Metabolomics Pathway analysis

Anatoly Sorokin

Metabolomics

- Metabolomics is the "systematic study of the unique chemical fingerprints that specific cellular processes leave behind", the study of their small-molecule metabolite profiles.

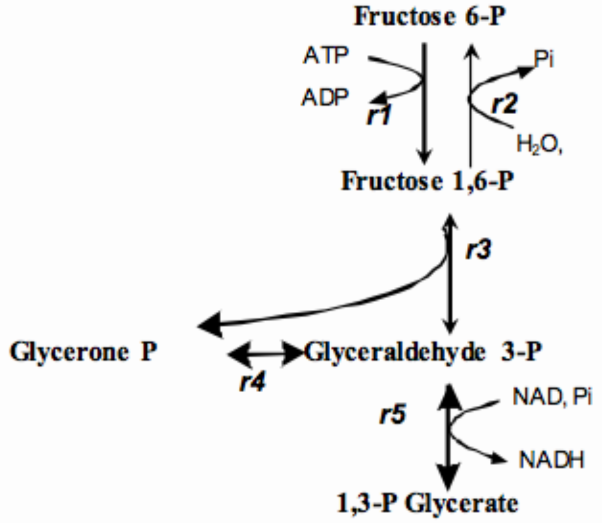
  Daviss, Bennett (2005) The Scientist 19 (8): 25–28
- Younger sister?:

  de Réaumur, RAF (**1752**). "Observations sur la digestion des oiseaux". Histoire de l'academie royale des sciences 1752: 266, 461.

Metabolic network

- Pathway is a series of reactions converting set of substrate into set of products
- Pathway definition is subjective and non-standard
- Pathways are overlapping
- Easier to talk about whole network
  - FBA
  - Extreme pathway etc

Network representations



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

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

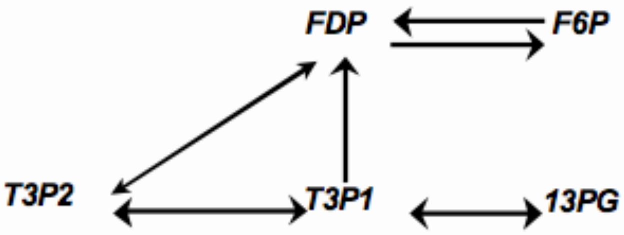
Stoichiometry matrix

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

Connectivity matrix

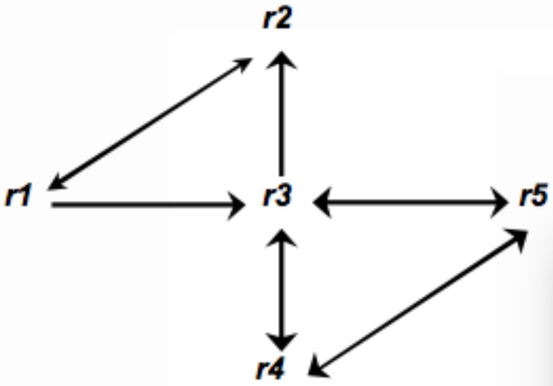
Matrix to the network

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



Metabolite graph

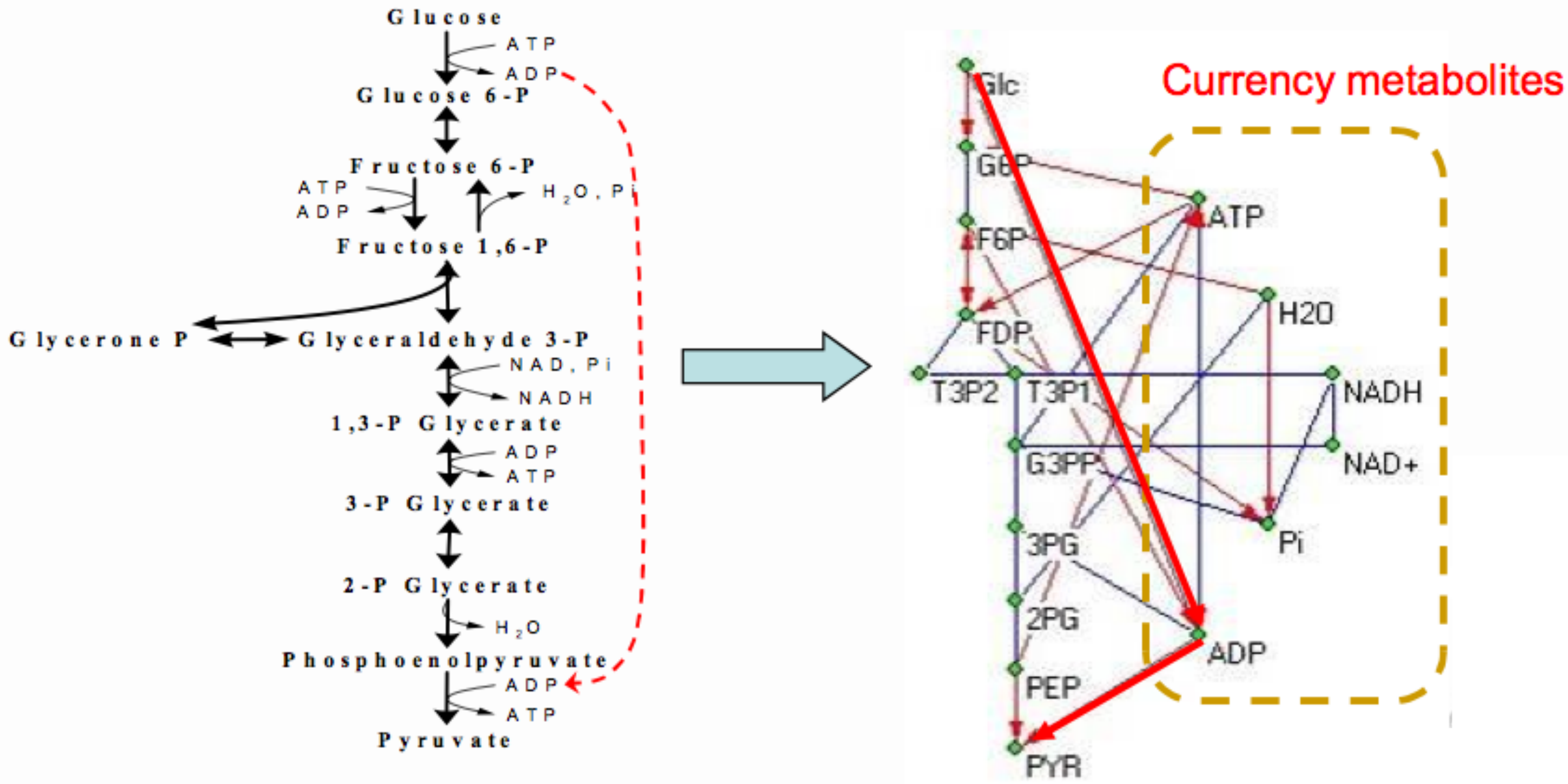
$$\begin{array}{c}r_1\\r_2\\r_3\\r_4\\r_5\end{array}\begin{pmatrix}0 & 1 & 1 & 0 & 0\\1 & 0 & 0 & 0 & 0\\0 & 1 & 0 & 1 & 1\\0 & 0 & 1 & 0 & 1\\0 & 0 & 1 & 1 & 0\end{pmatrix}$$



Connectivity (Adjacency) matrix

Reaction graph

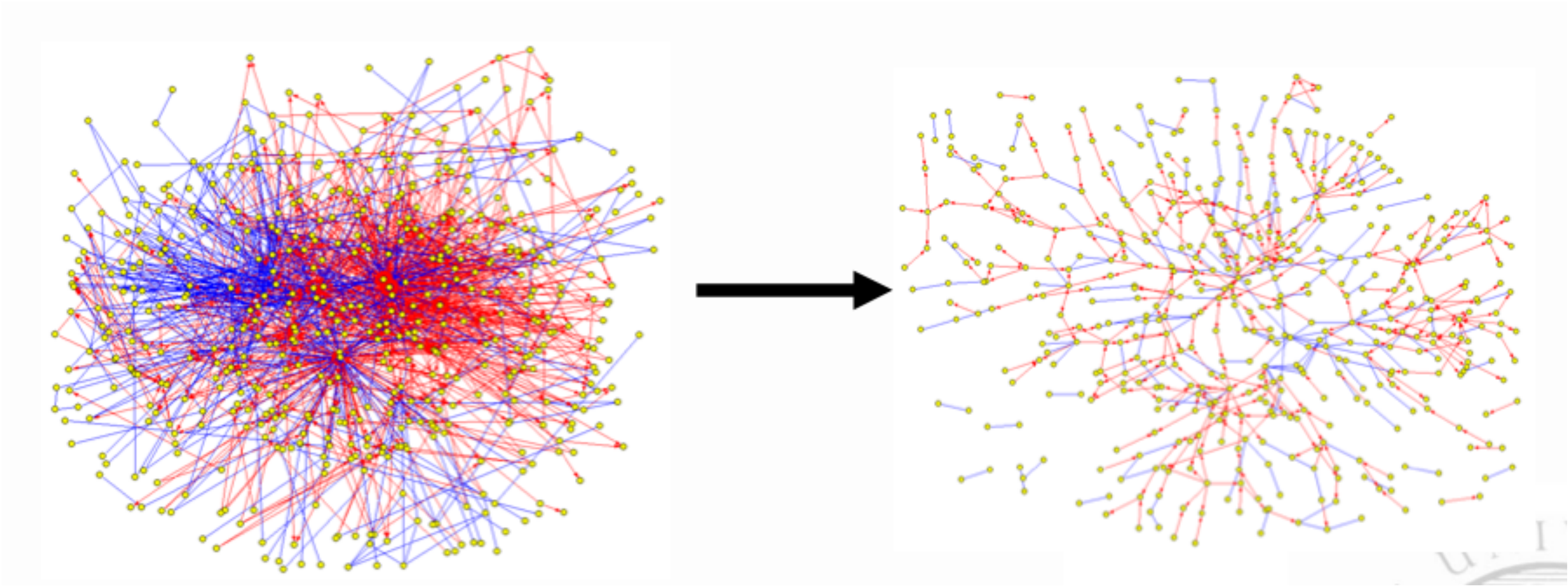
Currency metabolites



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)

Network metrics for metabolic network

- A typical genome scale metabolic network contains one thousand reactions/metabolites.
- We need to characterize importance of nodes and edges in the network

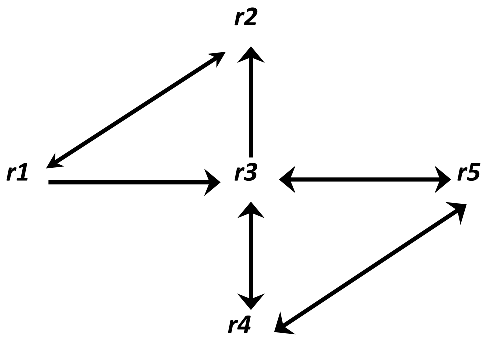
Neighbours and degree

Neighbours: directly linked nodes

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

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

For directed network: input and output degree.

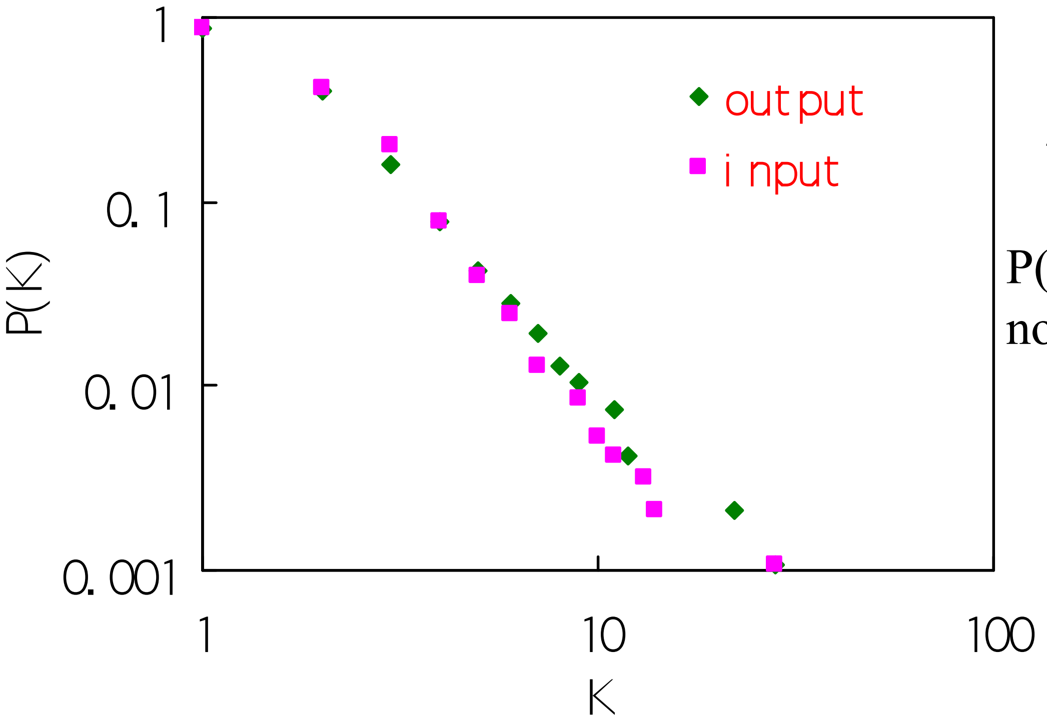


For r2, neighbours are 2, 2-neighbours are 4

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

Connection degree distribution

How node degrees distributed in a network.

