The structure of molecular & cellular networks

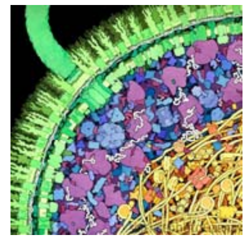
To be able to construct and analyze a cellular network, we need to clearly define what we identify as a node and what we represent with an edge.

The nodes and edges have to be at least similar to each other, e.g. represent the same type of cellular component (protein, chemical) or the same type of interaction (mass transfer, regulation).

We can, and often need to, define different types of nodes and edges.

Life at the cellular level

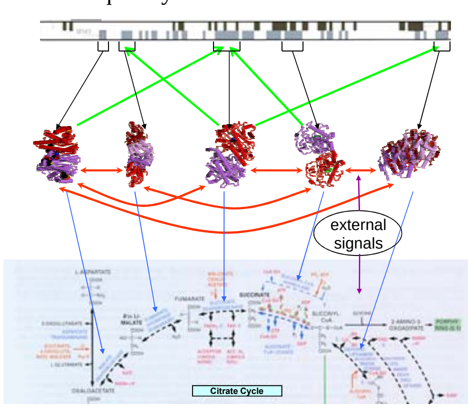
- Cellular functions rely on the coordinated action of interacting components.
- Proteins
  - provide structure to cells and tissues
  - work as molecular motors
  - sense chemicals in the environment
  - drive chemical reactions
  - regulate gene expression
- Interconnections between components are the essence of a living process.



receptor proteins, enzymes, ribosomes, DNA

David Goodsell/ Science Photo Library

Frequently defined molecular interaction networks



gene regulatory network

protein-protein interaction network

signal transduction network

bio-chemical reaction network

Examples of intracellular networks

1. Protein interaction networks
   nodes: proteins
   edges: protein-protein interactions (binding), modification of a protein
2. Biochemical reaction networks
   nodes:
   - reactants (substrates) or products of the reactions
   - enzymes – catalyze the reactions
   - reactant-enzyme complex (“reaction node”)

   edges:
   - reactions
   - catalysis (regulation)

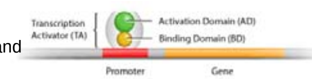
Examples of intracellular networks (cont.)

3. Gene regulatory networks
   nodes:
      gene, mRNA, protein
   edges:
      transcription, translation, regulation

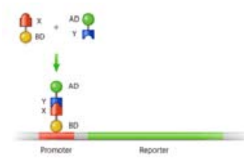
4. Signal transduction networks
   nodes:
      proteins, molecules
   edges:
      reactions and processes (e.g. ligand-receptor interaction)

Example of high-throughput experimental methods to map interactions

Transcription factors bind to the promoter regions of genes.
They have a DNA binding domain and an activation domain.



In the two-hybrid method the two domains are separated, and fused to two proteins.
If the two proteins interact by binding, the transcription factor activates the expression of a reporter gene.



Systematic experiments with all proteins in a given organism lead to genome-wide protein interaction maps.

Mapping of cellular interaction networks

Experimental advances allow the construction of genome-wide cellular interaction networks

- Protein networks:
  Individual studies:
  Uetz et al. 2000, Ito et al. 2001, Krogan et al. 2006, Yu et al. 2008 – *S. cerevisiae,*
  Giot et al. 2003 – *Drosophila melanogaster* , Li et al. 2004 – *C. elegans*
  Rual et al. 2005 - Human interactome

  High throughput methods:
  Co-affinity purification + mass spectrometry
  Yeast two hybrid

  Databases:
  Database of Interacting Protein (DIP), the Biomolecular Interaction Network (BIND), the Munich Information Center for Protein Sequences (MIPS), the Human Protein Reference Database (HPRD), and the Yeast Proteome Database (YPD)

Mapping of cellular interaction networks (cont.)

- Metabolic networks

  Experimental methods:
  Enzyme characterizations: Protein and DNA microarrays
  Metabolite identification: isotope labeling
  Flux quantification: Mass spectroscopy
  Databases: Kyoto Encyclopedia of Genes and Genomes (KEGG), Ecocyc, MetaCyc

- Transcriptional regulatory networks

  Individual studies: Shen-Orr et al. 2002 – *E. coli*, Guelzim et al 2002, Lee et al. 2002 - *S. cerevisiae,*
  Davidson *et al.* 2002 – sea urchin
  Experimental methods: DNA footprinting, chromatin immunoprecipitation (ChIP)
  Databases: Transcription Factor Database (TRANSFAC), Regulon Database (RegulonDB), KEGG

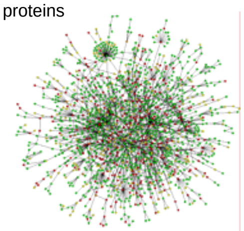
- Signal transduction networks
  Ma'ayan et al. 2005 – mammalian hippocampal neuron
  Databases: KEGG, Science STKE

Protein interaction maps now contain thousands of nodes and edges

Ito (yeast): 8868 interactions between 3280 proteins
Uetz (yeast): 4480 interactions, 2115 proteins
Giot (Drosophila): 4780 interactions among 4679 proteins
Li (C. elegans): 5534 interactions, 3024 proteins
Rual (human): 2800 interactions, 8300 proteins

- Although usually tested in a given bait/prey setting, protein interactions are considered symmetrical
- Many untested interactions – **problem**

- All networks have giant connected components.
- The topological properties of diverse protein interaction networks are similar.

