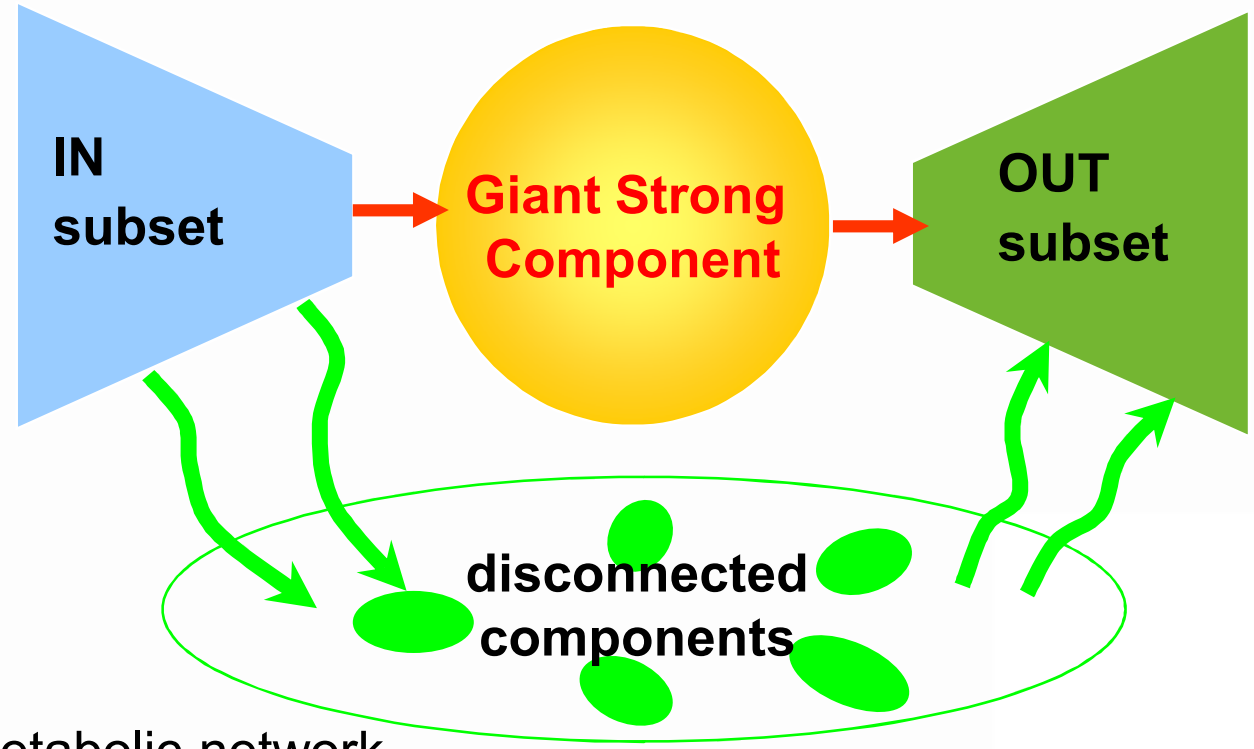
Product subset (161) produced from metabolites in GSC



Isolated subset (283)

