Pathways and networks

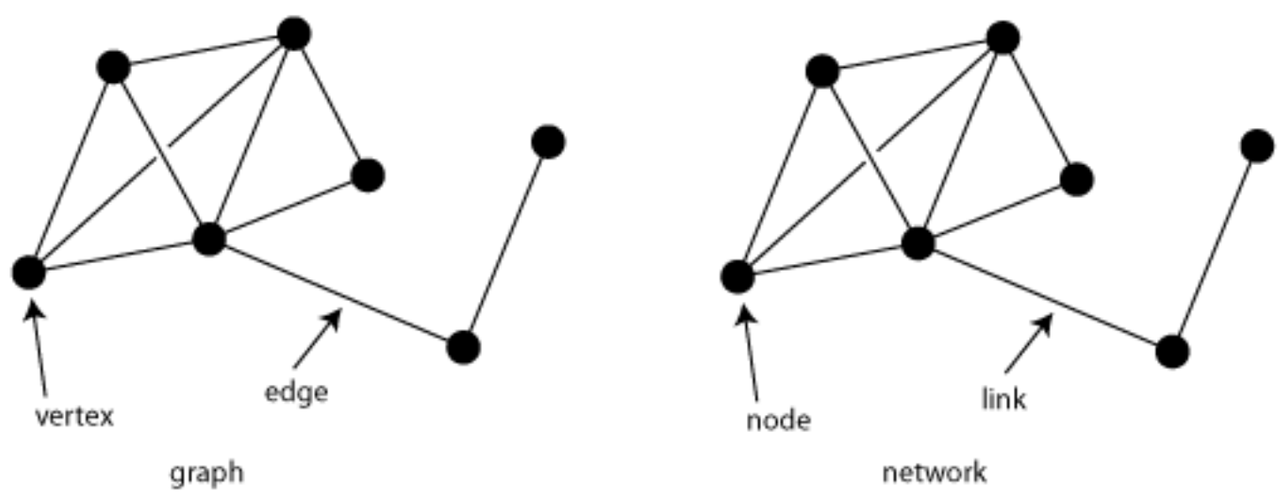
Jon Ambler
Nicky Mulder

Pathways and networks

- What is a network?
- Types of network
- How to get / make them
- What you can use them for

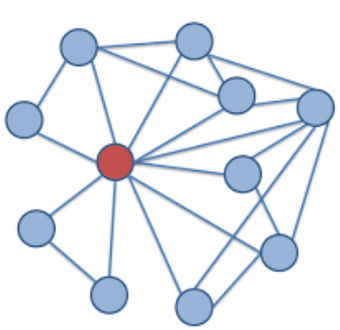
What is a network?

- Network vs graphs
  - “Graphs are mathematical structures that are used to model the pairwise interactions between objects” - Wikipedia
  - Maths: Part of discreet mathematics, are structures
  - Computer science: Abstract data type
- Nodes (vertices) and links (edges)
- Edges can have attributes such as weights (Nodes too)



What is a network?

- Useful for visualisation
- Way to structure data that is relational in nature
- Allow us to apply methods that require data to be in a relational structure
- Allow us to use methods from other disciplines to resolve challenges of working with complex datasets



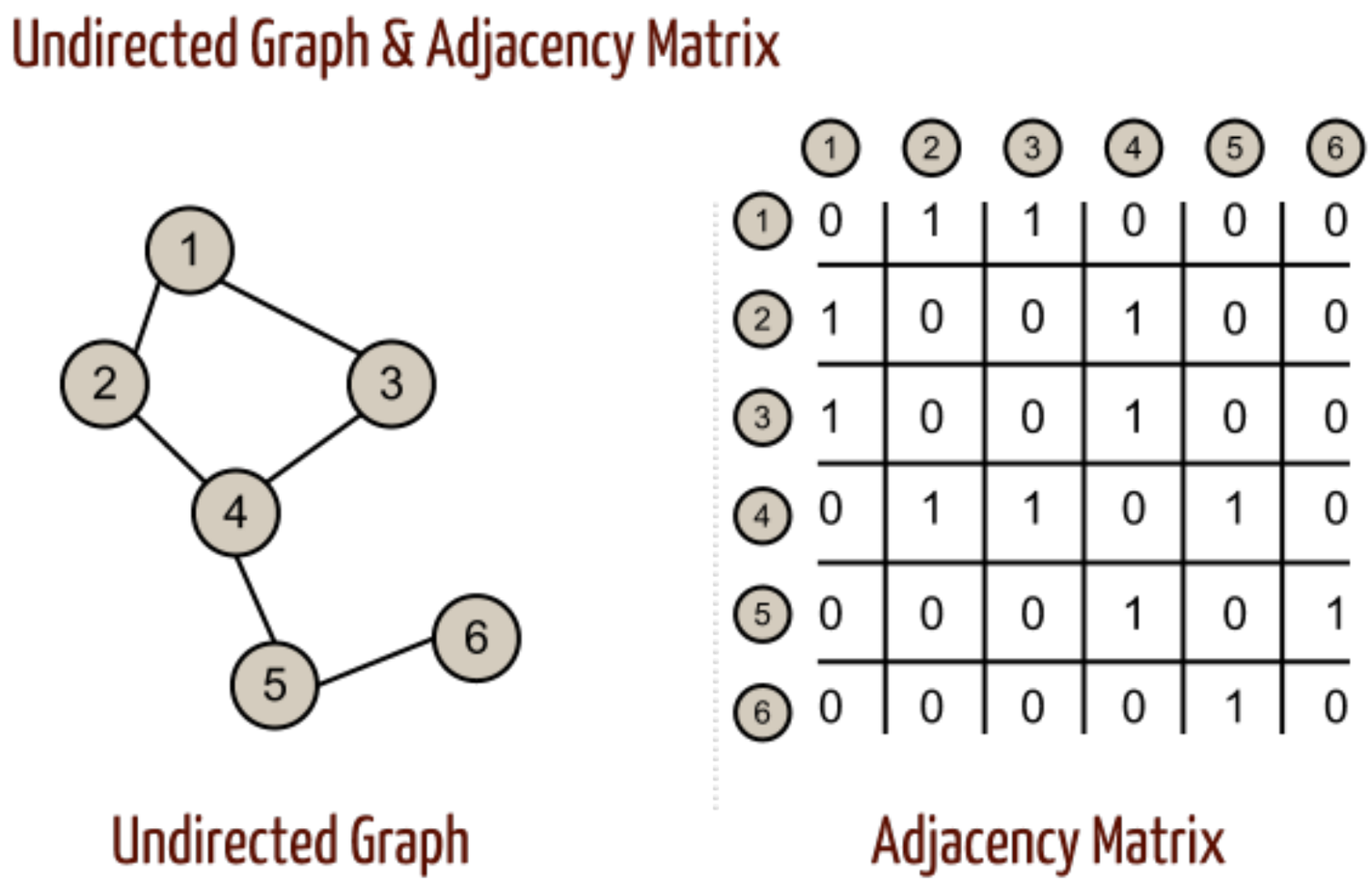
Types of networks

- Protein-protein interaction networks
- Gene regulatory networks
- Gene co-expression networks
- Metabolic networks
- Signalling networks

Graph notation

- A graph (G) is represented as an ordered pair
  - $G = (V, E)$
- Where V is as set of vertices / nodes
- E is a set of edges / lines

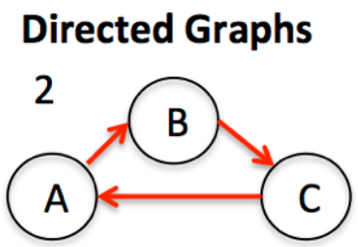
Adjacency matrix



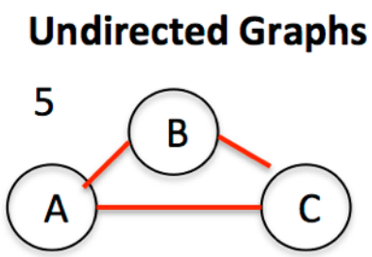
	1	2	3	4	5	6
1	0	1	1	0	0	0
2	1	0	0	1	0	0
3	1	0	0	1	0	0
4	0	1	1	0	1	0
5	0	0	0	1	0	1
6	0	0	0	0	1	0

Properties of a graph

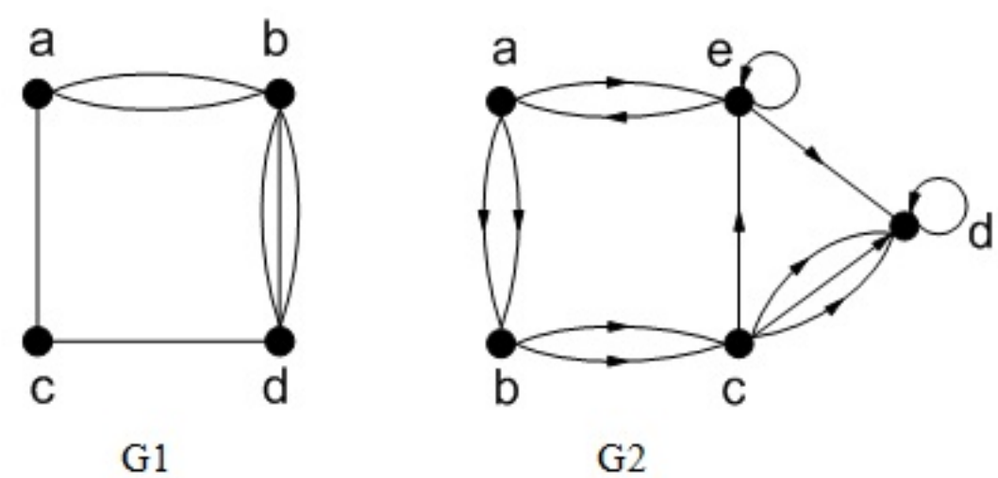
- A graph can be directed or undirected (Or mixed)



- Multigraph

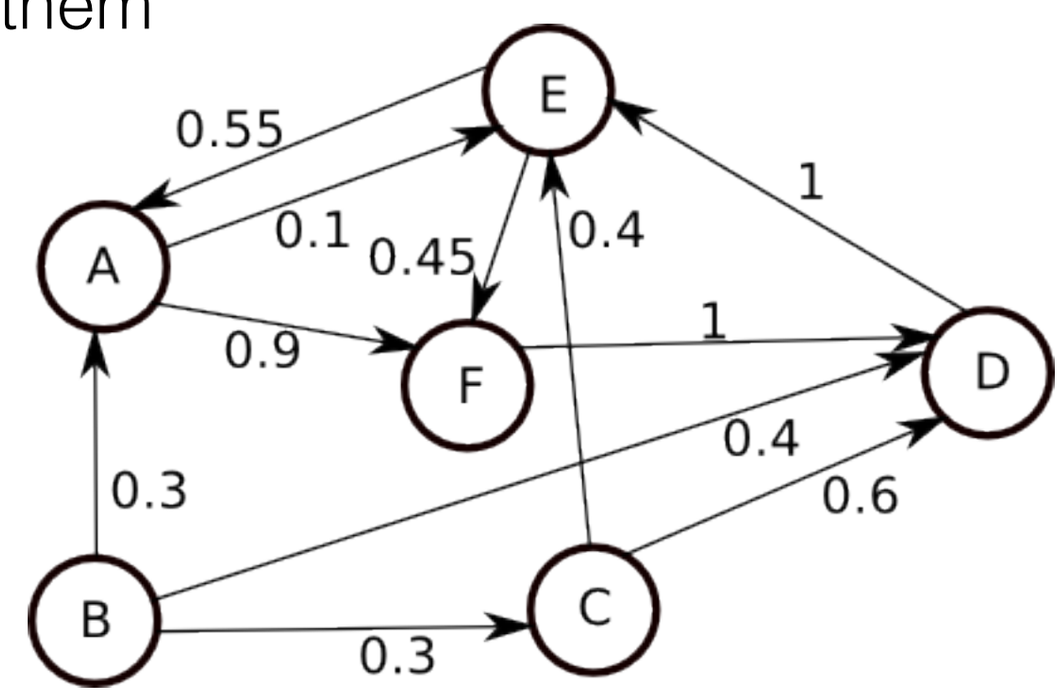


  - Allow two or more edges to connect to the same vertex
  - Allows for loops



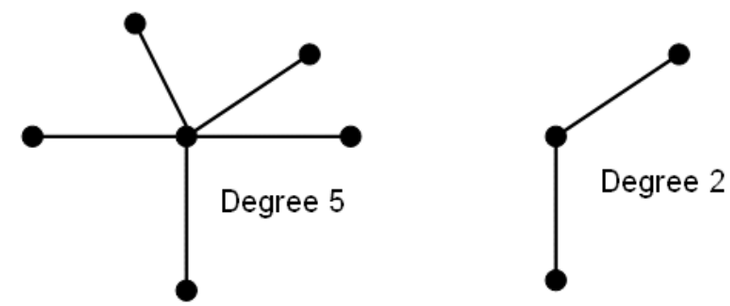
Properties of a graph

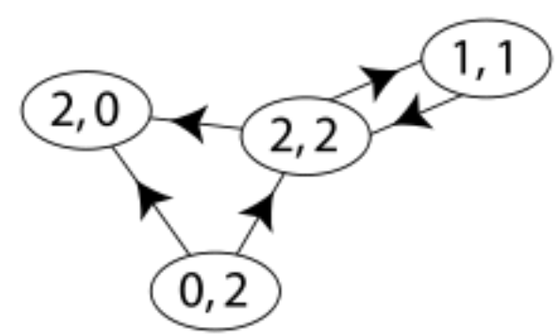
- Quiver
  - Directed multigraph
- Weighted graph
  - Edges have weights assigned to them



