Systems Biology

(2) Networks: Representation & static analysis

David Gilbert

Bioinformatics Research Centre

www.brc.dcs.gla.ac.uk

Department of Computing Science, University of Glasgow

Module outline

- ‘Putting it all together’ - Systems Biology
- Motivation
- Biological background
- Modelling
  - Network Models
  - Data models
- Analysis:
  - Static
  - Dynamic
- Standardisation (sbml & sbw)
- Technologies
- Current approaches
- Systems robustness

Admin

- Term 2; 2006-2007 Fri 23/2, Mon 26/2, Wed 28/2, Fri 2/3
  - Lectures: 10.30-12.00, A230 Joseph Black
  - Labs: 13.00-15.00, 101 Davidson

- Module information, resources & reading list: **www.brc.dcs.gla.ac.uk/~drg/courses/sysbiomres**

- Assessment: 1 Coursework + Exam question
- Summer project - optional

- Course staff
  - Lecturer: Professor David Gilbert
  - Demonstrator: Ms Xu Gu

- Additional: www.brc.dcs.gla.ac.uk/seminars (Fridays 11-12, BRC)

Note: Text-mining lecture

- ‘Text-mining for Bioinformatics & Systems Biology’, lecturer: Tamara Polajnar
- Part of the ‘Bioinformatics’ module in Computing Science
  www.brc.dcs.gla.ac.uk/~drg/courses/bioinformaticsHM
- Tuesday 27/2, 9-10 Modern Languages Room 208
  - Plus possible lab: 10-11

Resources

- DRG’s handouts
- www.brc.dcs.gla.ac.uk/~drg/bioinformatics/resources.html
- www.ebi.ac.uk/2can
  - Bioinformatics educational resource at the EBI
- International Society for Computational Biology: www.iscb.org
  - very good rates for students, and you get on-line access to the Journal of Bioinformatics.
- **Broder S, Venter J C, Whole genomes: the foundation of new biology and medicine, Curr Opin Biotechnol. 2000 Dec;11(6):581-5.**
- **Kitano H. Looking beyond the details: a rise in system-oriented approaches in genetics and molecular biology. Curr Genet. 2002 Apr;41(1):1-10.**
- Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. Network motifs: simple building blocks of complex networks. Science. 2002 Oct 25;298(5594):824-7.
- Yuri Lazebnick. Can a biologist fix a radio? - Or, What I learned while studying Apoptosis. Cancer Cell september 2002 vol 2 179-182.
- Post Genome Informatics Kanehisa. Publisher OUP. Year 2000. Isbn 0198503261. Category background

Lecture outline

- Data models for Networks, pathways
- Sets
- Graphs
- Analysis
  - Some algorithms over graphs
  - Paths, circuits, searching
  - Network motifs
  - Network properties

Motivation

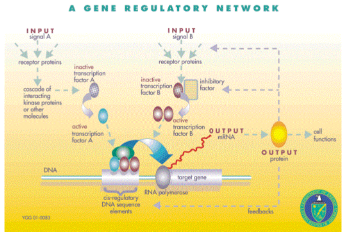
- We need to ***model*** aspects of an organism in order to be able to ***analyse*** its behaviour and function.
- In systems biology we are interested in the way in which biological ***components*** are composed so that they interact together in some way.
- Often the way in which a ***network of interactions*** can be modelled is by a ***graph***.
- We can then use techniques from graph theory to analyse some features and properties of these networks.
- We will also often need to visualise these networks somehow.
- We will also need to store the biological data in a database whose schema may be interpreted as a graph.

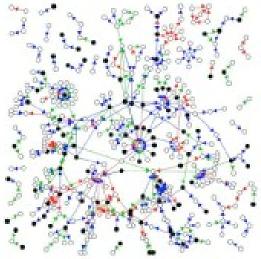
Terminology: Pathways or Networks?

- Pathways implies ‘paths’ - sequences of objects
- Networks - more complex connectivity
- Both are represented by ***graphs***
- Networks: generic; Pathways: specific (?)
  - ‘Signal transduction networks’
  - ‘The ERK signal transduction pathway’

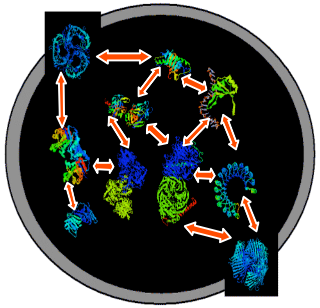
Networks

- Gene regulation

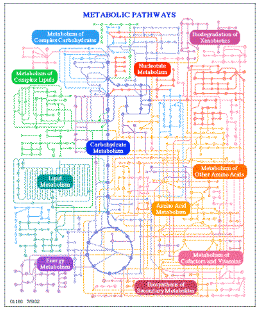




- Protein-protein interaction

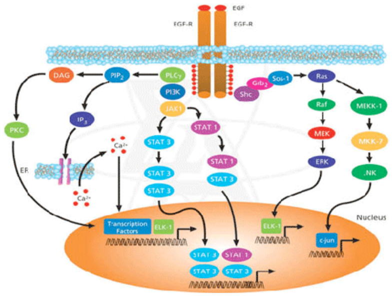


- Metabolic

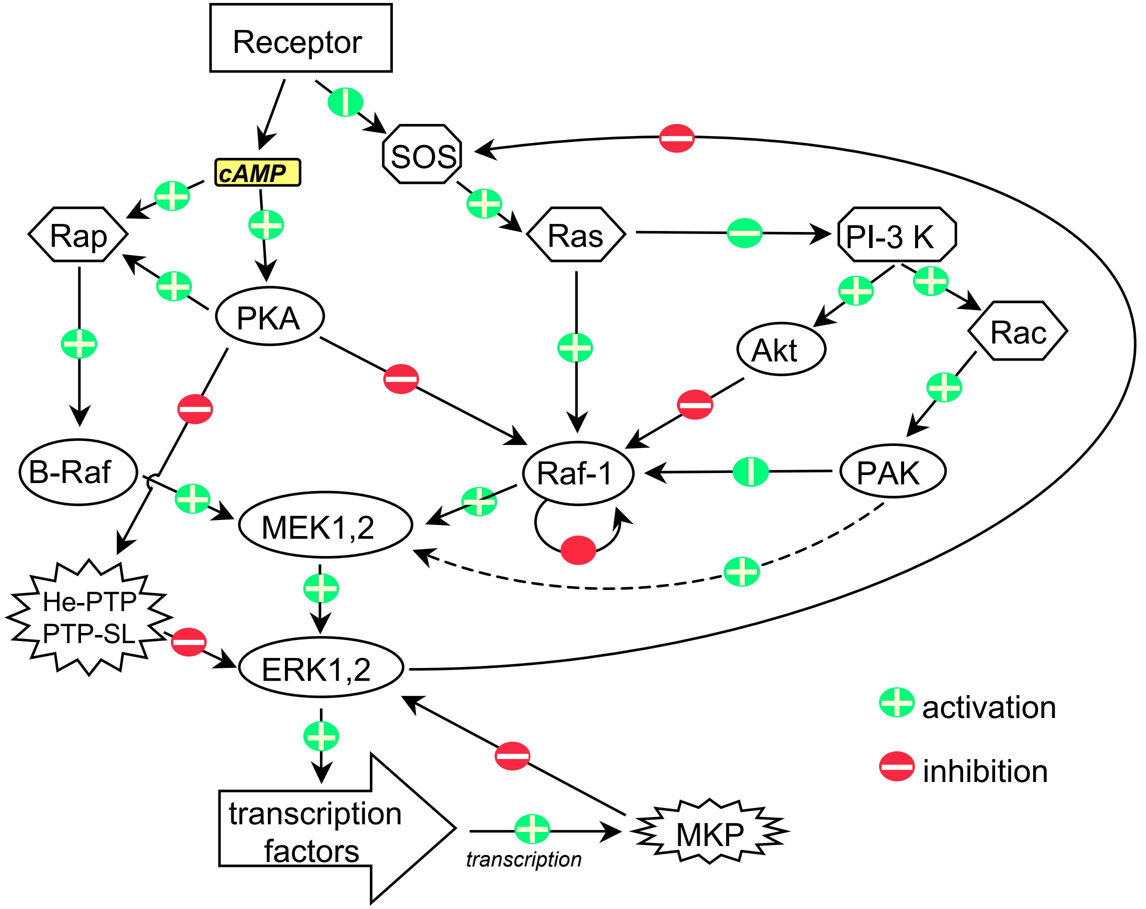


- Developmental

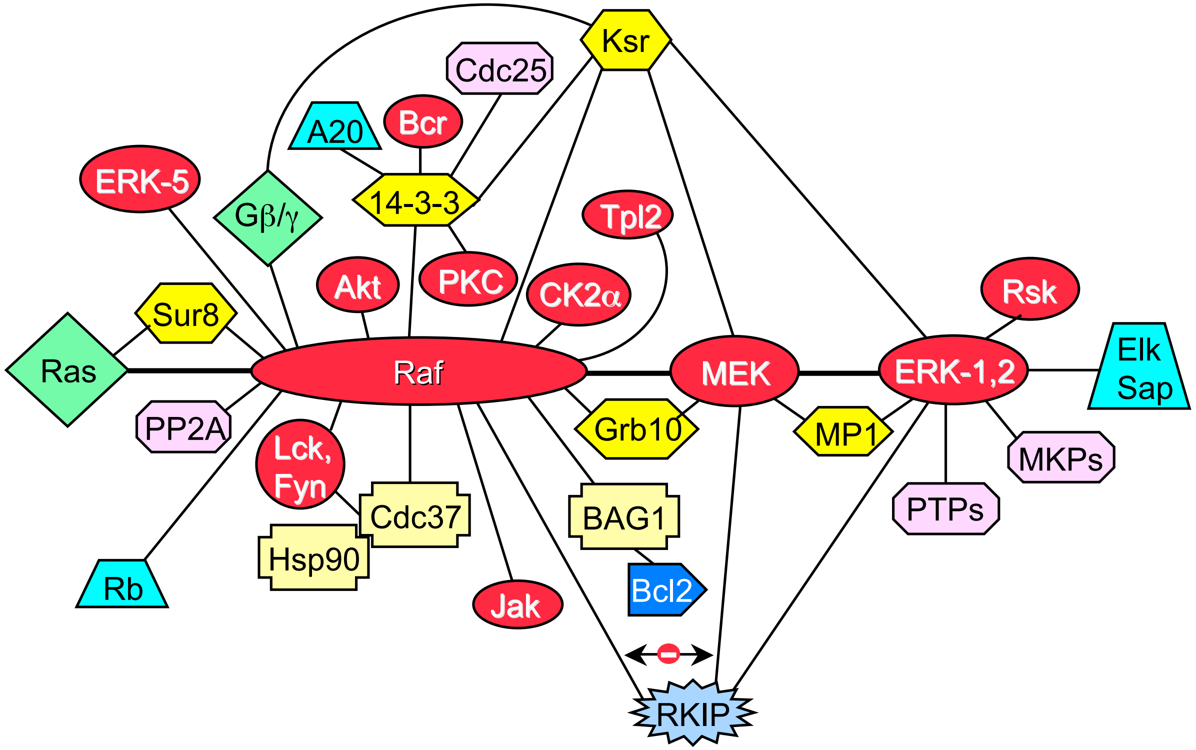
- Signalling



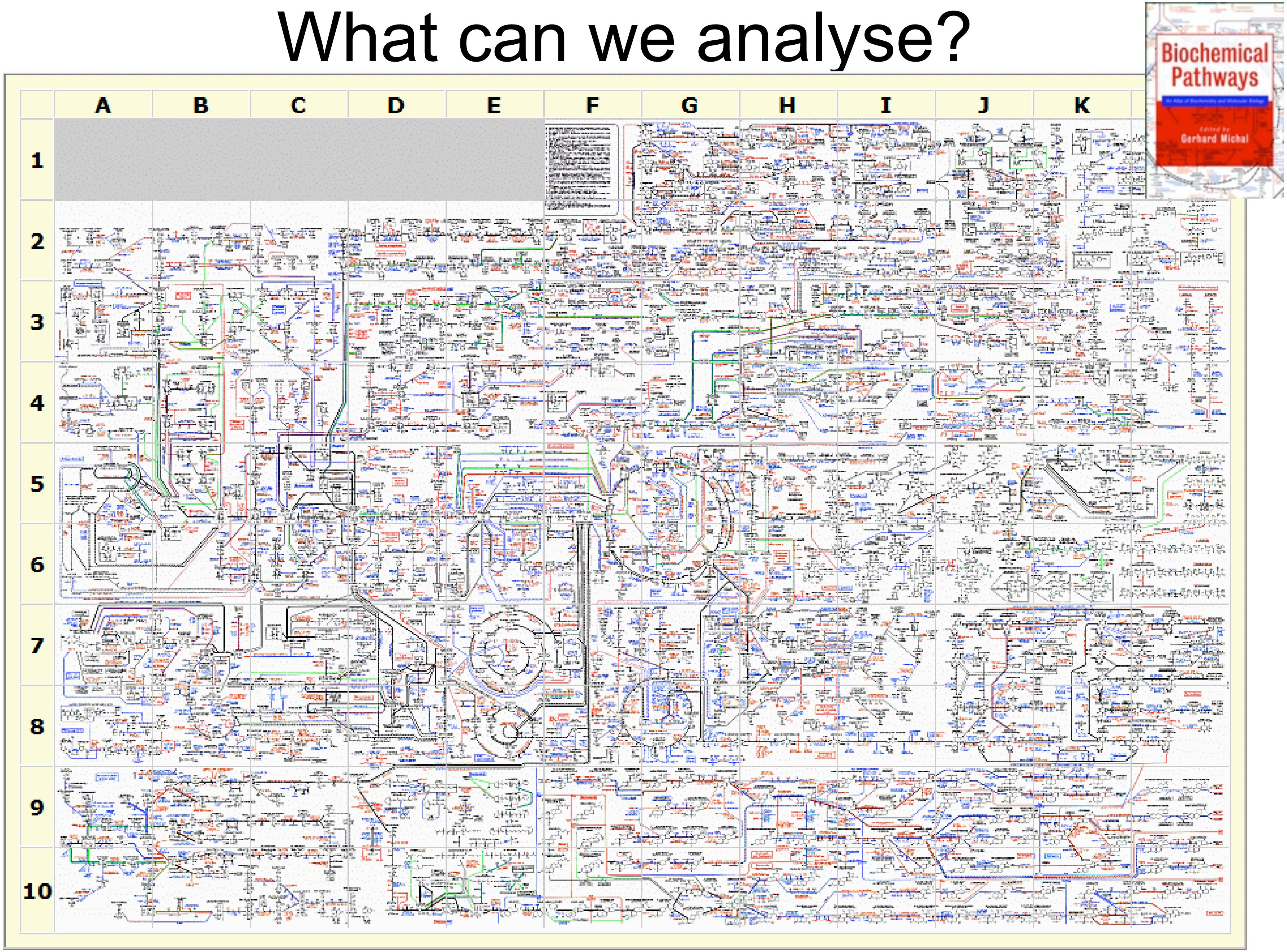
This pathway looks nice and linear, but it is embedded in a network…



… is regulated by protein:protein interactions



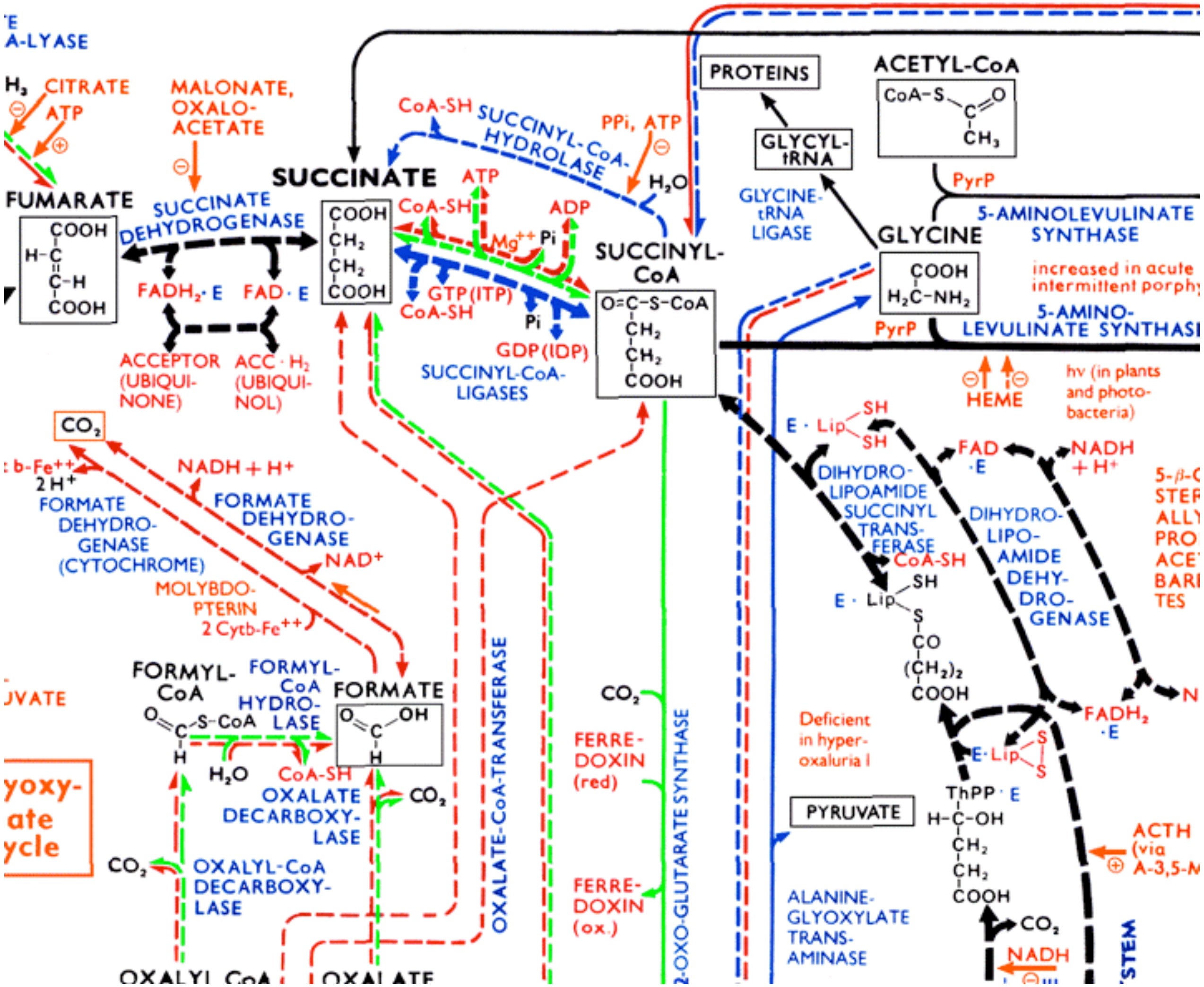
What can we analyse?



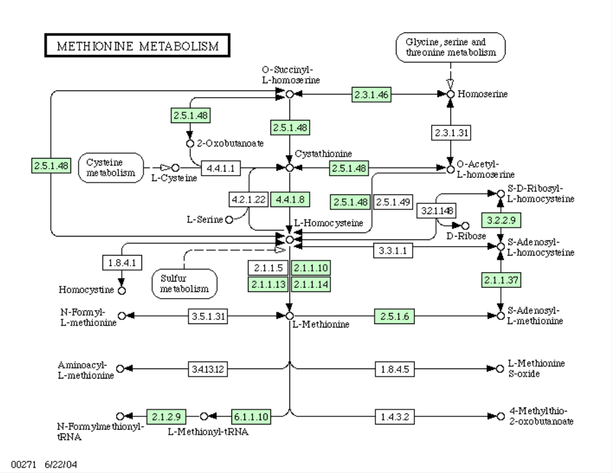
http://ca.expasy.org/tools/pathways/

Pathway templates & variations

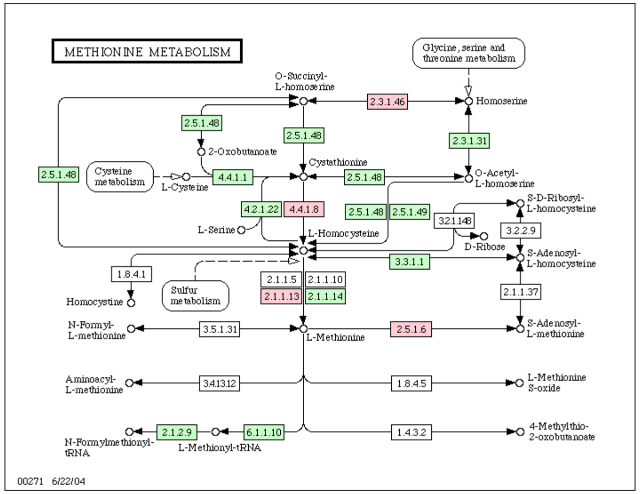
→ general biochemical pathways, → animals,
→ higher plants, → unicellular organisms