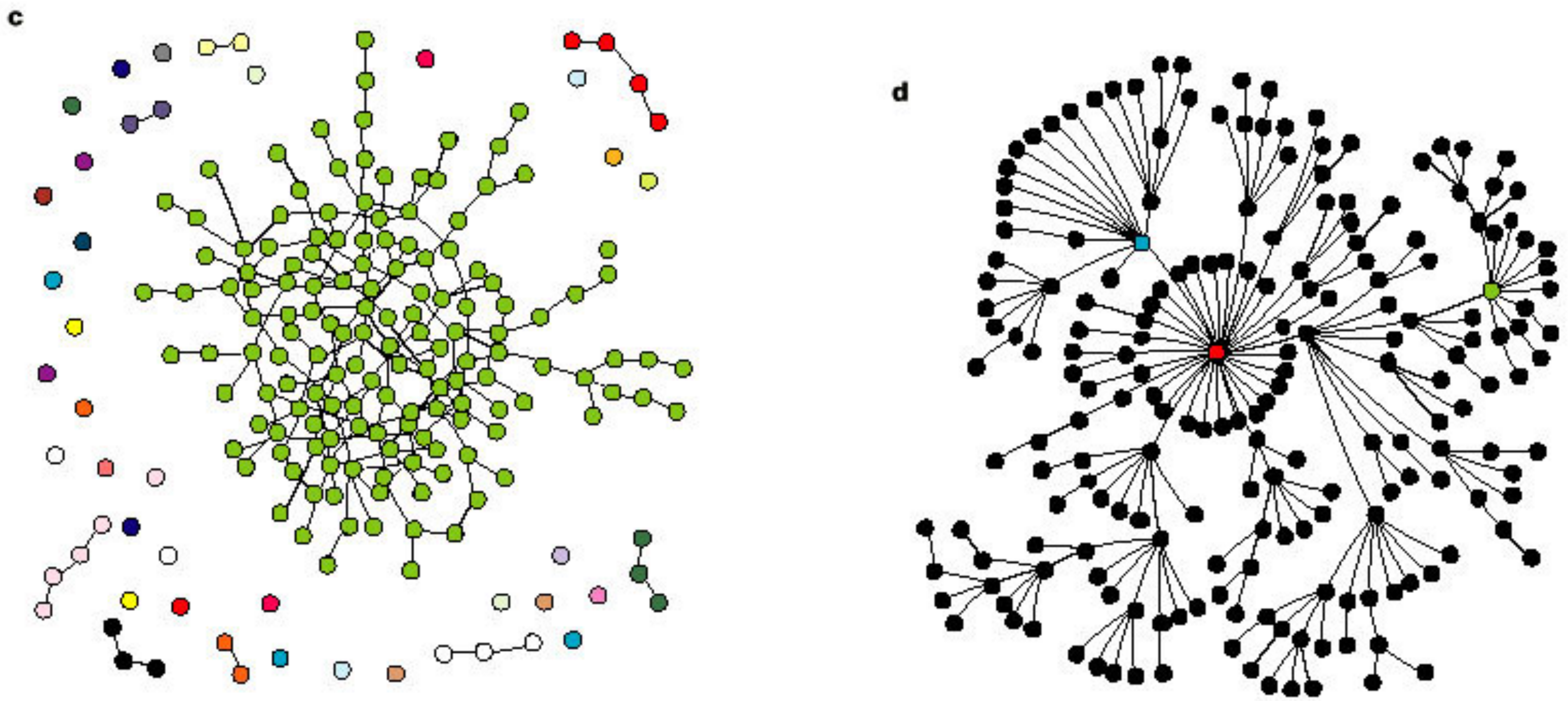


$$P(k) = ak^{-\gamma}$$

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 at the highly connected nodes (hubs). Scale free feature is the result of evolution (rich get richer generative model, like web)

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

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***

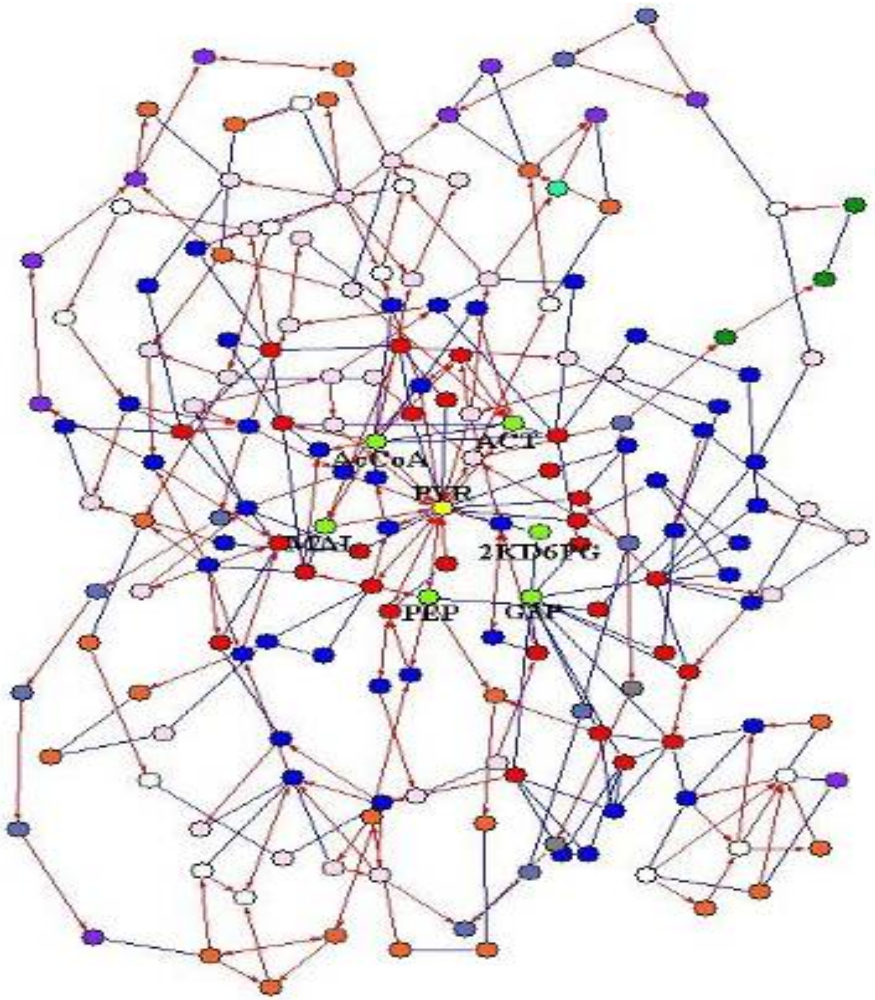
U **the set of all nodes**

$\overline{d}$ **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*

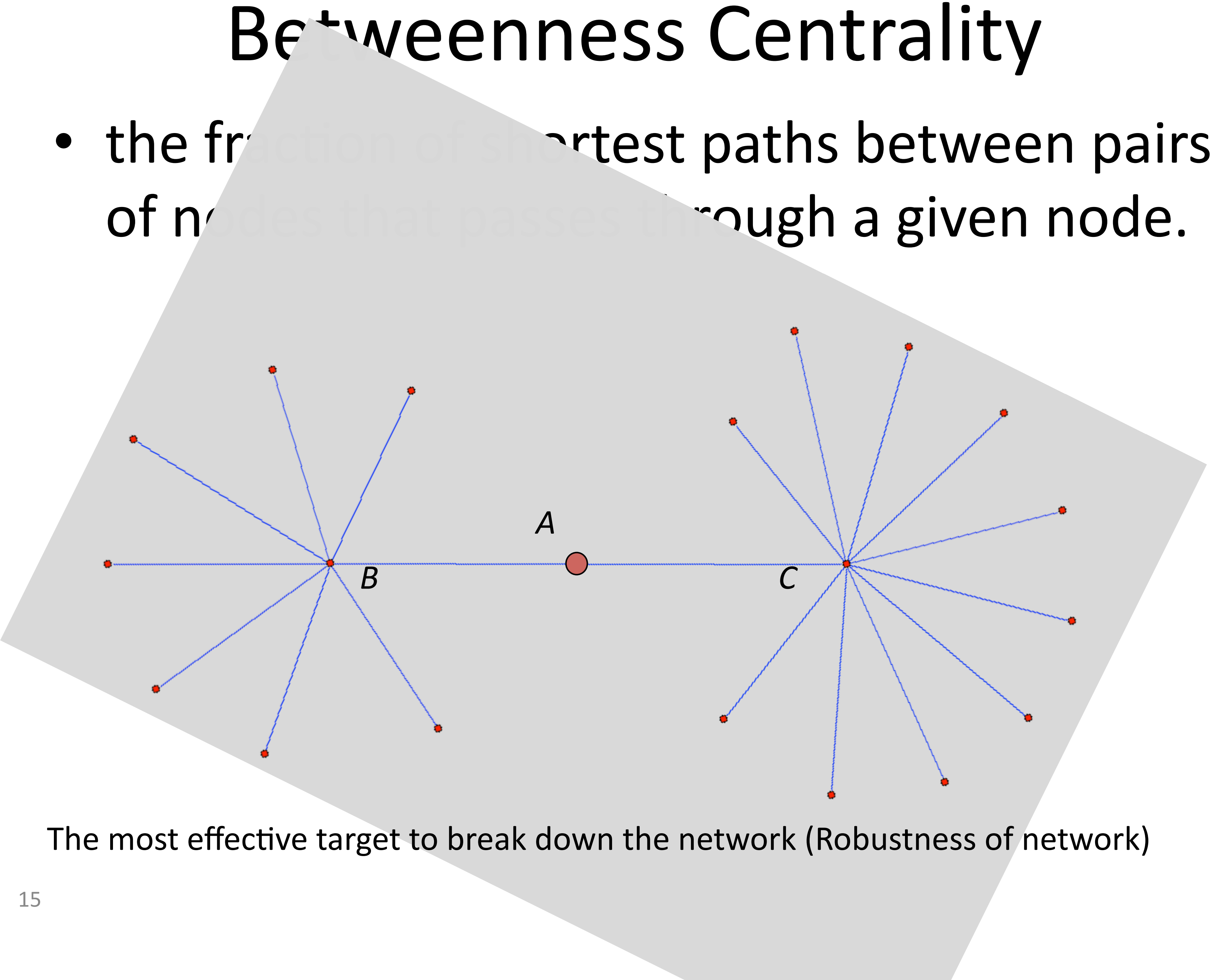
Metabolite	**Centrality**
Pyruvate	0.225
Actyl-CoA	0.210
Malate	0.204
2KD6PG	0.203
Acetate	0.201
Acetaldehyde	0.199
G3P	0.198



most central nodes ≠ highly connected nodes

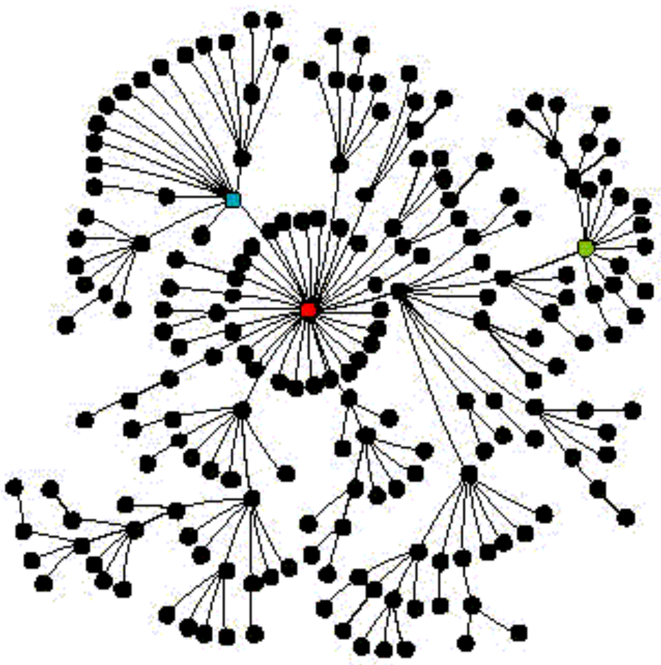
Betweenness Centrality

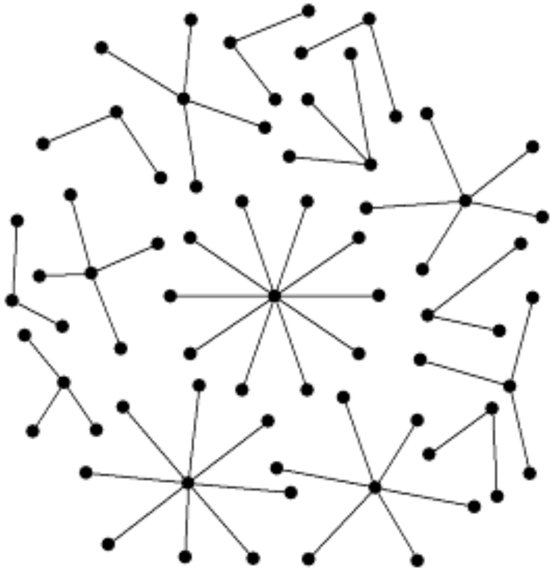
- the fraction of shortest paths between pairs of nodes that passes through a given node.



The most effective target to break down the network (Robustness of network)