Properties of a network

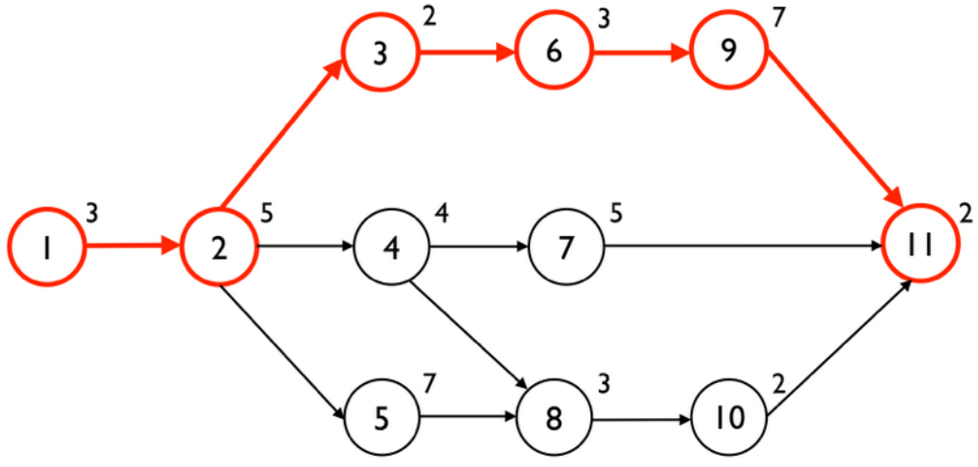
- The degree of a vertex / node
  - The number of other nodes connected to it by and edge
  - The number of neighbours
  - In-degree and out-degree (the degree is the combination of these)

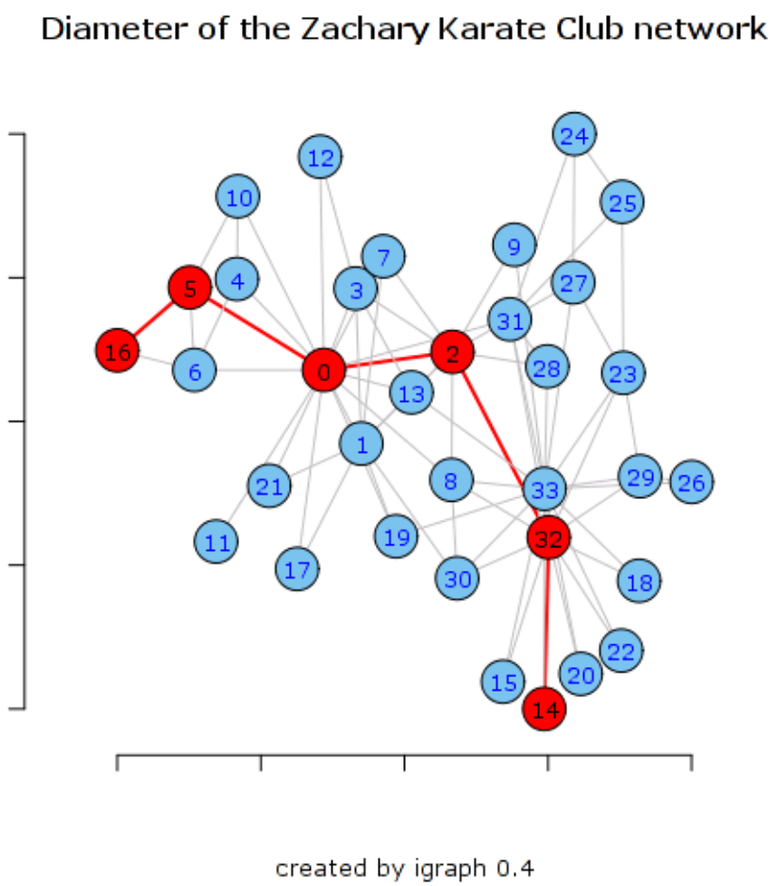




Properties of a network

- Network paths:
  - Series of steps from node to node along an edge
- The distance between two nodes
  - The length of shortest path between them
- The diameter of a network
  - Average distance between pairs of nodes
  - Gives an idea of how easy it is for information to move in the network

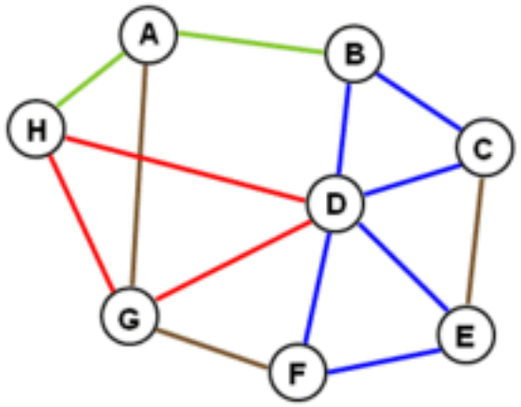




Properties of a network

- Paths
  - Eulerian path:
    - A trail in a graph that visits each **edge** exactly once
  - Hamiltonian path:
    - A trail in a graph that visits each **vertex** exactly once
  - De Bruijn graph
    - Both Eulerian and Hamiltonian
    - Used in genome assembly

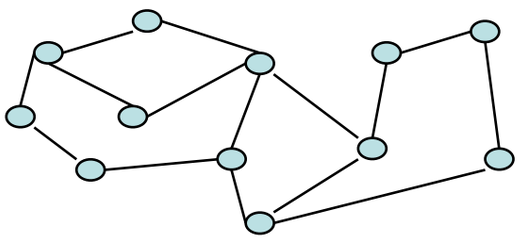
Properties of a network



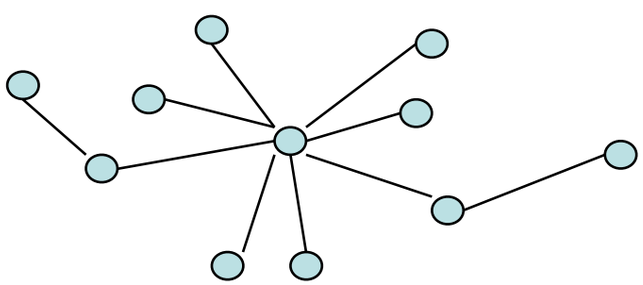
- Cycles
  - Closed walk
    - A sequence of verities that describes a path, moving from one vertex to the next along an edge, that returns to the original vertex
    - (F, D, B, C, D, E, F)
    - In directed graphs, the orientation of the edges must be respected
  - Simple cycle
    - Closed walk with no repetition of edges or vertices
    - (H, D, G, H)

Properties of a network

- Network types:
  - Random network model –random connections between nodes



  - Scale-free hierarchical model – most nodes have few connections and some have many, e.g. regulatory networks



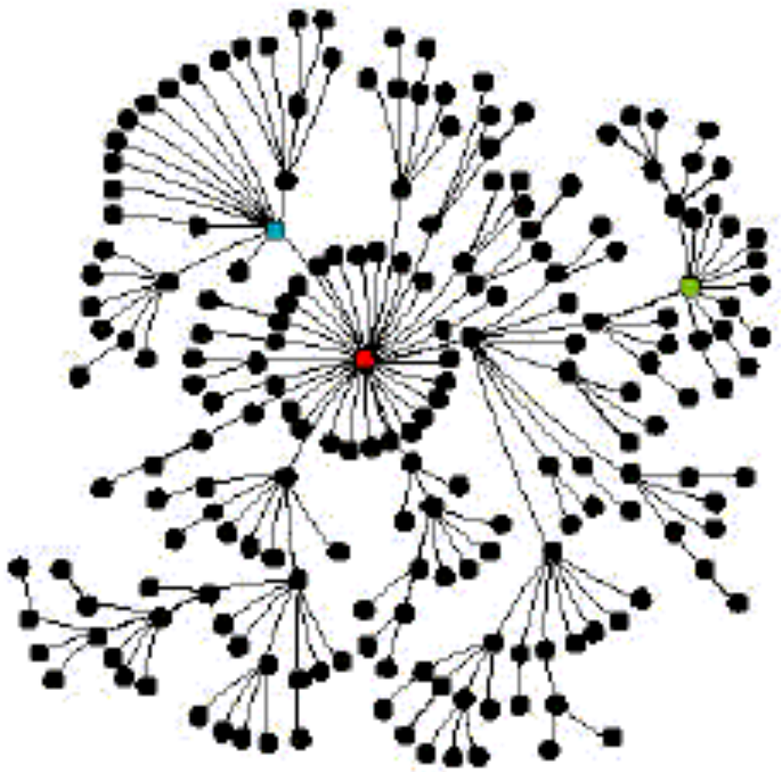
Properties of a network

- Also known as network topology
- Can be used to identify features of networks
  - “Important” nodes / edges
  - Communities within the network
  - Robustness or vulnerability of a network

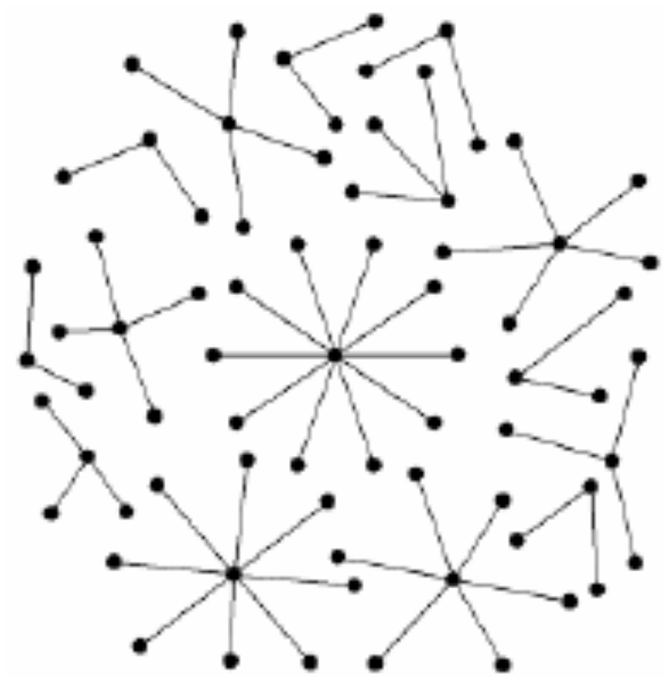
Properties of a network

- Centrality
  - Used to identify “important” vertices
  - Characterised by centrality indices
  - Network walk structures
  - Network flows