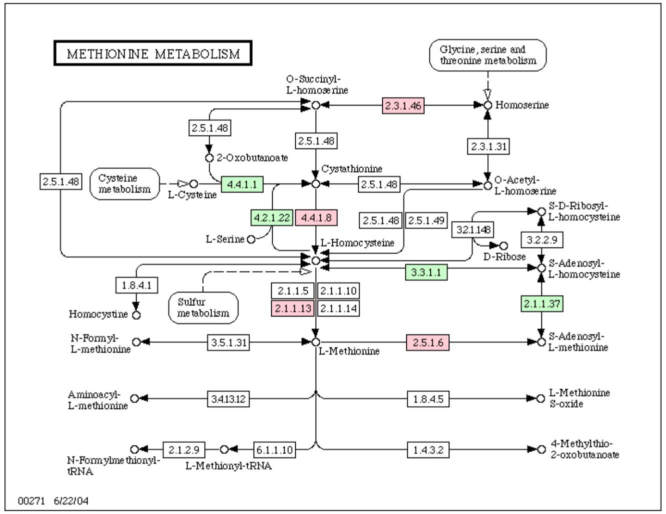
Pathway orthologues



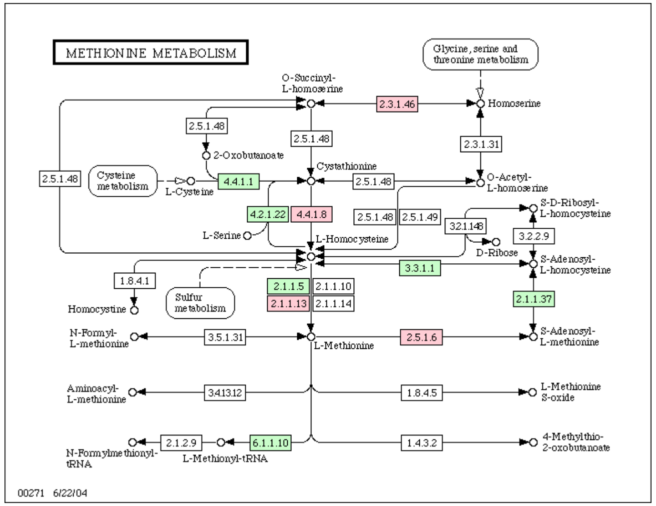
Escherichia coli K-12 MG1655



Yeast



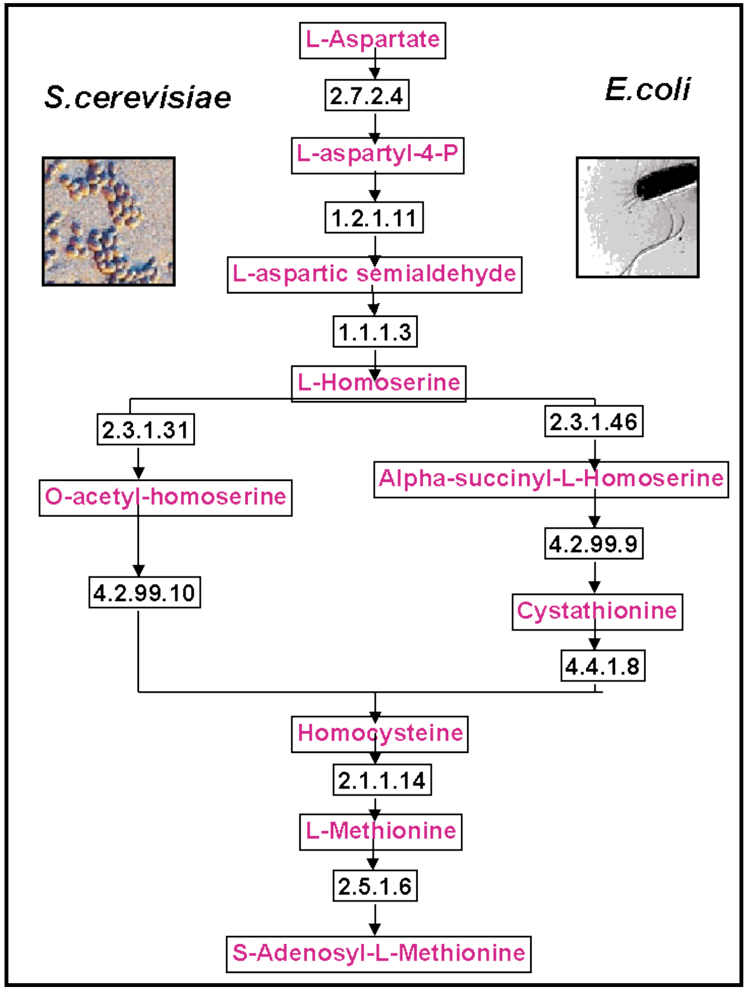
Fly



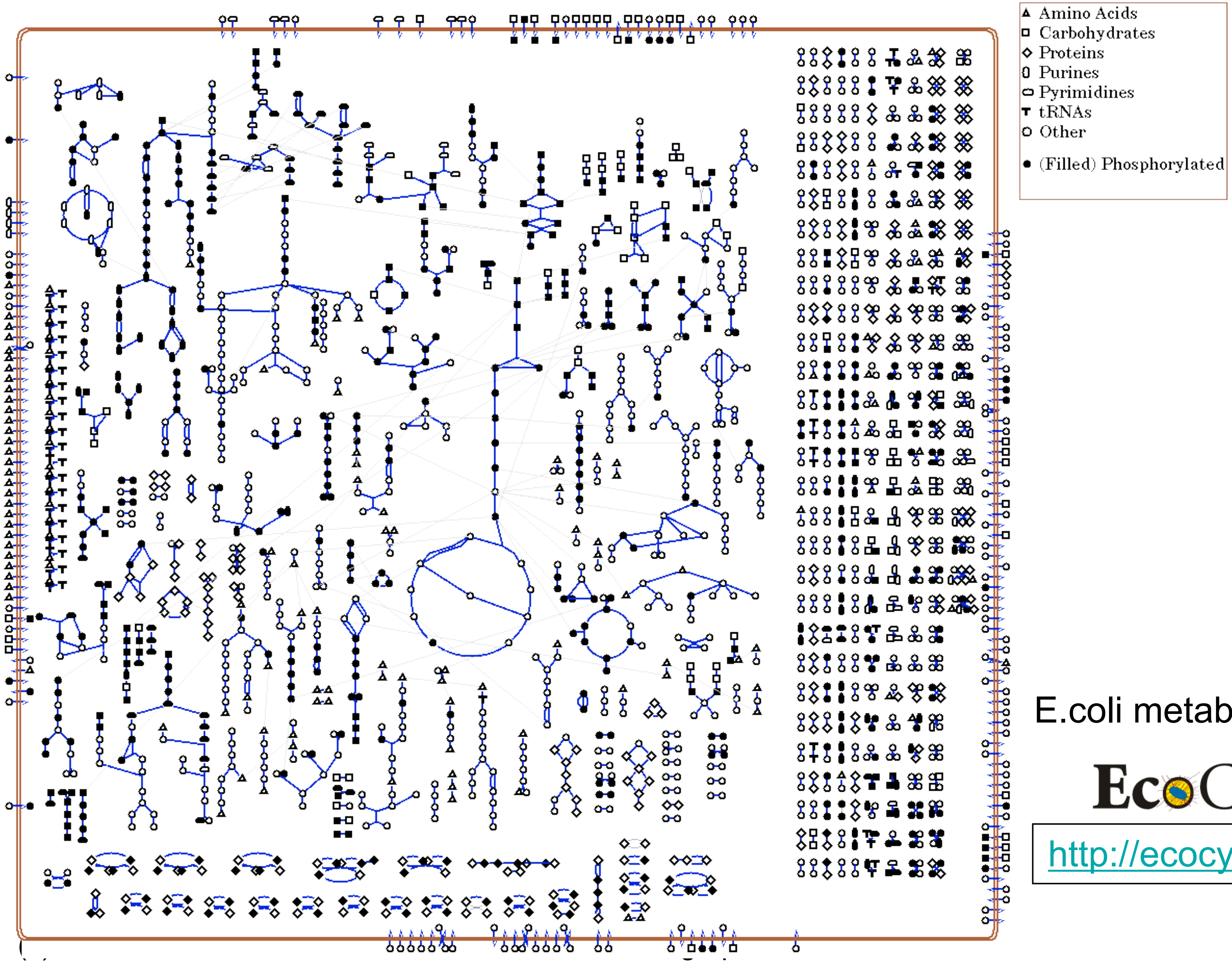
Human

Alternative Pathways

- Genome evolution
  - compare with known genome
  - infer for unknown genome
  - Find missing enzymes
- Biotechnology
  - identification of alternative enzymes
  - identification of alternative pathways
  - identification of alternative substrates
  - identification of alternative products
- Pharmacology
  - non-homologous gene displacement
  - species-specific drug targets
- Identification of previously unknown genes



Network features (motifs)

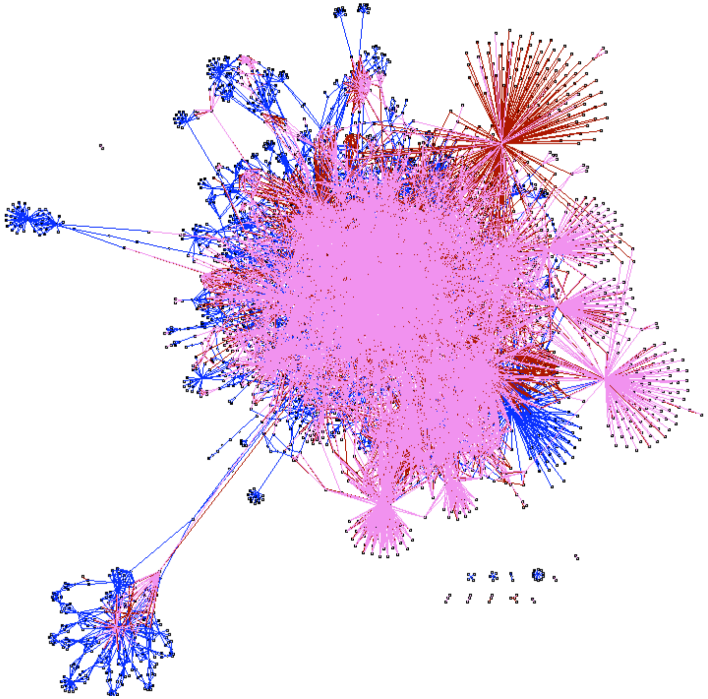


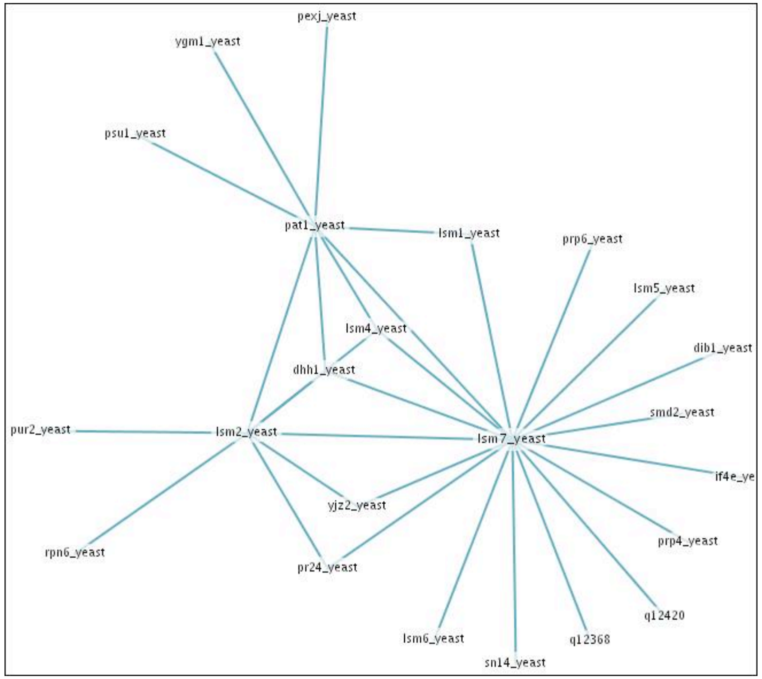
E.coli metabolic map

EcoCyc

http://ecocyc.org/

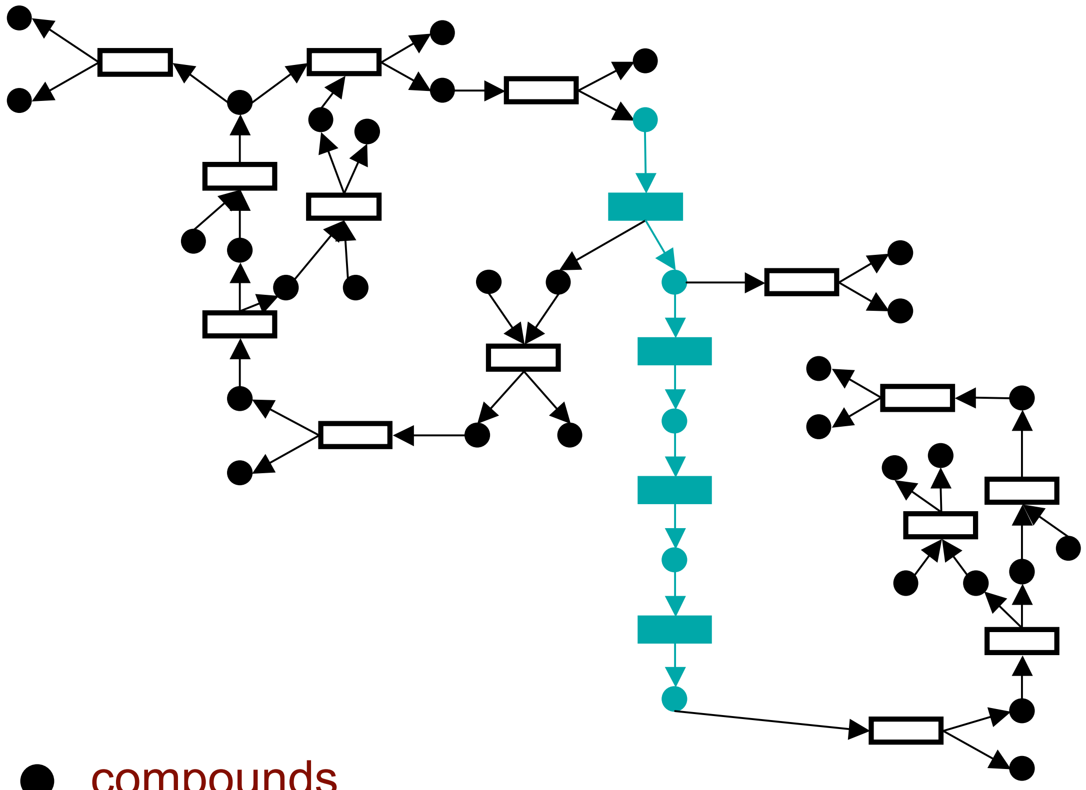
Network characteristics





Protein-protein interaction

Reactions and compounds as graphs



● compounds

▭ reactions

→ substrate → reaction

→ reaction → product

Slide from Jacques van Helden

What do network representations have in common?

- They consist of objects connected by lines or arrows
- The objects can be molecules, reaction labels,…
- Mathematically they can be modelled as ***graphs***

Some notation: set theory

- A set is any collection of ***distinct*** objects {,,,}
  Fruit = {apple, pear, orange, tomato}
  Veg = {carrot, potato, tomato}
- Member: *object* $\in$ *set*
  Apple $\in$ Fruit , Apple $\notin$ Veg, X $\in$ Fruit **and** X $\in$ Veg?
- Set equality: A = B
  {carrot, potato, tomato} = {tomato, carrot, potato}
- Subset: A $\subset$ B, A $\subseteq$ B
  {potato} $\subset$ Veg
  {tomato, carrot, potato} $\subseteq$ Veg
  {tomato, carrot} $\subseteq$ Veg
- Intersection: A $\cap$ B, (objects in common)
- Union: A $\cup$ B (all objects)
  Veg $\cap$ Fruit =
  Veg $\cup$ Fruit =
- Set subtraction: A \ B , A - B
  Fruit - C = {apple, pear, orange}
- Size (cardinality): |A|
  |Fruit| = ? |Fruit $\cap$ Veg| = ? , |Fruit $\cup$ Veg| = ?
- Empty set, cardinality: {} or $\varnothing$
  | $\varnothing$ | = ?

Graphs

- A graph G is an ordered pair (V, E)
  V = set of vertices (nodes), E = set of edges

  - *Dense* graph: $|E| \approx |V|^2$; *Sparse* graph: $|E| \approx |V|$

  - *Undirected graph: edge pairs are unordered*
  edge (u,v) = edge (v,u)

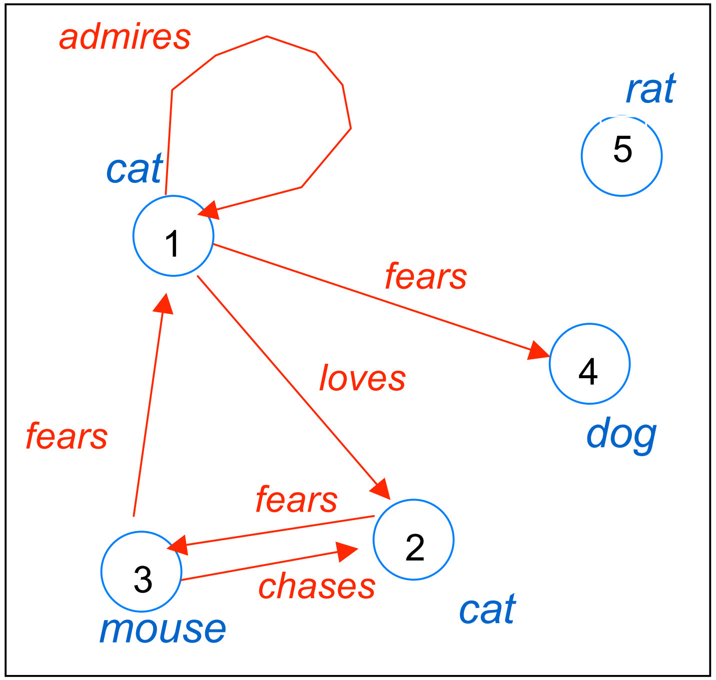
  - *Directed* graph: nodes & arcs
  Arc: i.e. directed edge (u,v) from initial vertex u to terminal vertex v, notation $u \rightarrow v$
  Two vertices u,v *adjacent* if $u \neq v$ and $u \rightarrow v$ or $u \rightarrow v$

  - Directed Acyclic Graph (DAG): directed graph with no cycles

  - A *weighted graph* associates weights with either the edges or the vertices

  - Input (output) degree of a node: number of input (output) arcs associated with the node

Graph Theory (simple!)



Graph = (V,A)

V = { 1 , 2 , 3, 4 , 5 }

A = {1→2, 2→3, 3→2, 3→1, 1→4 , 1→1}

Optionally ***label*** vertices & arcs

Graph = (V,A)

V = {cat:1, cat:2 , mouse:3, dog:4 , rat:5 }

A = {loves:1→2, fears:2→3, chases:3→2, fears:3→1, fears:1→4 , admires:1→1}

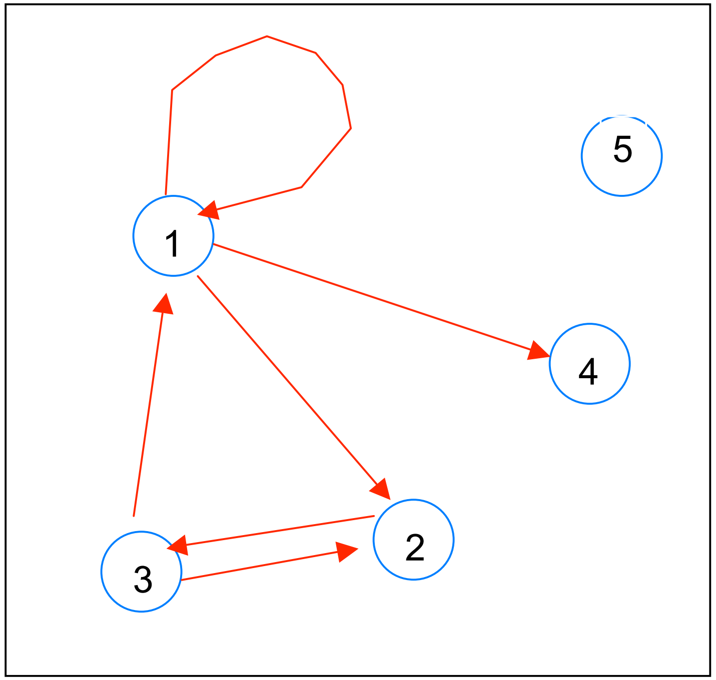
Pathway analysis

- What are the possible paths from entity A to entity B?
- How many paths, and of what lengths, lead from A to B?
- What is the average path distance between entities?
- Find all paths including a given set of entities
- Which genes are affected by a specific compound?
- Which pathways are affected if a given entity is missing or switched off?
- Compare pathways between two organisms or tissues, find common features or missing elements

Paths and Circuits of a Graph

- *Path* = sequence of arcs
  $(x_1 \rightarrow x_2 , x_2 \rightarrow x_3 , x_3 \rightarrow x_4 , \ldots x_{k-1} \rightarrow x_k)$
- Also can write $[x_1,x_2 ,x_3,\ldots, x_k]$
- *Simple* if does not use the same arc twice, else *composite*
- *Elementary* if does not use same vertex twice
- Can be *finite* or *infinite*
- *Circuit* = path $[x_1,x_2 ,x_3,\ldots, x_k]$ where initial vertex x_1= terminal vertex x_k
- *Elementary* circuit if all vertices distinct apart from $x_1= x_k$
- *Length* of path $(x_1 \rightarrow x_2 , \ldots x_{k-1} \rightarrow x_k)$ is K-1
- *Loop* is circuit length=1, I.e. $(x_1 \rightarrow x_1)$

Example

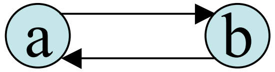


Paths - find these!

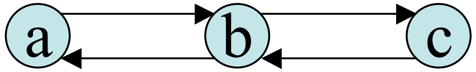
Circuits - find these!

Circuits & paths

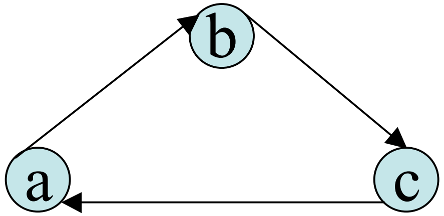
G1



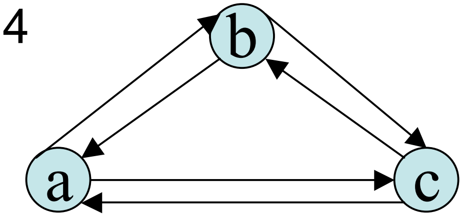
G2



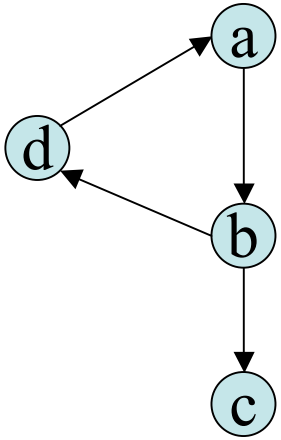
G3



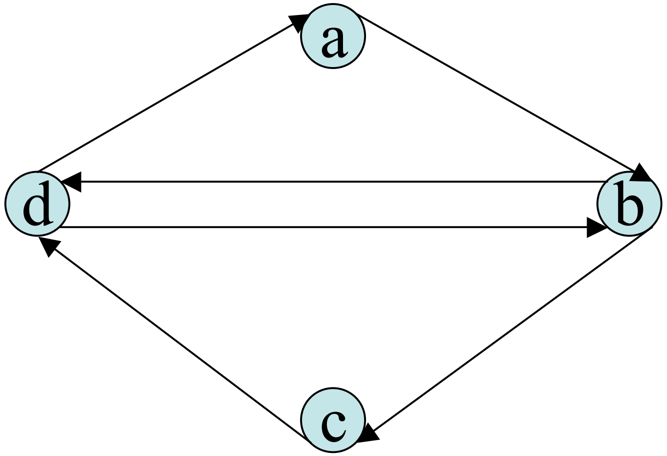
G4



G5



G6



Representing Graphs

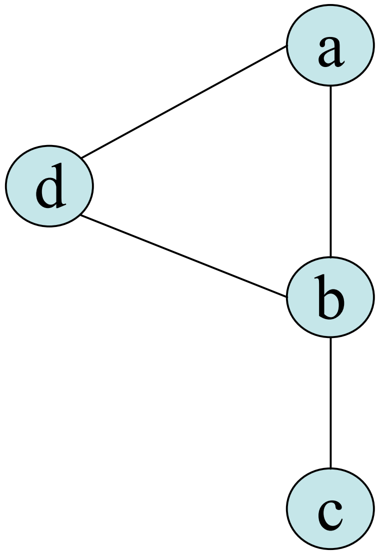
- Assume V = {1, 2, …, *n*}
- An *adjacency matrix* represents the graph as a *n*x*n* matrix M:
  - M[*i*, *j*] = 1 if edge $(i, j) \in E$ (or weight of edge)
    = 0 if edge $(i, j) \notin E$
  - Storage requirements: $O(V^2)$
    - A dense representation
  - But, can be very efficient for small graphs
    - Especially if store just one bit/edge
    - Undirected graph: only need one diagonal of matrix

Adjacency matrix - undirected graph

8	4	2	1

	a	b	c	d
a	0	1	0	1
b				
c				
d				

Val
5



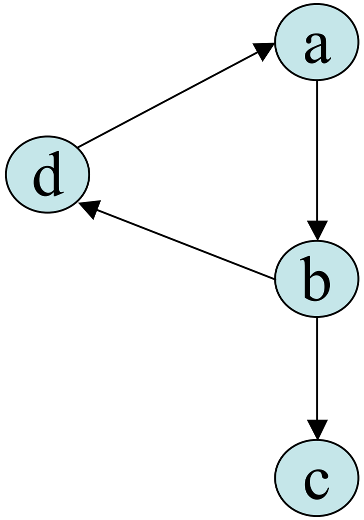
Binary ⇒ Base 10: Compact representation in a computer!

… but what if large number of vertices?

Adjacency matrix - directed graph

Outgoing: 1, Incoming: -1

	a	b	c	d
a	0	1	0	-1
b				
c				
d				



How to represent in binary?

Adjacency lists

- Associate each node with list of edges

- Undirected

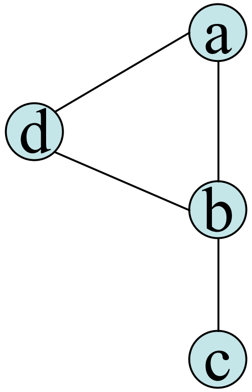
Vertex : Edges

a : {b,d}

b : {a,c,d}

c : {b}

d : {a,b}



- Directed

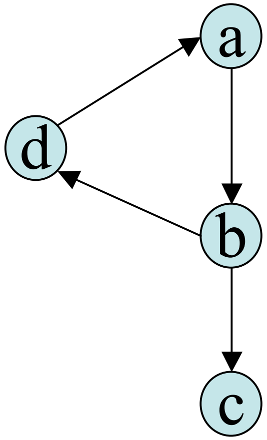
Ins : Vertex : Outs

{d} : a : {b}

{a}: b : {c,d}

{b}: c : {}

{b}: d : {a}

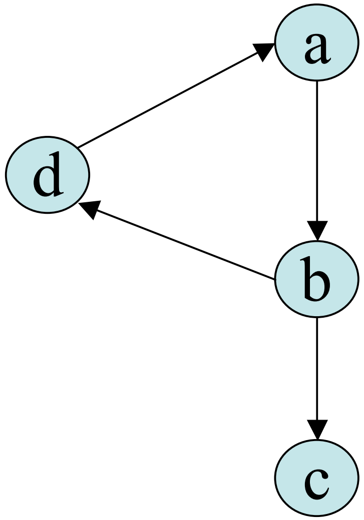


Less compact representation in a computer!
… but what if large number of vertices and few edges?