Network Global Connectivity



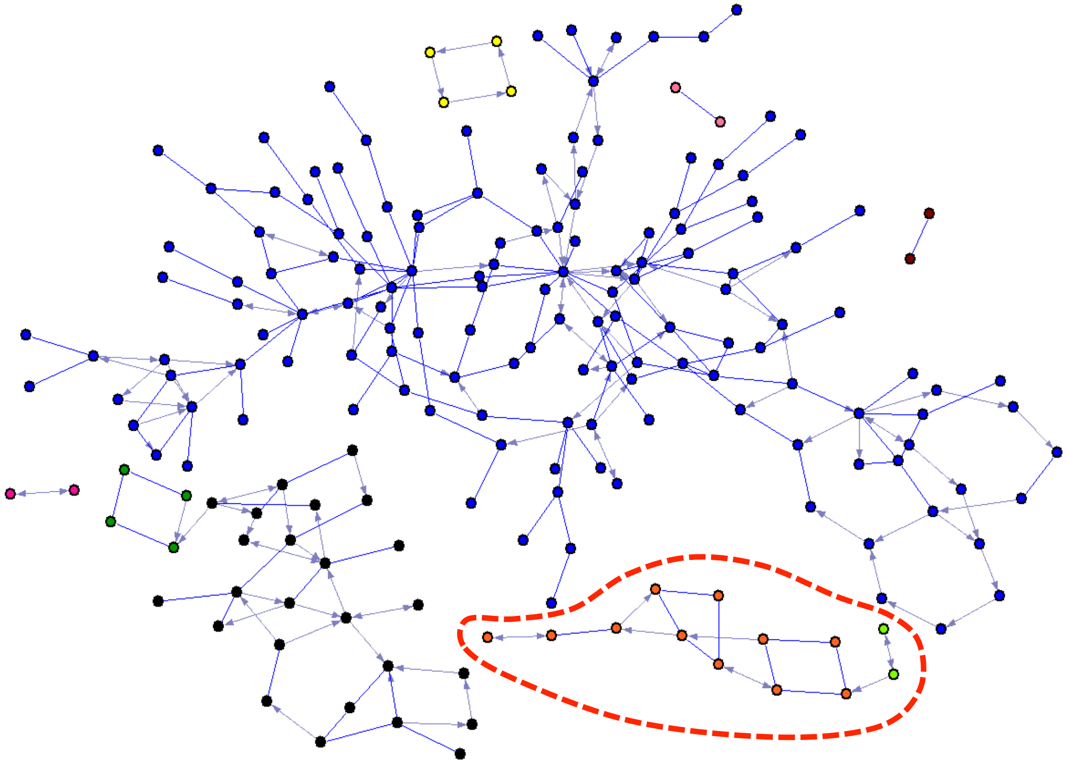


Degree distribution tells nothing about global connectivity

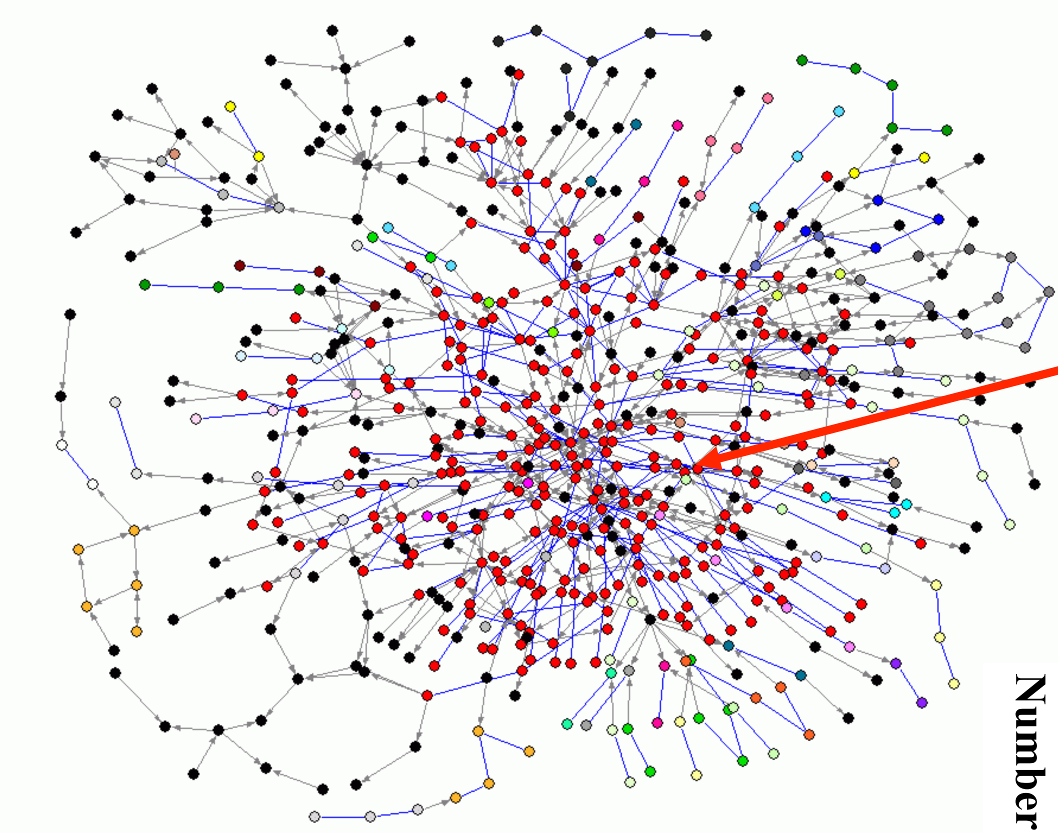
The right network can have short average path length though not connected at all

Strongly or weakly connected components

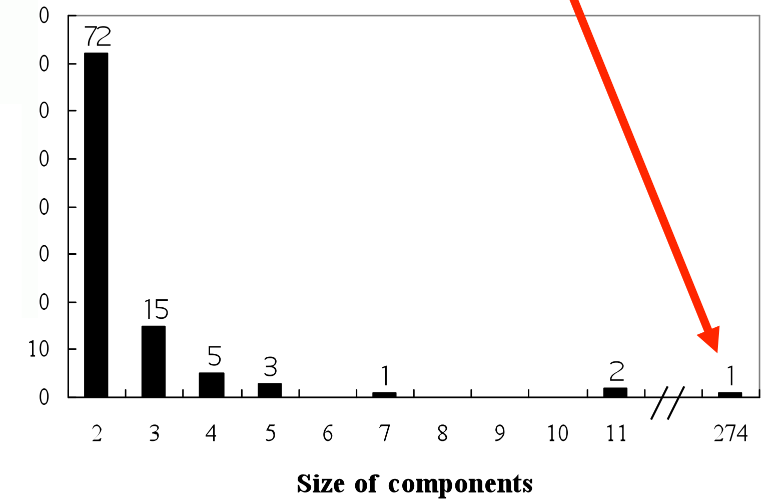
a **connected component** is a maximal connected subgraph. Two nodes are defined to be in the same connected component if there exists a path between them.
If link direction is considered it is strongly connected, otherwise weakly connected.



SC distribution in a metabolic network

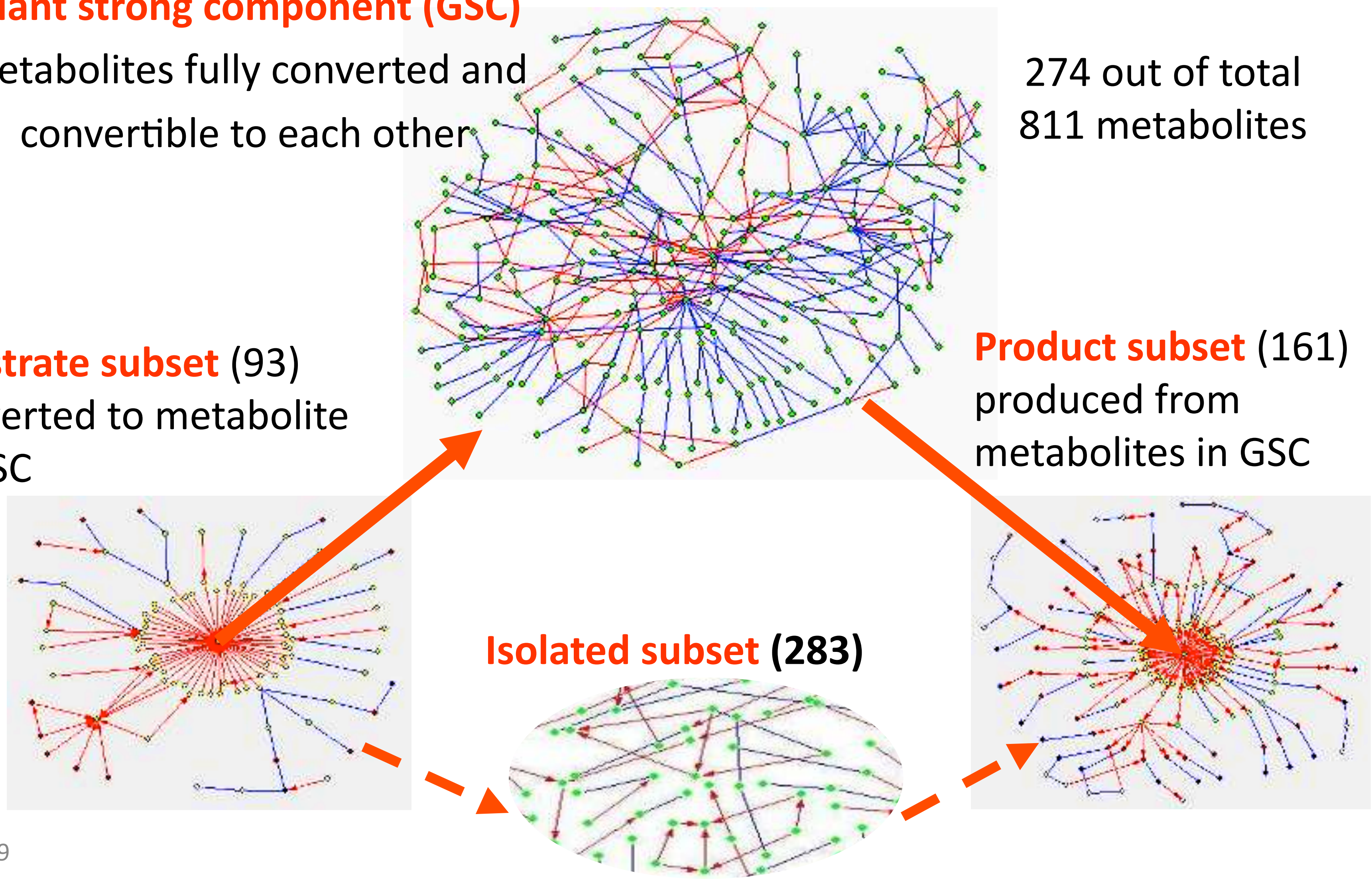


GSC: Giant strong component



One big SC and many small SCs

Connectivity structure of MN



Giant strong component (GSC)
metabolites fully converted and
convertible to each other

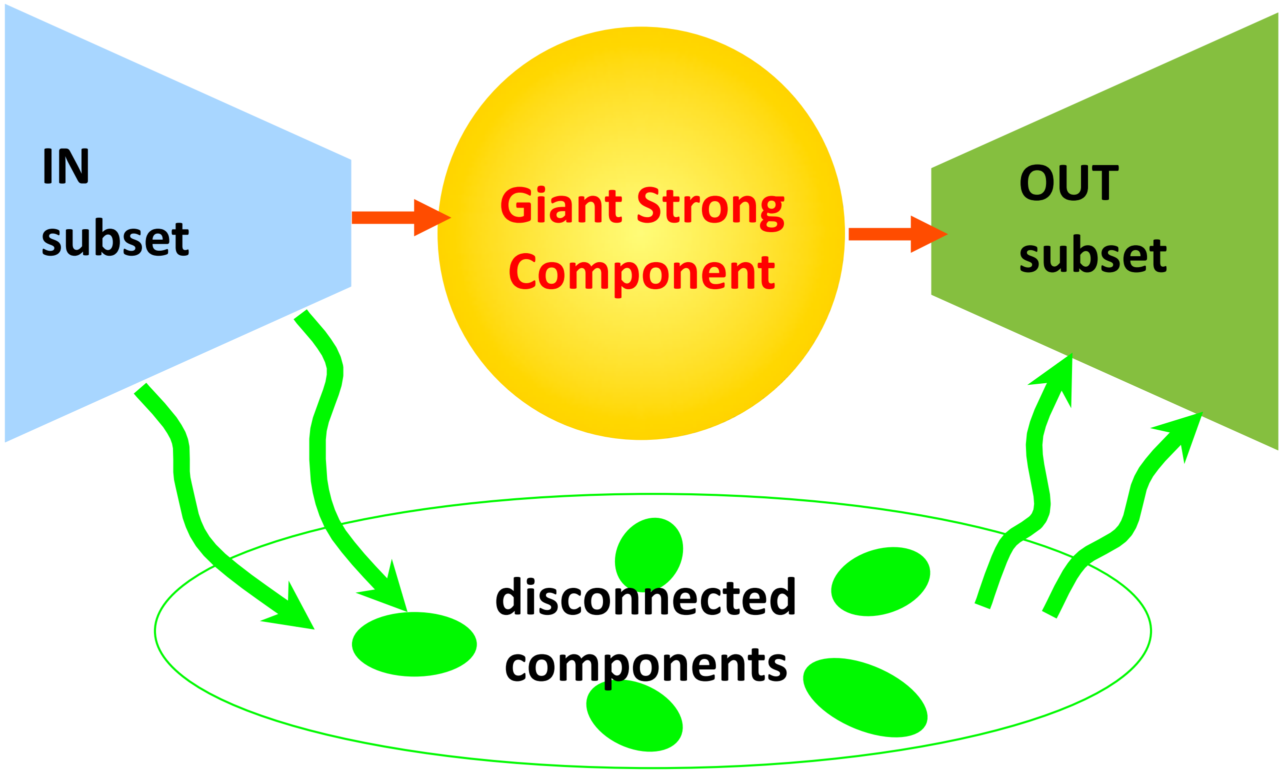
274 out of total
811 metabolites

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 pages network
- Material processing and other tech. systems

Tools for network analysis

- KNEVA http://csb.inf.ed.ac.uk/kneva
- Pajek (good manual and book): http://pajek.imfm.si/doku.php
- Cytoscape http://www.cytoscape.org/ (for Biological networks, mapping data), many plugins
- Bioconductor and R (SNA)
- Java and Python packages (NetworkX)

Network databases

- KEGG
- Metacyc
- Yeast (http://www.comp-sys-bio.org/yeastnet/)
- Human-specific networks
  - Recon1 (Palsson group, 1496 ORFs, 2004 proteins, 2766 metabolites and 3311 reactions)
  - EHMN (Edinburgh group, 2671 compounds, 2322 genes, 2823 reactions 66 pathways)

Metabolomics

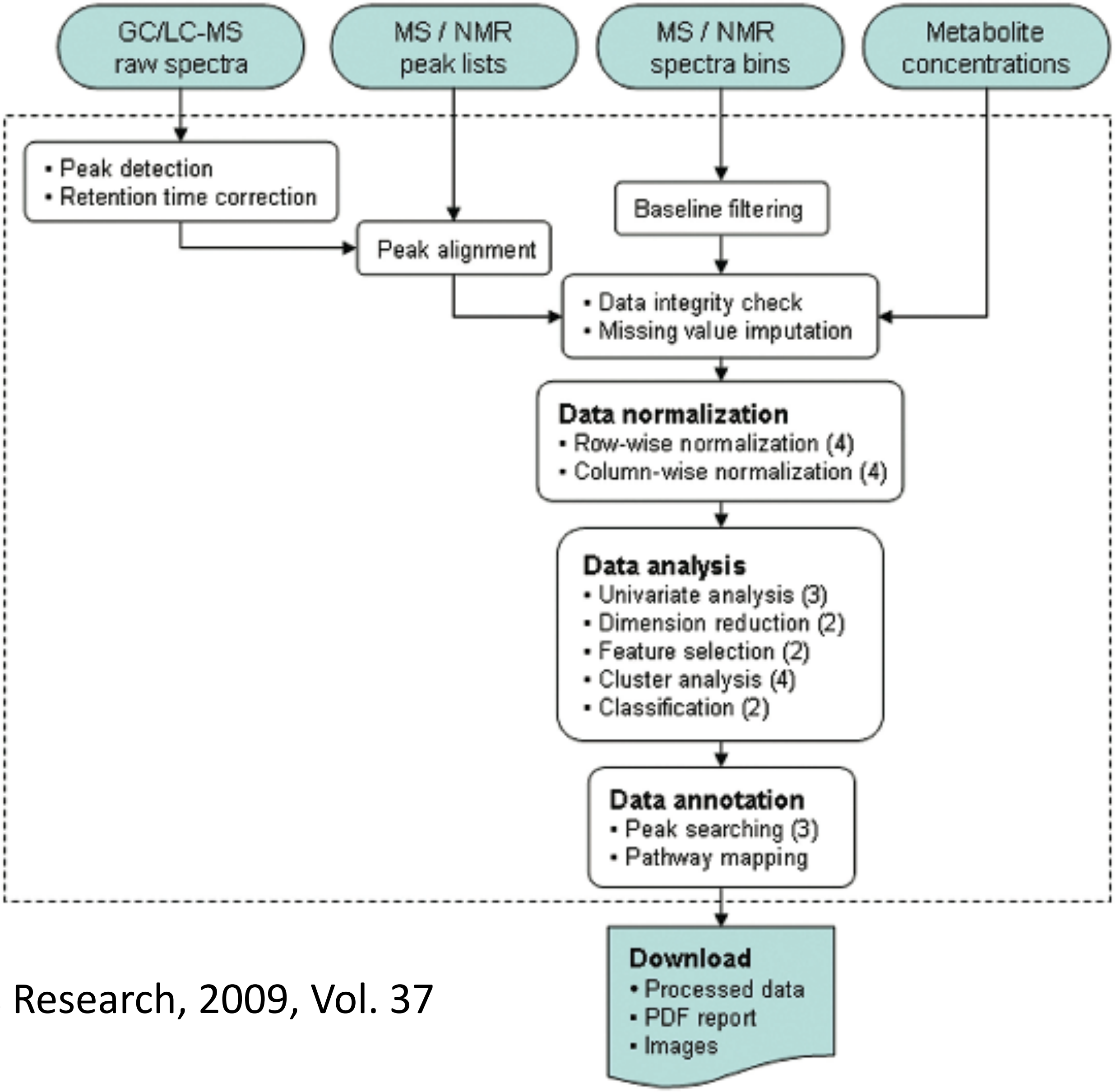
- Chemometric
  - Spectral method based
  - Compounds are not defined
  - Feature extraction
  - Qualitative