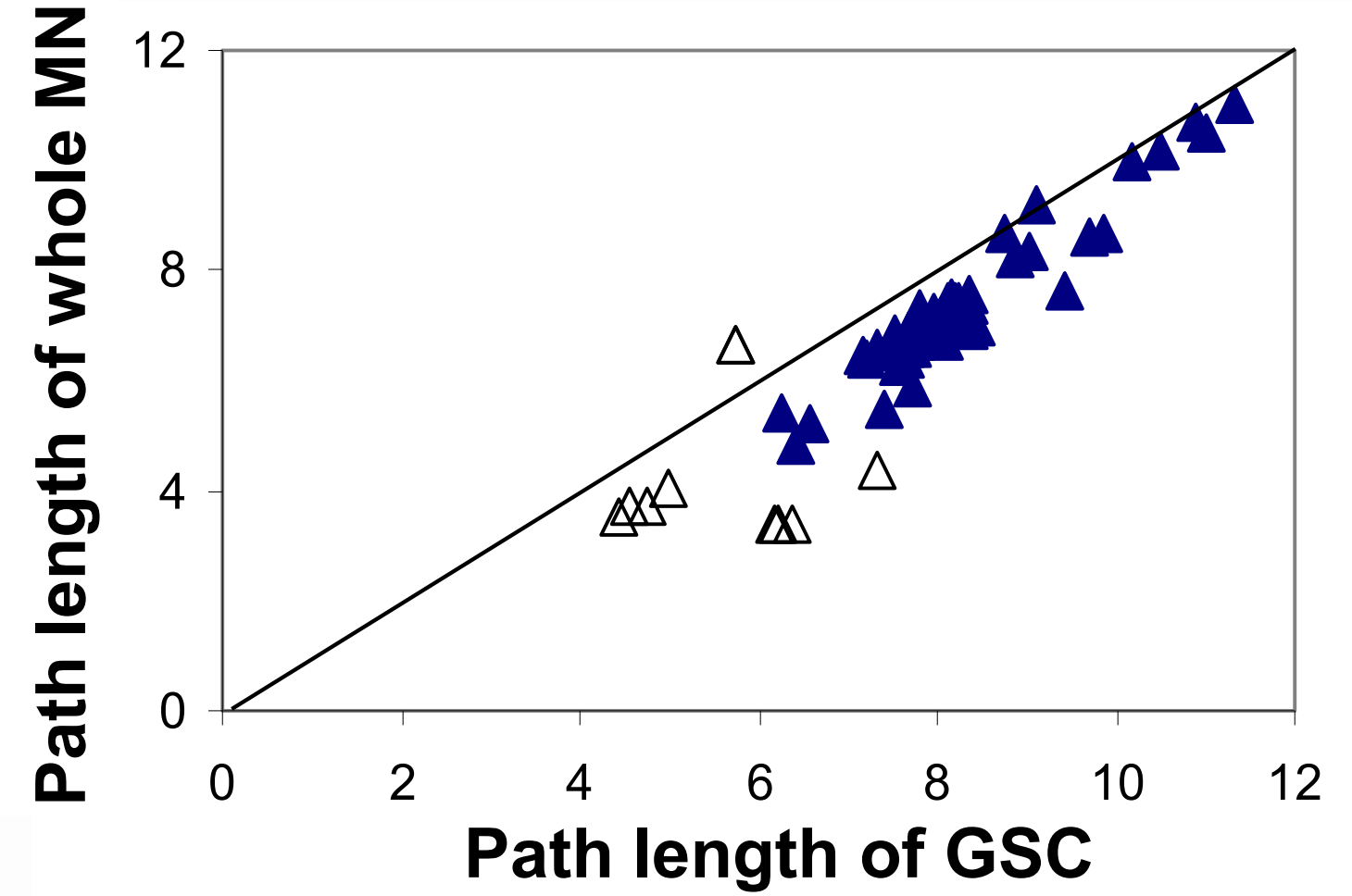


Bow-tie: a general structure of biological and physical networks