Adjacency matrix - directed graph

Outgoing: 1, Incoming: -1

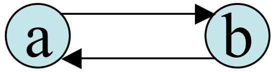
	a	b	c	d
a	0	1	0	-1
b				
c				
d				



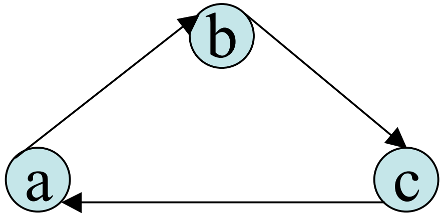
How to represent in binary?

Construct adjacency matrices for

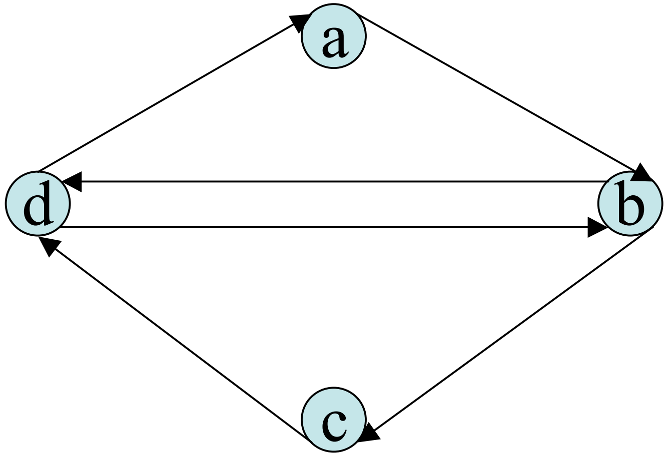
G1



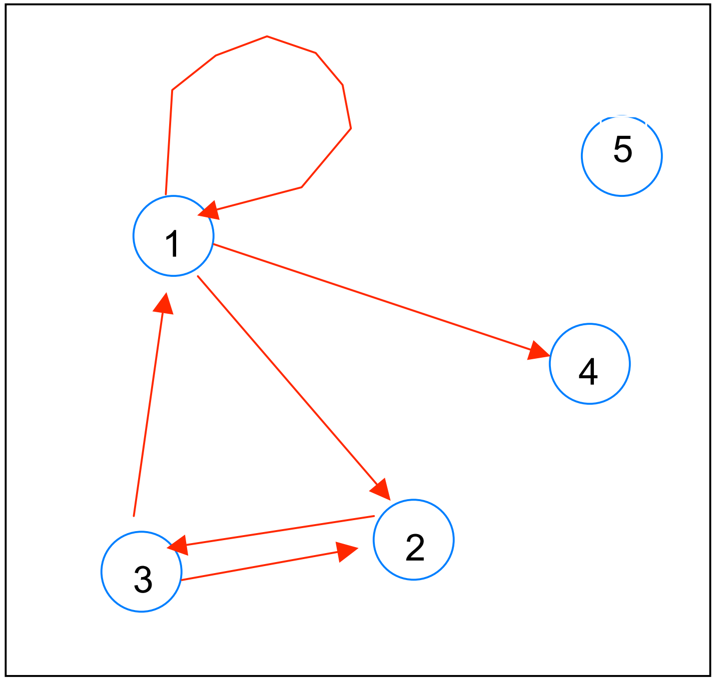
G3



G6



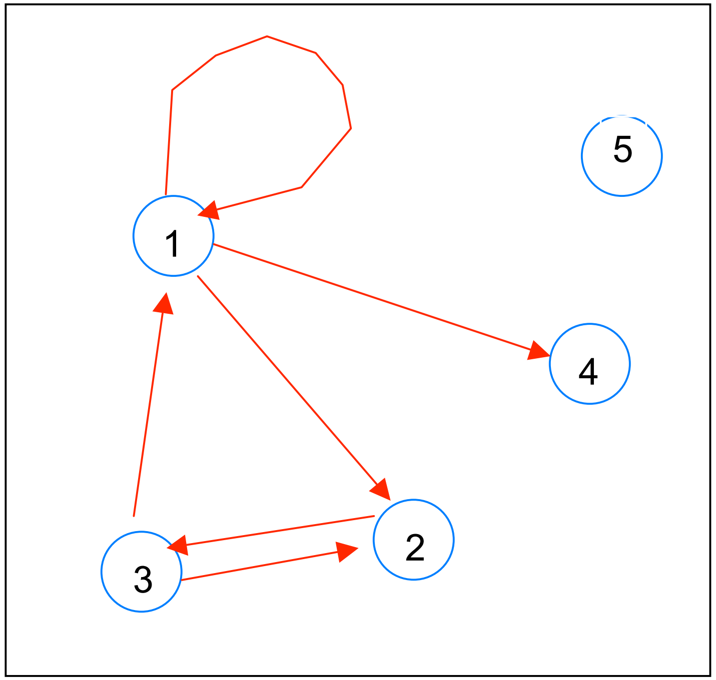
G0



Input & output degrees

- Compute the input and output degrees for the nodes in

G0



Search strategies

- A search strategy is defined by picking the order of node expansion

- Strategies are evaluated along the following dimensions:
  - completeness: does it always find a solution if one exists?
  - time complexity: number of nodes generated
  - space complexity: maximum number of nodes in memory
  - optimality: does it always find a least-cost solution?

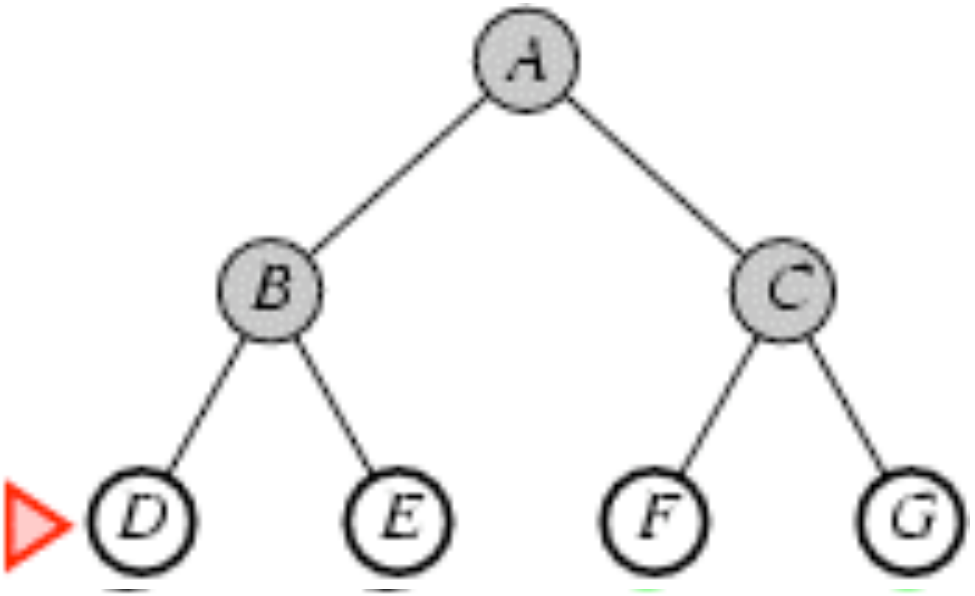
- Time and space complexity are measured in terms of
  - *b:* maximum branching factor of the search tree
  - *d:* depth of the least-cost solution
  - *m*: maximum depth of the state space (may be ∞)

Breadth-First Search

- “Explore” a graph, turning it into a tree
  - One vertex at a time
  - Expand frontier of explored vertices across the *breadth* of the frontier
- Builds a tree over the graph
  - Pick a *source vertex* to be the root
  - Find (“discover”) its children, then their children, etc.

Breadth-first search

- Expand shallowest unexpanded node
- Implementation:
  - *fringe* is a FIFO queue, i.e., new successors go at end



Breadth-First Search: Properties

- BFS calculates the *shortest-path distance* to the source node
  - Shortest-path distance δ(s,v) = minimum number of edges from s to v, or ∞ if v not reachable from s
- BFS builds *breadth-first tree*, in which paths to root represent shortest paths in G
  - Thus can use BFS to calculate shortest path from one vertex to another in O(V+E) time

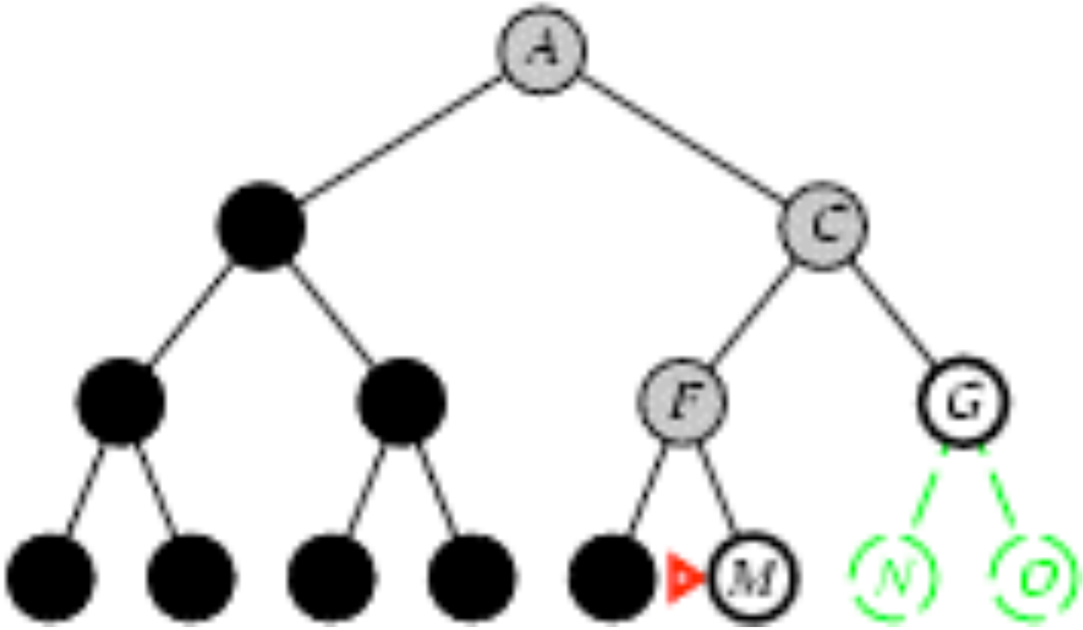
Properties of breadth-first search

- Complete? Yes (if b is finite)
- Time? $1+b+b^2+b^3+\dots+b^d + b(b^d-1) = O(b^{d+1})$
- Space? $O(b^{d+1})$ (keeps every node in memory)
- Optimal? Yes (if cost = 1 per step)

- Space is the bigger problem (more than time)

Depth-first search

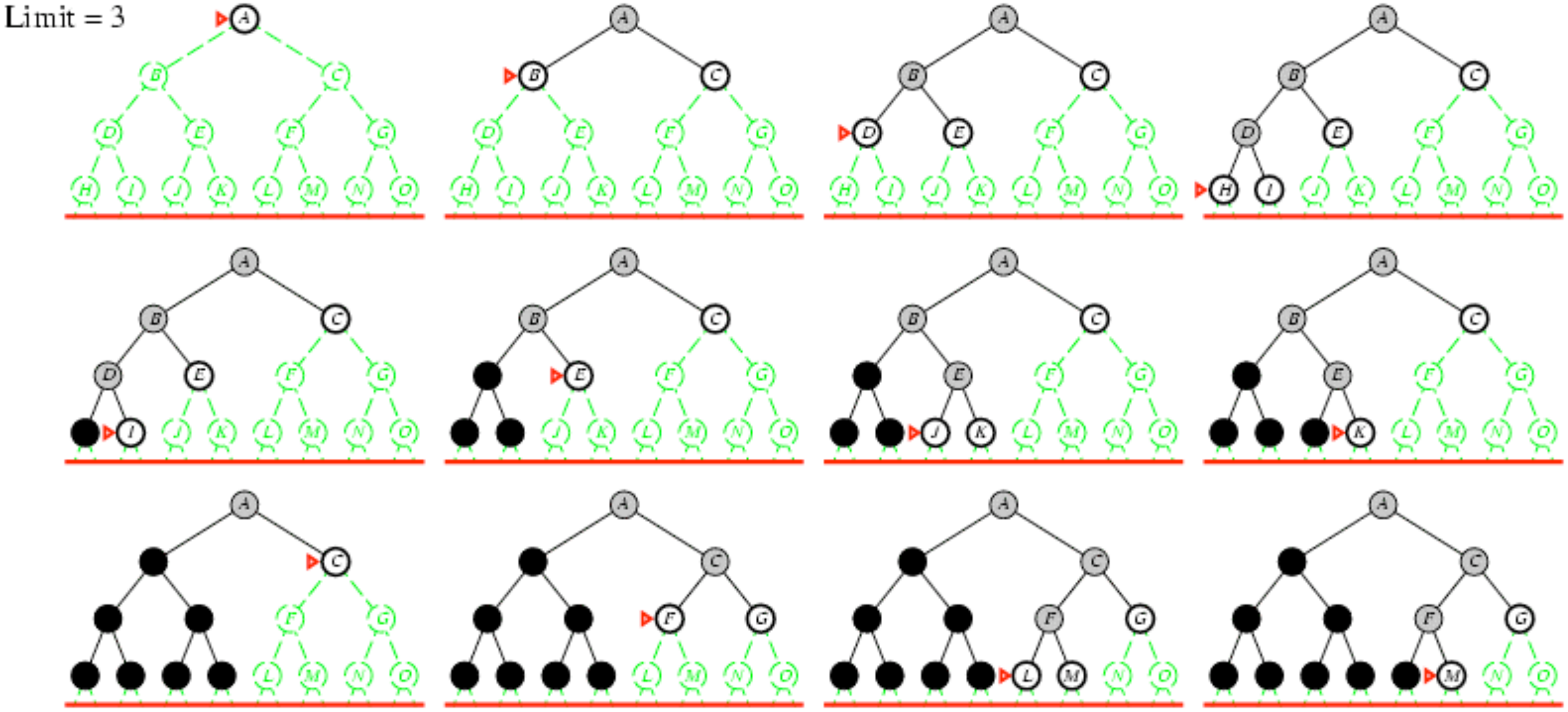
- Expand deepest unexpanded node
- Implementation:
  - *fringe* = LIFO queue, i.e., put successors at front



Properties of depth-first search

- Complete? No: fails in infinite-depth spaces, spaces with loops
  - Modify to avoid repeated states along path
    - → complete in finite spaces
- Time? $O(b^m)$: terrible if m is much larger than d
  - but if solutions are dense, may be much faster than breadth-first
- Space? $O(bm)$, i.e., linear space!
- Optimal? No

Iterative deepening search *l* =3



Properties of iterative deepening search

- Complete? Yes
- Time? $(d+1)b^0 + d\, b^1 + (d-1)b^2 + \ldots + b^d = O(b^d)$
- Space? $O(bd)$
- Optimal? Yes, if step cost = 1

Simple path search algorithm

Search path From … To

```
Given G=(V,A)
Initialise: Path:= [From]

While $(From \rightarrow Next) \in A$ and $Next \neq To$
        Path := Path + [Next]
        From := Next, Next:=NewNext

If $(From \rightarrow To) \in A$ then
        Path:= Path + [To]
        Return Path
Else Return 'Fail'
```

Depth-first or Breadth-first?

Circuit detection algorithm

- Do this…!

KEGG

- http://www.genome.ad.jp/kegg/ Institute for Chemical Research, Kyoto University (part of the Japanese Human Genome Program).
- Repository of metabolic pathways for organisms whose genome is completely sequenced. Also regulatory information.
- For many of these organisms, the body of experimental data is very restricted. Protein function inferred from sequence similarity with proteins characterised experimentally in other organisms.
- Pathways represented as diagrams, manually created & stored as static gif files.
- Upon selection of an organism, the reactions for which an enzyme is known in that organism are highlighted in colour in the generic pathway diagrams.

KEGG - search & compute

- KEGG pathways searched by EC numbers (enzymes), compound numbers, & by gene accessions.
- Combine search with KEGG grouping or the hierarchical classification. (e.g. EC numbers from a specific group in the superfamily table (or SCOP table) & searching against pathway diagrams.
- Search KEGG pathways by sequence similarity. (identify orthologs & reconstruct pathways from the gene catalog).
- Given list of enzymes, automatically generate the organism specific pathways by marking the matching enzymes on the diagram. Missing elements imply either gene catalog wrong or unknown reaction pathway utilizing different enzymes in the catalog.
- Compute pathways from a given list of enzymes. Deduction from binary relations of substrates and products with optional use of query relaxation for functional hierarchies.

KEGG Query & result

Pathway Search Result

- map00271 Methionine metabolism
    - EC 2.1.1.13
    - EC 2.3.1.46
    - EC 2.5.1.6
    - EC 4.4.1.8
- map00260 Glycine, serine and threonine metabolism
- map00300 Lysine biosynthesis
- map00450 Selenoamino acid metabolism
- map00920 Sulfur metabolism
- map00272 Cysteine metabolism
- map00670 One carbon pool by folate
- map00910 Nitrogen metabolism

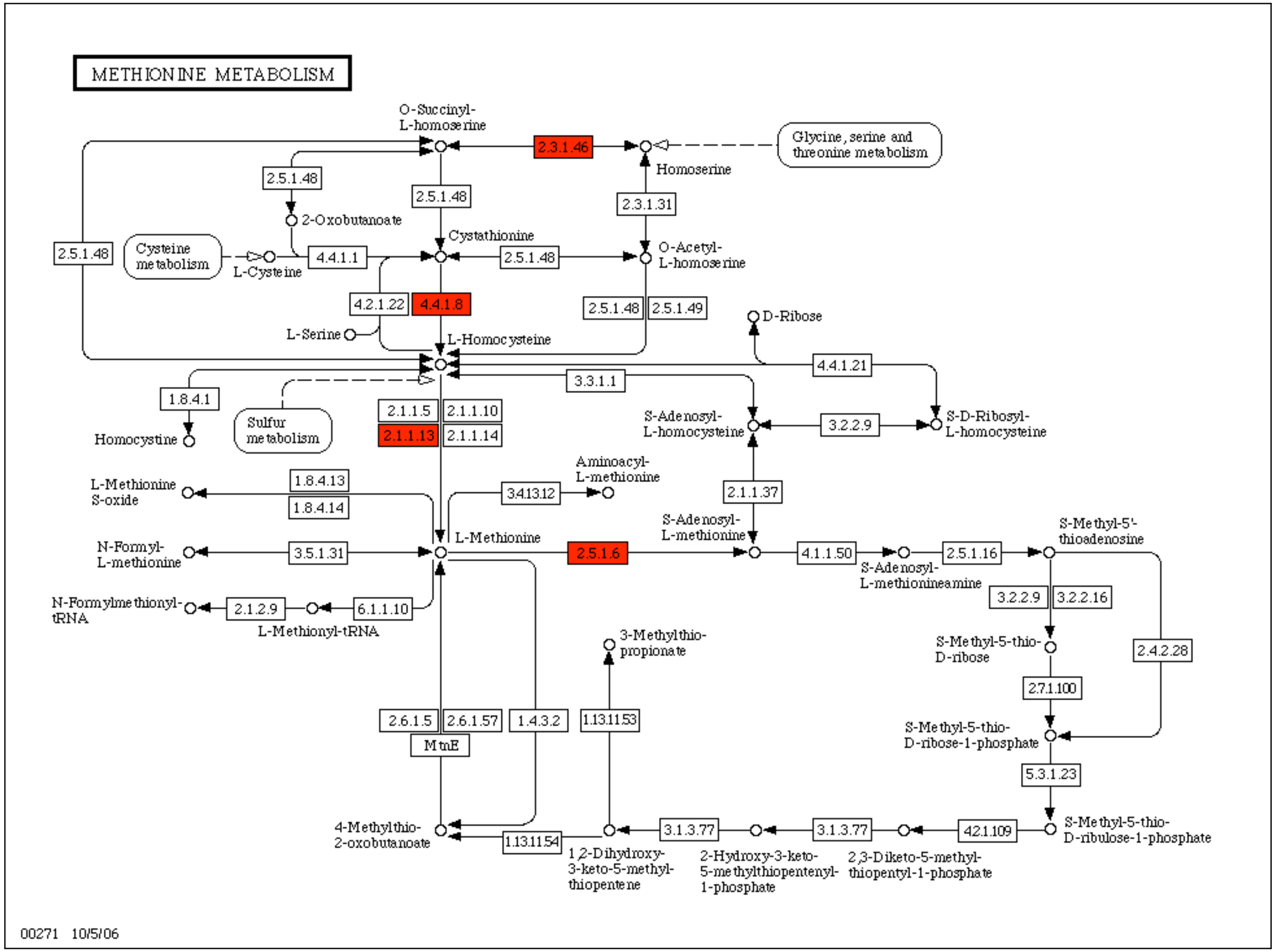
Query =

2.7.2.4
1.2.1.11
1.1.1.3
2.3.1.46
4.2.99.9
4.4.1.8
2.1.1.13
2.5.1.6

map00271 Methionine metabolism

http://www.genome.ad.jp/kegg-bin/mk_point_html?ec

KEGG Query & result



Query =

2.7.2.4
1.2.1.11
1.1.1.3
2.3.1.46
4.2.99.9
4.4.1.8
2.1.1.13
2.5.1.6

map00271
Methionine
metabolism

http://www.genome.ad.jp/kegg-bin/mk_point_html?ec

Pathway Hunter Tool

- Metabolic pathway analysis web service (Pathway Hunter Tool at CUBIC). S. A. Rahman , P. Advani , R. Schunk , R. Schrader and Dietmar Schomburg. Bioinformatics 2005 21(7):1189-1193

- Motivation: Pathway Hunter Tool (PHT), is a fast, robust and user-friendly tool to analyse the shortest paths in metabolic pathways. The user can perform shortest path analysis for one or more organisms or can build virtual organisms (networks) using enzymes. Using PHT, the user can also calculate the average shortest path, average alternate path and the top 10 hubs in the metabolic network. The comparative study of metabolic connectivity and observing the cross talk between metabolic pathways among various sequenced genomes is possible.

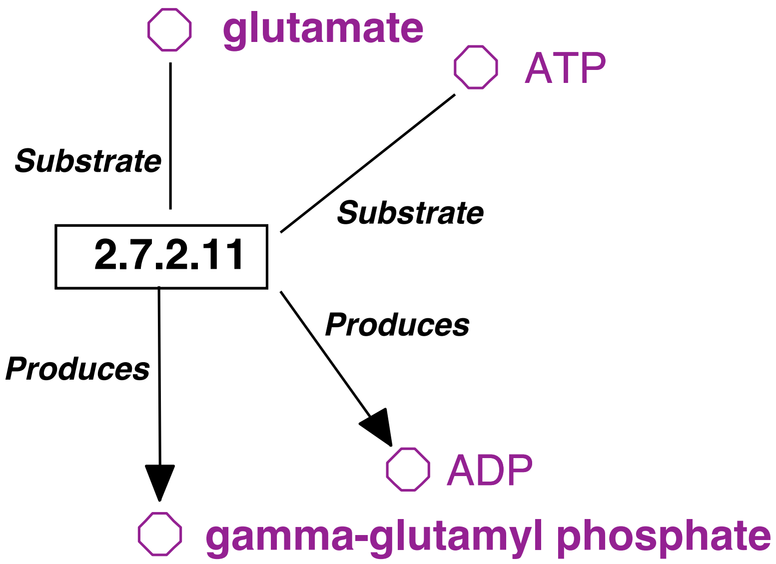
- Results: A new algorithm for finding the biochemically valid connectivity between metabolites in a metabolic network was developed and implemented. A predefined manual assignment of side metabolites (like ATP, ADP, water, CO2 etc.) and main metabolites is not necessary as the new concept uses chemical structure information (global and local similarity) between metabolites for identification of the shortest path.