- Quantitative
  - MS method based
  - Compounds are defined
  - Quantitative

Pathway analysis in metabolomics

- Quantitative metabolomics data is similar to microarray data
- Can be processed and understood in similar way
- MetaboAnalyst (www.metaboanalyst.ca) on-line tool for data analysis in metabolomics

MetaboAnalyst



GC/LC-MS raw spectra

MS / NMR peak lists

MS / NMR spectra bins

Metabolite concentrations

- Peak detection
- Retention time correction

Peak alignment

Baseline filtering

- Data integrity check
- Missing value imputation

Data normalization
- Row-wise normalization (4)
- Column-wise normalization (4)

Data analysis
- Univariate analysis (3)
- Dimension reduction (2)
- Feature selection (2)
- Cluster analysis (4)
- Classification (2)

Data annotation
- Peak searching (3)
- Pathway mapping

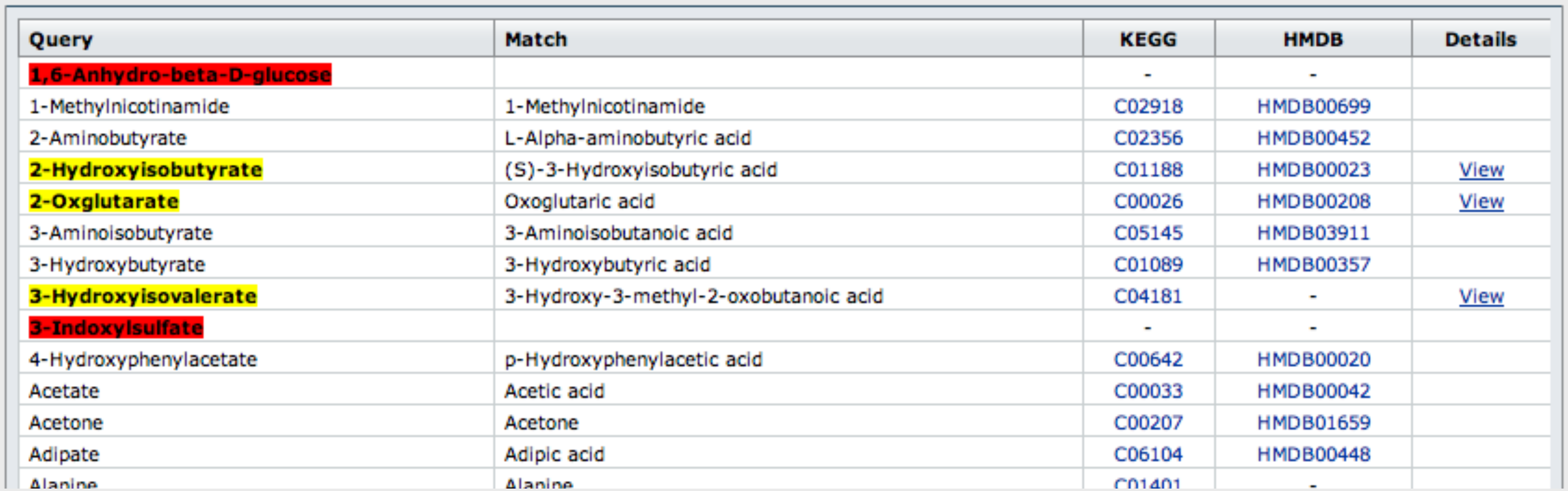
Download
- Processed data
- PDF report
- Images

Nucleic Acids Research, 2009, Vol. 37

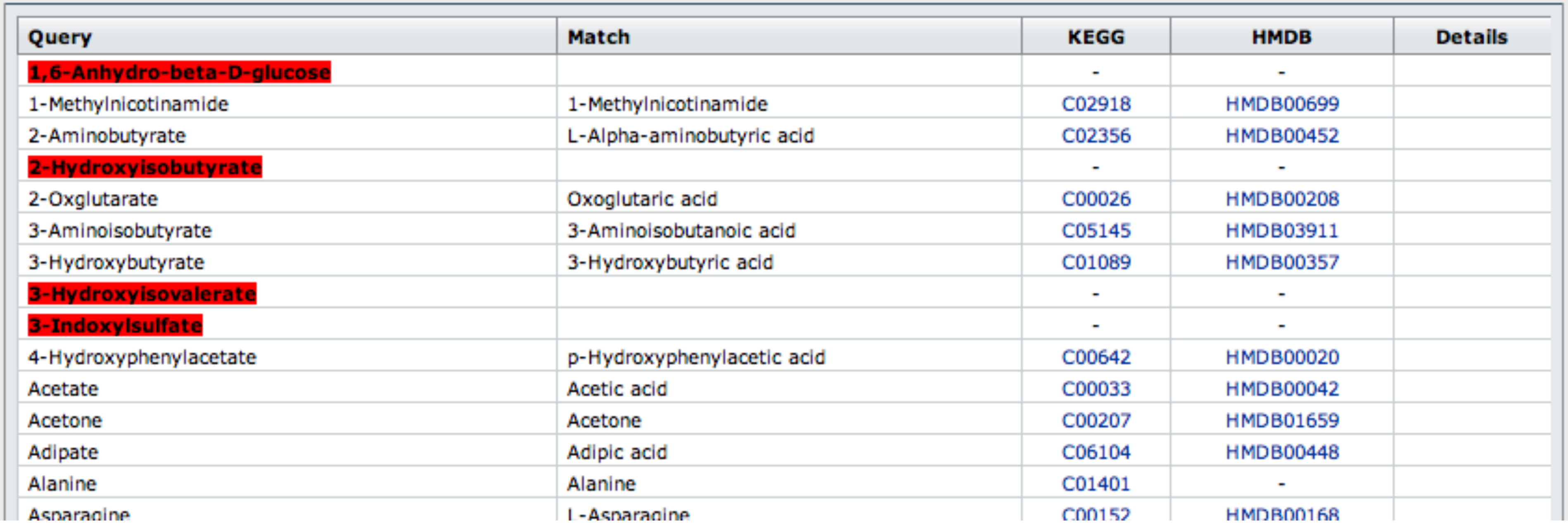
Pathway analysis in metabolomics

- We have data in “standard” format similar to transcriptomics and proteomics
- We have networks and pathways
- We can apply standard pathway analysis
  - Pure metabolomics
  - Metabolomics/transcriptomics
- MetPA (http://metpa.metabolomics.ca) online tool for metabolic pathway analysis

Compound mapping



Query	Match	KEGG	HMDB	Details
1,6-Anhydro-beta-D-glucose		-	-	
1-Methylnicotinamide	1-Methylnicotinamide	C02918	HMDB00699	
2-Aminobutyrate	L-Alpha-aminobutyric acid	C02356	HMDB00452	
2-Hydroxyisobutyrate	(S)-3-Hydroxyisobutyric acid	C01188	HMDB00023	View
2-Oxglutarate	Oxoglutaric acid	C00026	HMDB00208	View
3-Aminoisobutyrate	3-Aminoisobutanoic acid	C05145	HMDB03911	
3-Hydroxybutyrate	3-Hydroxybutyric acid	C01089	HMDB00357	
3-Hydroxyisovalerate	3-Hydroxy-3-methyl-2-oxobutanoic acid	C04181	-	View
3-Indoxylsulfate		-	-	
4-Hydroxyphenylacetate	p-Hydroxyphenylacetic acid	C00642	HMDB00020	
Acetate	Acetic acid	C00033	HMDB00042	
Acetone	Acetone	C00207	HMDB01659	
Adipate	Adipic acid	C06104	HMDB00448	
Alanine	Alanine	C01401	-	

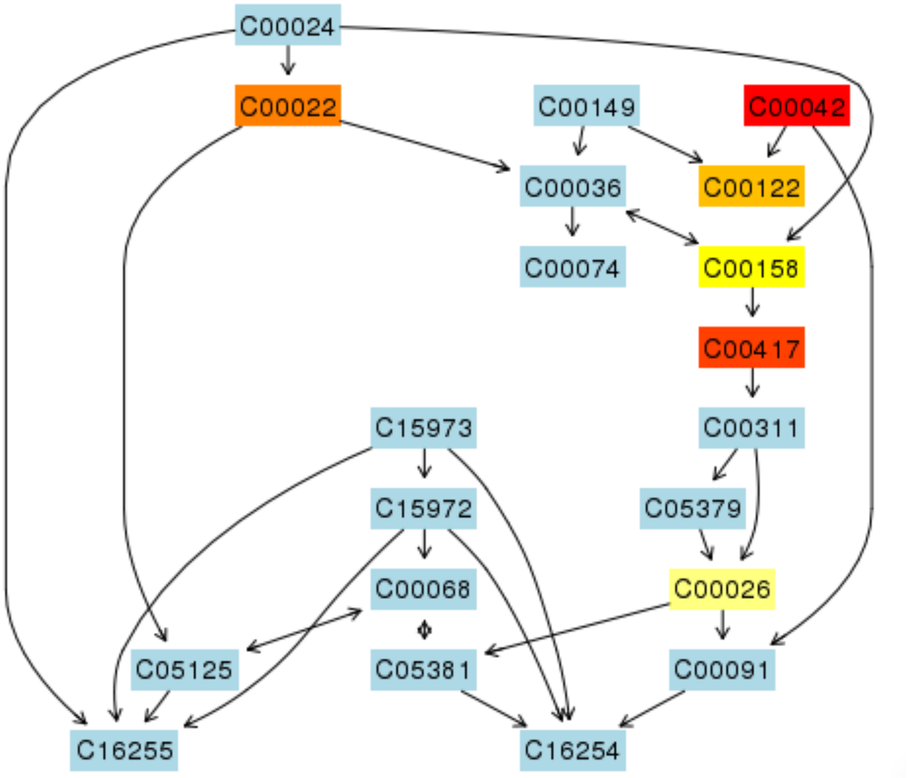


Query	Match	KEGG	HMDB	Details
1,6-Anhydro-beta-D-glucose		-	-	
1-Methylnicotinamide	1-Methylnicotinamide	C02918	HMDB00699	
2-Aminobutyrate	L-Alpha-aminobutyric acid	C02356	HMDB00452	
2-Hydroxyisobutyrate		-	-	
2-Oxglutarate	Oxoglutaric acid	C00026	HMDB00208	
3-Aminoisobutyrate	3-Aminoisobutanoic acid	C05145	HMDB03911	
3-Hydroxybutyrate	3-Hydroxybutyric acid	C01089	HMDB00357	
3-Hydroxyisovalerate		-	-	
3-Indoxylsulfate		-	-	
4-Hydroxyphenylacetate	p-Hydroxyphenylacetic acid	C00642	HMDB00020	
Acetate	Acetic acid	C00033	HMDB00042	
Acetone	Acetone	C00207	HMDB01659	
Adipate	Adipic acid	C06104	HMDB00448	
Alanine	Alanine	C01401	-	
Asparagine	L-Asparagine	C00152	HMDB00168	

Pathway impact

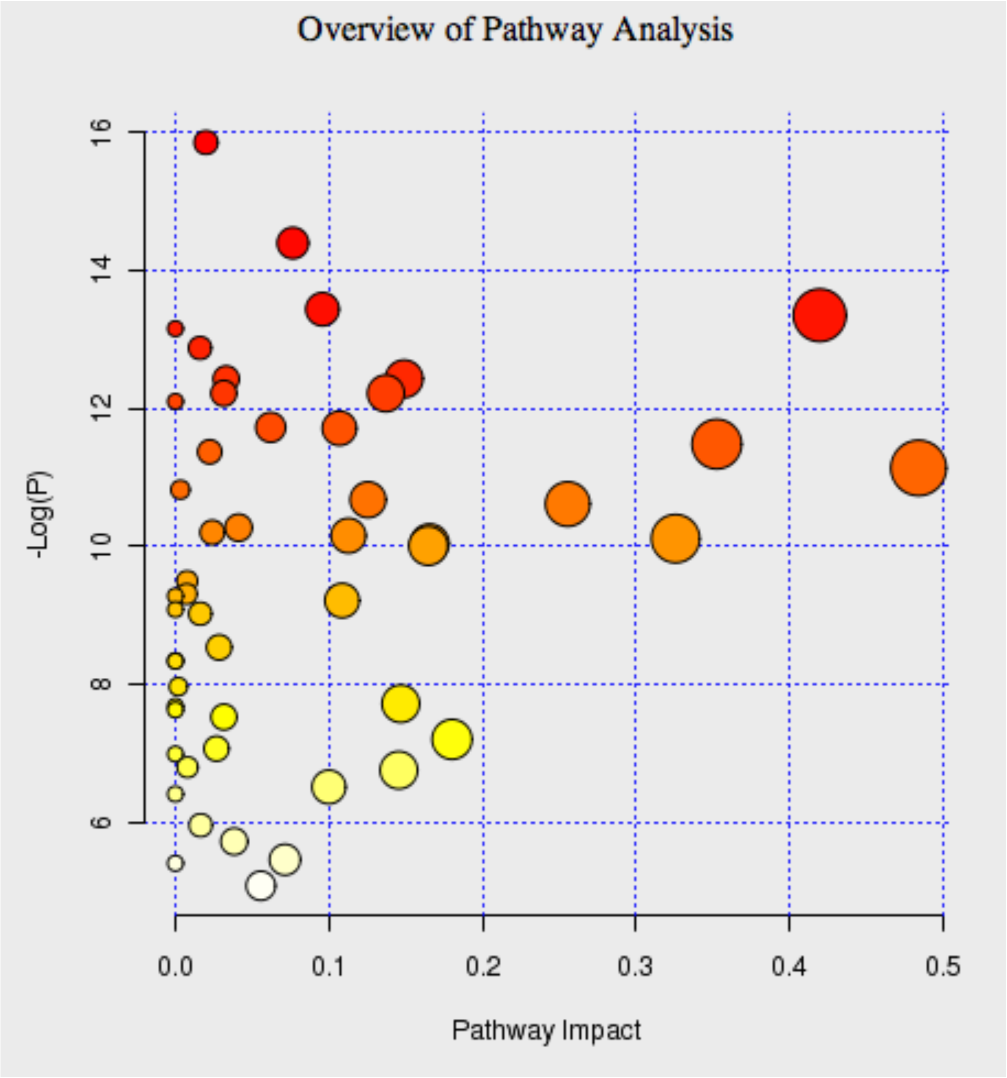
- Calculate importance of metabolites, found in the pathway
  - Degree
  - Betweenness