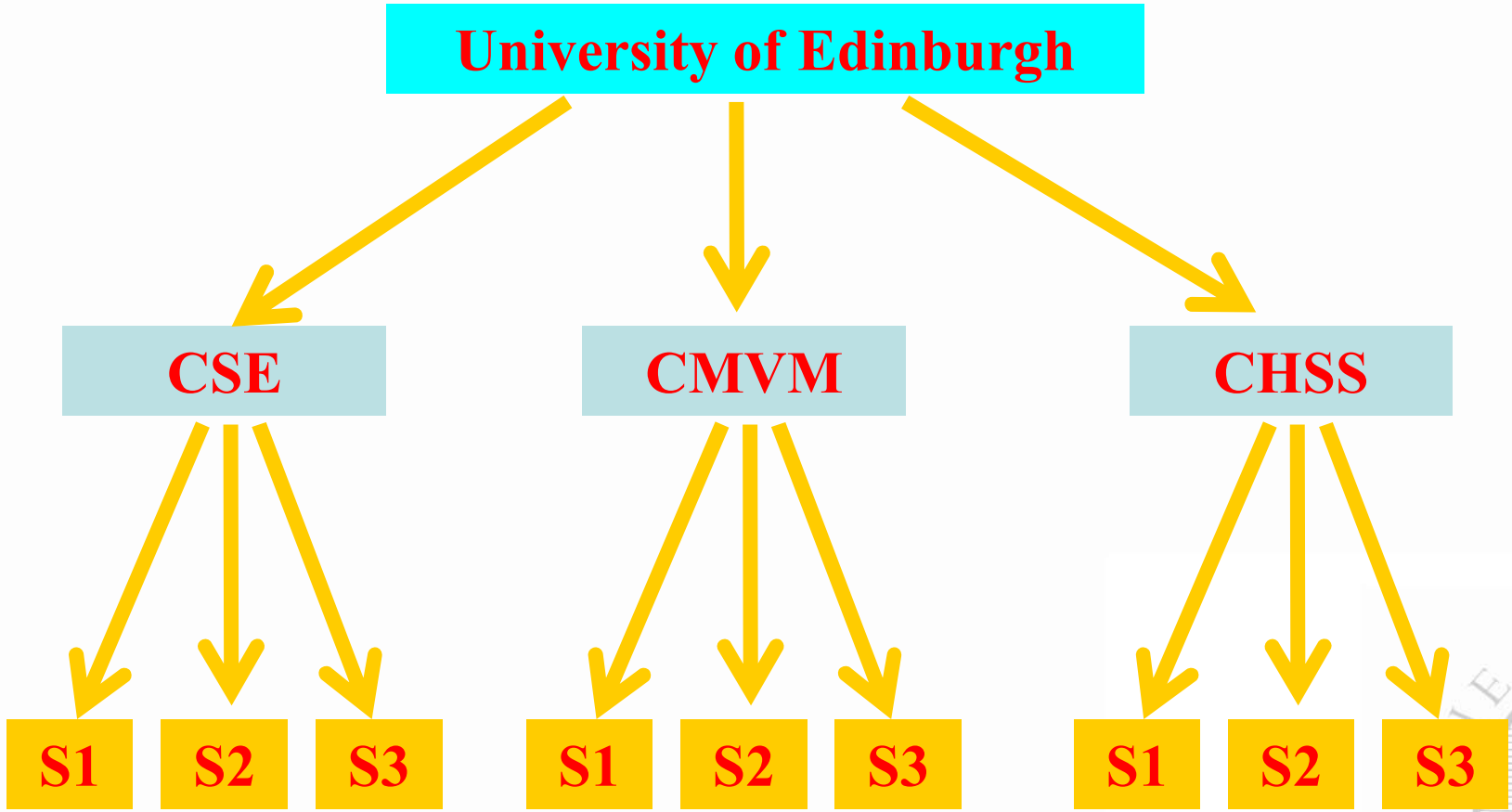
- Metabolic network
- Signal transduction
- Web page
- Material processing and other tech. systems