- Availability: PHT is accessible at http://www.pht.uni-koeln.de

A scheme for representing metabolic and regulatory networks

- Slides from Jacques van Helden

Chemical Reaction



glutamate

ATP

Substrate

Substrate

2.7.2.11

Produces

Produces

ADP

gamma-glutamyl phosphate

Set of Biochemical Entities (substrates)

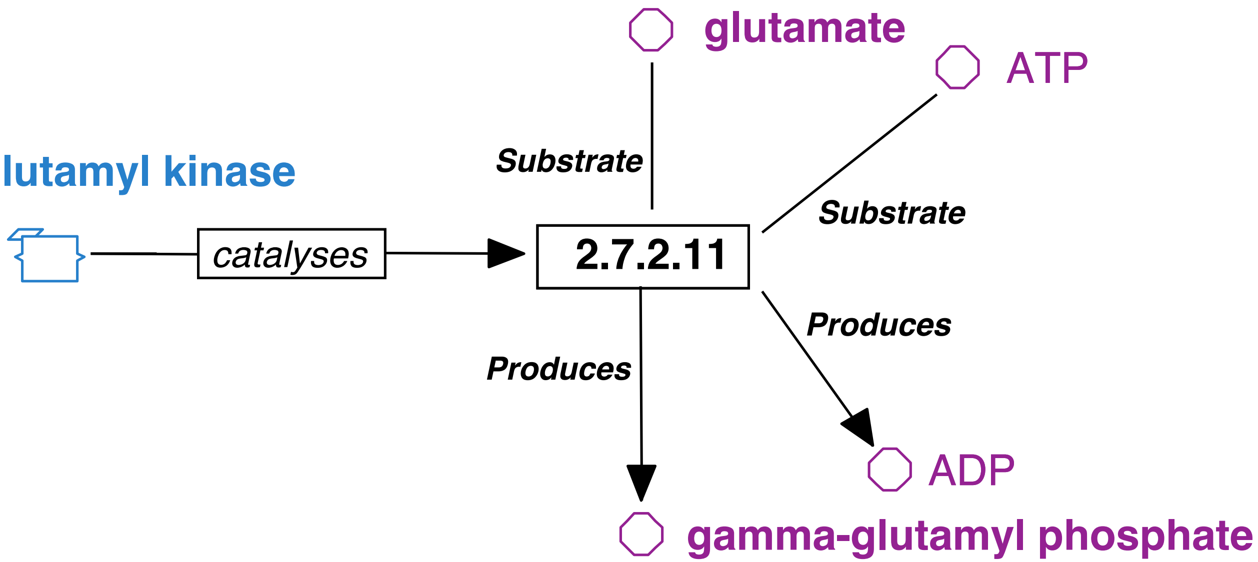
-o [Reaction] ->

Set of Biochemical Entities (products)

Slide from Jacques van Helden

1.5.1.2	**EC (reaction) number**
○	**compound**

Enzymatic catalysis



gamma-glutamyl kinase

catalyses

2.7.2.11

Substrate

Substrate

Produces

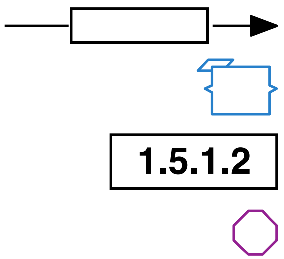
Produces

glutamate

ATP

ADP

gamma-glutamyl phosphate



Positive interaction

Protein

1.5.1.2 **EC (reaction) number**

compound

Protein (enzyme)

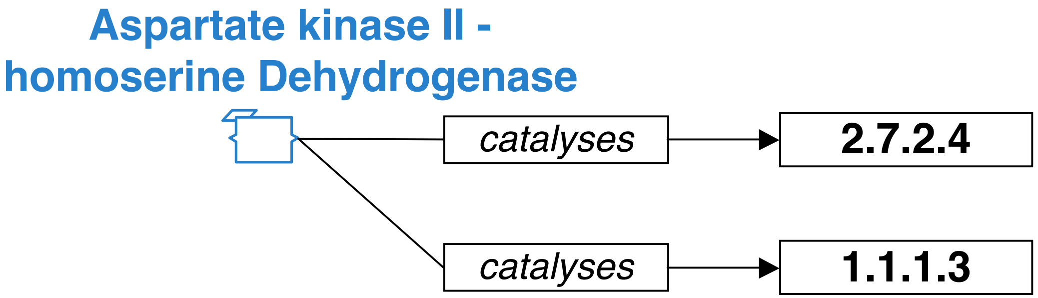
-o [Catalyses] ->

Reaction

Slide from Jacques van Helden

Enzymatic catalysis

Multifunctional enzyme



Aspartate kinase II - homoserine Dehydrogenase

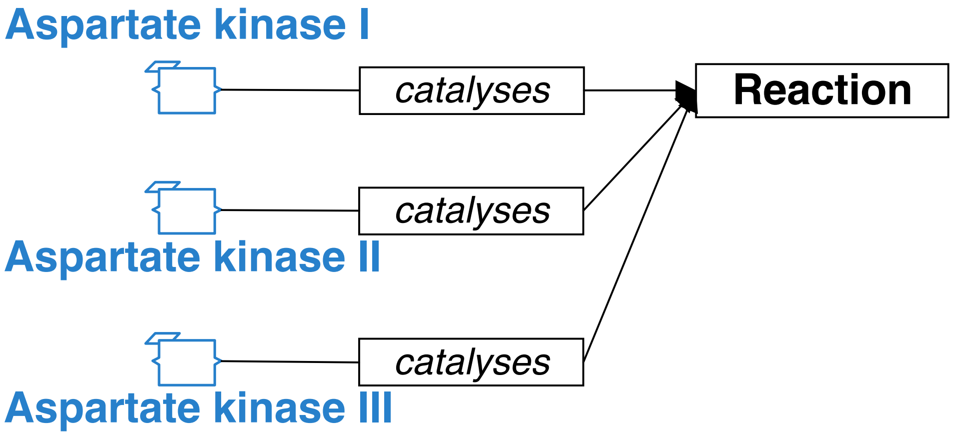
catalyses

2.7.2.4

catalyses

1.1.1.3

Isofunctional enzymes



Aspartate kinase I

catalyses

Reaction

catalyses

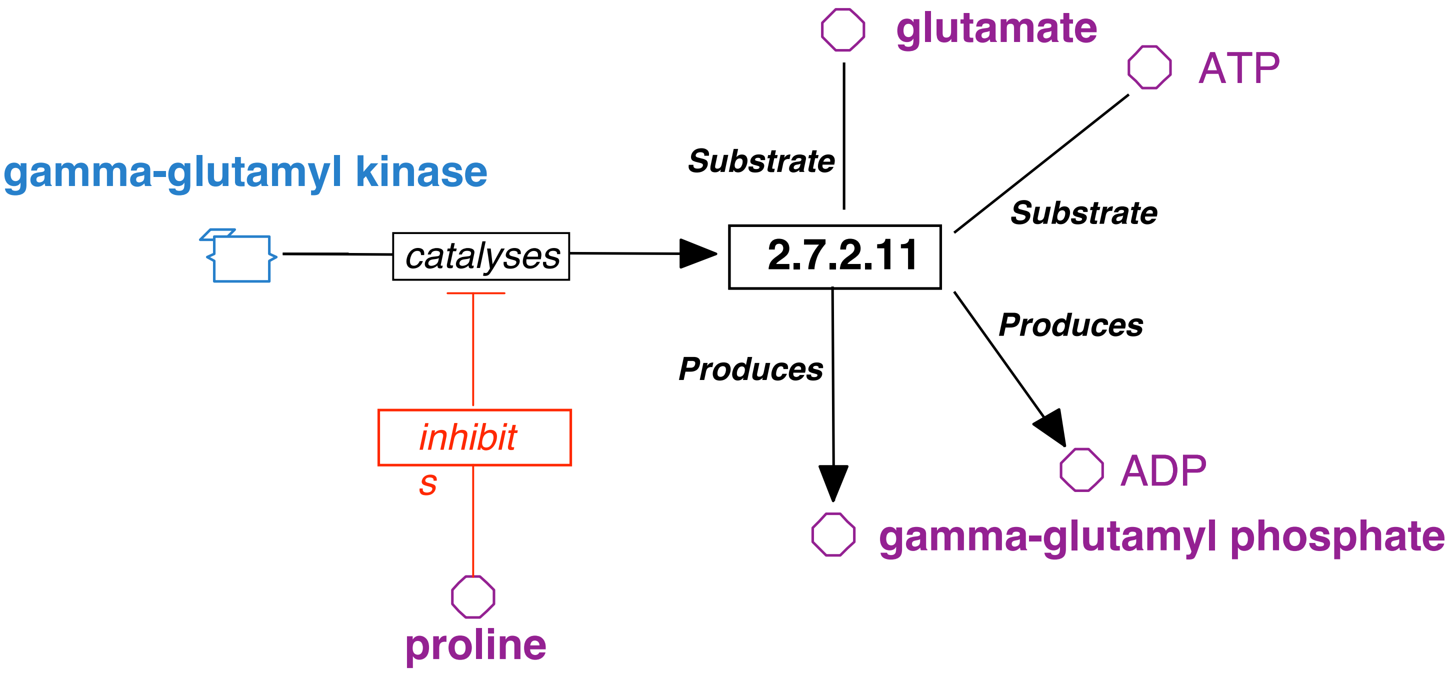
Aspartate kinase II

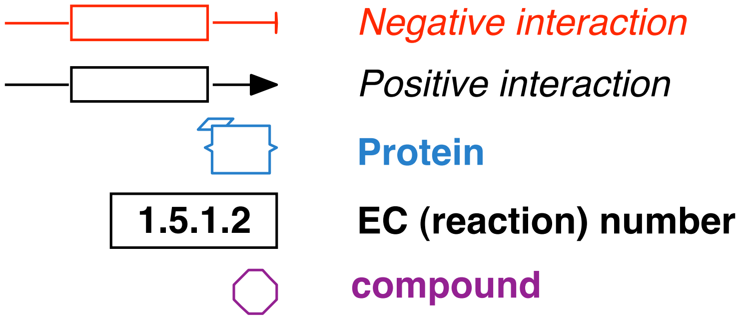
catalyses

Aspartate kinase III

Slide from Jacques van Helden

Inhibition/Activation





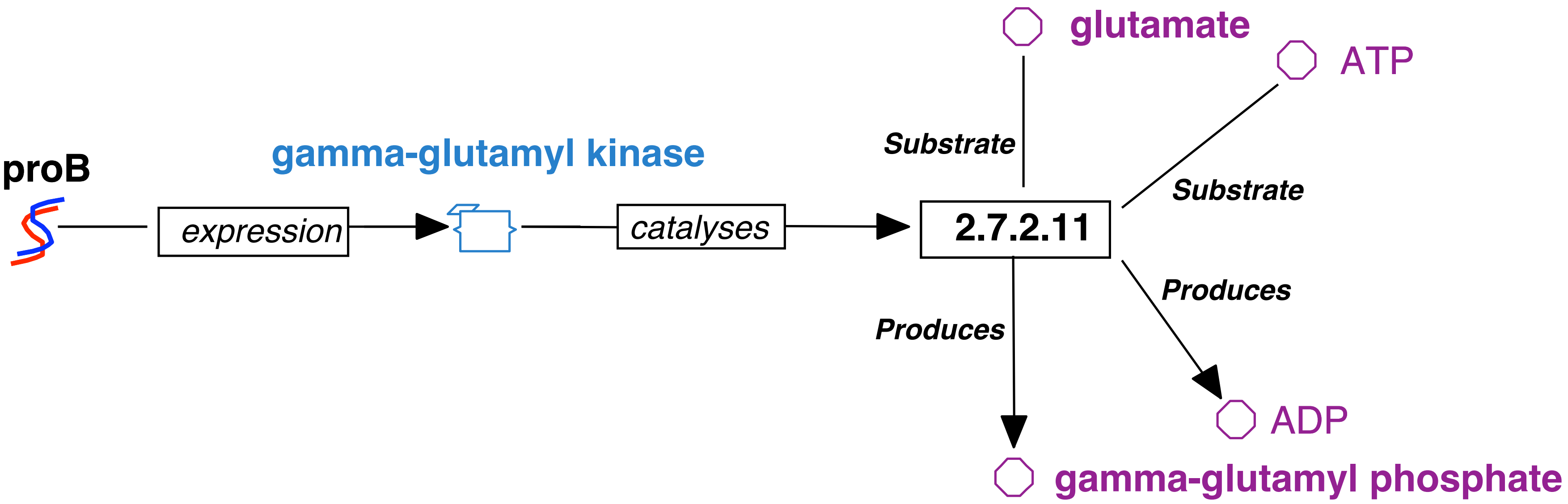
Biochemical Entity

-o [Inhibits] ->

Reaction Catalysis

Slide from Jacques van Helden

Gene expression



proB

gamma-glutamyl kinase

expression

catalyses

2.7.2.11

Substrate

Substrate

Produces

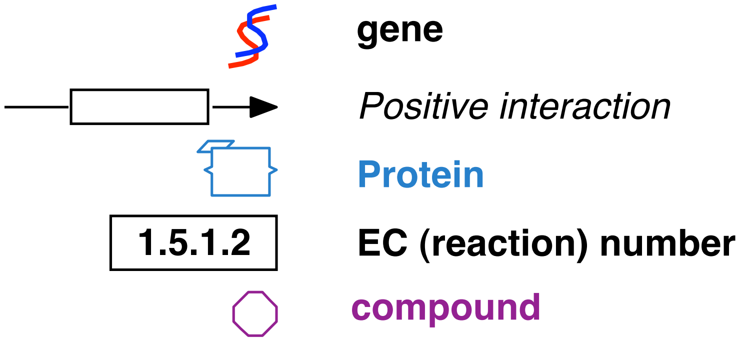
Produces

glutamate

ATP

ADP

gamma-glutamyl phosphate



gene

Positive interaction

Protein

1.5.1.2 **EC (reaction) number**

compound

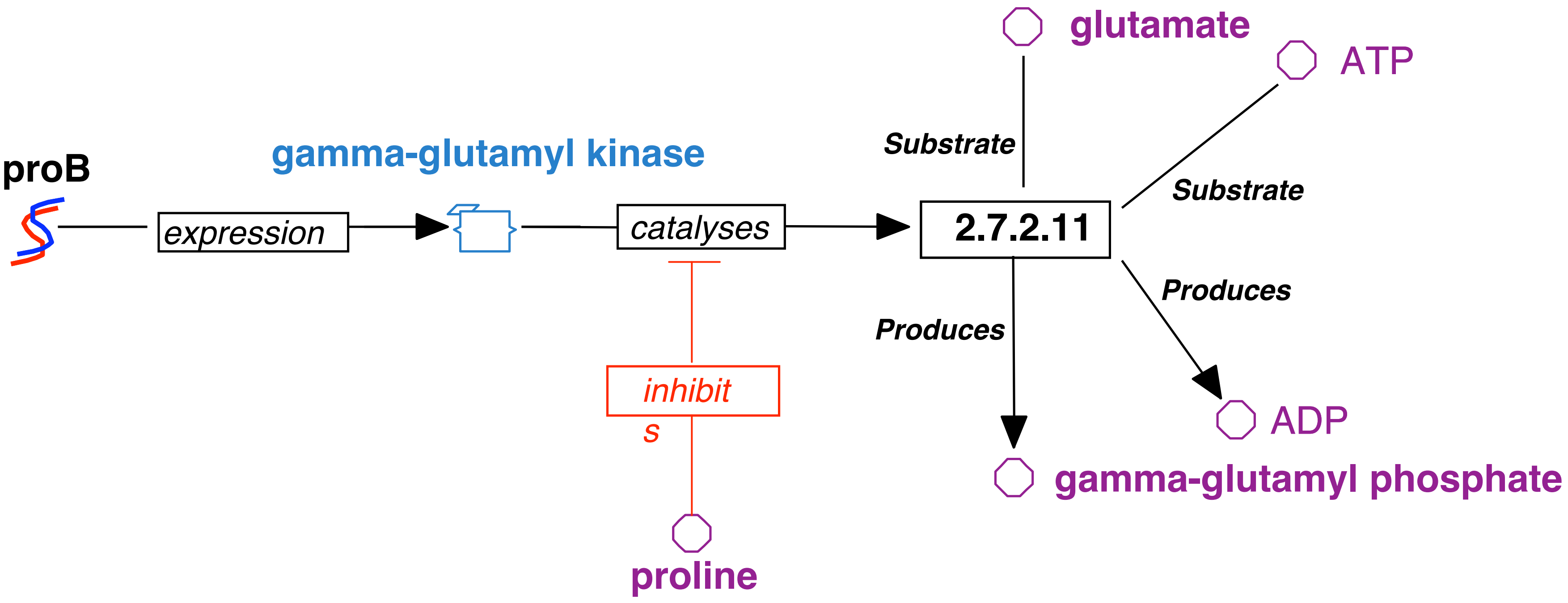
Gene

-o [Expression] ->

Protein

Slide from Jacques van Helden

Metabolic Step



proB

gamma-glutamyl kinase

expression

catalyses

inhibits

proline

2.7.2.11

glutamate

ATP

Substrate

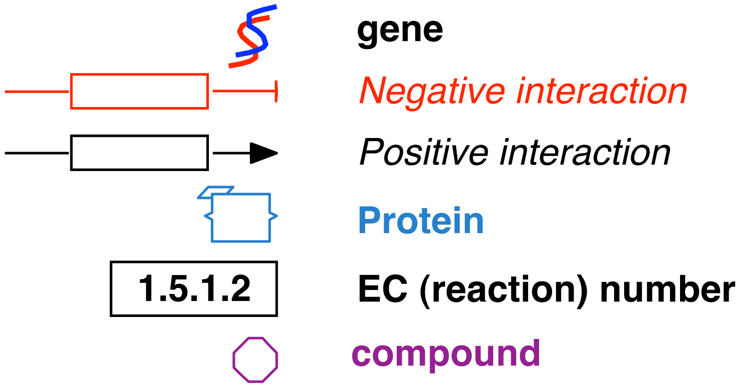
Substrate

Produces

Produces

ADP

gamma-glutamyl phosphate



gene

Negative interaction

Positive interaction

Protein

1.5.1.2

EC (reaction) number

compound

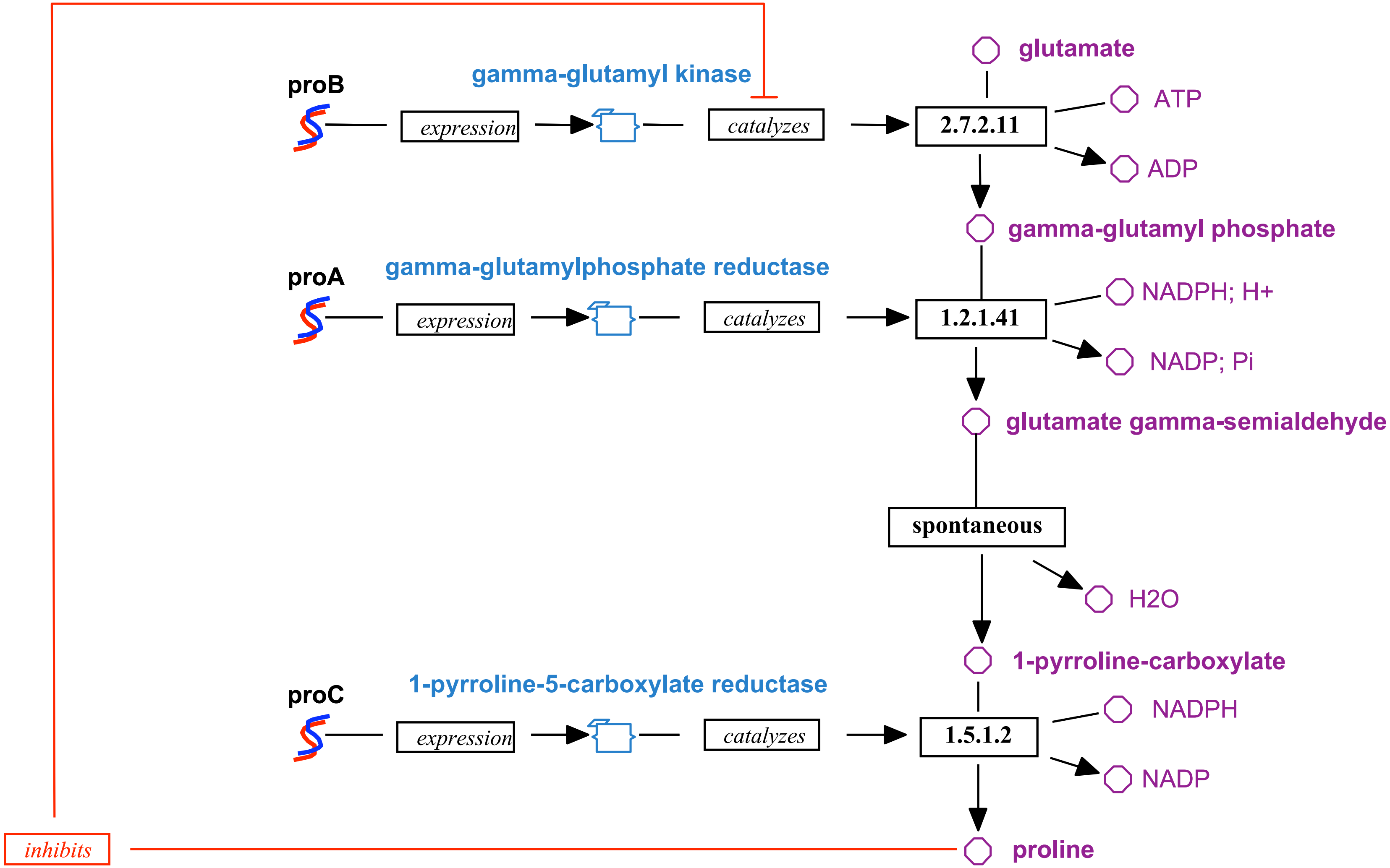
Biochemical Entity

-o [Inhibits] ->

Reaction Catalysis

Slide from Jacques van Helden

Metabolic Pathway: Proline Biosynthesis



Slide from Jacques van Helden

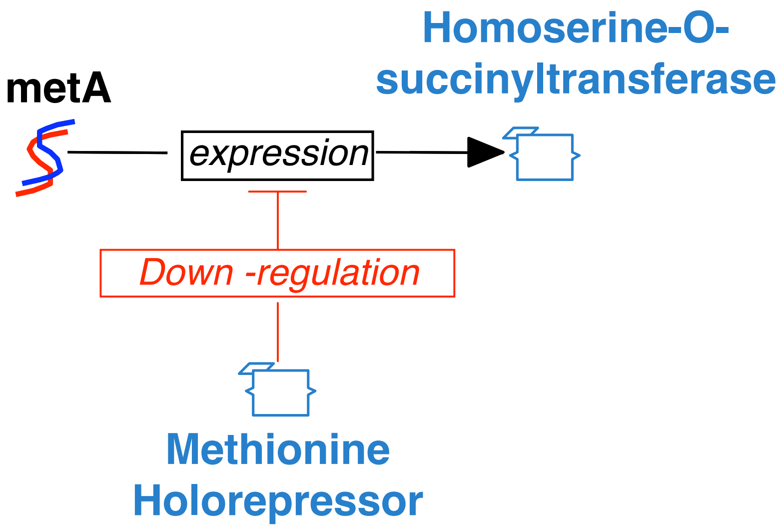
Transcriptional Regulation

Transcriptional repression (down-regulation)