Network centrality



Different points of centrality but each is connected

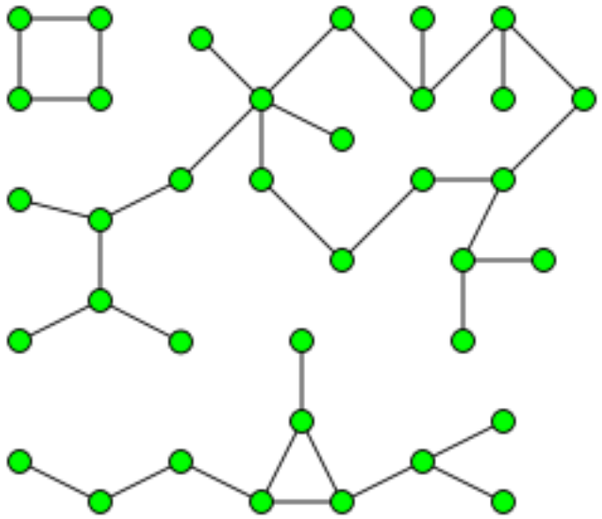


Different points of centrality, not necessarily connected

Degree distribution says nothing about connectivity

Properties of a network

- A graph is made up of components
  - This graph has 3 connected components
  - Normally refers to a undirected graph

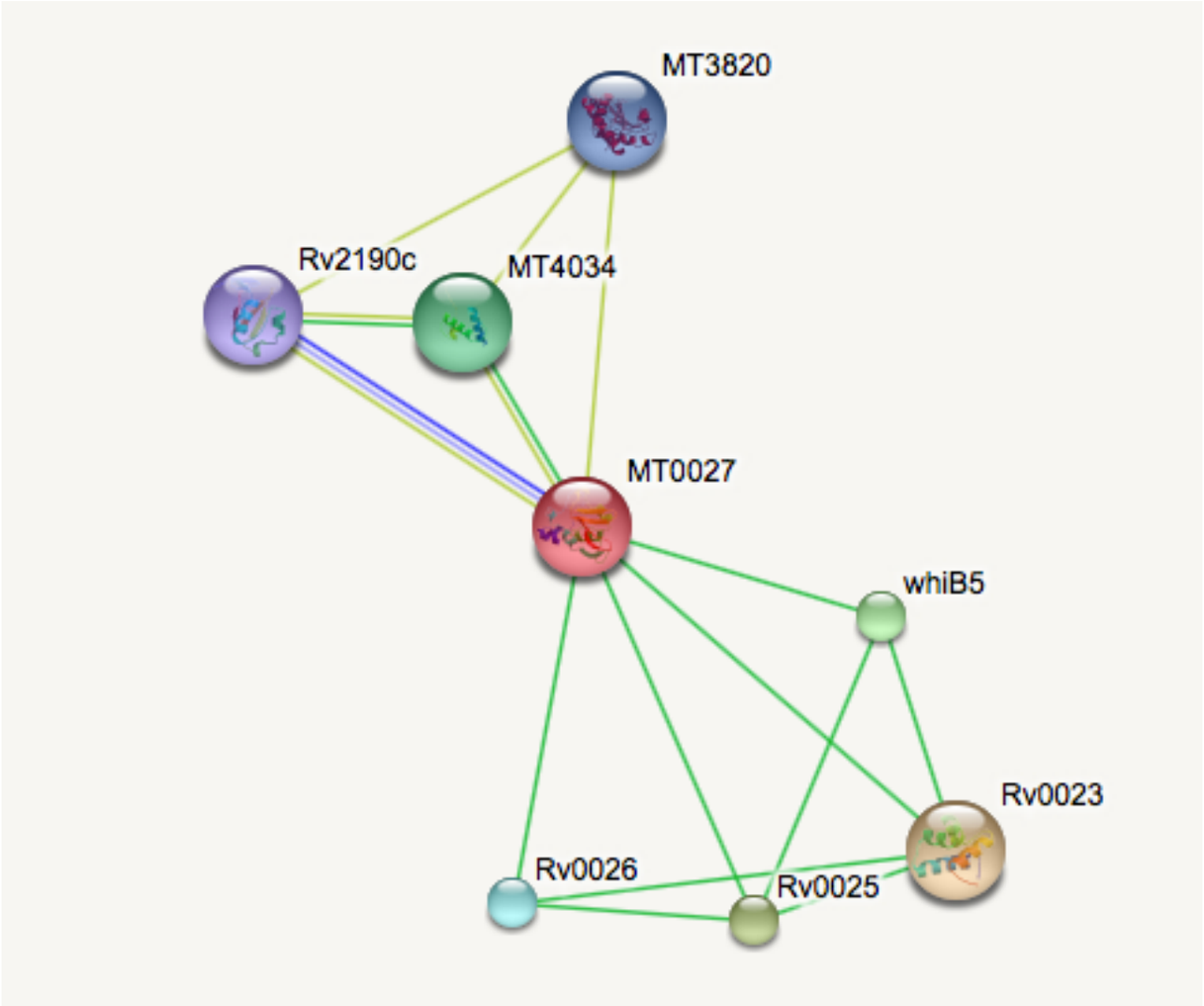


Protein-protein interaction networks

- Structures that represent the interaction between proteins
    - Refer to physical interactions
- Experiments to identify physical interactions between DNA and proteins or between two proteins:
    - Yeast two hybrid
    - Co-IP precipitation
    - Protein arrays
- Protein-protein interaction databases:
    - IntAct
    - DIP (Database of Interacting Proteins)
    - BIND (Biomolecular Interaction Network Database)

Protein-protein interaction networks

- STRING database
- ‘Search Tool for the Retrieval of Interacting Genes/Proteins’
- Data is fully pre-computed
- Includes known and predicted interactions
- Integrated scoring scheme with evidence providing confidence
- Confidence score is associated with each data set, benchmarked using KEGG data



Protein-protein interaction networks

Edges:

Edges represent protein-protein associations

associations are meant to be specific and meaningful, i.e. proteins jointly contribute to a shared function; this does not necessarily mean they are physically binding each other.

Known Interactions

- *from curated databases*
- *experimentally determined*

Predicted Interactions

- *gene neighborhood*
- *gene fusions*
- *gene co-occurrence*

Others

- *textmining*
- *co-expression*
- *protein homology*

Your Input:

MT0027 *NLP/P60 family protein (281 aa)*

Predicted Functional Partners:

		Neighborhood	*Gene Fusion*	*Cooccurence*	*Coexpression*	*Experiments*	*Databases*	*Textmining*	*[Homology]*	***Score***
Rv0023	*transcriptional regulatory protein (256 aa)*	●								0.859
Rv0025	*hypothetical protein (120 aa)*	●								0.728
whiB5	*transcriptional regulatory protein whib-like WhiB5; A transcription factor that is probably redox- responsive. Probably pl...*	●								0.606
MT4034	*N-acetymuramyl-L-alanine amidase-related protein (406 aa)*							●		0.538
Rv0026	*hypothetical protein (448 aa)*	●								0.492
MT3820	*hypothetical protein (241 aa)*							●		0.459
Rv2190c	*hypothetical protein (385 aa)*			●				●	●	0.404

Your Current Organism:

Mycobacterium tuberculosis H37Ra

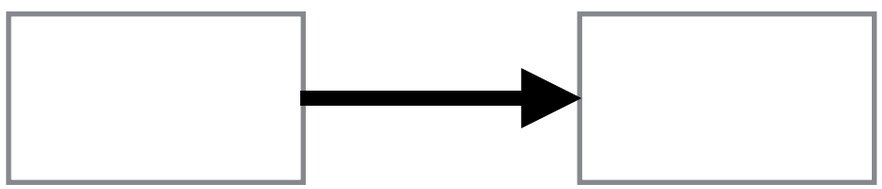
NCBI taxonomy Id: 419947

Other names: M. tuberculosis H37Ra, Mycobacterium tuberculosis ATCC 25177, Mycobacterium tuberculosis H37Ra, Mycobacterium tuberculosis str. H37Ra, Mycobacterium tuberculosis strain H37Ra