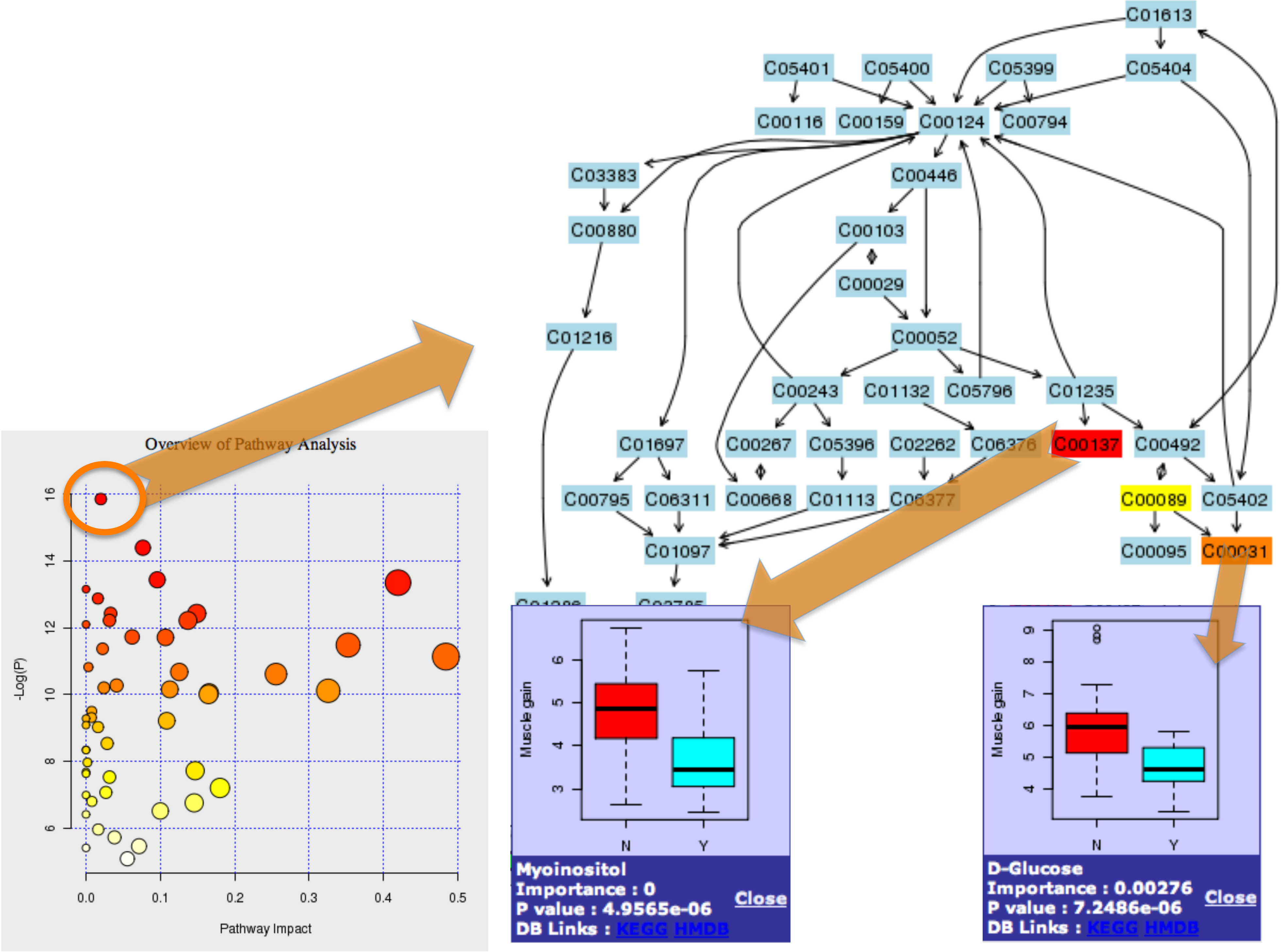
$$PI = \frac{\sum_{found} Imp_i}{\sum_{all} Imp_i}$$



Pathway Impact

- P is calculated from GSE analysis
- Most significant pathways has low impact



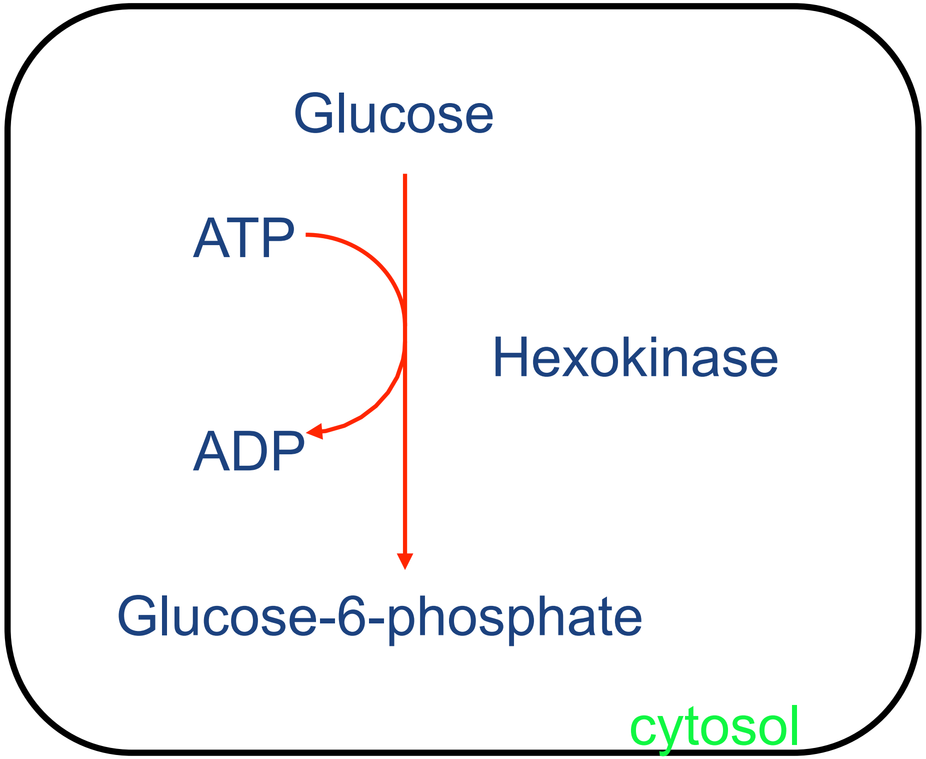


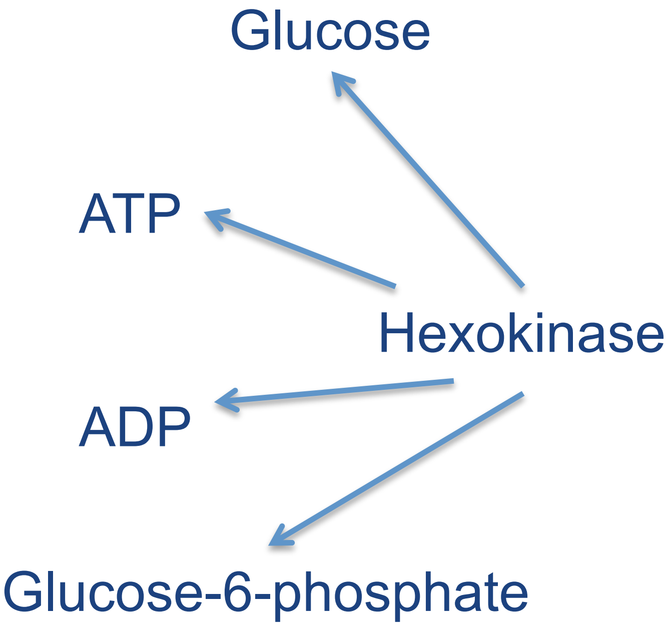
Myoinositol
Importance : 0
P value : 4.9565e-06
DB Links : KEGG HMDB
Close

D-Glucose
Importance : 0.00276
P value : 7.2486e-06
DB Links : KEGG HMDB
Close

Metabolome/Transcriptome

- Patil and Nielsen 2005
- Convert metabolic network into compound-enzyme





Identify reporter metabolite

- Calculate Z-score for each enzyme

$$Z_{ni} = \theta^{-1}(1 - p_i)$$

- Calculate Z-score for metabolite

$$Z_{\mathrm{metabolite}} = \frac{1}{\sqrt{k}} \sum Z_{ni}$$

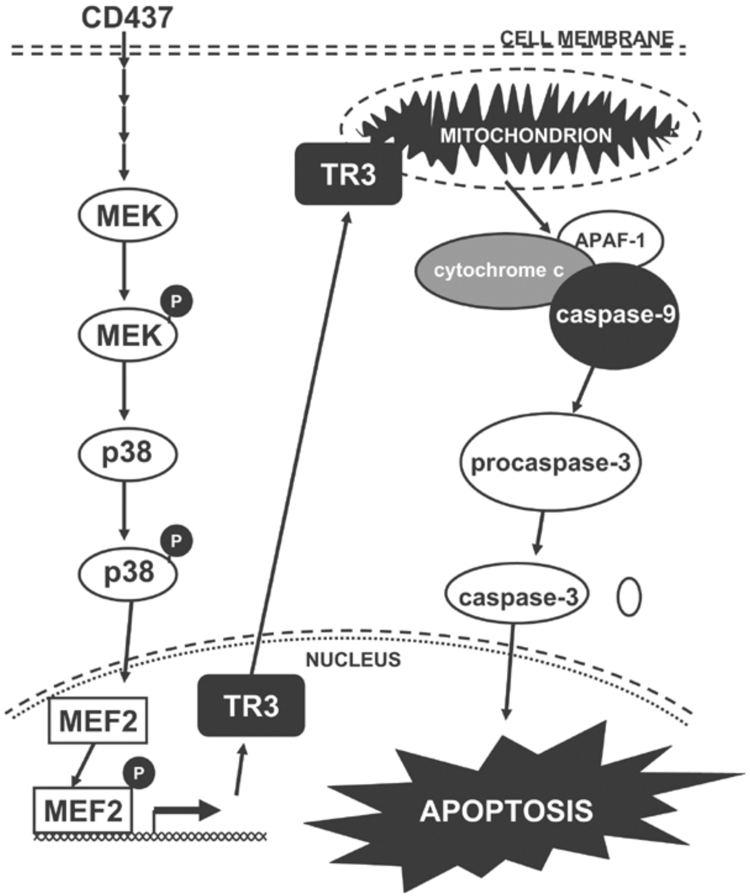
Tools and database

- Experiment repository: www.metabolome-express.org
- Metscape2 metscape.ncibi.org
- Vanted vanted.ipk-gatersleben.de/
- MetPA metpa.metabolomics.ca/
- MetaboAnalyst

SBGN

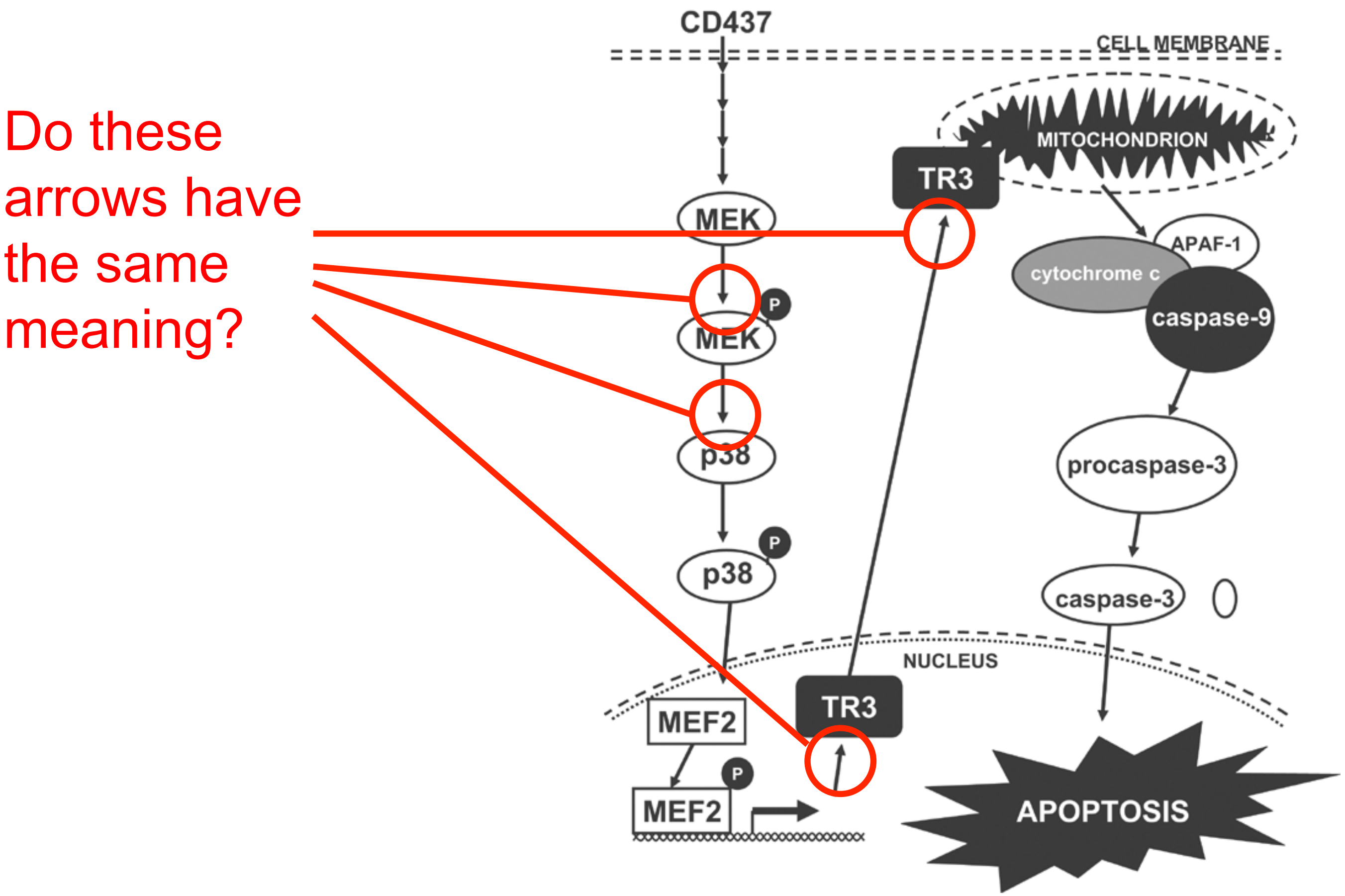
- To analyse
- To discuss
- To share

Can a Biologist Understand This Diagram?