Protein

-o [down-regulates] ->

expression



Homoserine-O-succinyltransferase

metA

expression

Down -regulation

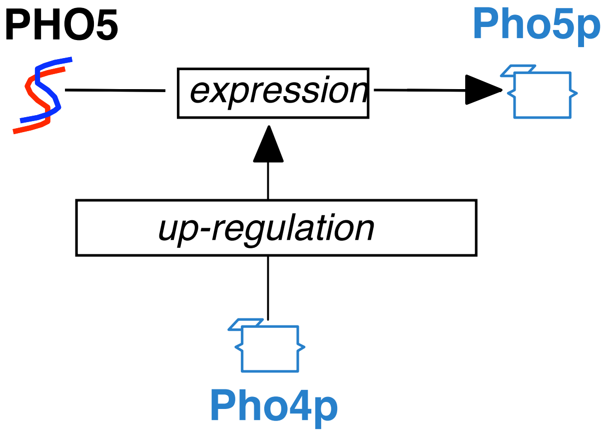
Methionine Holorepressor

Transcriptional activation (up-regulation)

Protein

-o [up-regulates] ->

expression



PHO5

Pho5p

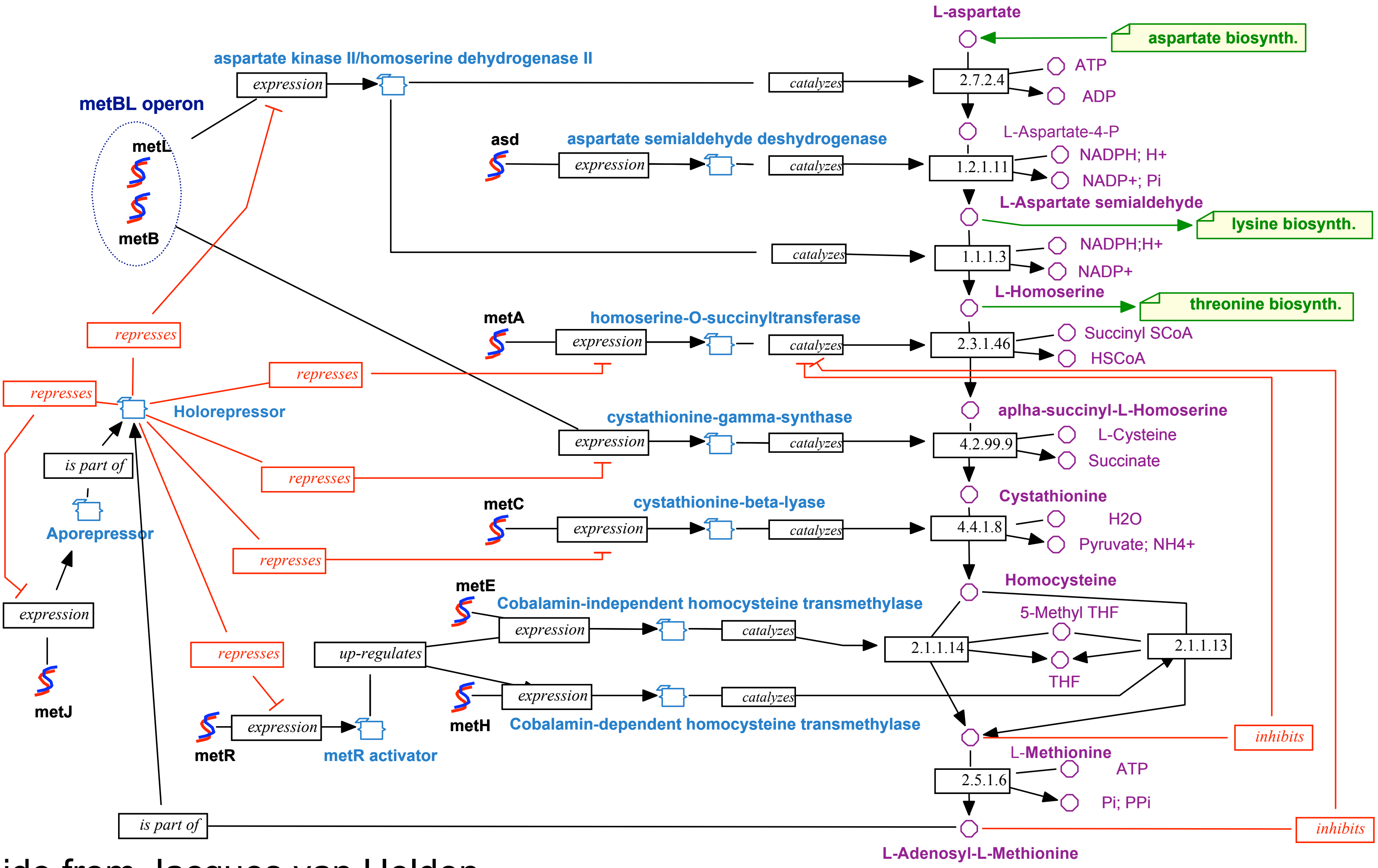
expression

up-regulation

Pho4p

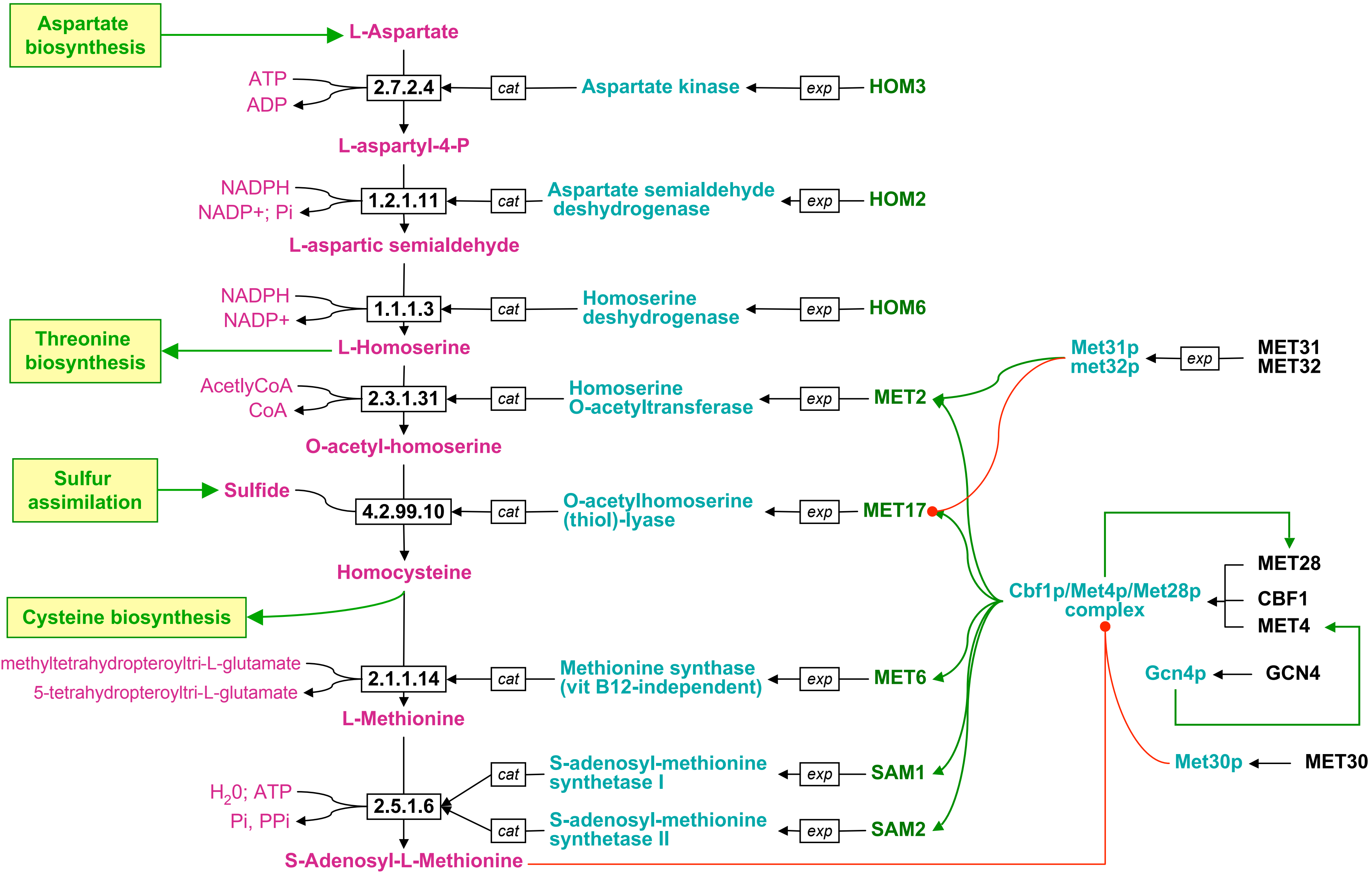
Slide from Jacques van Helden

Methionine Biosynthesis in E.coli



Slide from Jacques van Helden

Methionine Biosynthesis in S.cerevisiae

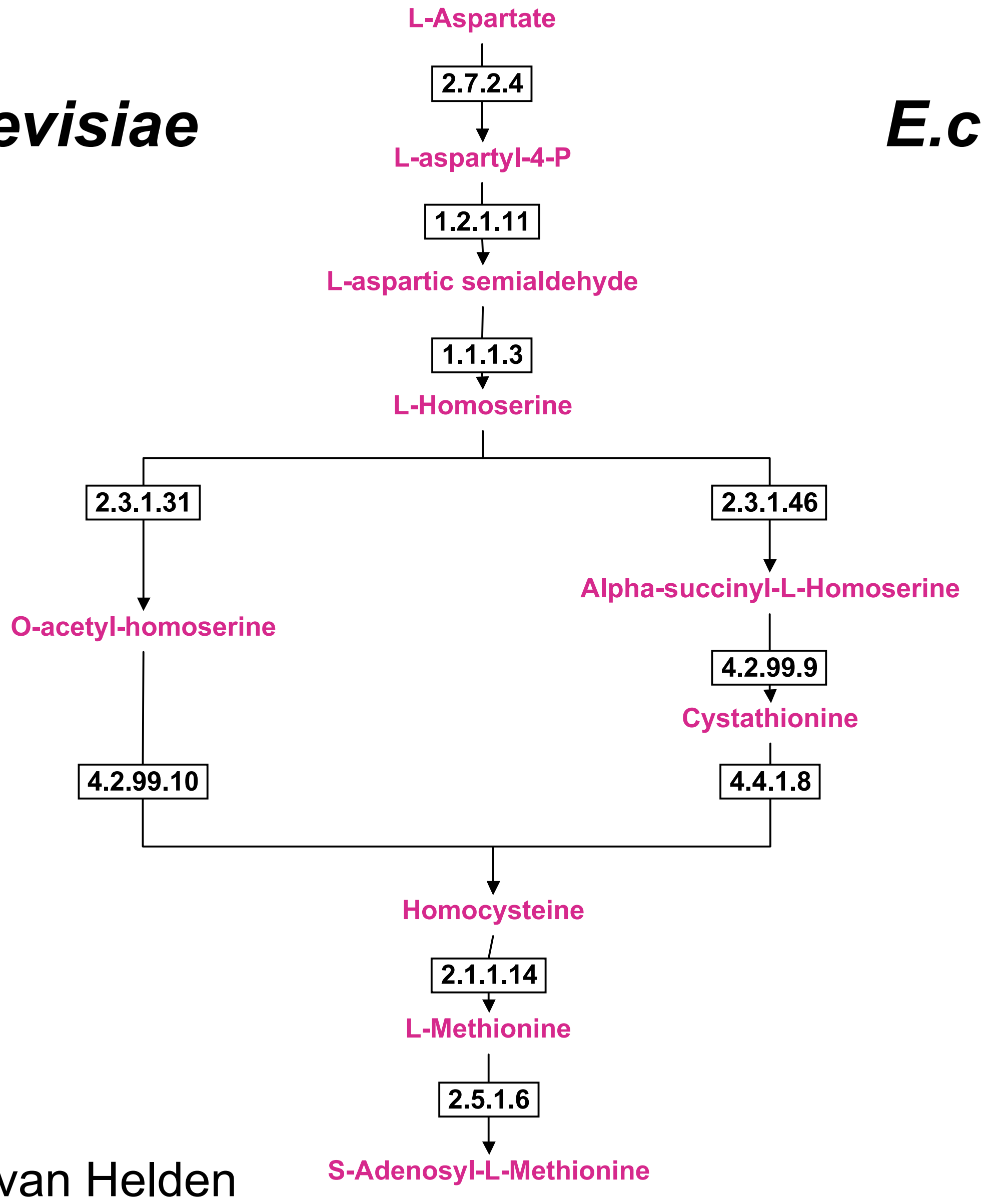


Slide from Jacques van Helden

Alternative methionine pathways

S.cerevisiae

E.coli



L-Aspartate

2.7.2.4

L-aspartyl-4-P

1.2.1.11

L-aspartic semialdehyde

1.1.1.3

L-Homoserine

2.3.1.31

O-acetyl-homoserine

4.2.99.10

2.3.1.46

Alpha-succinyl-L-Homoserine

4.2.99.9

Cystathionine

4.4.1.8

Homocysteine

2.1.1.14

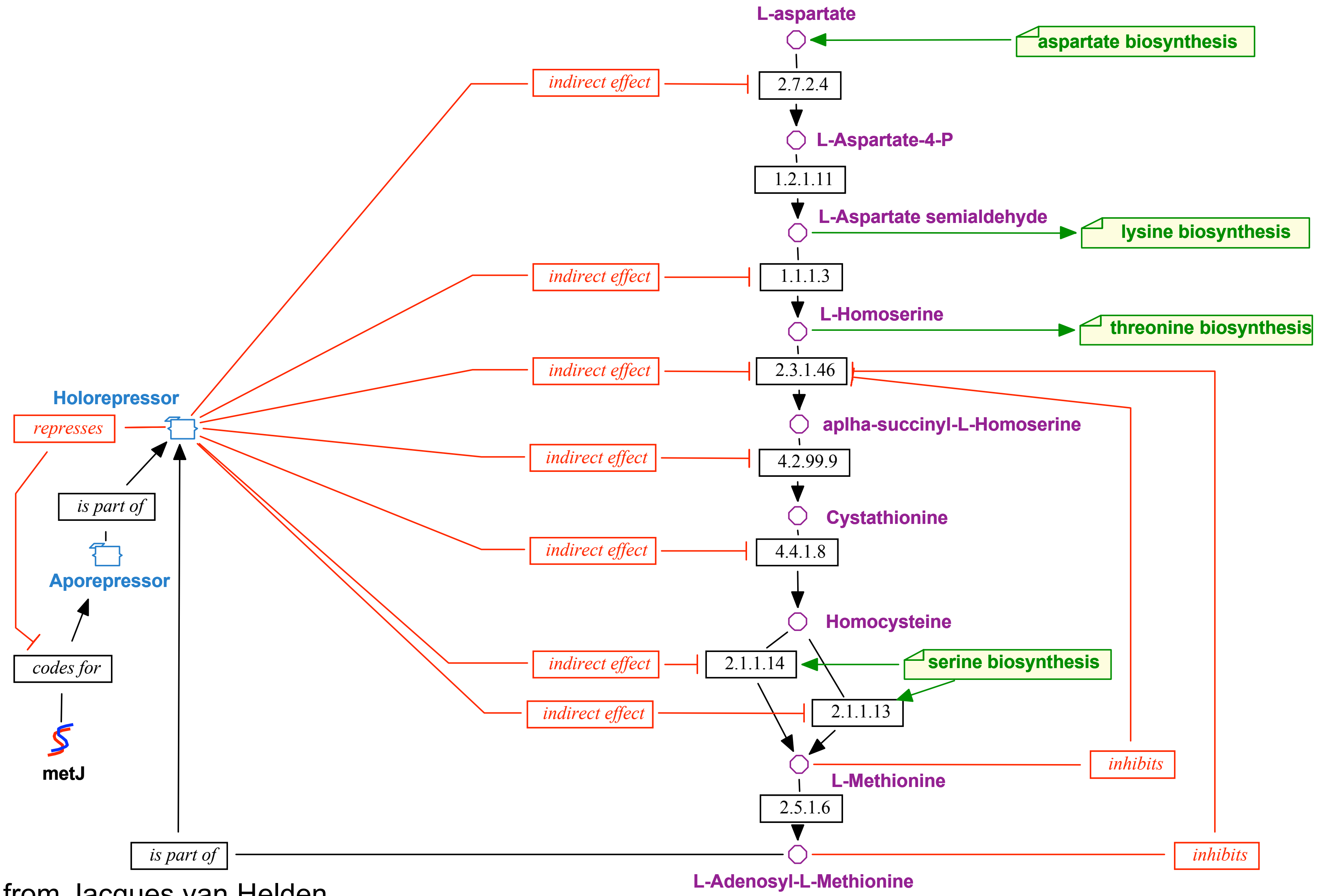
L-Methionine

2.5.1.6

S-Adenosyl-L-Methionine

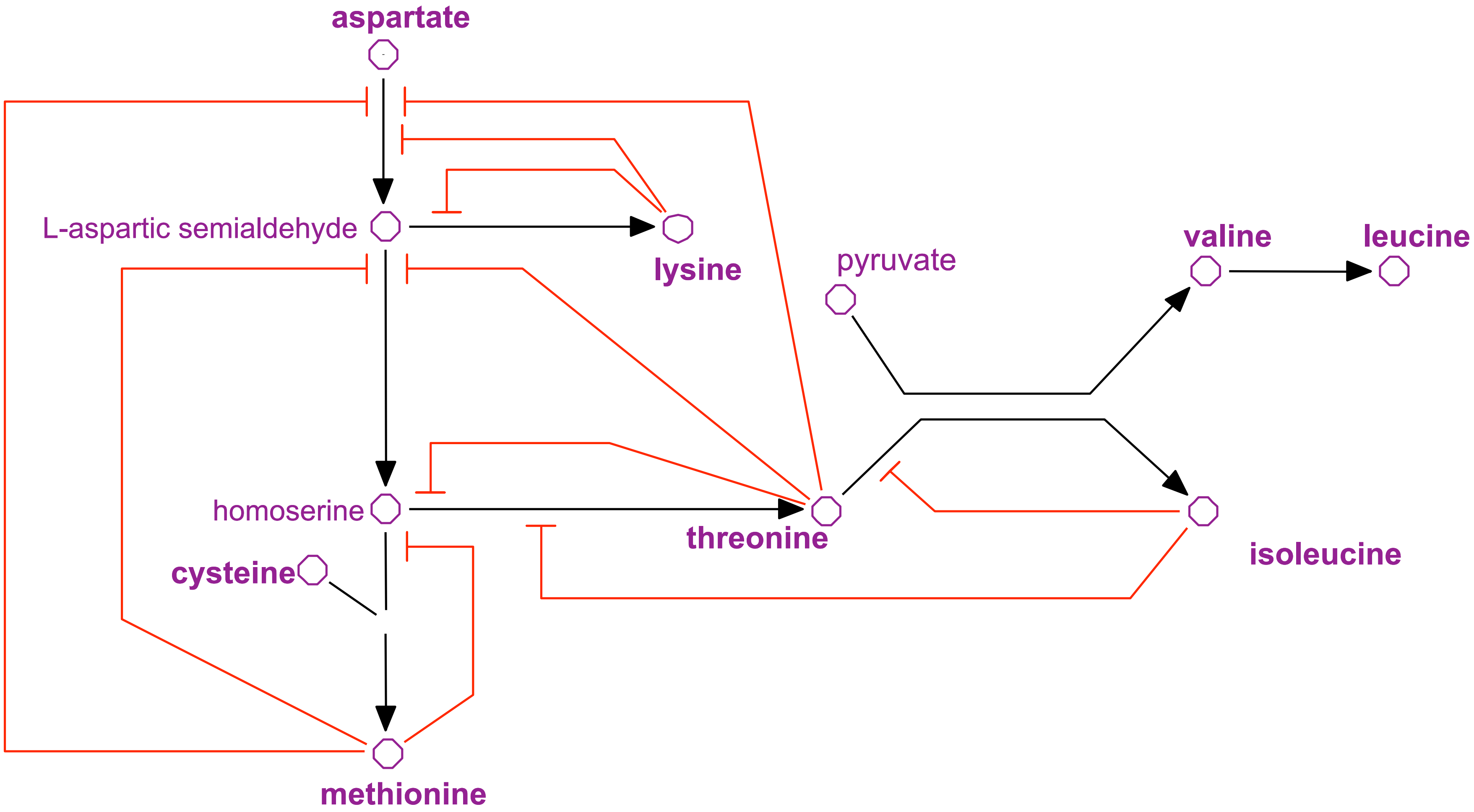
Slide from Jacques van Helden

Shortcut Representation



Slide from Jacques van Helden

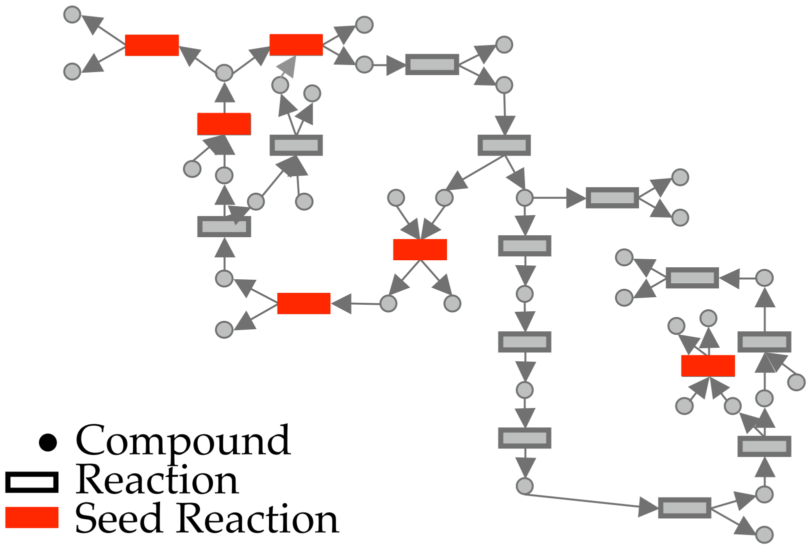
High-level Abstraction



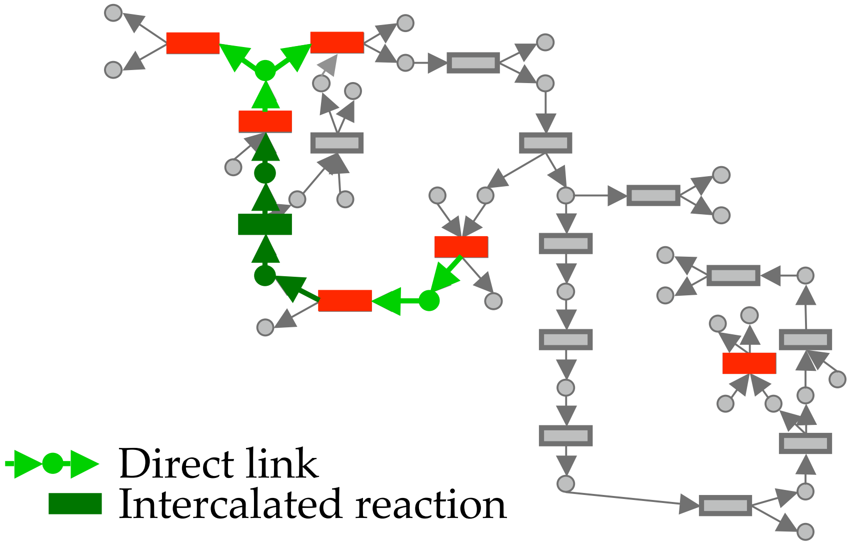
Slide from Jacques van Helden

Queries - subgraph extraction

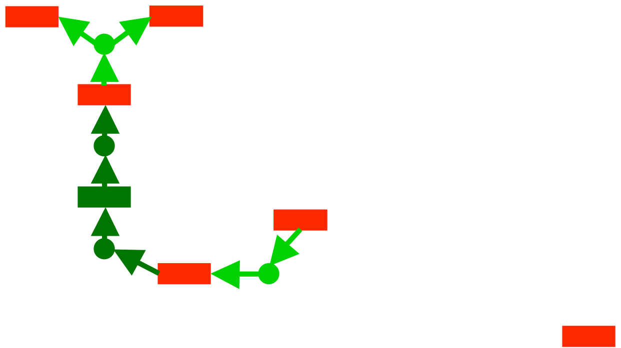
A. Seed reactions



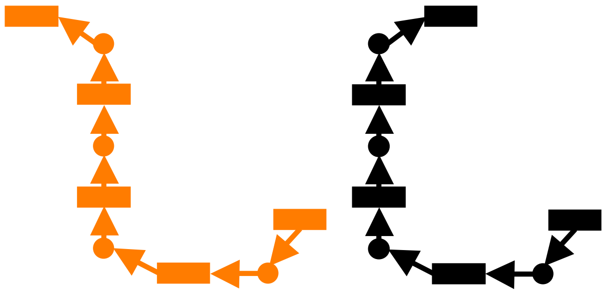
B. Reaction linking



C. Subgraph extraction

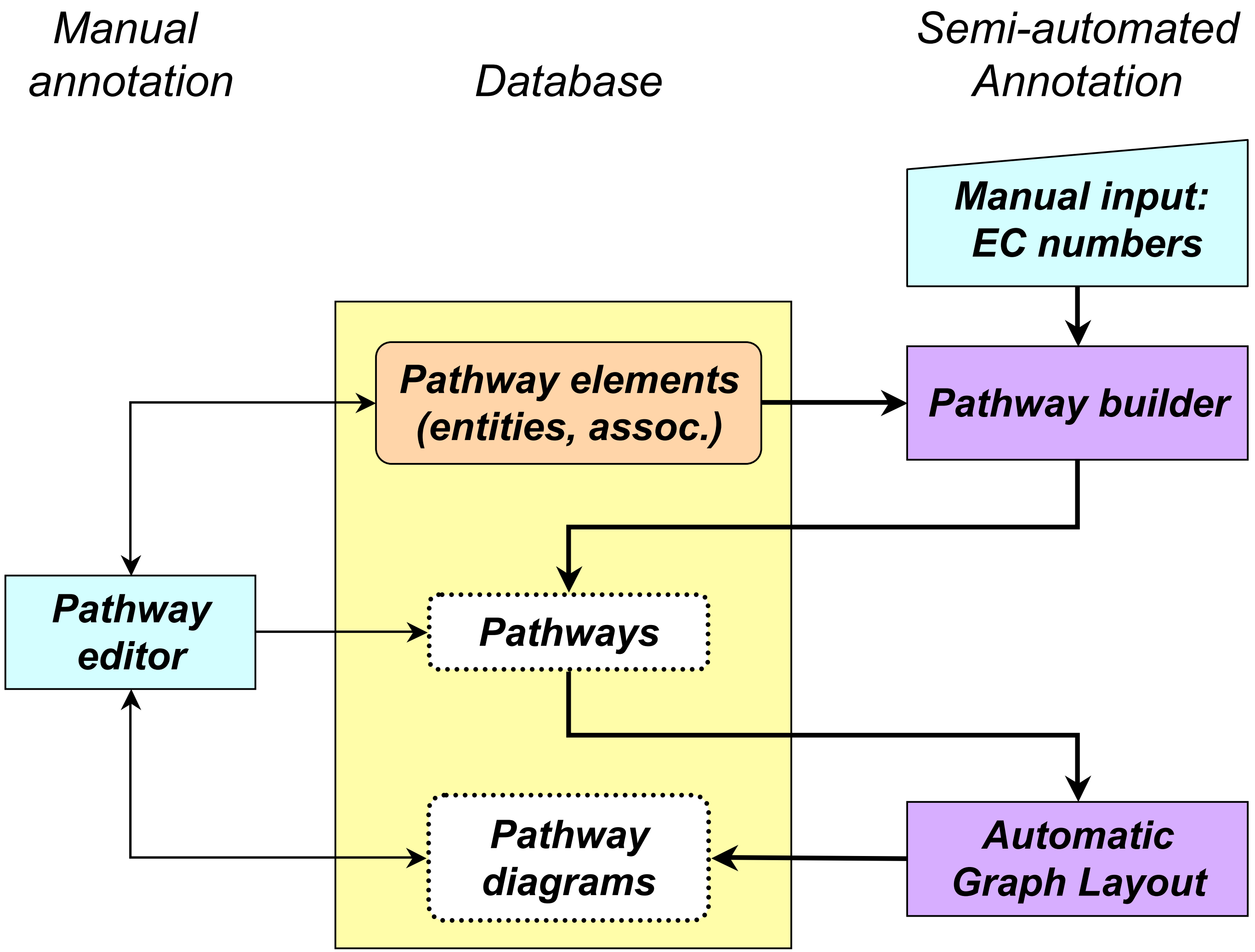