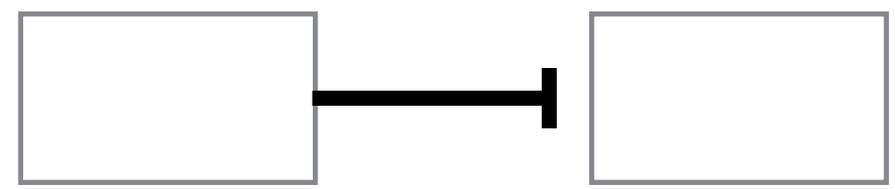
Metabolic network

- Specific symbolism involved

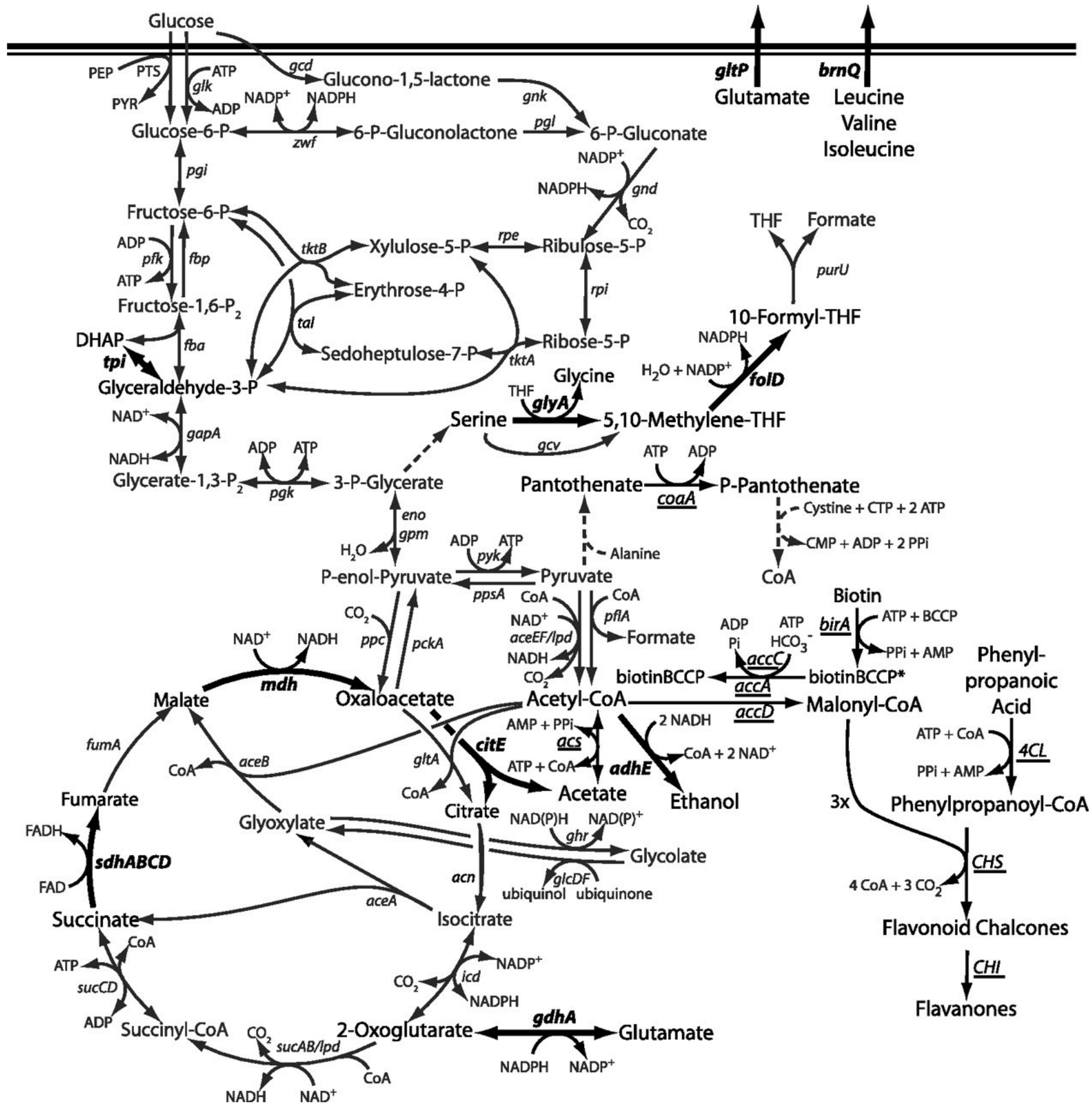
Activates



Inhibits



- Nodes referred to as metabolites

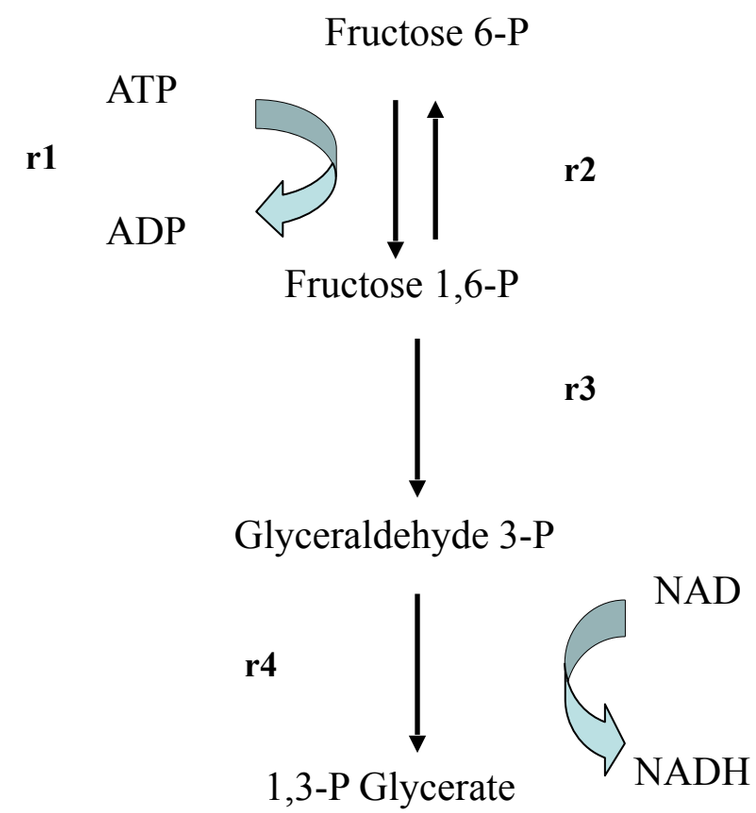

Glucose
PEP
PTS
PYR
ATP
glk
ADP
gcd
Glucono-1,5-lactone
NADP+
NADPH
gnk
gltP
brnQ
Glutamate
Leucine
Valine
Isoleucine
Glucose-6-P
zwf
6-P-Gluconolactone
pgl
6-P-Gluconate
pgi
NADP+
NADPH
gnd
CO2
Fructose-6-P
ADP
pfk
ATP
fbp
tktB
Xylulose-5-P
rpe
Ribulose-5-P
THF
Formate
purU
Erythrose-4-P
rpi
Fructose-1,6-P2
tal
10-Formyl-THF
DHAP
fba
Sedoheptulose-7-P
Ribose-5-P
NADPH
tpi
tktA
Glycine
H2O + NADP+
folD
Glyceraldehyde-3-P
THF
glyA
Serine
5,10-Methylene-THF
NAD+
gapA
gcv
NADH
ADP
ATP
ATP
ADP
Glycerate-1,3-P2
pgk
3-P-Glycerate
Pantothenate
coaA
P-Pantothenate
Cystine + CTP + 2 ATP
eno
gpm
CMP + ADP + 2 PPi
H2O
ADP
ATP
pyk
Alanine
CoA
P-enol-Pyruvate
Pyruvate
ppsA
CoA
CoA
Biotin
CO2
NAD+
pflA
ADP
ATP
birA
ATP + BCCP
NAD+
NADH
ppc
pckA
aceEF/lpd
Formate
Pi
HCO3-
PPi + AMP
NADH
accC
Phenyl-
propanoic
Acid
CO2
biotinBCCP
biotinBCCP*
mdh
accA
Malate
Oxaloacetate
Acetyl-CoA
Malonyl-CoA
accD
AMP + PPi
2 NADH
ATP + CoA
4CL
fumA
acs
citE
CoA + 2 NAD+
PPi + AMP
aceB
gltA
ATP + CoA
adhE
CoA
Acetate
Ethanol
3x
Phenylpropanoyl-CoA
Fumarate
CoA
Citrate
NAD(P)H
NAD(P)+
Glyoxylate
ghr
FADH
Glycolate
sdhABCD
CHS
FAD
acn
glcDF
4 CoA + 3 CO2
ubiquinol
ubiquinone
aceA
Succinate
Isocitrate
Flavonoid Chalcones
CoA
ATP
NADP+
sucCD
icd
CHI
CO2
NADPH
ADP
Succinyl-CoA
2-Oxoglutarate
gdhA
Glutamate
Flavanones
CO2
sucAB/lpd
CoA
NADPH
NADP+
NADH
NAD+

Metabolic network

- The network may:
  - Have compartments
  - Be used to simulate change in the system over time
- Nodes can have properties such as:
  - Concentration
- Links representing reactions may have:
  - Rates

Metabolic network

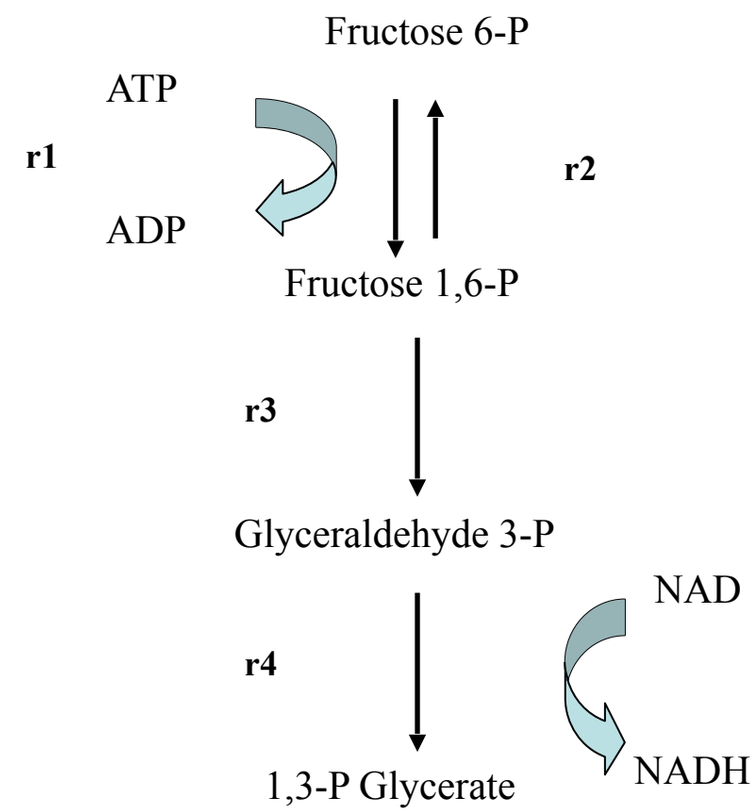
- Can be represented mathematically
- Stoichiometric matrix
- Where S is the vector of concentration values:
  - S = (S1, S2, S3, etc…)
- v is the vector of reaction rates
  - v = (v1, v2, v3, etc…)



Metabolic network

- Can be represented mathematically
- Stoichiometric matrix

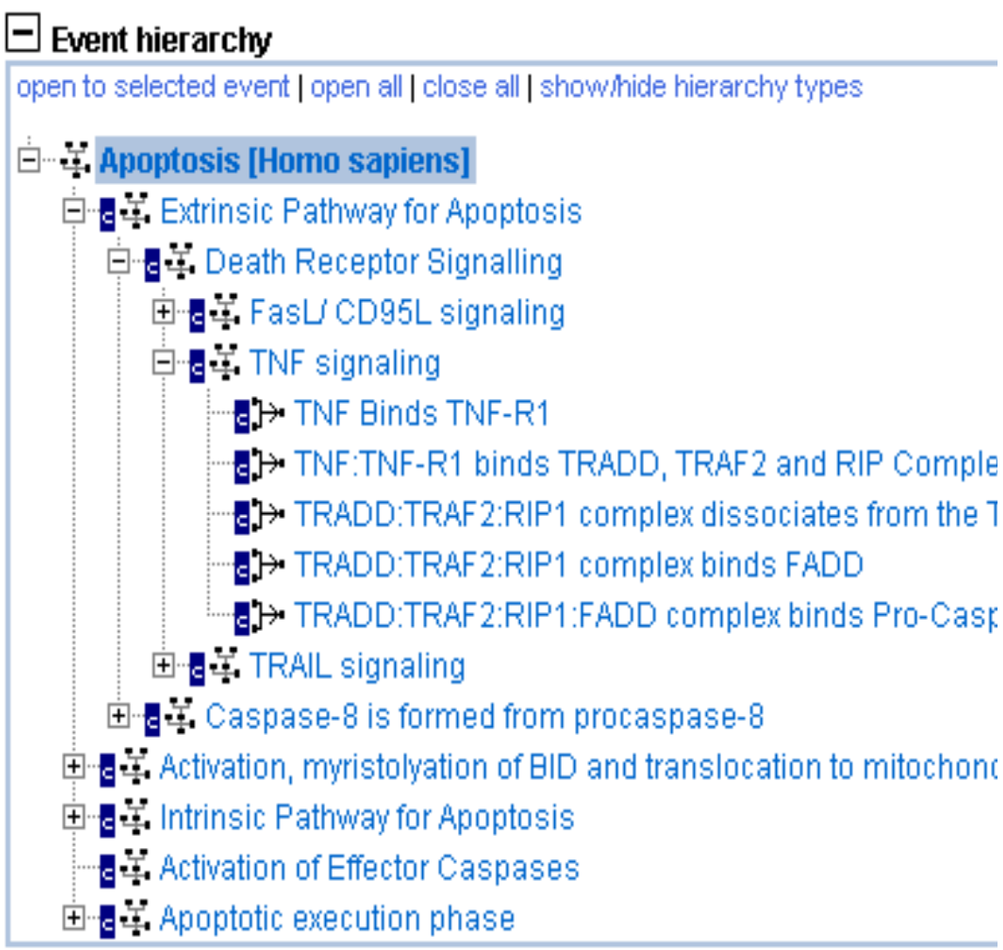
$$
\begin{array}{c}
\begin{array}{ccccc} & \text{F6P} & \text{F16P} & \text{G3P} & \text{13PG} \end{array} \\
\begin{pmatrix}
\text{r1} & -1 & 1 & 0 & 0 \\
\text{r2} & 1 & -1 & 0 & 0 \\
\text{r3} & 0 & -1 & 1 & 0 \\
\text{r4} & 0 & 0 & -1 & 1
\end{pmatrix}
\end{array}
$$