The average path length of the GSC determines that of the whole network



Most connected pairs are between nodes in GSC

From structure analysis to function analysis: top-down approach



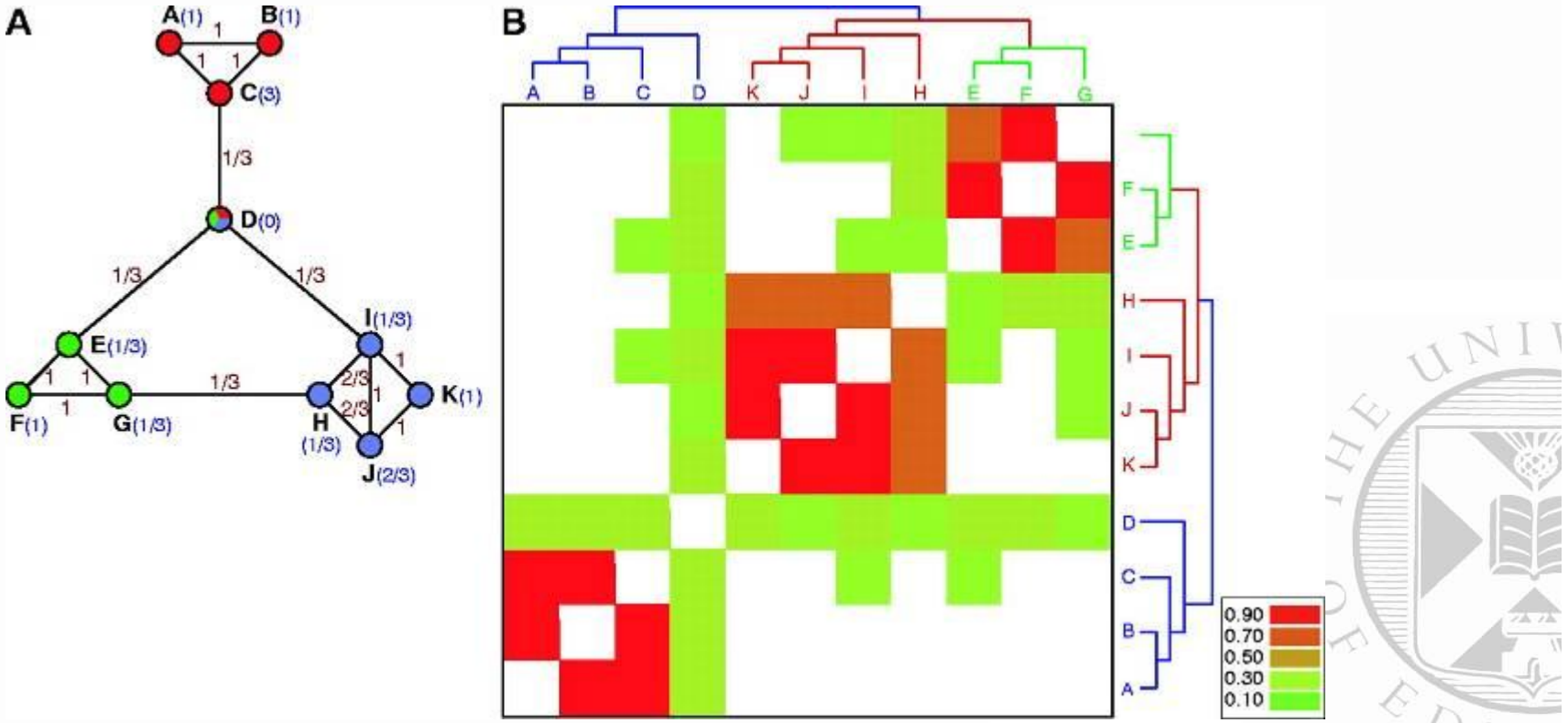
Manageability, robustness

Network decomposition

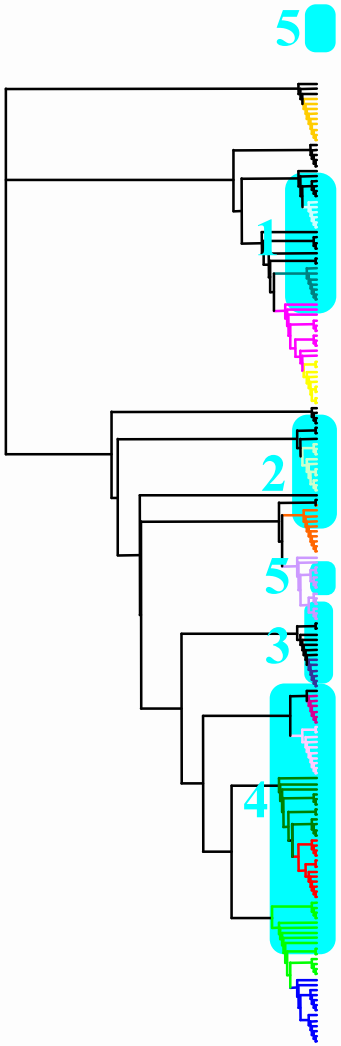
- Complex real networks are hierarchically organized (ex, pathways in MN).
- Identifying somehow independent subnetworks (modules) from such complex networks for biological function analysis or developing more detailed kinetic model
- There are different methods available