From Holmes WF *et al.* (2003) Early events in the induction of apoptosis in ovarian carcinoma cells by CD437: activation of the p38 MAP kinase signal pathway. *Oncogene* 22: 6377–6386.

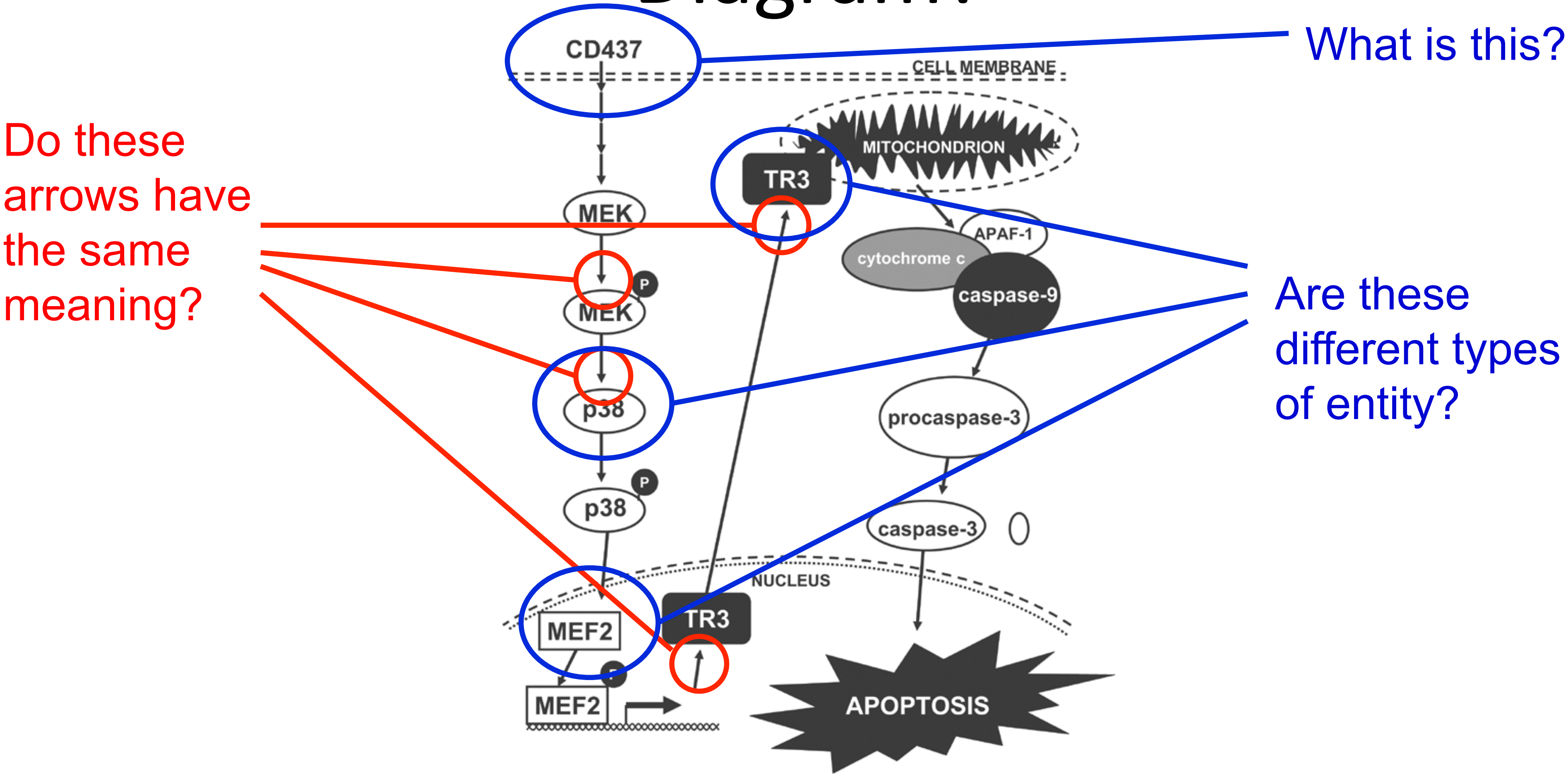
Can a Biologist Understand This Diagram?



Do these arrows have the same meaning?

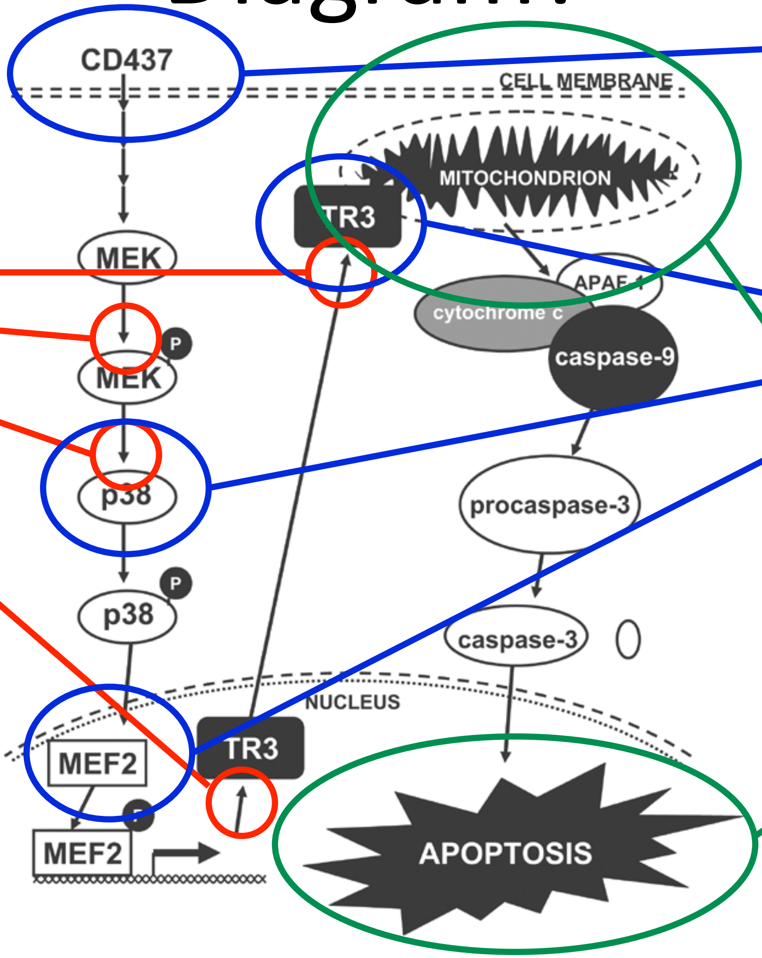
From Holmes WF *et al.* (2003) Early events in the induction of apoptosis in ovarian carcinoma cells by CD437: activation of the p38 MAP kinase signal pathway. *Oncogene* 22: 6377–6386.

Can a Biologist Understand This Diagram?



From Holmes WF *et al.* (2003) Early events in the induction of apoptosis in ovarian carcinoma cells by CD437: activation of the p38 MAP kinase signal pathway. *Oncogene* 22: 6377–6386.

Can a Biologist Understand This Diagram?



What is this?

Do these arrows have the same meaning?

Are these different types of entity?

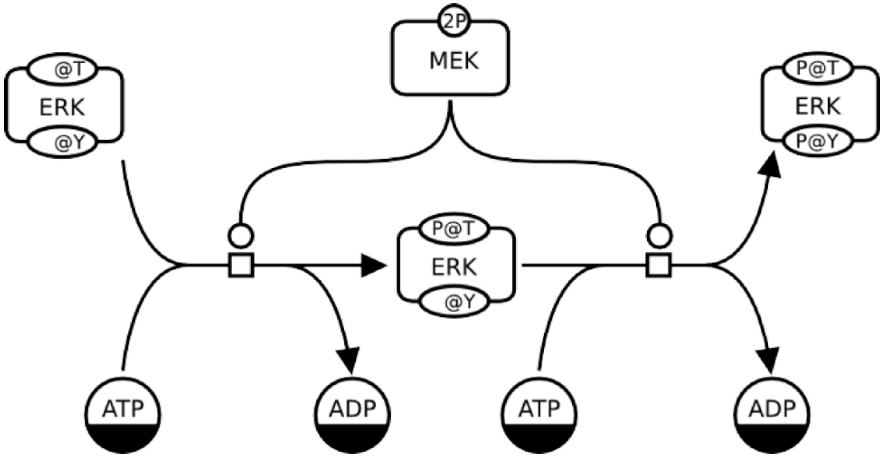
How about these?

From Holmes WF *et al.* (2003) Early events in the induction of apoptosis in ovarian carcinoma cells by CD437: activation of the p38 MAP kinase signal pathway. *Oncogene* 22: 6377–6386.

What Happens if one Cannot Read the Blueprint

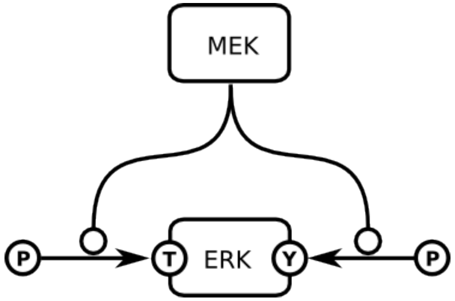
Graph Trinity: Three Languages in One

Process Description maps



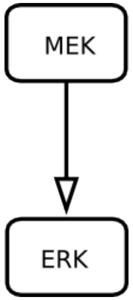
- Unambiguous
- Mechanistic
- Sequential
- Combinatorial explosion

Entity Relationship maps



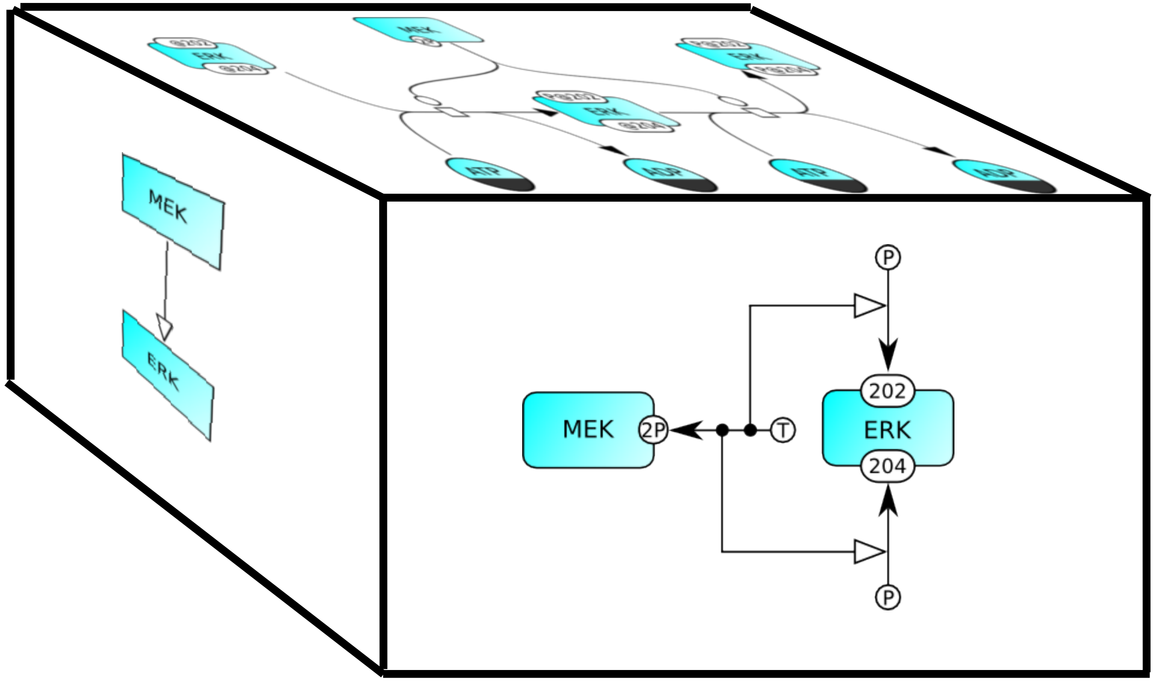
- Unambiguous
- Mechanistic
- Non-Sequential

Activity Flow maps



- Ambiguous
- Conceptual
- Sequential

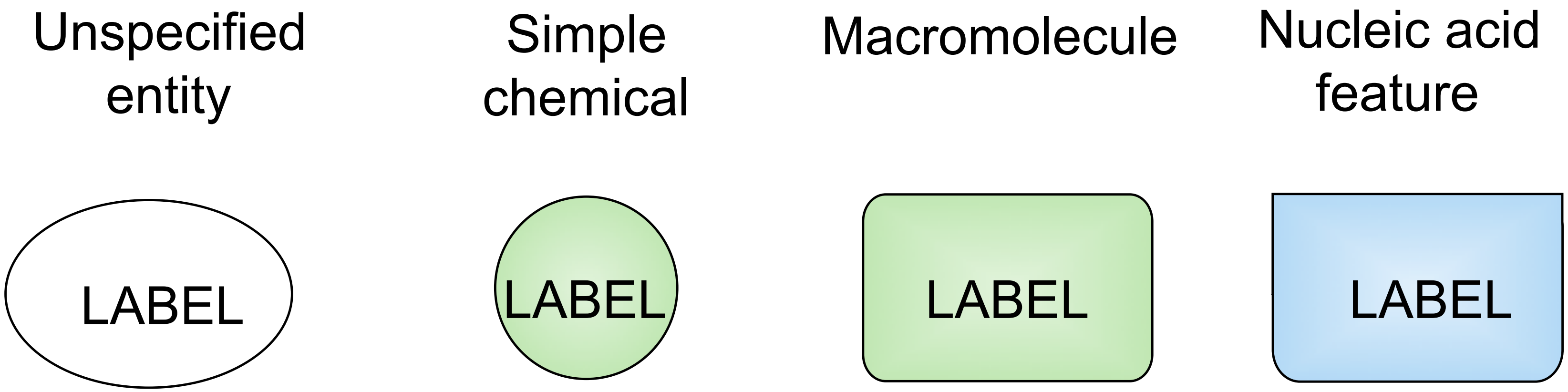
Three Orthogonal Projections of Biology



SBGN Process Description Language

- Inspired and based on Kitano's Process Diagram Notation
- A Process Description (PD) Diagram represents all molecular processes and interactions occurring between various biochemical entities
- It depicts how entities transition forms as a result of biochemical reactions (including non-covalent modifications such as binding)
- Most of the classic metabolic pathways (e.g., glycolysis and TCA cycle) in biochemistry textbooks were drawn in this approach
- Though not the conventional approach for drawing signaling pathways, this approach captures the details of biochemical reactions within the pathway network and provides, in most cases, unambiguous interpretation of pathway mechanisms