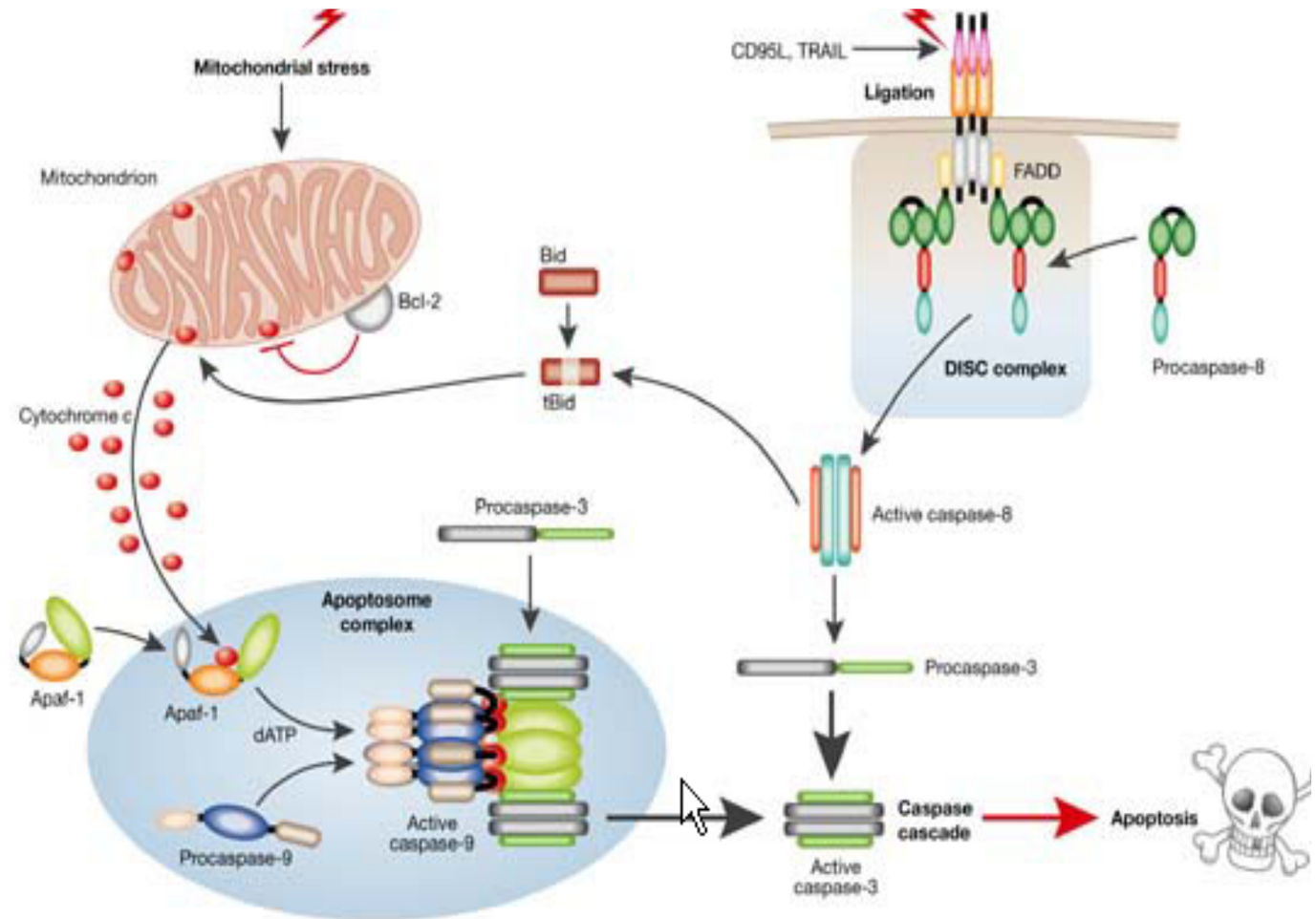


Pathway databases

- PATHGUIDE >200 pathway databases: http://www.pathguide.org
- KEGG
  - specific coverage of metabolism, some other networks too (e.g. regulatory)
  - Well-curated and quite specific
- MetaCyc, EcoCyc, BioCyc etc.
- Reactome –higher eukaryotes, manually curated
- GenMAPP –pathways contributed by users

Pathway example in Reactome

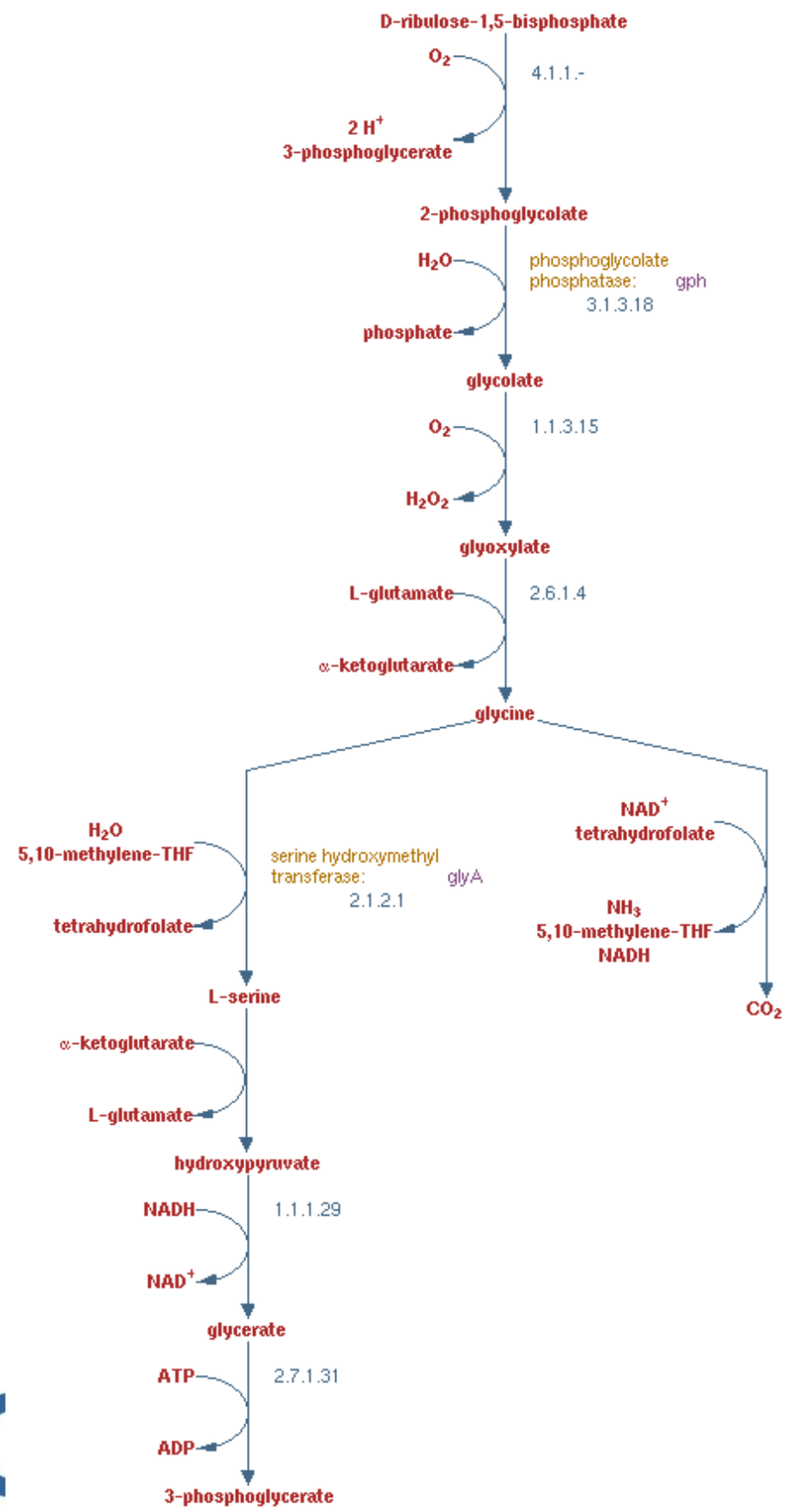




Apoptosis and disease: a life or death decision.
EMBO Rep. 2004 Jul;5(7):674-8. Epub 2004 Jun

A. aeolicus Pathway: photorespiration

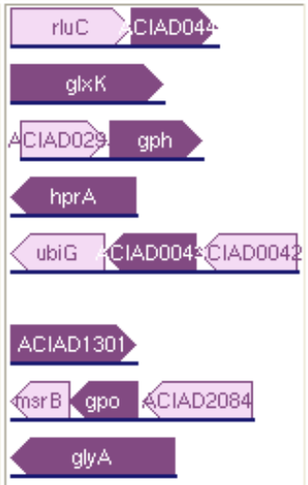
More Detail | Less Detail | Cross-Species Comparison | Download Genes | BioPAX format



Pathway in BioCyc

- http://www.bioyc.org
- Set of curated pathways for many organisms
- Can overlay 'omics' data, e.g. gene expression
- Can add GO terms or other annotation

Comparison of pathway across 2 organisms

Organism	Evidence Glyph	Enzymes and Genes for photorespiration		Operons
A. sp ADP1		EC# 4.1.1.-	None	
		EC# 3.1.3.18	putative phosphoglycolate phosphatase protein : ACIAD0443 phosphoglycolate phosphatase, contains a phophatase-like domain : gph putative phosphoglycolate phosphatase 2 : ACIAD0043	
		EC# 1.1.3.15	None	
		EC# 2.6.1.4	None	
		GCVMULTI-RXN	glutathione peroxidase / homoserine O-succinyltransferase / dolichyl-phosphate mannose synthase / O-succinylbenzoic acid-CoA ligase / 2-oxoglutarate decarboxylase / glutathione peroxidase 4 / dihydroxyacetone kinase / sorbitol-6-phosphate dehydrogenase / 2-dehydro-3-deoxygluconokinase / adenine deaminase / gcv system / trimethylamine N-oxide reductase III / fumarate reductase : gpo	
		EC# 2.1.2.1	serine hydroxymethyltransferase : glyA	
		RXN-974	None	
		EC# 1.1.1.29	putative glycerate dehydrogenase : ACIAD1301 glycerate dehydrogenase : hprA	
		EC# 2.7.1.31	glycerate kinase : glxK	
B. subtilis 168		**This pathway is not marked as present in this organism.**		glyA
		EC# 4.1.1.-	None	
		EC# 3.1.3.18	None	
		EC# 1.1.3.15	None	
		EC# 2.6.1.4	None	
		GCVMULTI-RXN	None	
		EC# 2.1.2.1	serine hydroxymethyltransferase : glyA	
		RXN-974	None	
		EC# 1.1.1.29	None	
		EC# 2.7.1.31	None	

Key to pathway evidence glyph edge colors:

- green: reactions for which a candidate enzyme has been identified in this organism
- black: reactions for which a candidate enzyme has not been identified in this organism
- orange: reactions which do not appear, or whose enzyme does not appear in any other pathway in this database
- magenta: reactions that are spontaneous, or edges that do not represent reactions at all (e.g. in polymerization pathways)

Change Organisms

Incomplete pathways -finding the pieces

- To completely model metabolic pathways you need all the enzymes and what reactions they catalyze
- From a genome –get list of ORFs, assign enzyme functions by sequence similarity
- Check that enzymes balance out –input metabolites = output metabolites
- Look for the missing ones

Pathway tools

- Software available for finding pathways in a whole genome and creating pathway/genome databases (PGDB) –integrates genome data with functional annotations
- Predicts enzymes first based on GenBank annotation then sequence similarity
- http://bioinformatics.ai.sri.com/ptools/
- Allows you to compare pathways across organisms

Assembling nodes

Genome –set of genes

Gene1
Gene2
Gene3
Gene4
Gene5
Gene6

→

Enzyme sets

EC 1.1.1.3
EC 1.2.3.1
EC 1.1.2.6
EC 2.1.2.1
……

→

Reactions

D-glucose + ATP = …
Fructose + ADP =….
……
……

→

Metabolic network

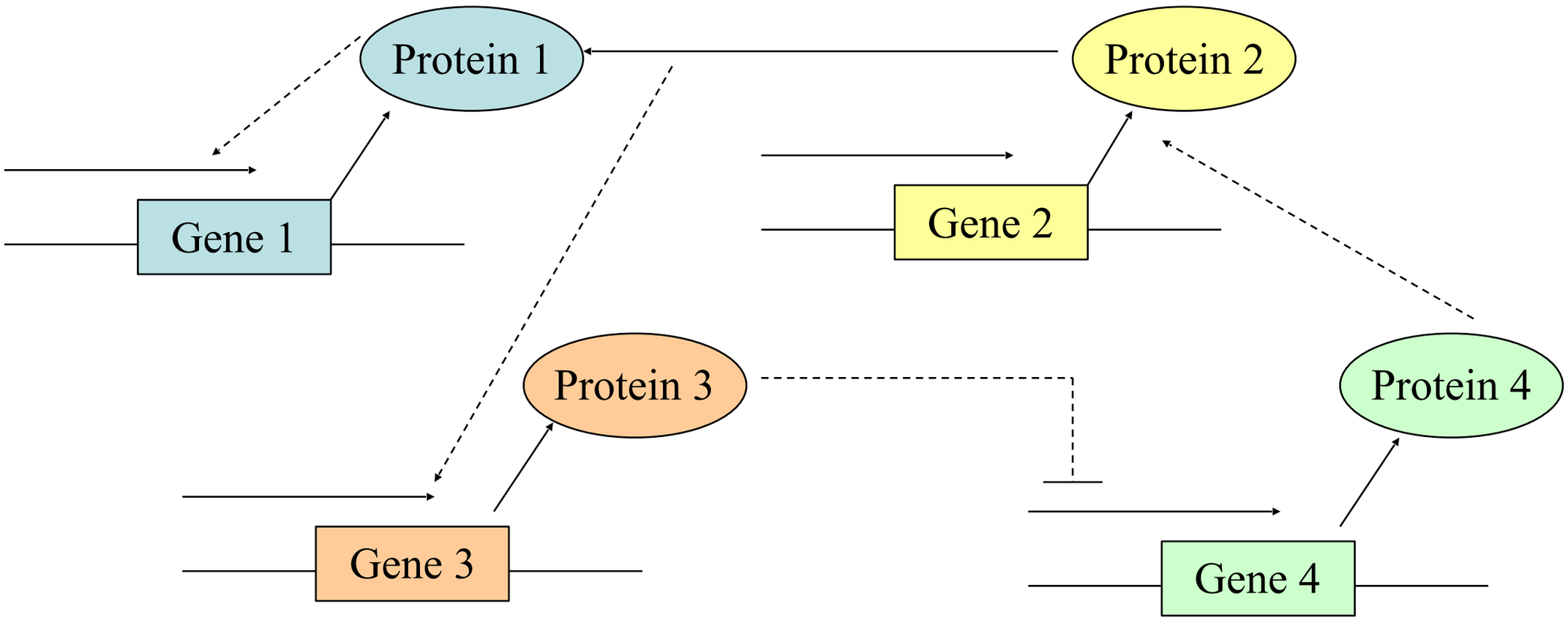
Applications of metabolic networks

- Biomarker discovery –identifying which metabolites are indicative of disease
  - Derive metabolic networks for the complete metabolome
  - Get healthy and diseased cells and do metabolic profiling –experimental procedure to identify and measure all metabolites
  - Generate a model to relate metabolite data with disease state
  - Look for statistical differences in metabolites for healthy *vs* diseased