D. Linear Path Enumeration



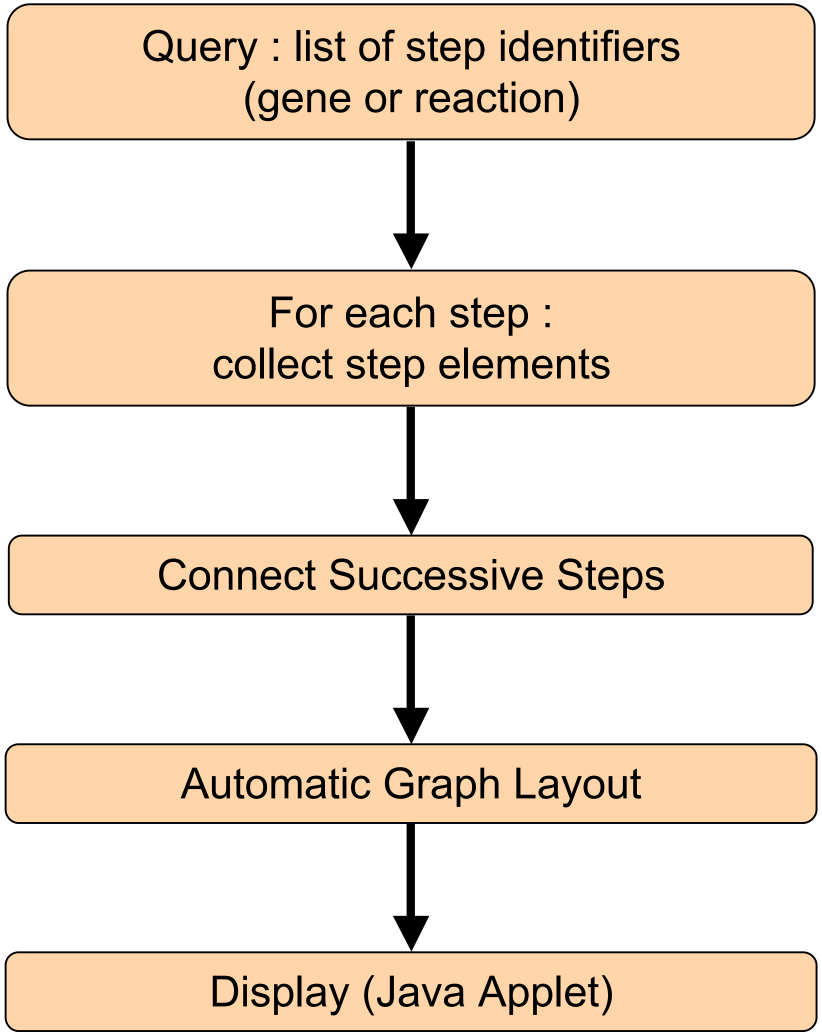
Slide from Jacques van Helden

Pathway Building : semi-automated annotation



Slide from Jacques van Helden

Pathway builder program



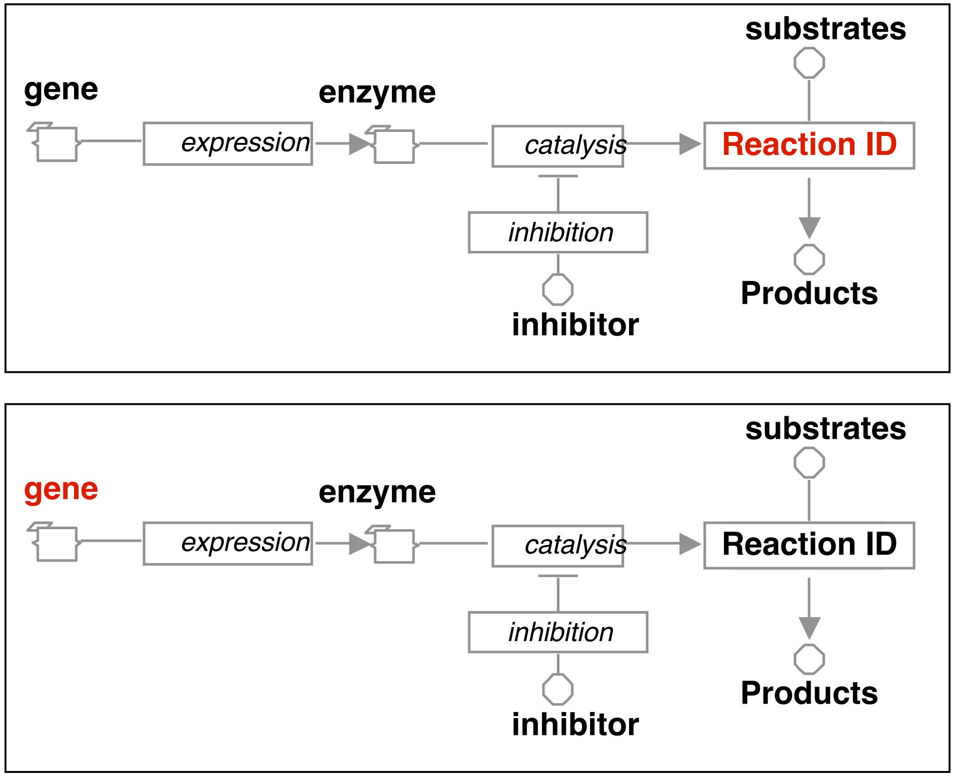
Query : list of step identifiers (gene or reaction)

For each step : collect step elements

Connect Successive Steps

Automatic Graph Layout

Display (Java Applet)

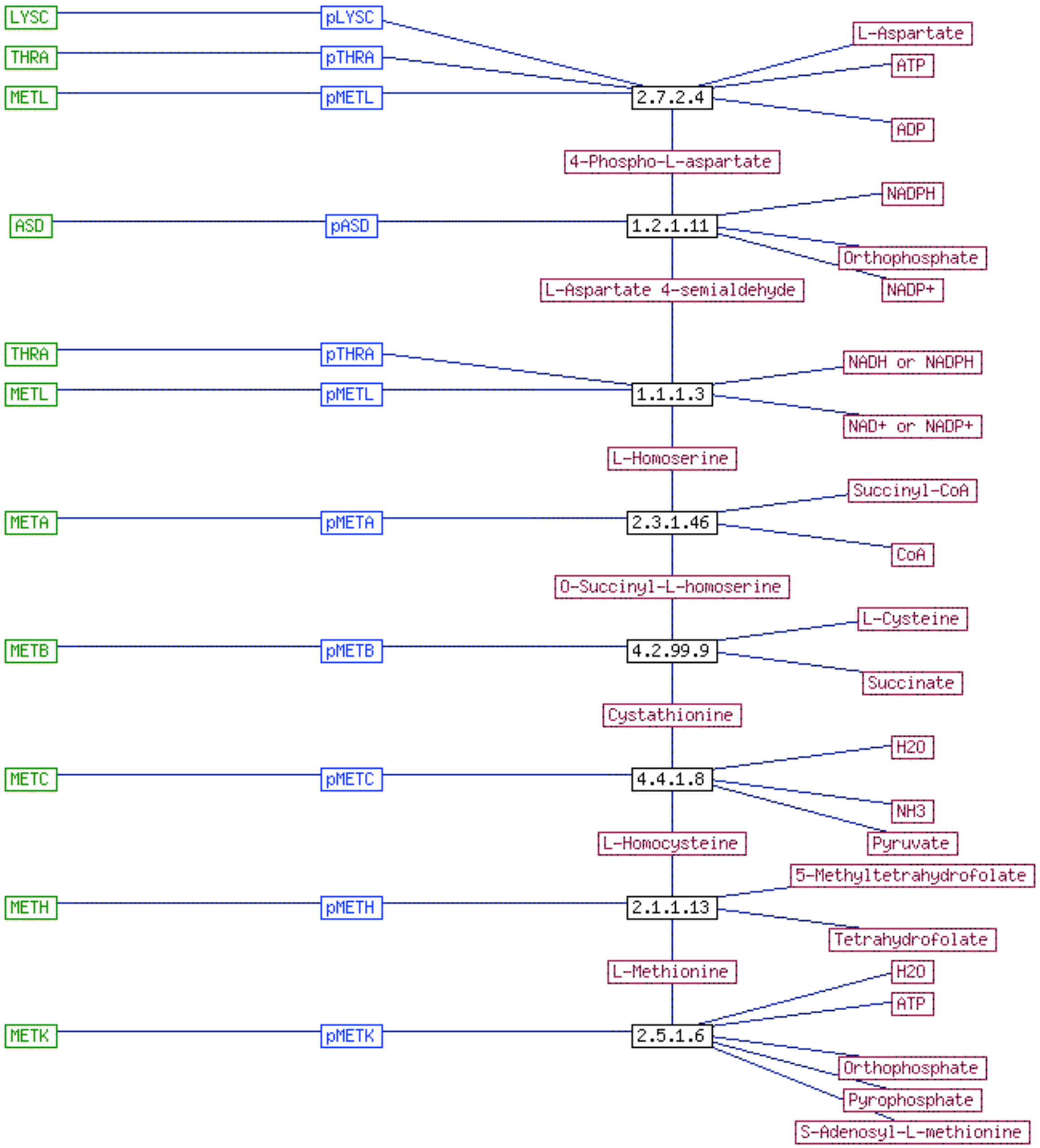


Slide from Jacques van Helden

Metabolic pathway: Query on EC numbers:

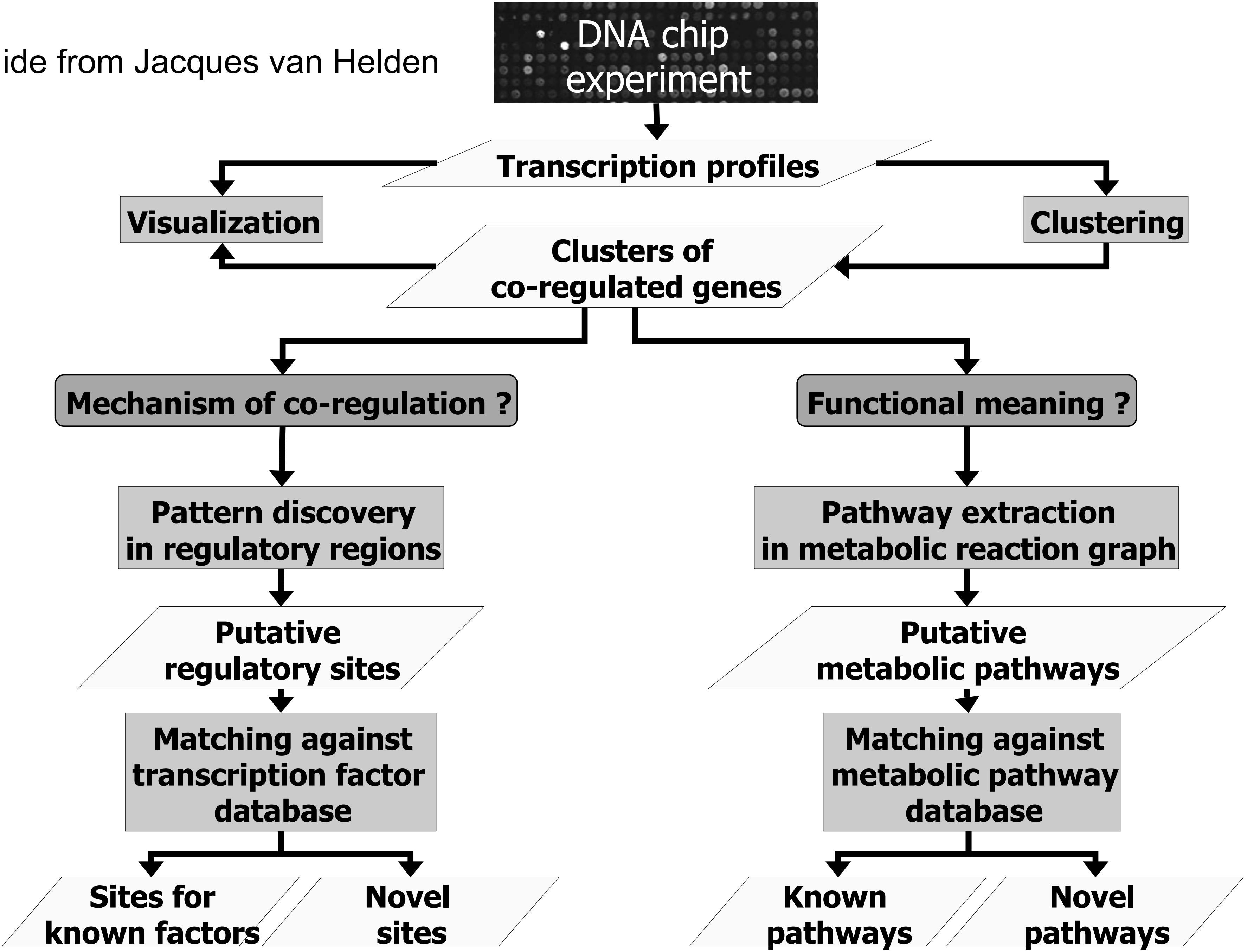
E.coli, methionine biosynthesis

2.7.2.4
1.2.1.11
1.1.1.3
2.3.1.46
4.2.99.9
4.4.1.8
2.1.1.13
2.5.1.6



BRC

Slide from Jacques van Helden


DNA chip experiment

Transcription profiles

Visualization

Clustering

Clusters of co-regulated genes

Mechanism of co-regulation ?

Functional meaning ?

Pattern discovery in regulatory regions

Pathway extraction in metabolic reaction graph

Putative regulatory sites

Putative metabolic pathways

Matching against transcription factor database

Matching against metabolic pathway database

Sites for known factors

Novel sites

Known pathways

Novel pathways

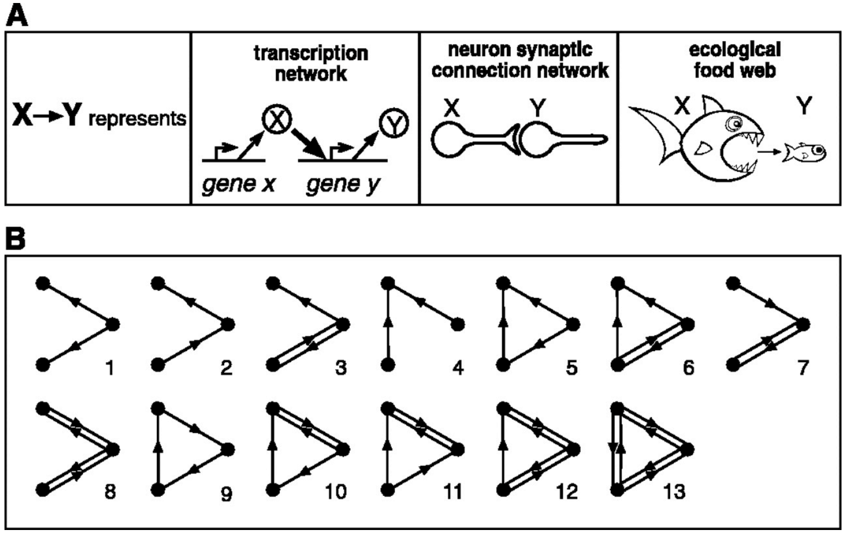
Further graph operations

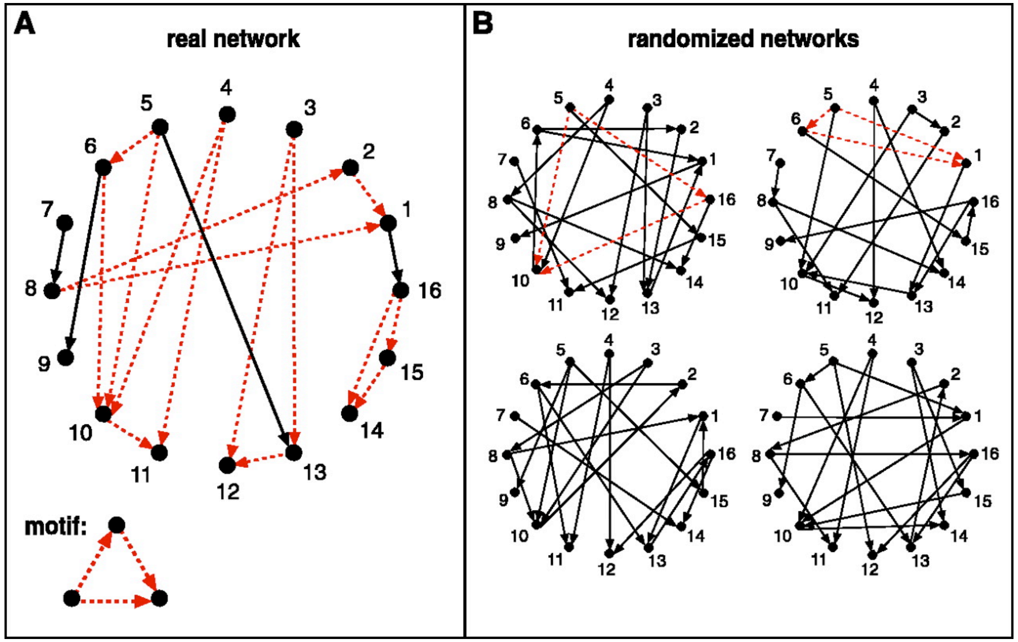
- Sub-graph matching
  - Pattern (graph motif) matching
- Pattern discovery
  - common motif repeated in 1 graph or
  - across many graphs
- Graph comparison

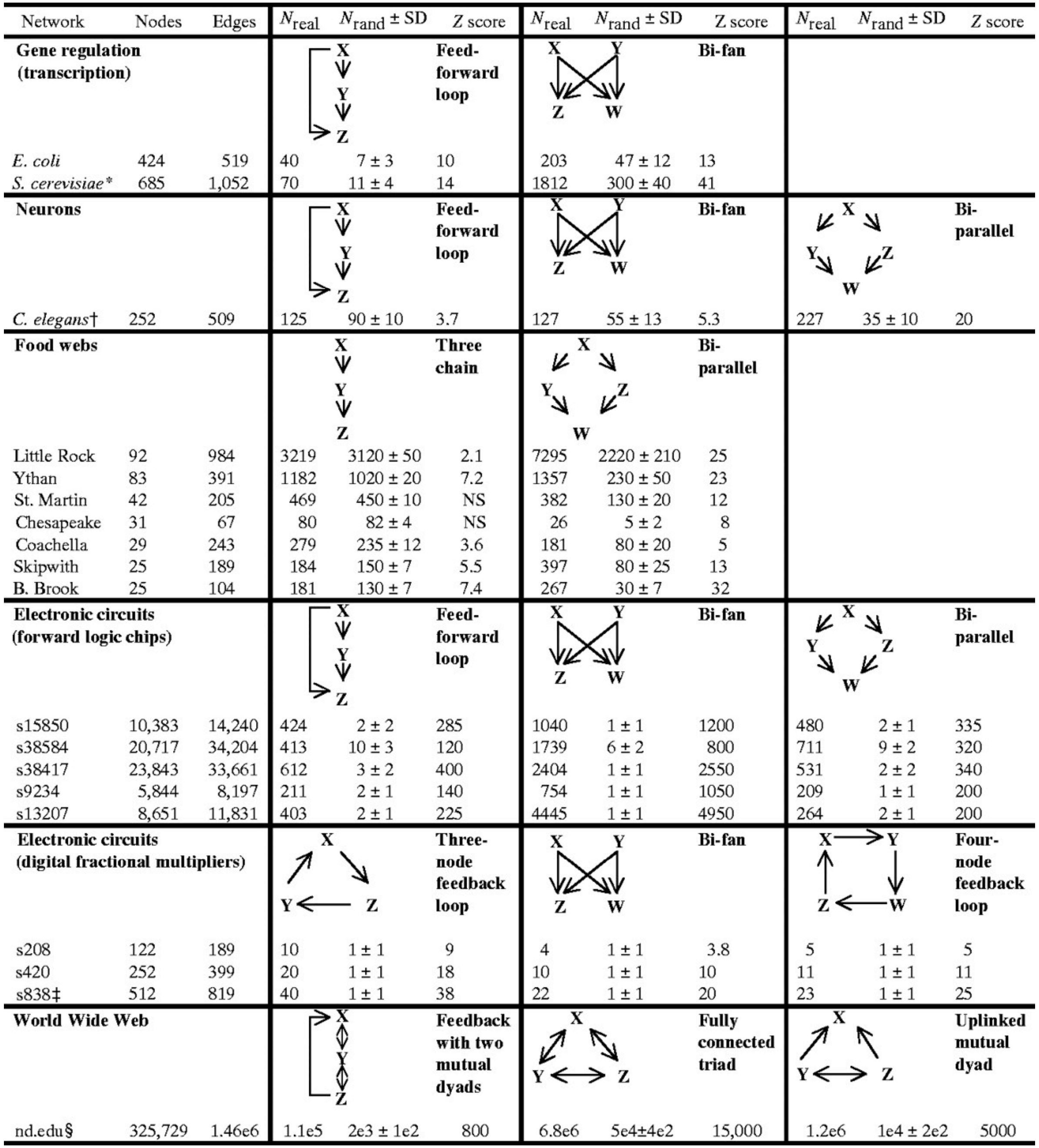
What are the uses?