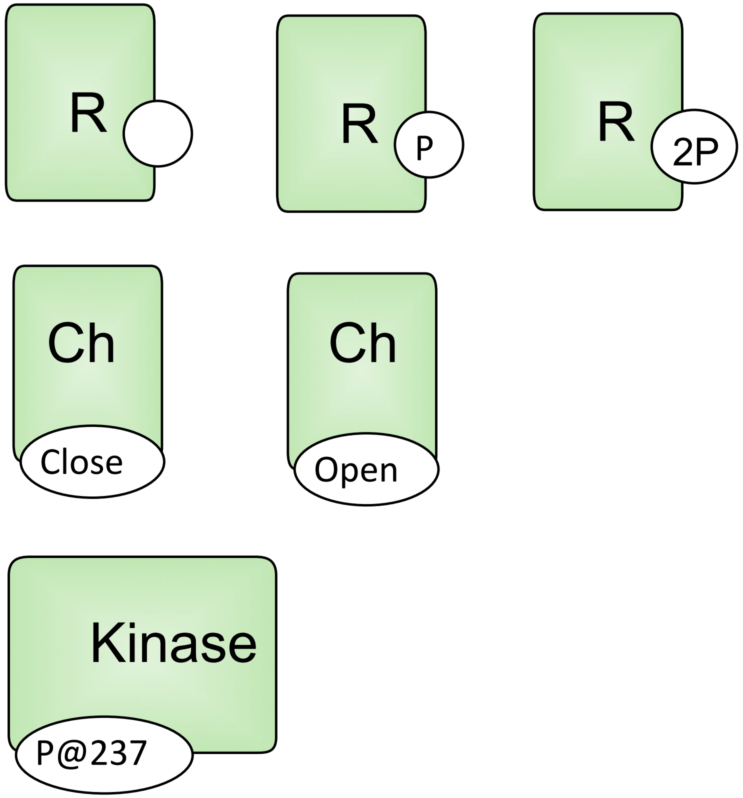
Entity Types



Macromolecules: biochemical substances that are built up from the covalent linking of pseudo-identical units. Examples of macromolecules include proteins, nucleic acids (RNA, DNA), and polysaccharides (glycogen, cellulose, starch, etc.).

Macromolecular Pools: State Variables

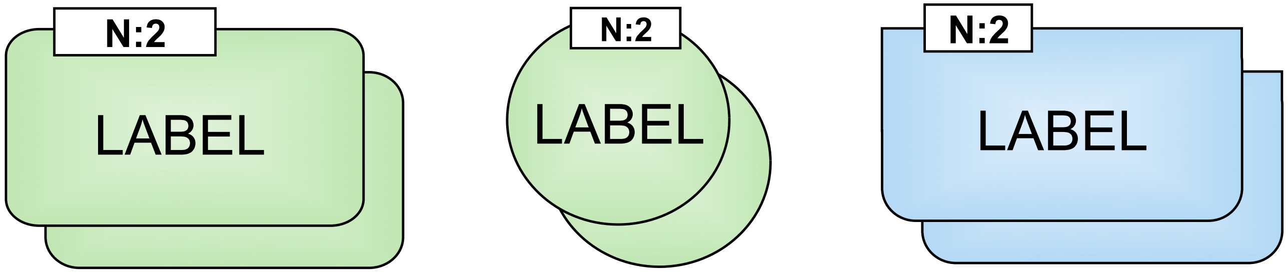
- Pool is set of molecules somehow undistinguishable
- Molecules can be in different state
  - (Non)phosphorylated
  - Open/close channel
  - Modified at some state



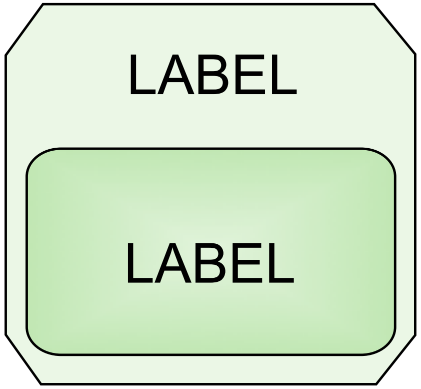
Complex and Multimer

- Represents complexes of molecules held together by non-covalent bonds
- Multimer require cardinality
- Can have state variables
  - In multimer it means that all monomers have same state
  - Use complex if not the same states

Multimers

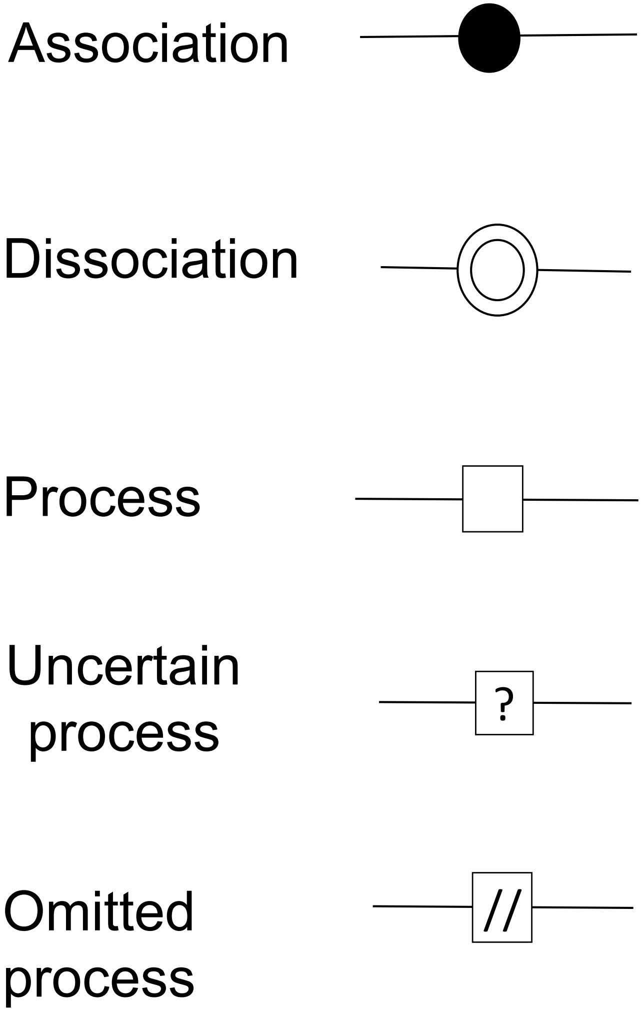


Complex



Key Concept: Process

- Process: conversion of element of one pool to another
- Special cases:
  - Non-covalent binding
    - Association
    - Dissociation
  - Incompleteness
    - Uncertain process
    - Omitted process



Association

Dissociation

Process

Uncertain process

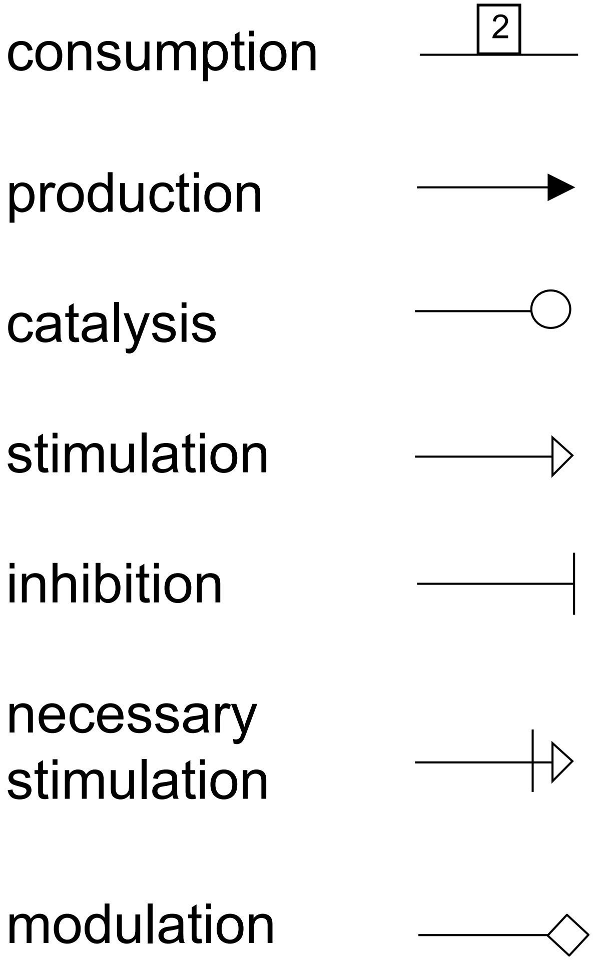
?

Omitted process

//

Arcs

- Using pools by process
  - Consumption/production
  - Stoichiometry (optional)
- Regulating process rate
  - Stimulation
  - Inhibition
  - Catalysis
- Requirement for process
  - Necessary stimulation



consumption

production

catalysis

stimulation

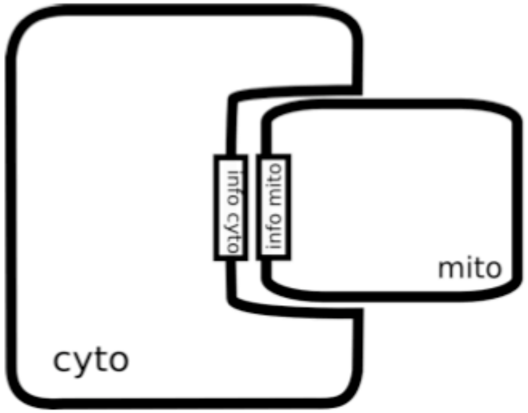
inhibition

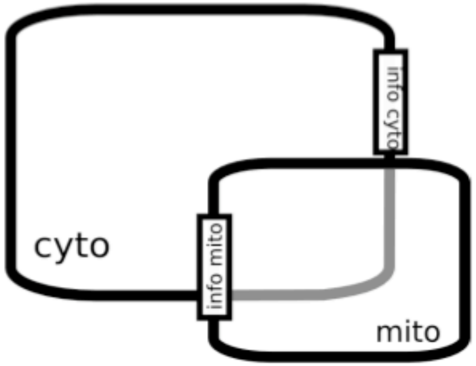
necessary stimulation

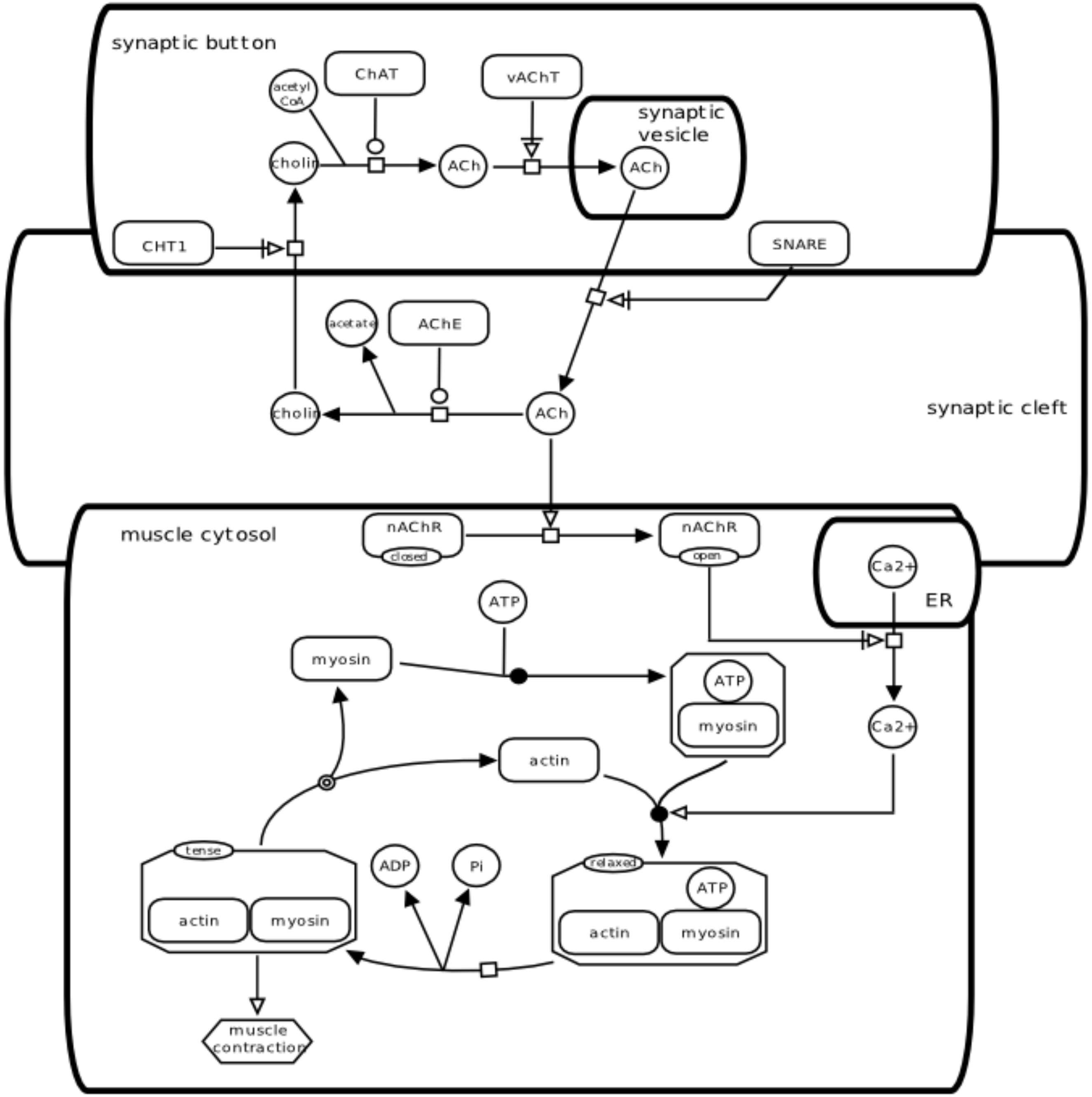
modulation

Compartments

- Container to represent physical or logical structure
  - Free form
  - Visually thicker line
- The same entity pools in different compartments are different
- Compartments are independent
- Overlapping do not mean containment