Gene regulatory networks

- Set of DNA fragments that interact with each other as well as other components, e.g. proteins, to regulate gene expression
- Input of network is genes or proteins, output is gene expression
- Includes additional factors such as external signalling, feedback loops etc.

An example network

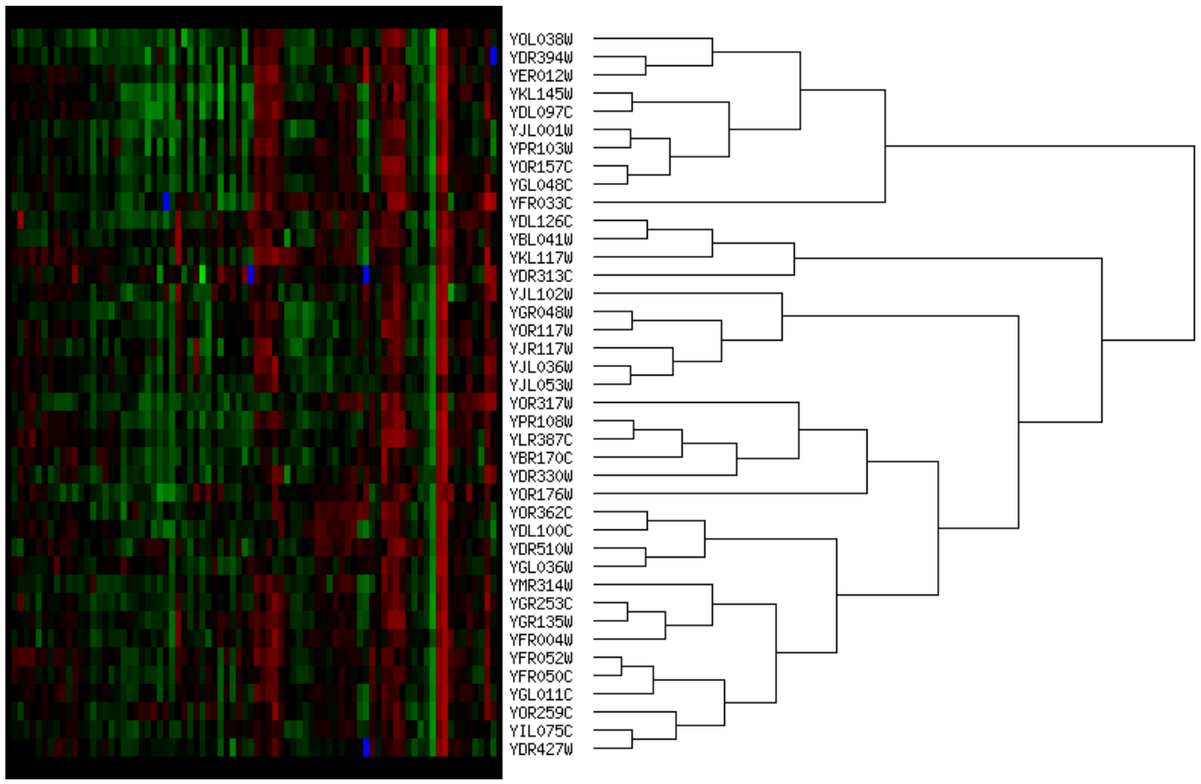


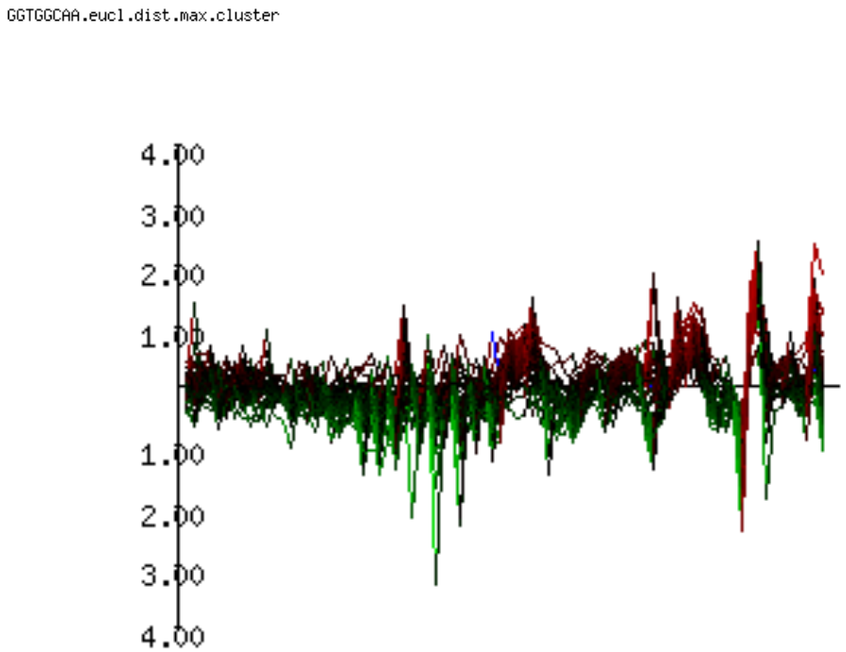
Protein 1 regulates itself and complexes with protein 2 to regulate expression of gene 3.

Protein 3 inhibits gene 4 expression and protein 4 activates translation from gene 2 to protein 2

Where is the data from?

- Gene expression experiments, e.g. microarrays



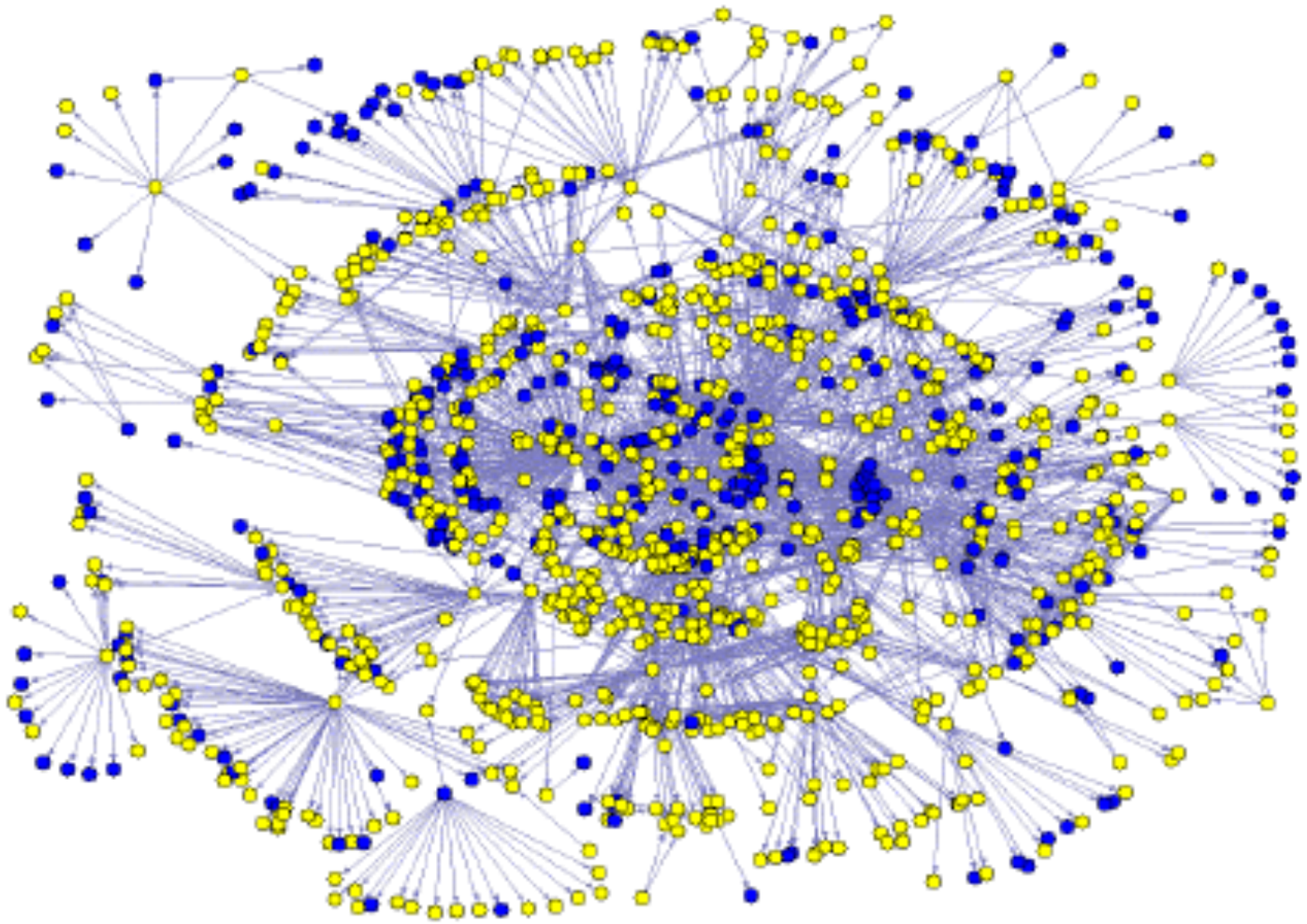


Genetic interactions

- **Mutants** –Knock out a gene and see response
- **Synthetic interaction** - delete gene A, keep B, get wild-type and *vice versa,* delete both, if non-WT – A & B have synthetic relationship

Example of *E. coli* gene regulatory network

1278 genes
2724 interactions
157 genes for TFs
382 metabolic enzyme genes (Blue)