Network motifs

- Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. Network motifs: simple building blocks of complex networks. Science. 2002 Oct 25;298(5594):824-7.





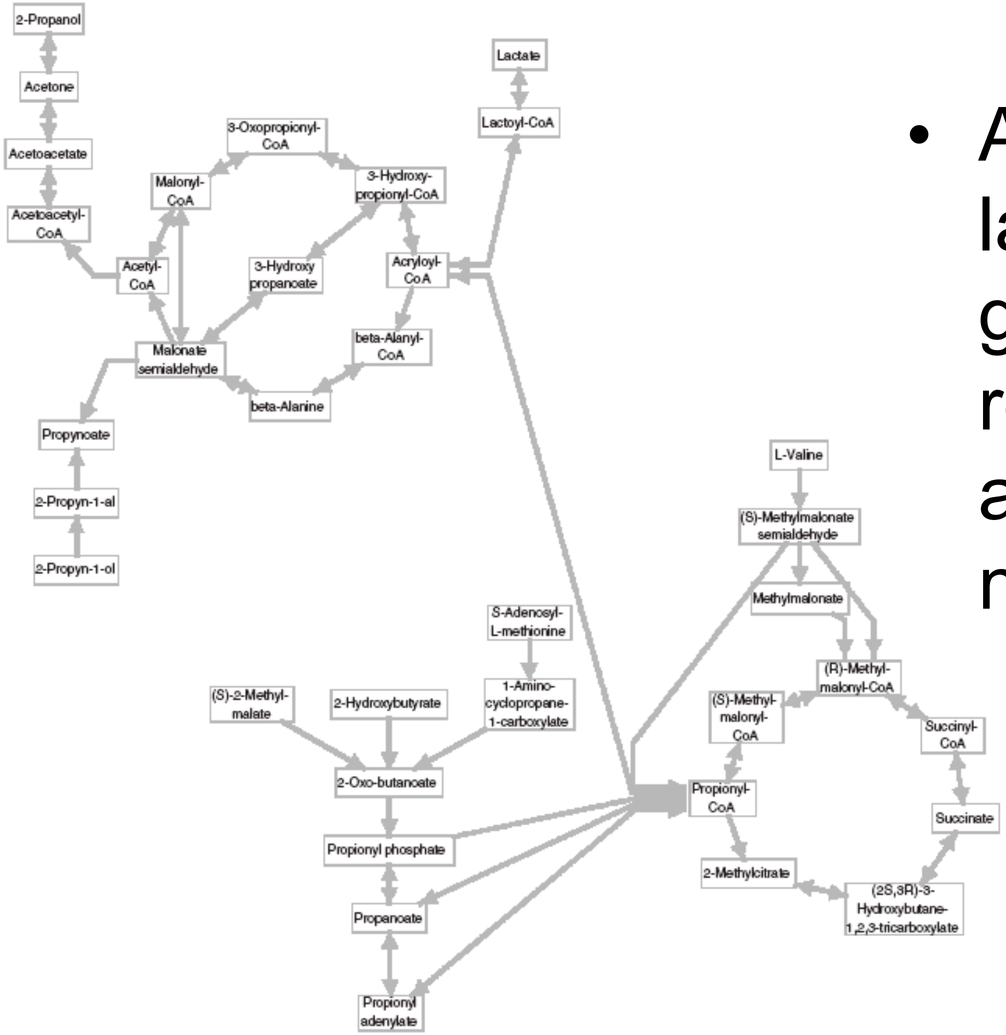


Network	Nodes	Edges	N_{real}	N_{rand} ± SD	Z score	N_{real}	N_{rand} ± SD	Z score	N_{real}	N_{rand} ± SD	Z score
Gene regulation (transcription)					**Feed-forward loop**			**Bi-fan**			
E. coli	424	519	40	7 ± 3	10	203	47 ± 12	13			
*S. cerevisiae**	685	1,052	70	11 ± 4	14	1812	300 ± 40	41			
Neurons					**Feed-forward loop**			**Bi-fan**			**Bi-parallel**
C. elegans†	252	509	125	90 ± 10	3.7	127	55 ± 13	5.3	227	35 ± 10	20
Food webs					**Three chain**			**Bi-parallel**			
Little Rock	92	984	3219	3120 ± 50	2.1	7295	2220 ± 210	25			
Ythan	83	391	1182	1020 ± 20	7.2	1357	230 ± 50	23			
St. Martin	42	205	469	450 ± 10	NS	382	130 ± 20	12			
Chesapeake	31	67	80	82 ± 4	NS	26	5 ± 2	8			
Coachella	29	243	279	235 ± 12	3.6	181	80 ± 20	5			
Skipwith	25	189	184	150 ± 7	5.5	397	80 ± 25	13			
B. Brook	25	104	181	130 ± 7	7.4	267	30 ± 7	32			
Electronic circuits (forward logic chips)					**Feed-forward loop**			**Bi-fan**			**Bi-parallel**
s15850	10,383	14,240	424	2 ± 2	285	1040	1 ± 1	1200	480	2 ± 1	335
s38584	20,717	34,204	413	10 ± 3	120	1739	6 ± 2	800	711	9 ± 2	320
s38417	23,843	33,661	612	3 ± 2	400	2404	1 ± 1	2550	531	2 ± 2	340
s9234	5,844	8,197	211	2 ± 1	140	754	1 ± 1	1050	209	1 ± 1	200
s13207	8,651	11,831	403	2 ± 1	225	4445	1 ± 1	4950	264	2 ± 1	200
Electronic circuits (digital fractional multipliers)					**Three-node feedback loop**			**Bi-fan**			**Four-node feedback loop**
s208	122	189	10	1 ± 1	9	4	1 ± 1	3.8	5	1 ± 1	5
s420	252	399	20	1 ± 1	18	10	1 ± 1	10	11	1 ± 1	11
s838‡	512	819	40	1 ± 1	38	22	1 ± 1	20	23	1 ± 1	25
World Wide Web					**Feedback with two mutual dyads**			**Fully connected triad**			**Uplinked mutual dyad**
nd.edu§	325,729	1.46e6	1.1e5	2e3 ± 1e2	800	6.8e6	5e4±4e2	15,000	1.2e6	1e4 ± 2e2	5000

Databases, data structures

- Adjacency matrix
- Relational Database models
- ?Can you construct a simple database?

Visualisation of pathways



- Automatic graph layout algorithms are good for visualising relational information and work for small networks

Summary

- Data models for Networks, pathways
- (Sets)
- (Trees)
- Graphs
  - Paths, circuits, searching
  - Breadth-first search
  - Depth-first search

- Analysis
  - Some algorithms over graphs

Scale-free networks

- Using a Web crawler, physicist Albert-Laszlo Barabasi and his colleagues at the University of Notre Dame in Indiana, USA, in 1999 mapped the connectedness of the Web. To their surprise, the web did not have an even distribution of connectivity (so-called "random connectivity").

- Instead, a very few network nodes (called "hubs") were far more connected than other nodes.

- In general, they found that the probability $p(k)$ that a node in the network connects with k other nodes was, in a given network, proportional to $k-\gamma$.

- The degree exponent γ is not universal and depends on the detail of network structure. Numerical values of the exponent γ for various systems are diverse but most of them are in the range $2 < \gamma \leq 3$.
- At the same time a similar observation was obtained to the Internet by the Faloutsos brothers (1999). In this form, essentially all graphs with a power law degree distribution were grouped together as "scale-free". Several revisions of this definition have been suggested.

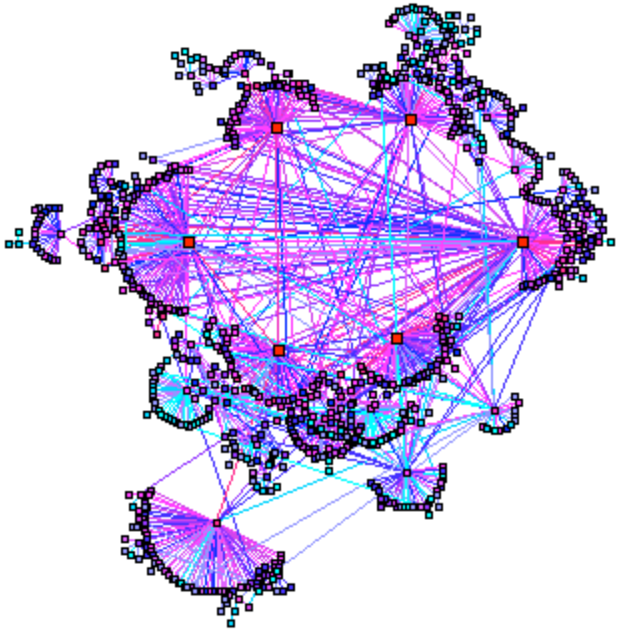
[Wikipedia]

World Wide Web

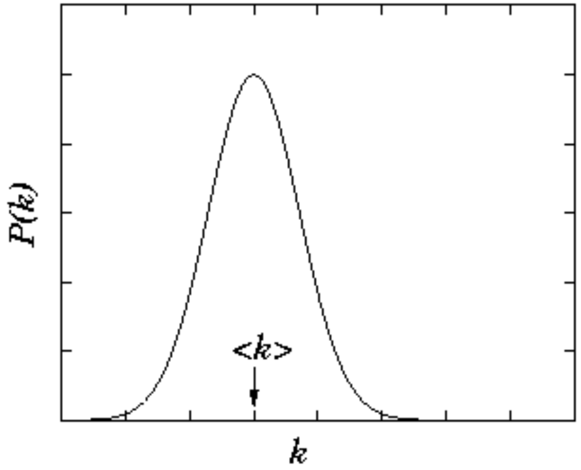
Nodes: WWW documents
Links: URL links

Over 1 billion documents

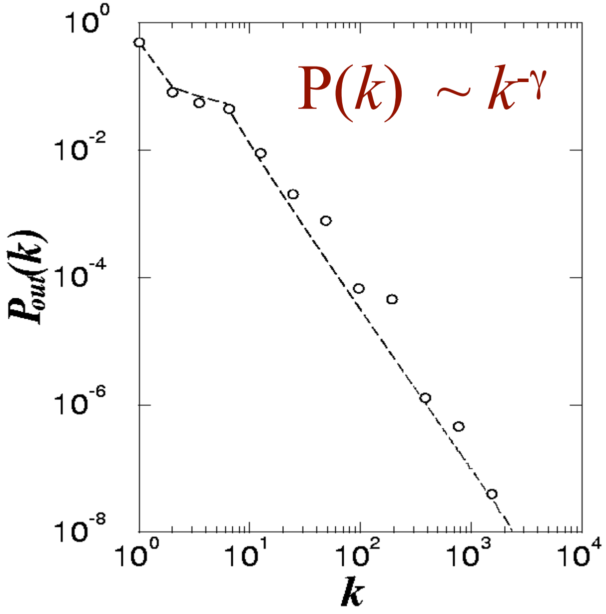
ROBOT: `collects all URL's found in a document and follows them recursively`



R. Albert, H. Jeong & A.-L. Barabasi, **Nature**, 1999



Expected



Found

Stefan Wuchty
www.nd.edu/~swuchty/Download/pisa.pp

Scale-free networks

- Tend to contain centrally located and interconnected high degree "hubs", which dramatically influences the way a network operates.
  - For example, random node failures have very little effect on a scale-free network's connectivity or effectiveness
  - Deliberate attacks on such a network's hubs can dismantle a network with alarming ease. Thus, the realization that certain networks are scale-free is important to security.

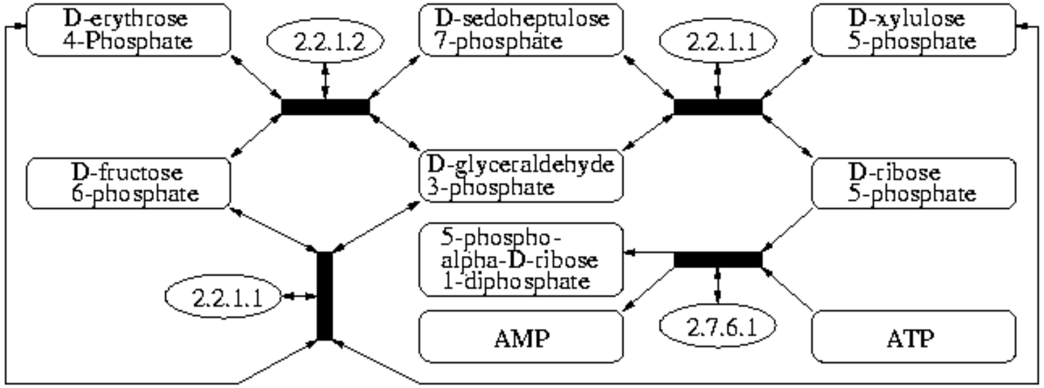
- SCF also exhibit the Small world phenomenon: two average nodes are separated by a very small number of connections.
-
- Also, scale-free networks generally have high clustering coefficients.

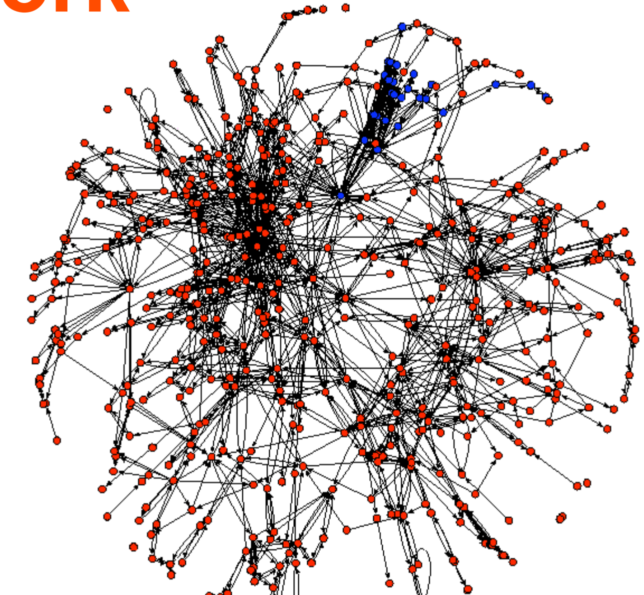
- A multitude of real-world networks have been shown to be scale-free, including:
  - Social networks, including collaboration networks. An example that have been studied extensively is the collaboration of movie actors in films.
  - Protein-Protein interaction networks.
  - Sexual partners in humans, which affects the dispersal of sexually transmitted diseases.
  - Many kinds of computer networks, including the World Wide Web.

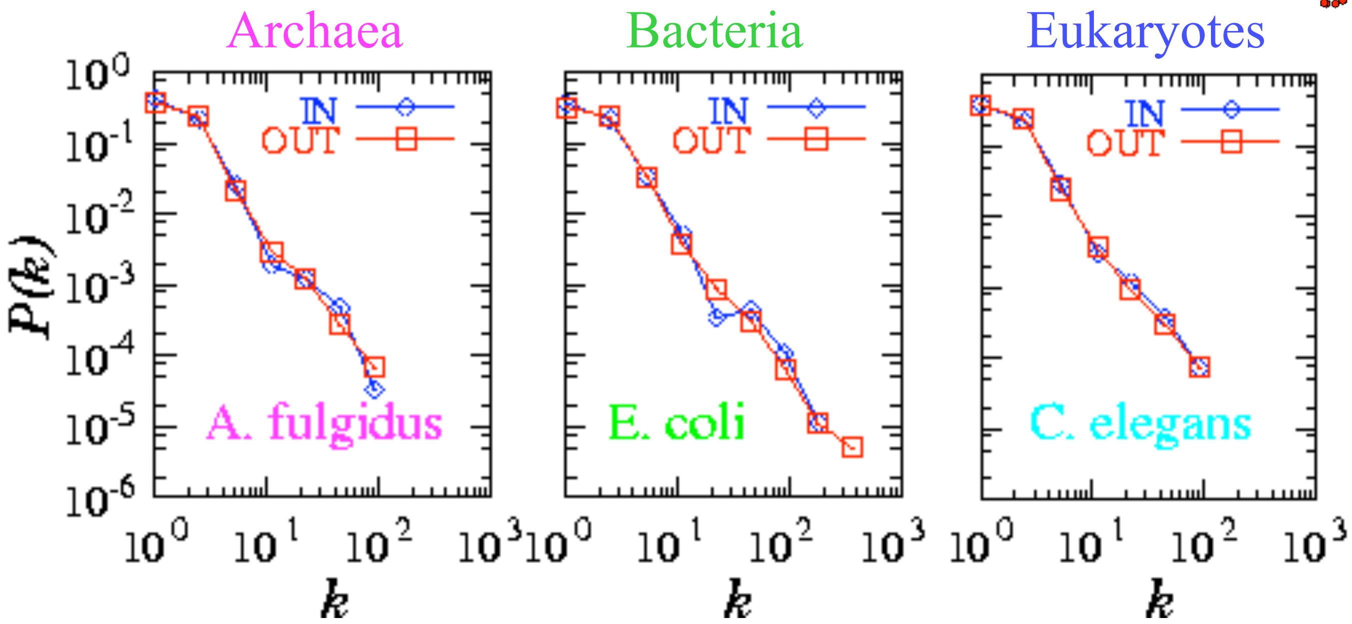
[Wikipedia]

Stefan Wuchty
www.nd.edu/~swuchty/Download/pisa.ppt

Metabolic network







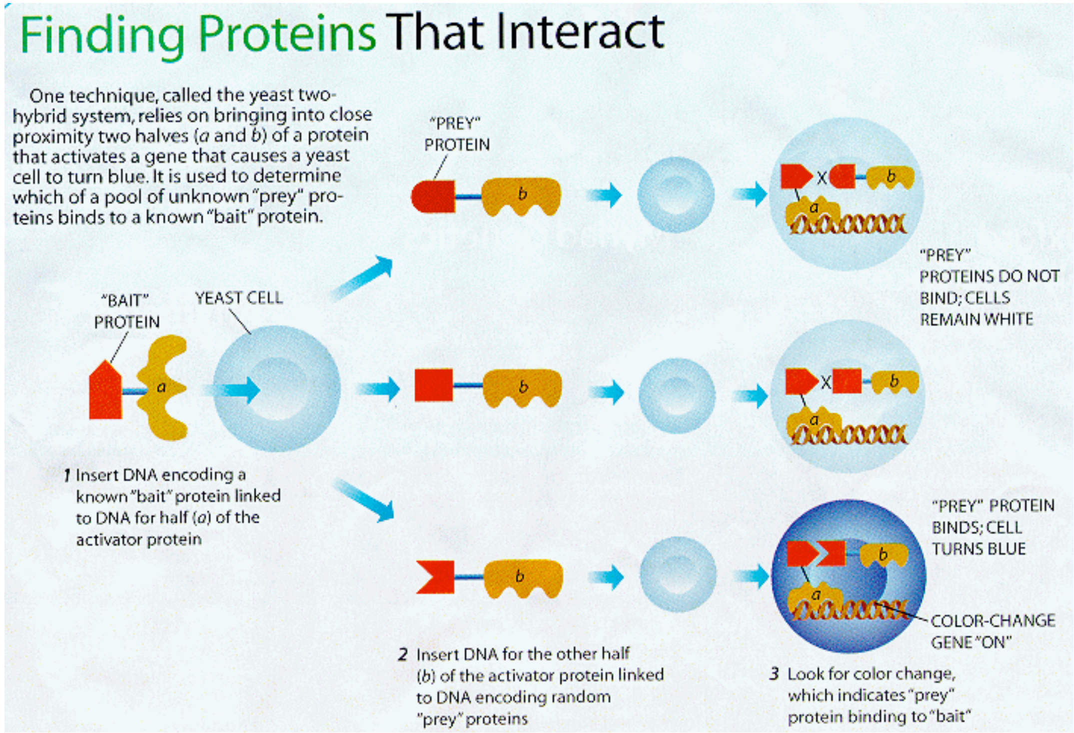
Organisms from all three domains of life are **scale-free** networks!

H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai, and A.L. Barabasi, **Nature**, 2000

Yeast protein network

Nodes: proteins

Links: physical interactions (binding)



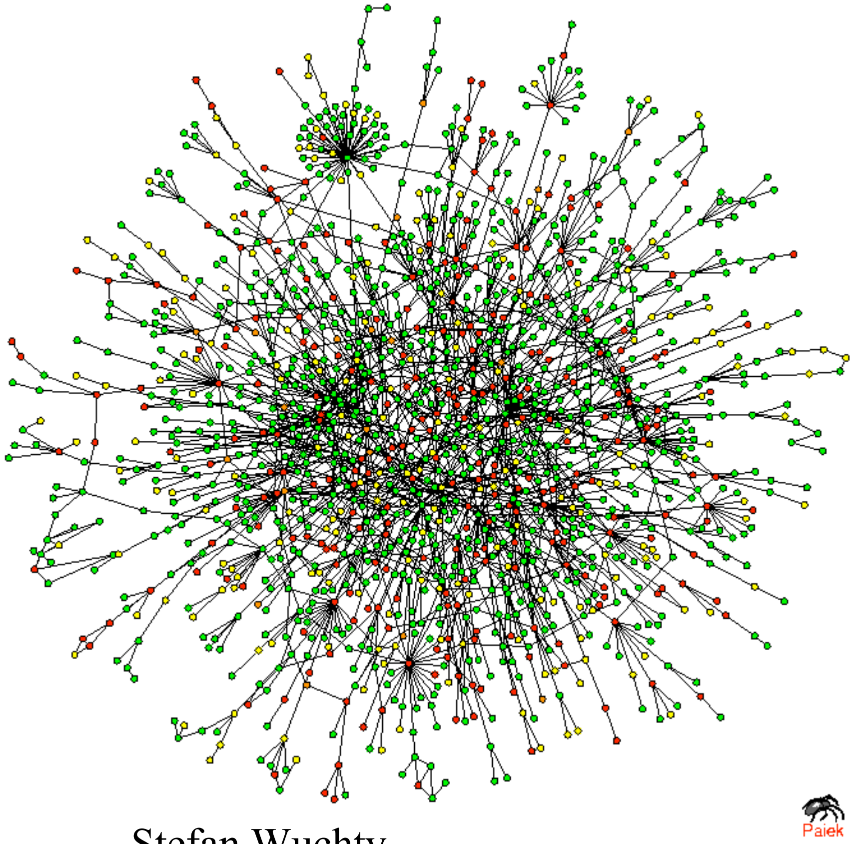
Finding Proteins That Interact

One technique, called the yeast two-hybrid system, relies on bringing into close proximity two halves (*a* and *b*) of a protein that activates a gene that causes a yeast cell to turn blue. It is used to determine which of a pool of unknown "prey" proteins binds to a known "bait" protein.