A Method

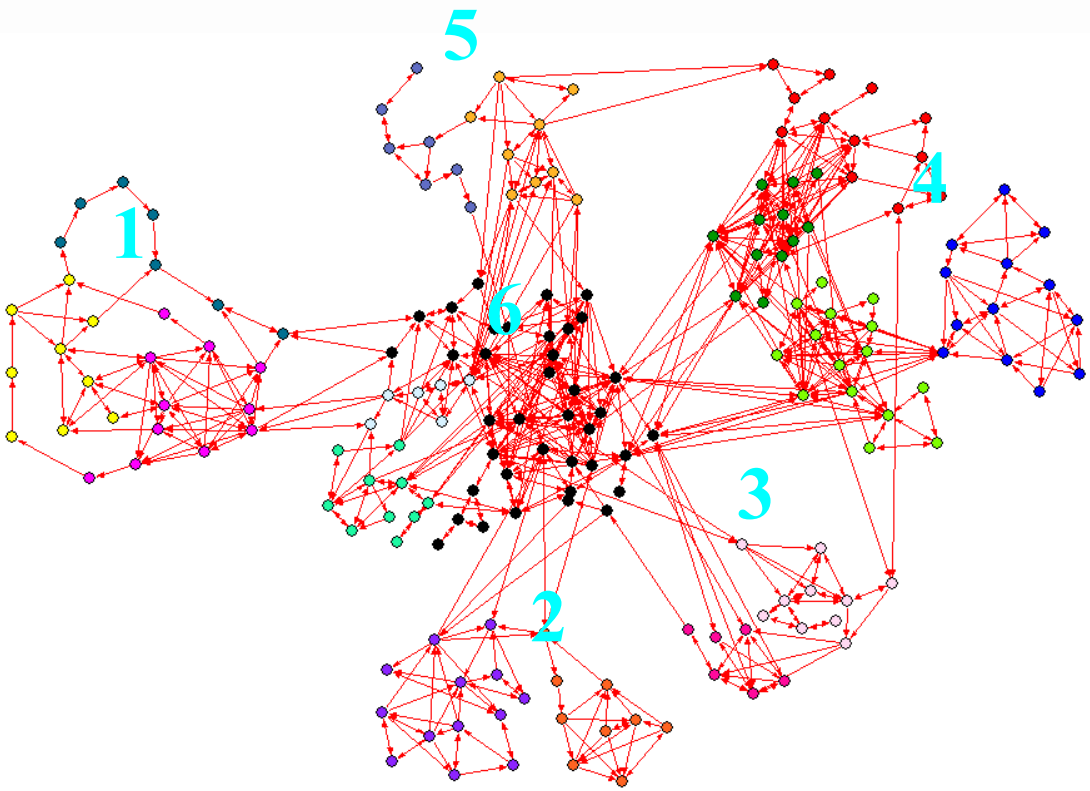
- **Define a distance between two nodes: the shortest path between node *A* and node *B***
- **Distance matrix → a hierarchical classification tree (same methods used for constructing evolutionary tree from DNA sequences)**