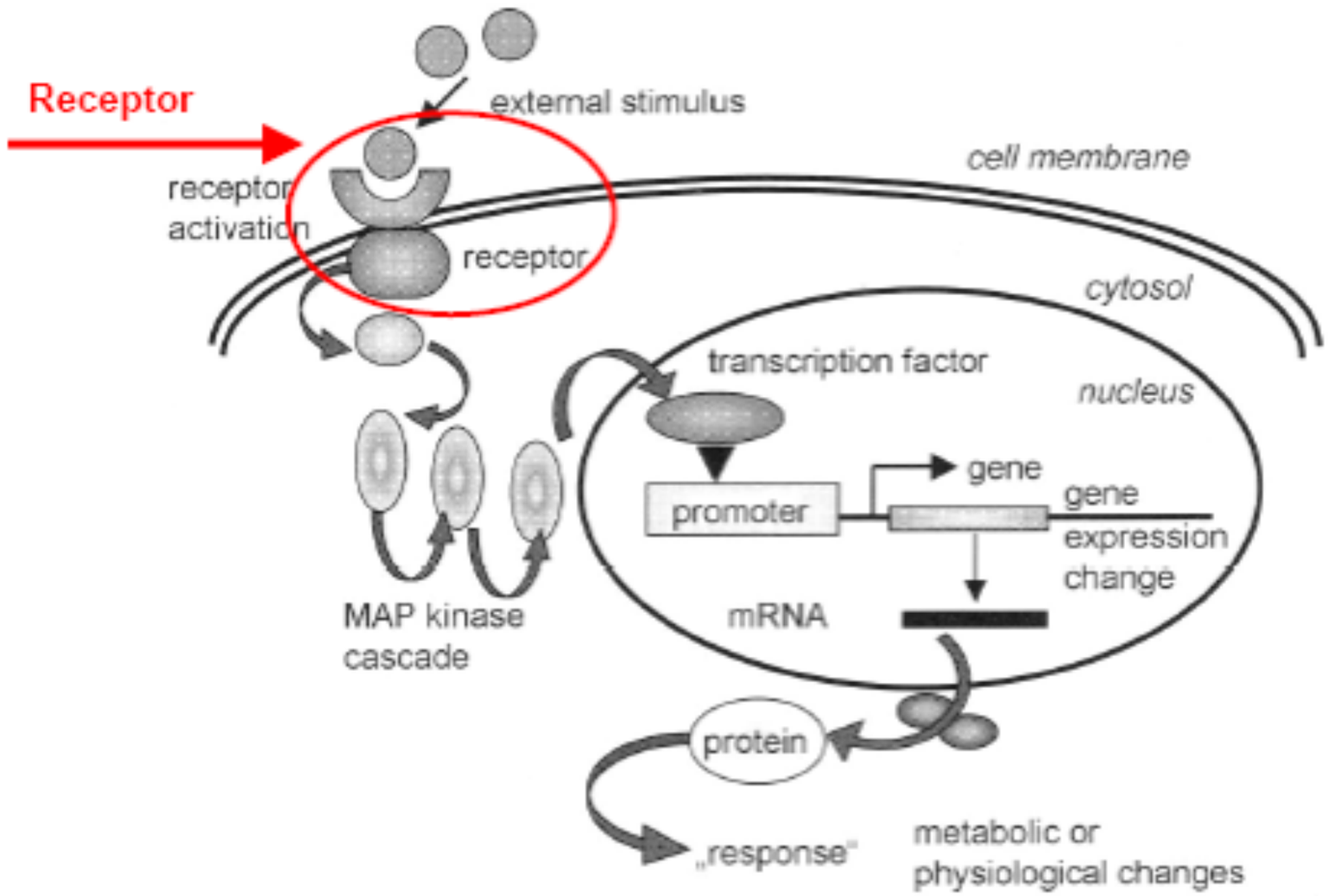


Ma et al. (2004) Nucleic Acid Research 32, 6643

Signalling networks

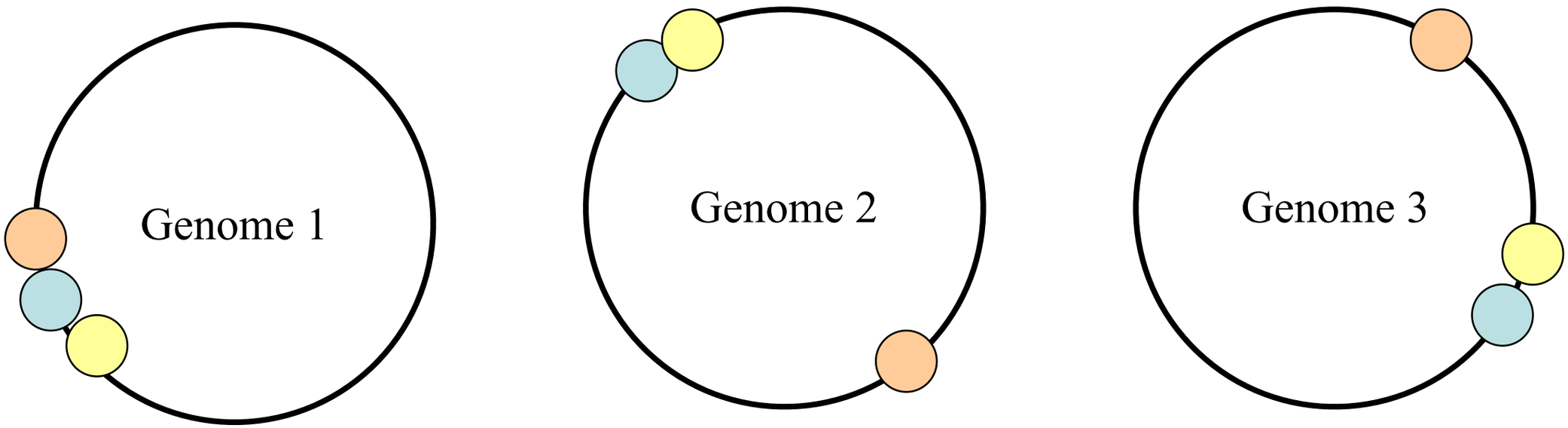
- Signal transduction is a response to an external signal
- Usually involves receptor on cell surface to identify signals then internal proteins that respond to the signals
- Signal transduction is usually by modification of proteins –phosphorylation
- Whereas metabolism provides mass transfer, signalling provides information transfer

Signal cascade



E. Kipp, Systems Biology in practice

Gene location



Two genes always found together on different genomes

Can infer functional linkage

Phylogenetic profiles

Protein	E. coli	S. aureus	H. pylori	Y. pestis
P1	1	0	0	1
P2	1	1	1	0
P3	0	1	1	0
P4	1	0	0	1
P5	0	1	1	0
P6	1	0	0	1
P7	0	1	0	1

P1	1	0	0	1
P4	1	0	0	1
P6	1	0	0	1

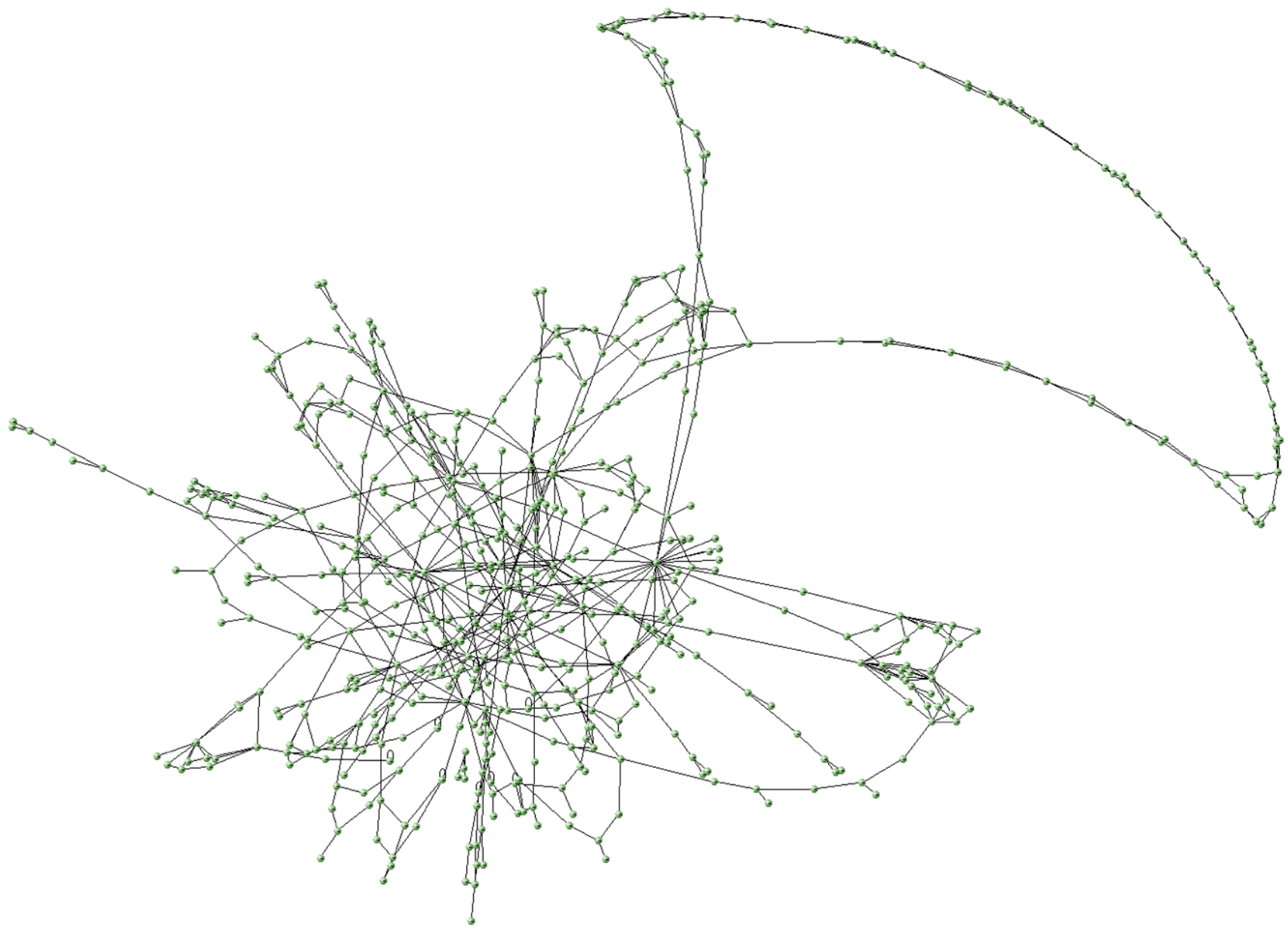
P2	1	1	1	0

P3	0	1	1	0
P5	0	1	1	0

P7	0	1	0	1

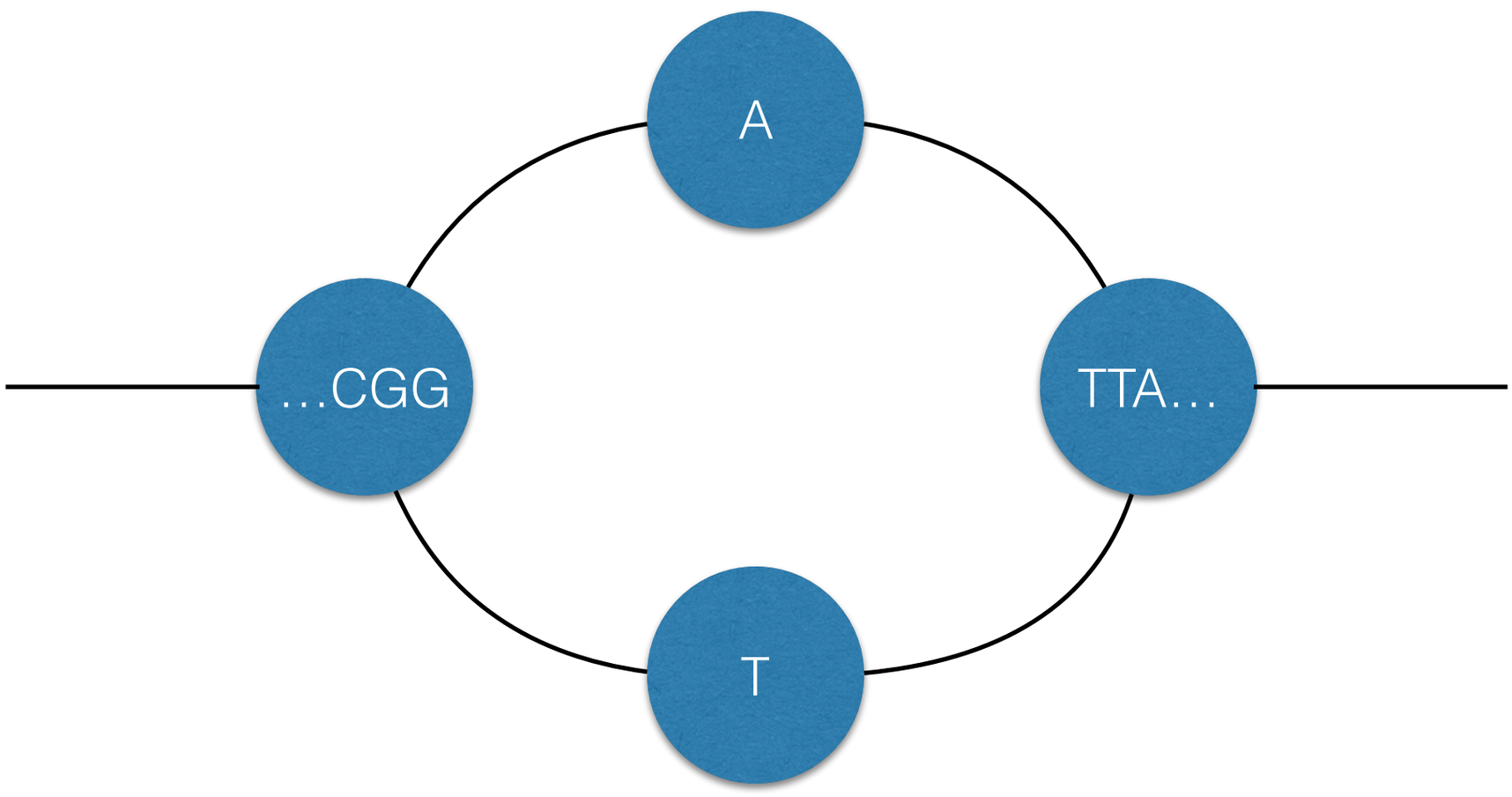
Can predict that P1, P4, P6 and P3, P5 are functionally linked