1 Insert DNA encoding a known "bait" protein linked to DNA for half (*a*) of the activator protein

2 Insert DNA for the other half (*b*) of the activator protein linked to DNA encoding random "prey" proteins

3 Look for color change, which indicates "prey" protein binding to "bait"



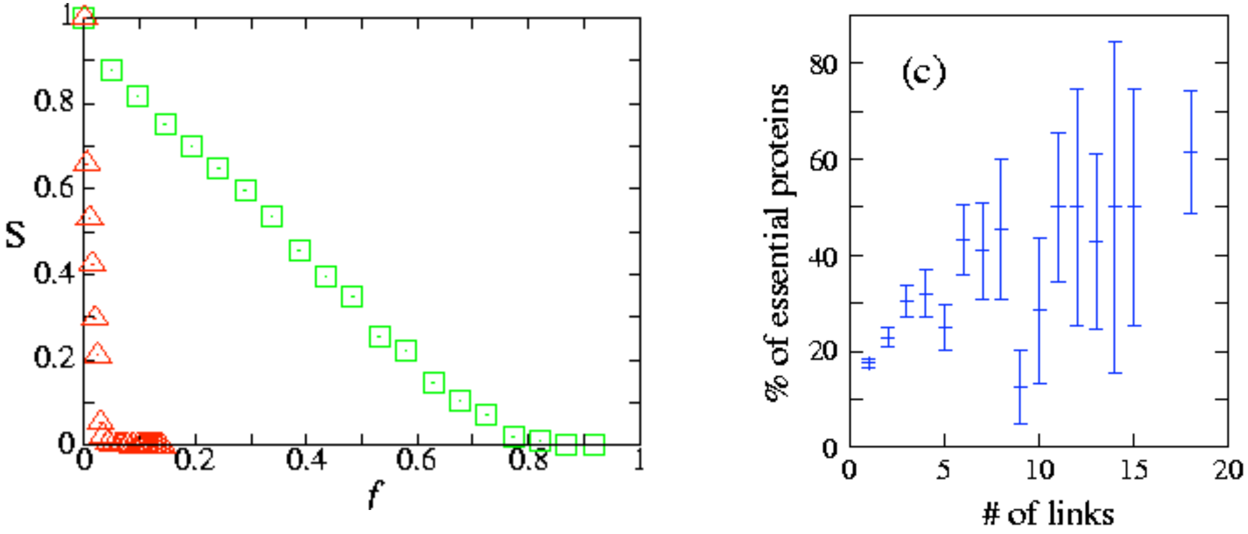
P. Uetz, et al. **Nature**, 2000

Stefan Wuchty
www.nd.edu/~swuchty/Download/pisa.pp

Stefan Wuchty
www.nd.edu/~swuchty/Download/pisa.ppt

Yeast protein network

- lethality and topological position -

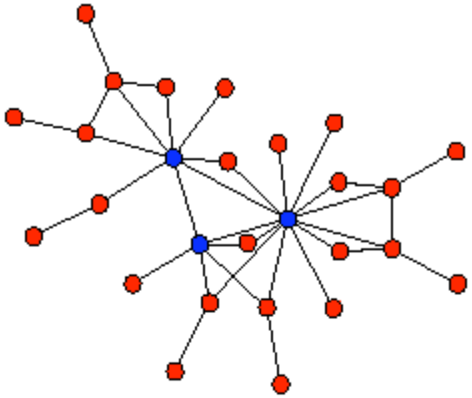


Highly connected proteins are more **essential (lethal)**...

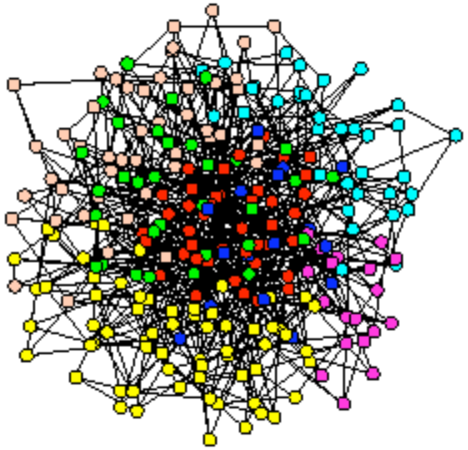
H. Jeong, S.P. Mason, A.-L. Barabasi &Z.N. Oltvai, **Nature**, 2001

Modular vs. Scale-free Topology

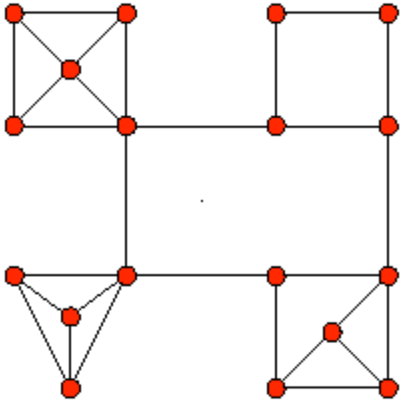
(a)



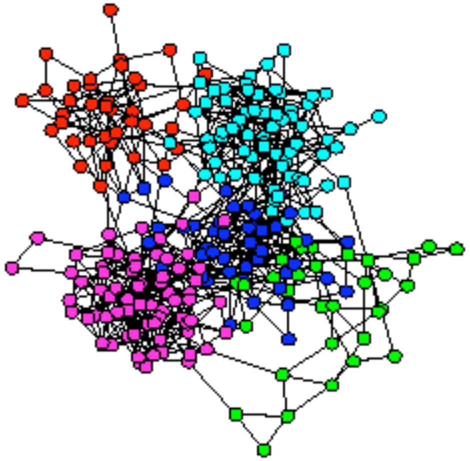
Scale-free



(b)



Modular

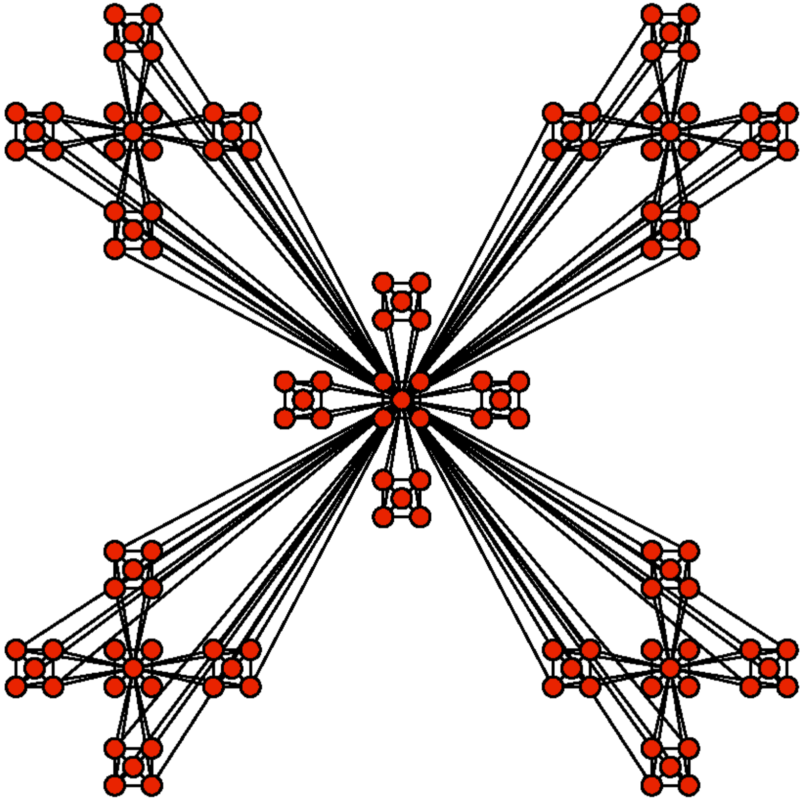


Stefan Wuchty
www.nd.edu/~swuchty/Download/pisa.pp

Real Networks Have a Hierarchical Topology

What does it mean?

Many highly connected small clusters
combine into
few larger but less connected clusters
combine into
even larger and even less connected clusters



Stefan Wuchty
www.nd.edu/~swuchty/Download/pisa.pp

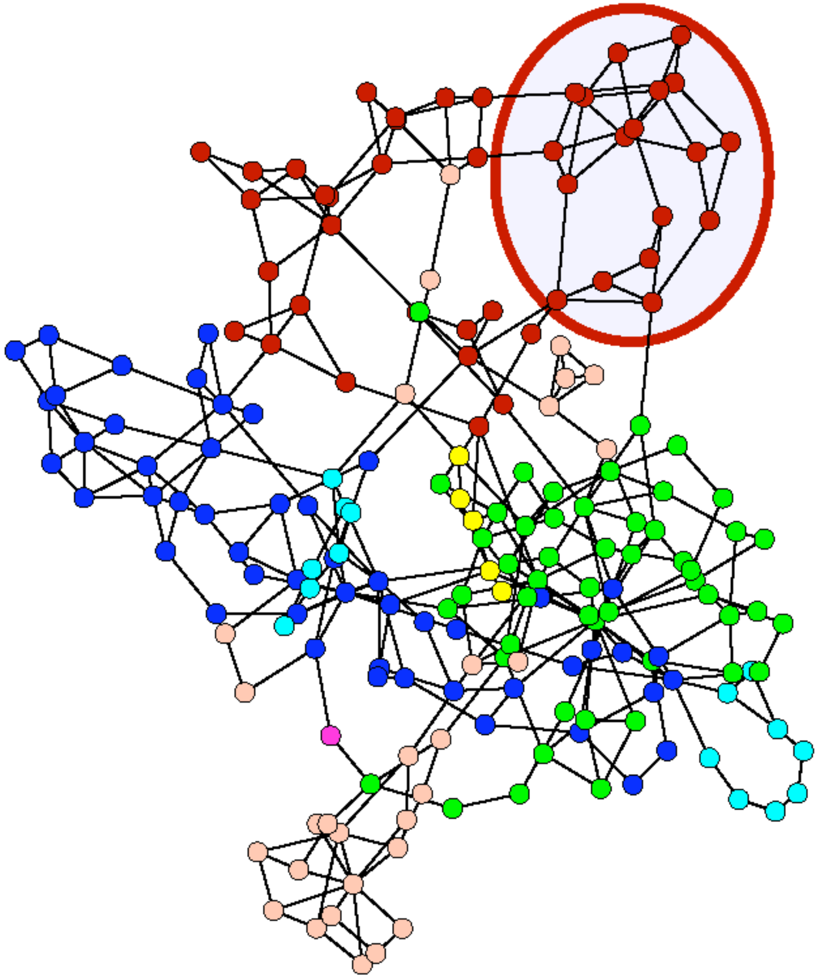
- The degree of clustering follows:

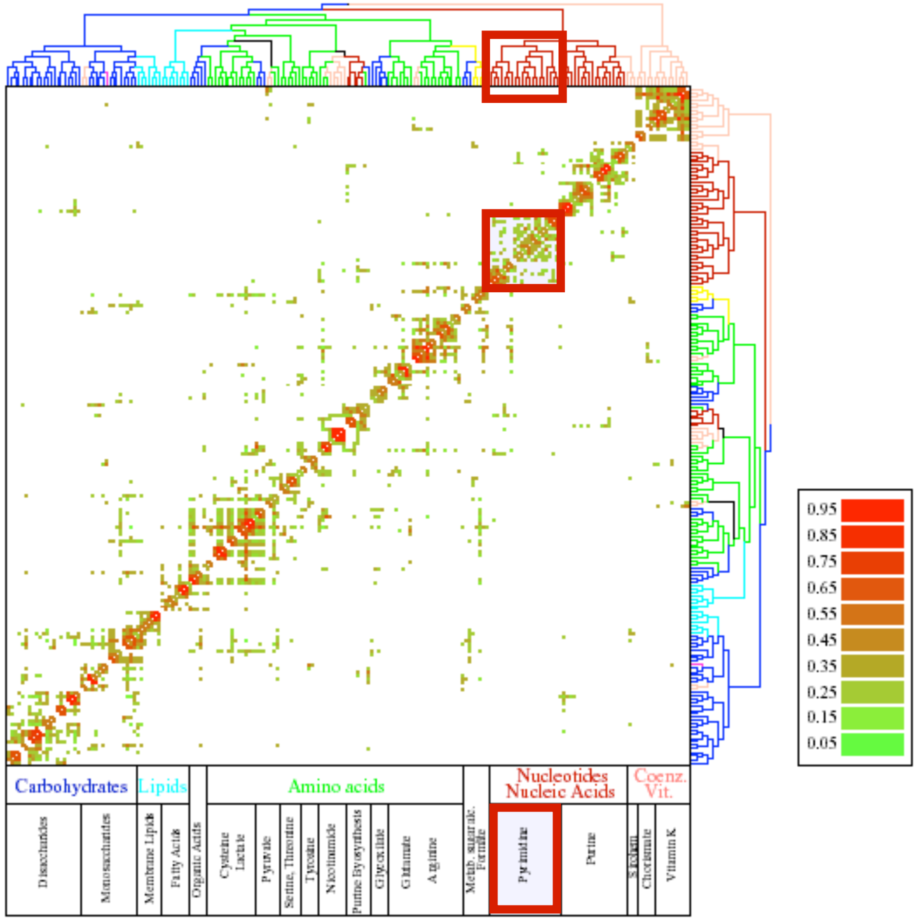
$$C(k) \sim k^{-\beta}$$

Modules in the *E. coli* metabolism

E. Ravasz et al., **Science**, 2002

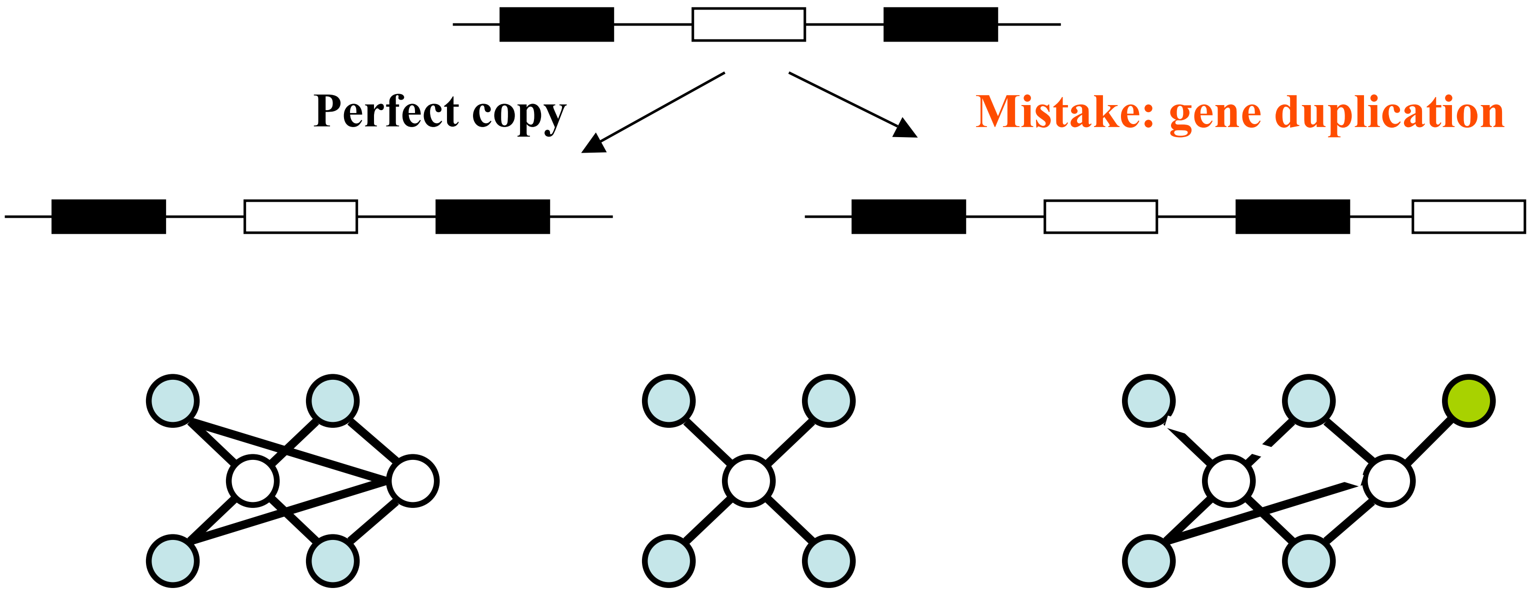
Stefan Wuchty
www.nd.edu/~swuchty/Download/pisa.ppt





Origin of scaling in protein interaction networks

Stefan Wuchty
www.nd.edu/~swuchty/Download/pisa.ppt

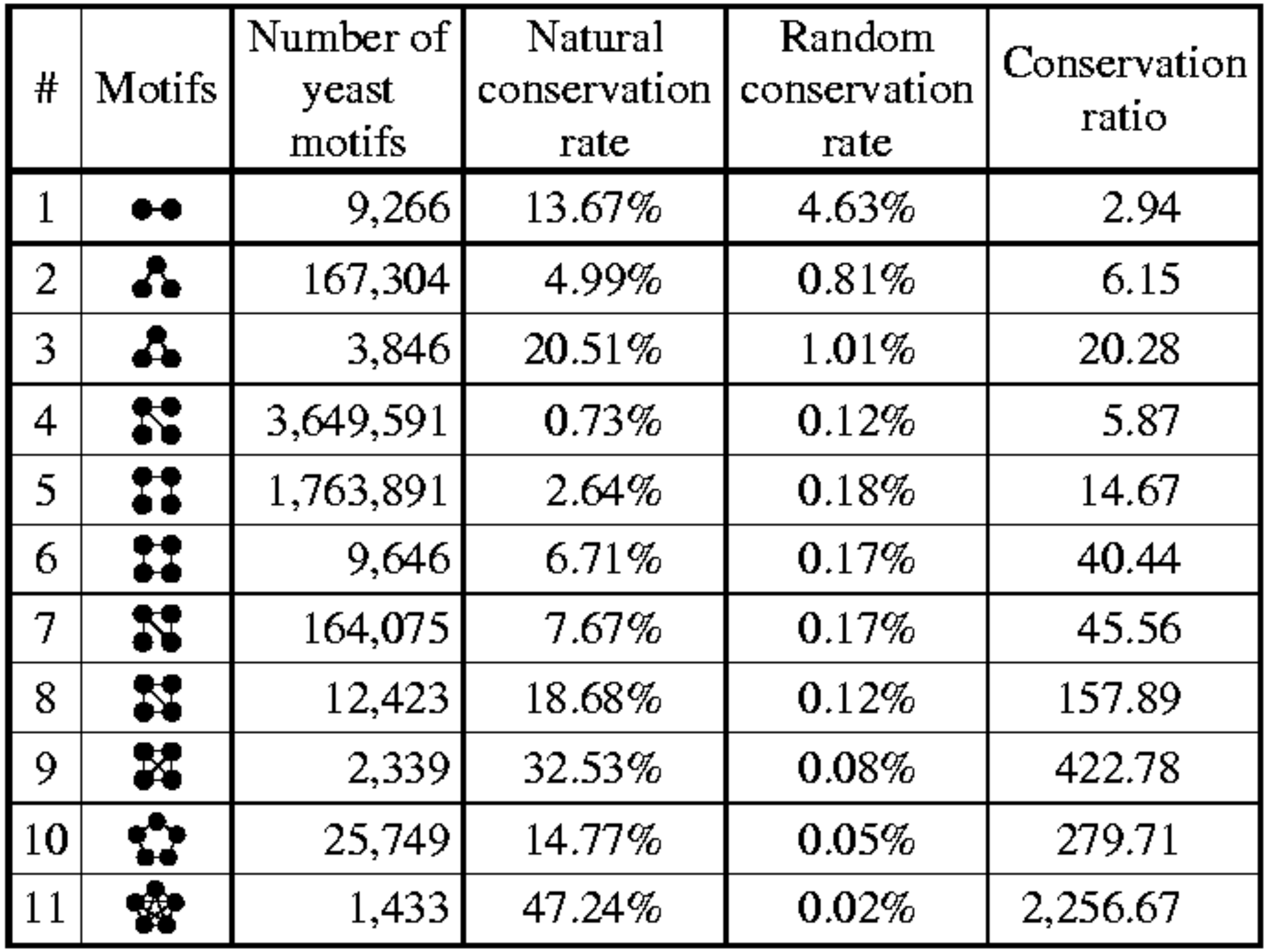


Proteins with more interactions are more likely to get a new link:
Π(k)~k
(preferential attachment)

Vazquez et al., ***cond-mat/0108043***
Sole et al., ***Adv. Compl. Syst***., 2001

Stefan Wuchty
www.nd.edu/~swuchty/Download/pisa.ppt

Topology and Evolution

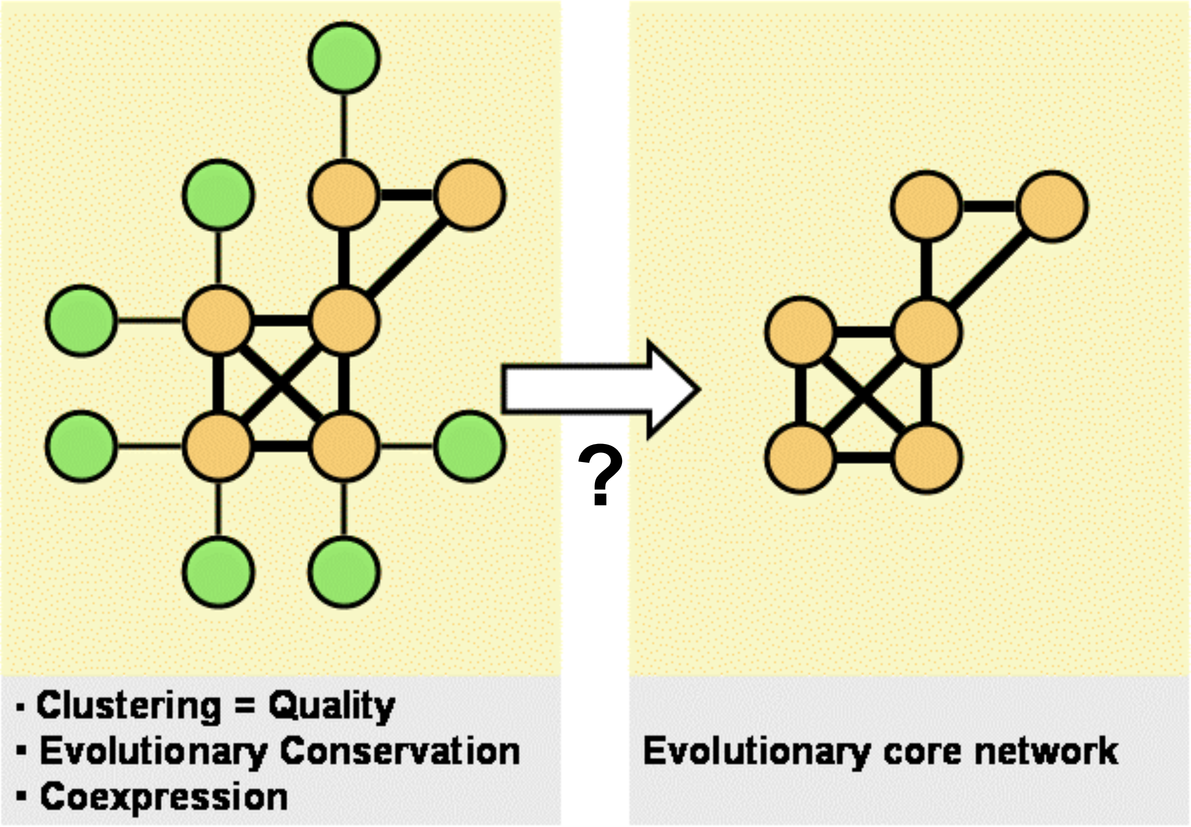


#	Motifs	Number of yeast motifs	Natural conservation rate	Random conservation rate	Conservation ratio
1		9,266	13.67%	4.63%	2.94
2		167,304	4.99%	0.81%	6.15
3		3,846	20.51%	1.01%	20.28
4		3,649,591	0.73%	0.12%	5.87
5		1,763,891	2.64%	0.18%	14.67
6		9,646	6.71%	0.17%	40.44
7		164,075	7.67%	0.17%	45.56
8		12,423	18.68%	0.12%	157.89
9		2,339	32.53%	0.08%	422.78
10		25,749	14.77%	0.05%	279.71
11		1,433	47.24%	0.02%	2,256.67

S. Wuchty, Z. Oltvai & A.-L. Barabasi, **Nature Genetics**, 2003

Stefan Wuchty
www.nd.edu/~swuchty/Download/pisa.ppt

Open questions



- Clustering = Quality
- Evolutionary Conservation
- Coexpression

Evolutionary core network

What is the meaning of clustering in other systems (quality measure)?