Hierarchical decomposition of GSC



For reaction graph

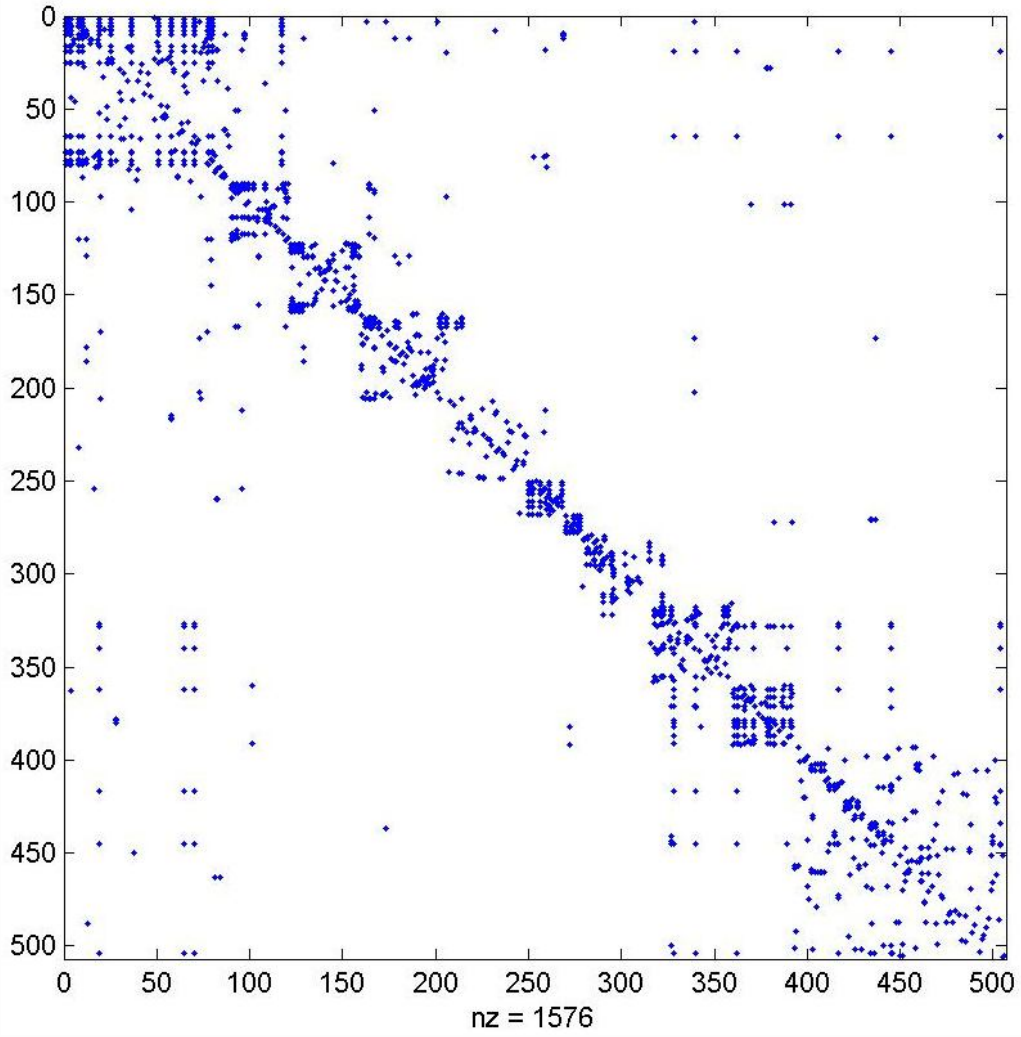


Biological function of these modules

1. Aspartate metabolism, Thr, Lys synthesis
2. Glutathione and THF metabolism
3. Glutamate metabolism, Arg, Pro synthesis
4. galacterate, glycerate metabolism, Ser synthesis
5. PP pathway, upper part of glycolysis pathway, sugar, aminosugar and glycerol metabolism
6. TCA cycle and glyoxylate cycle, pyruvate metabolism, AcCoA, AcAcCoA metabolism