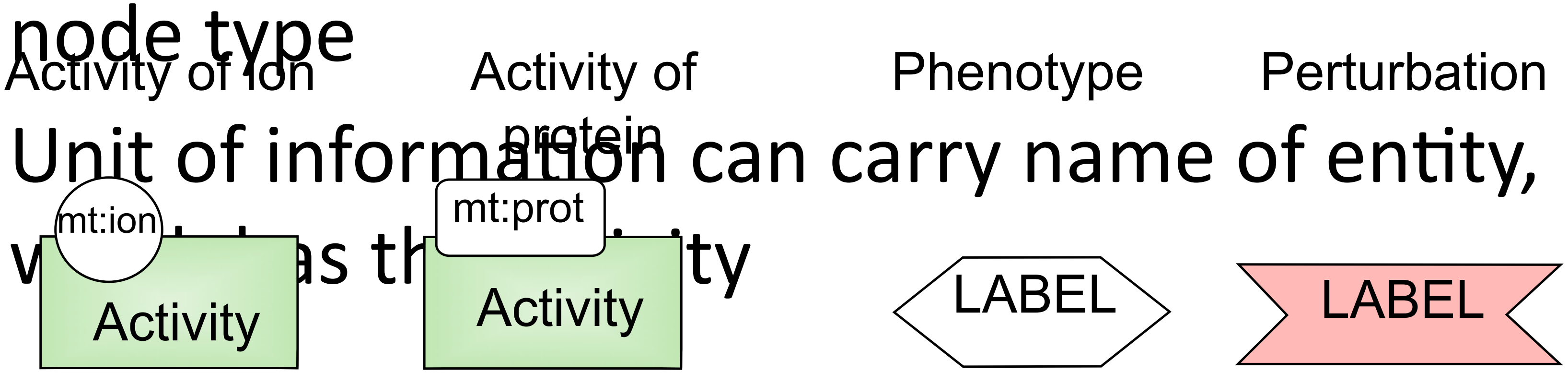
Neuro-muscular Junction

Activity Flow: Abstraction

- Main concept is Biological Activity
  - Each node represents an activity, but not the entity
  - Multiple nodes can be used to represent activities from one entity (e.g., receptor protein kinase)
  - One node can be used to represent activities from a group of entities (e.g., a complex, generics etc.)

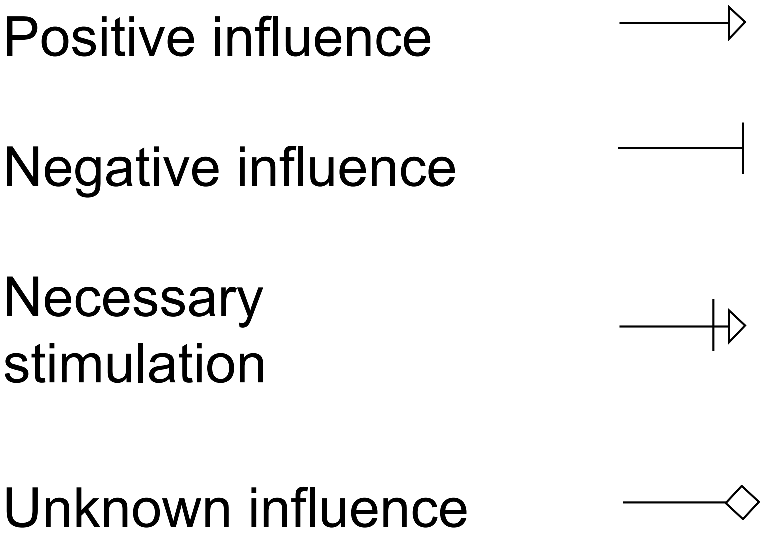
Material and Conceptual Types in AF

- Activity node is rectangular to emphasize similarity to reaction
- Unit of information has shape according to node type
- Unit of information can carry name of entity, which has the activity



Regulatory Arcs

- Operates on activities
- Shows influences
  - – Positive
    - Catalysis
    - Stimulation
  - – Negative
    - Inhibition
  - – Required
    - Necessary stimulation



Positive influence

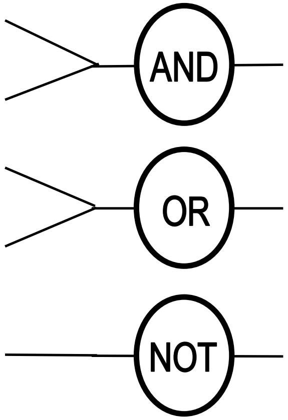
Negative influence

Necessary stimulation

Unknown influence

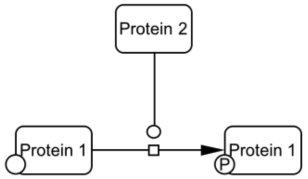
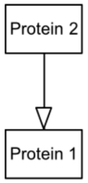
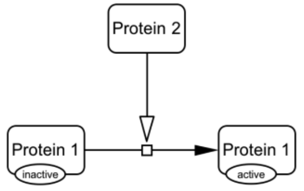
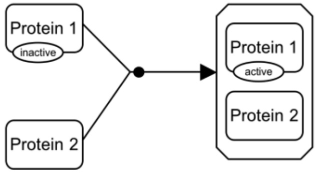
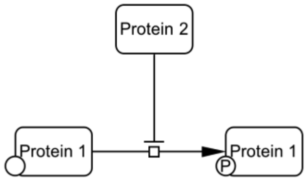
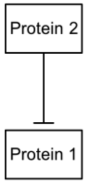
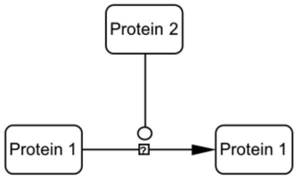
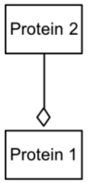
Logical Gates

- Three main logic operations
  - AND: all are required
  - OR: any combination is required
  - NOT: prevent influence
- Crucial for AF
  - No complex
  - No outcome
  - No modifications

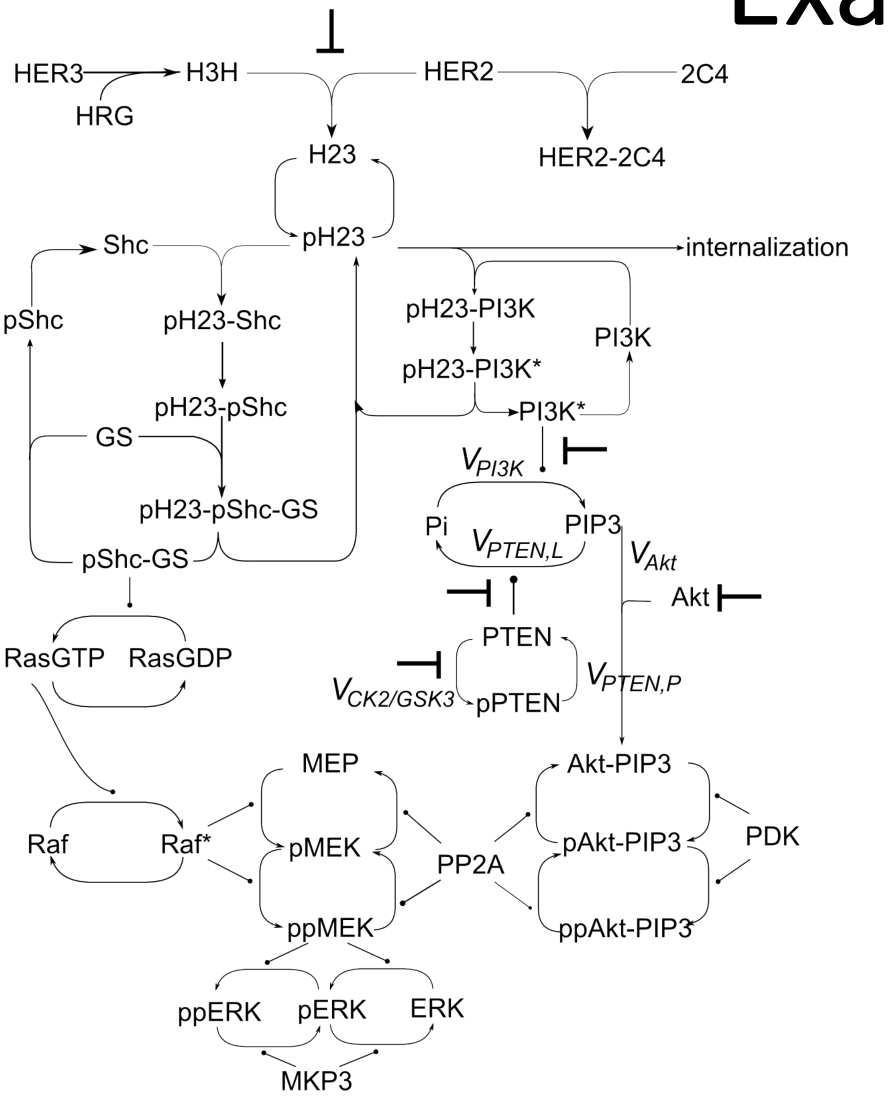


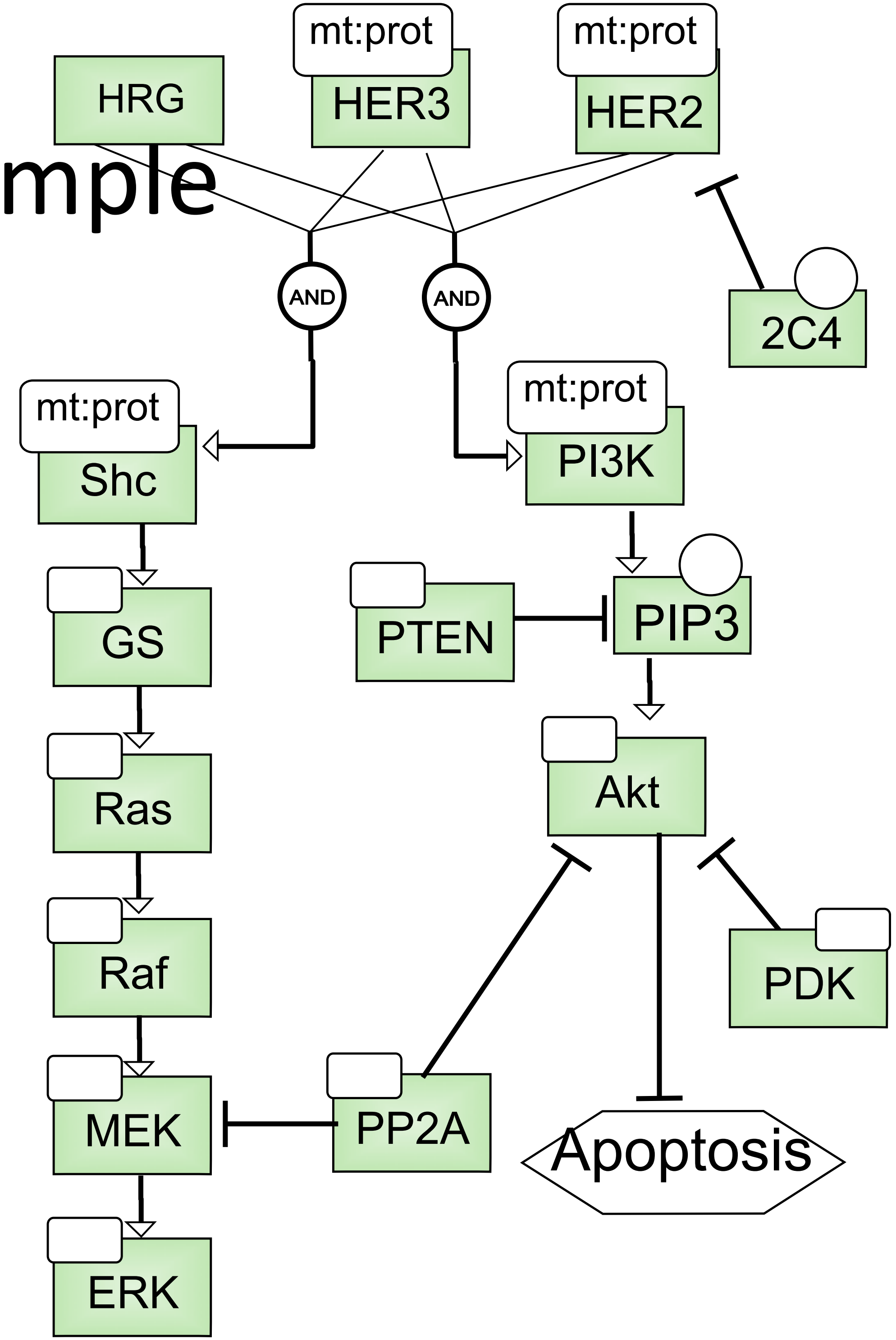
Activity Flow Map is Ambiguous

- AF diagrams are ambiguous
- An AF diagram should be associated with either a PD or ER diagram, if possible
- Automatic conversion between PD and/or ER to AF is planed

Process Diagram		Activity Flow	
	Protein 2 catalyzes the transition of Protein 1 from non-phosphorylated form to phosphorylated form.		Protein 2 positively influences Protein 1.
	Protein 2 stimulates the conversion of Protein 1 from an inactive state to a active state.		
 Protein 1 inactive Protein 2 Protein 1 active Protein 2	Inactive form of Protein 1 associates with Protein 2 to form a Protein 1/Protein 2 complex, and Protein 1 becomes active through the association.		
	Protein 2 inhibits the transition of Protein 1 from non-phosphorylated form to phosphorylated form.		Protein 2 negatively influences Protein 1.
	Protein 2 catalyzes an unknown transition of Protein 1		Protein 2 influences Protein 1 with unknown mechanism.

Example


HER3
H3H
HRG
HER2
2C4
H23
HER2-2C4
pH23
internalization
Shc
pShc
pH23-Shc
pH23-PI3K
PI3K
pH23-PI3K*
pH23-pShc
PI3K*
GS
V_{PI3K}
pH23-pShc-GS
Pi
PIP3
$V_{PTEN,L}$
V_{Akt}
pShc-GS
Akt
PTEN
RasGTP
RasGDP
$V_{CK2/GSK3}$
pPTEN
$V_{PTEN,P}$
MEP
Akt-PIP3
Raf
Raf*
pMEK
PP2A
pAkt-PIP3
PDK
ppMEK
ppAkt-PIP3
ppERK
pERK
ERK
MKP3


HRG
mt:prot
HER3
mt:prot
HER2
AND
AND
2C4
mt:prot
Shc
mt:prot
PI3K
GS
PTEN
PIP3
Ras
Akt
Raf
PDK
MEK
PP2A
Apoptosis
ERK