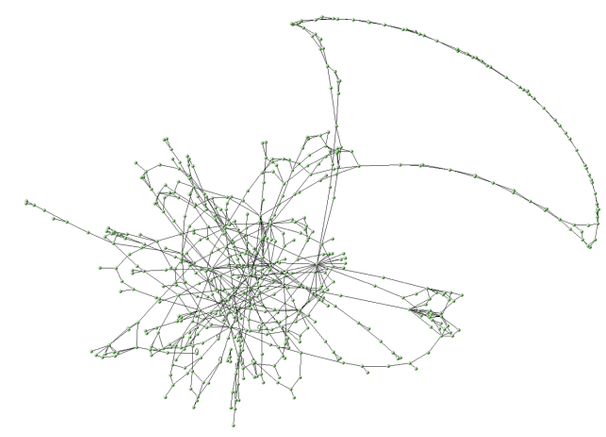
Genome graphs



Genome graphs

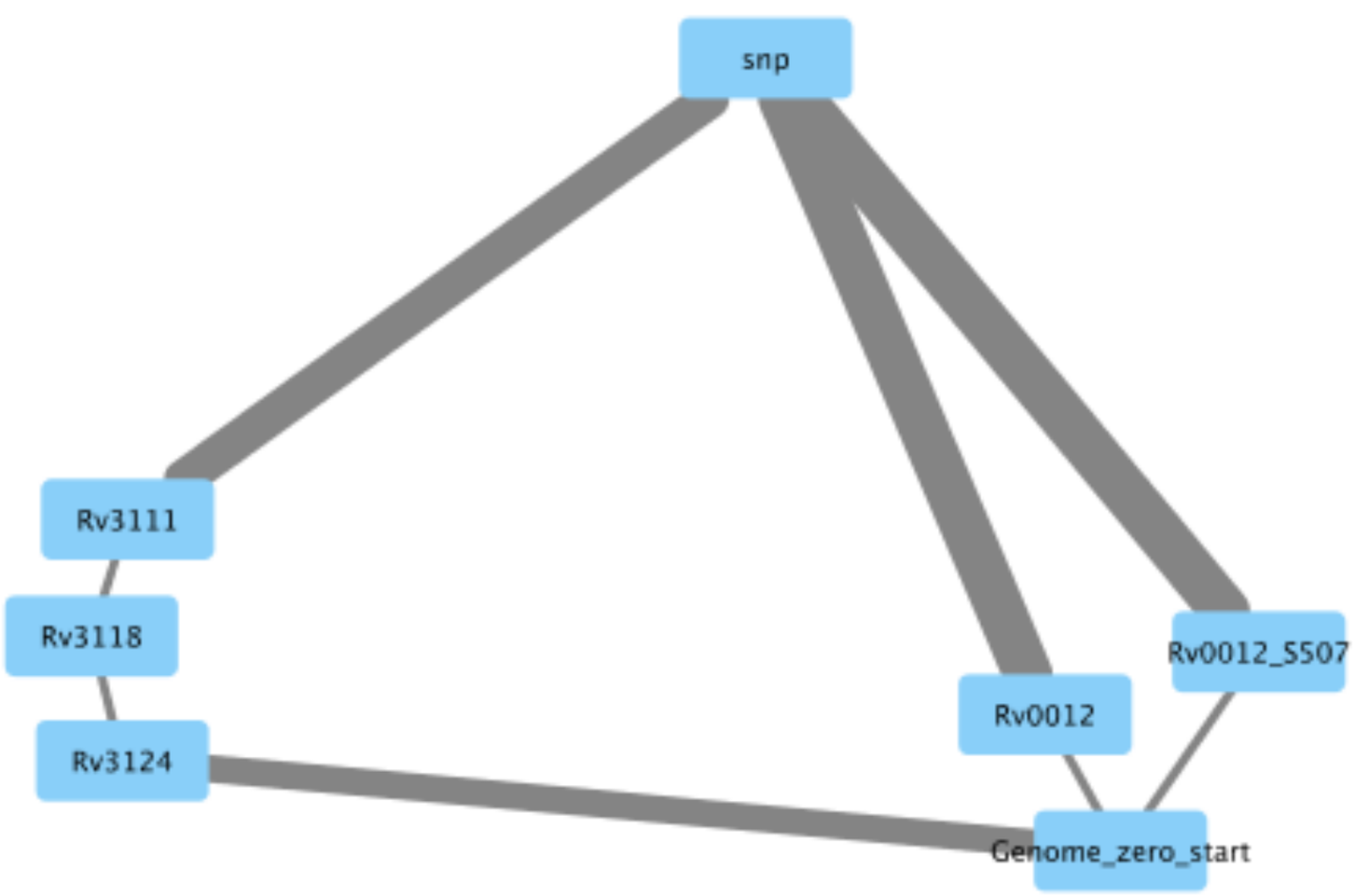
- Allows multiple genomes to be represented in one structure

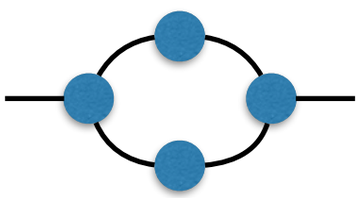


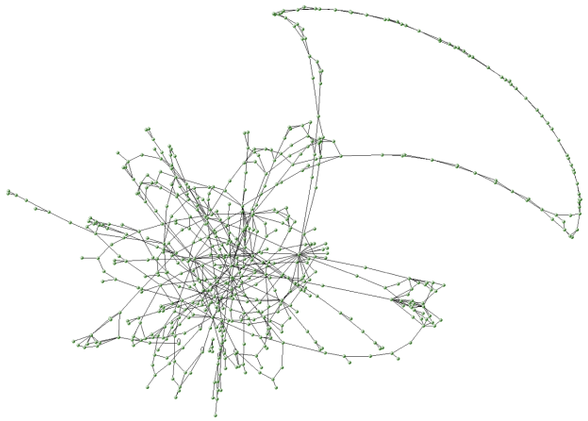


Genome graphs

- Allows multiple genomes to be represented in one structure





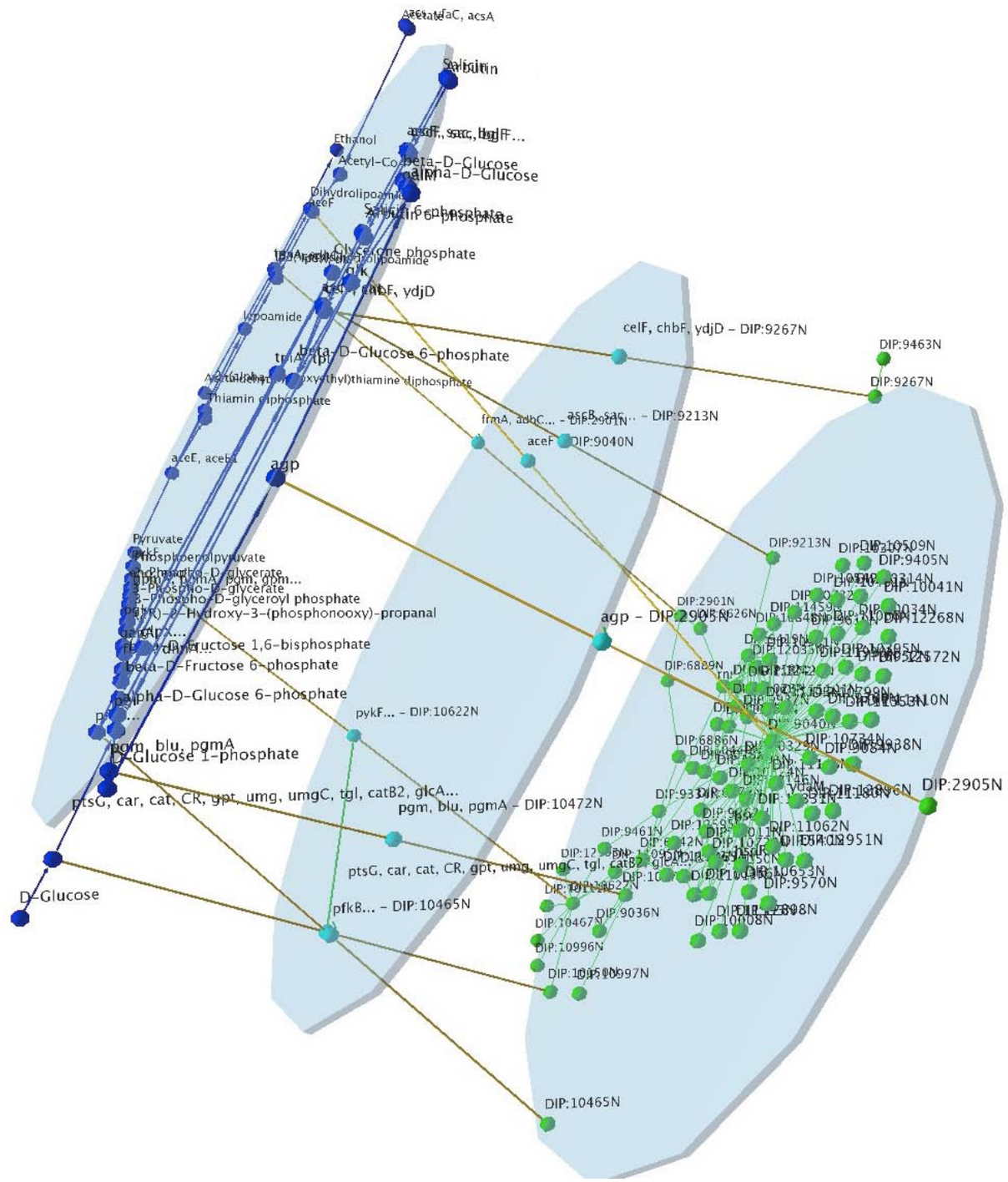


Predicting interactions from literature

- Text mining of literature abstracts
- Mechanical Turking
- Look for co-occurrences of genes/proteins in same text
- Assume a functional relationship
- Issues with gene/protein naming

Data integration

- Overlay PPI networks, metabolic networks, genome networks, etc…



Issues with data integration

- Data is noisy
- Datasets are incomplete
- Data is heterogeneous
- Some data is more trustworthy than others
- Data comes in different formats
- Need some ways of evaluating accuracy of integration

Ranking data for integration

- Some data is more likely to give false positives
- Can assign more weight to evidence from some experimental data types and less to others
- This ensures most trustworthy data scores highest
- Evaluating network –GO annotation (gold standard)

Data integration

- Visualisation and manipulation tools:
  - Cytoscape
  - PINV
  - NetworkX

Conclusions

- Relational way to structure data
- Networks can identify important nodes (proteins / genes / bottlenecks)
- Can be used for analysis of functional genomics data
- Many additional factors to consider in data integration, e.g. scoring, noise, etc.