Modules identified by the decomposition method have true biological meaning

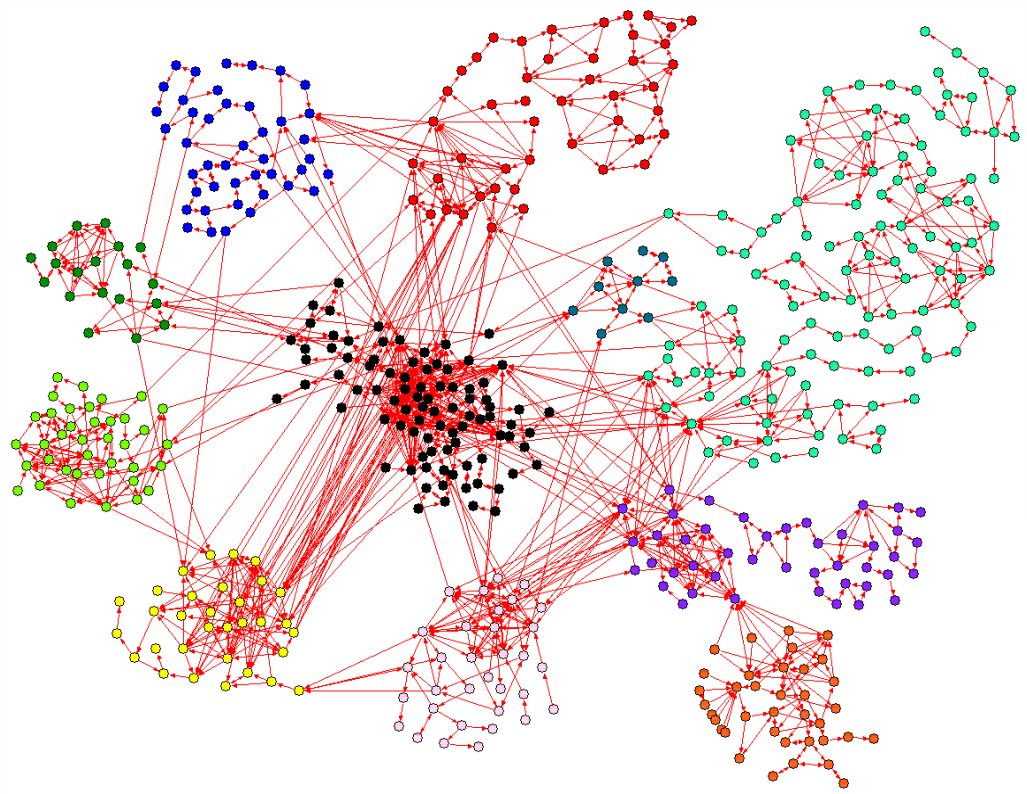
The reaction connectivity map



Dots represent links

Dots distributing around diagonal indicates modularity

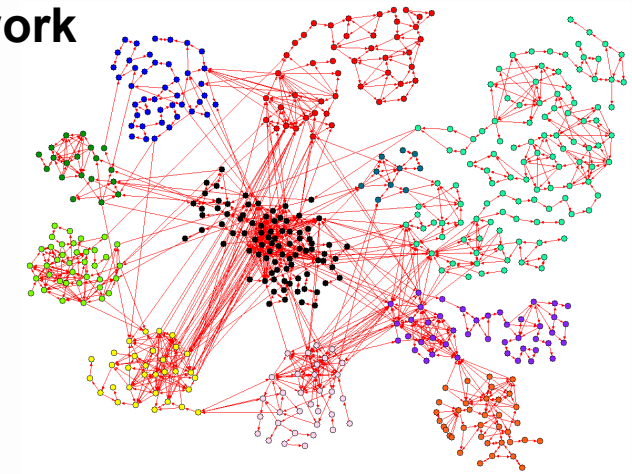
Core-periphery structure: network organization at module level



A central core module to connect other modules with specific functions.

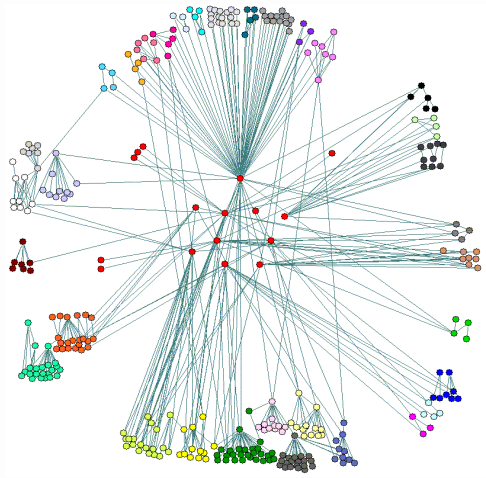
Core-periphery structure: a common organization principle of bionetworks?

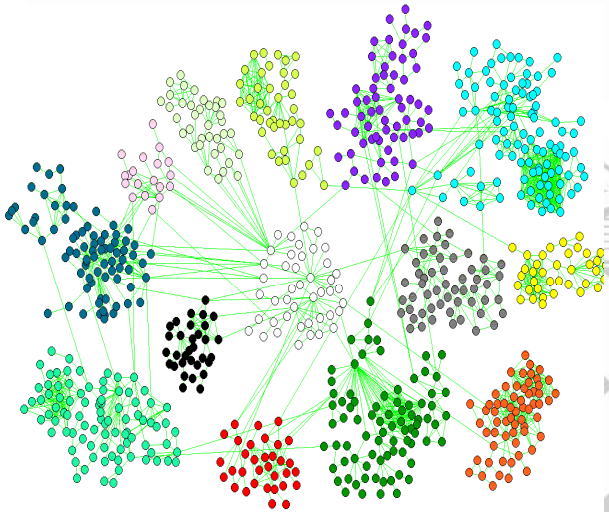
Metabolic network



Yeast protein-protein interaction network
(Han et al. 2004, Nature
Date hub and party hub)

Gene regulatory network





Ma et al. 2004, BMC Bioinformatics

Network analysis with NetworkX

- http://networkx.lanl.gov
- Python package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks.

```
import networkx as nx
G=nx.XDiGraph()
G.add_edge(a[0],a[1])
G.add_node(n)
G.degree()
G.neighbors(n)
nx.clustering(G)
nx.connected_component_subg
raphs(G)
nx.closeness_centrality(G)
nx.all_pairs_shortest_path_leng
th(G)
```

KNEVA for network structure analysis

- Degree distribution
- Centrality analysis
- Connectivity analysi
- Network visualization
- Pathway analysis

Pajek Demo

- Input file
- Calculate degree distribution, domain, clustering coefficient, path length, centrality etc.
- Visualization, layout