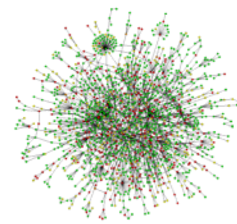


H. Jeong et al.Nature 411, 41-42 (2001)

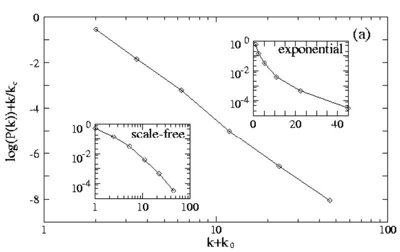
S.-H Yook, Z.N. Oltvai, A.-L. Barabasi, Proteomics 4, 928 (2004)

Exercise

- Which graph theoretical measures will be useful to analyze these networks?
- What information is not incorporated in these protein-protein interaction maps?



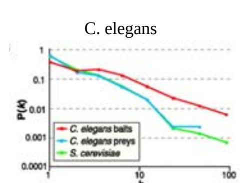
Degree distribution of the yeast protein network

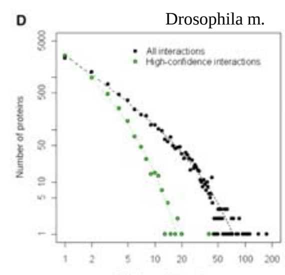


$$P(k) \sim (k+k_0)^{-\gamma} \exp\left(-\frac{k+k_0}{k_\tau}\right)$$

H. Jeong, S.P. Mason, A.-L. Barabasi, Z.N. Oltvai, Nature 411, 41-42 (2001)

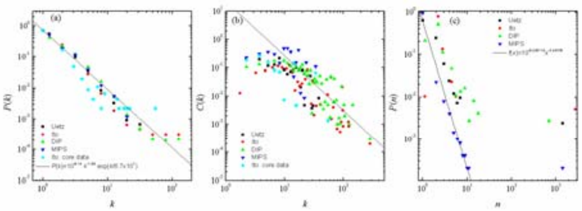
Degree distribution of *C. elegans* and *D. melanogaster* protein networks





$$P(k) = Ak^{-\gamma} \exp(-\beta k)$$

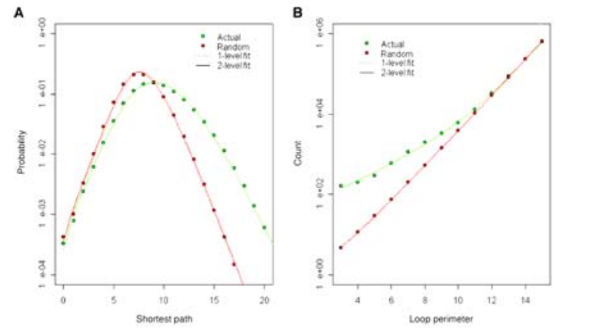
Comparison of yeast interaction networks



Degree distribution	Clustering coefficient	Connected components
$P(k) \sim k^{-2.5}$	$C(k) \sim k^{-2}$	$p(n) \sim n^{-3.4}$

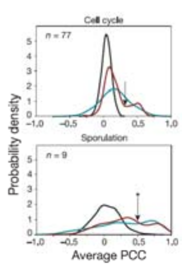
Yook, Oltvai and Barabási, Proteomics 4, 928 (2004)

Average path length larger, short cycles more abundant than in randomized networks



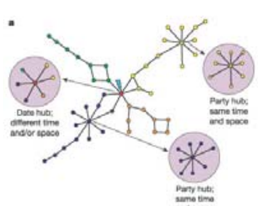
Randomization: swap the endpoints of two edges, node degrees stay the same.

Not all interactions are simultaneously active



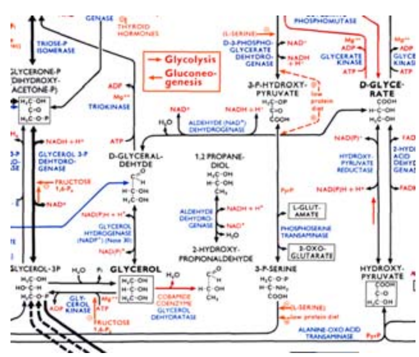
Calculate the correlation between the expression time-course of genes encoding the first neighbors of hub proteins.

Two peaks – two different types of hubs. **Party hubs** are inside connected modules that interact simultaneously. **Date hubs** connect different modules.



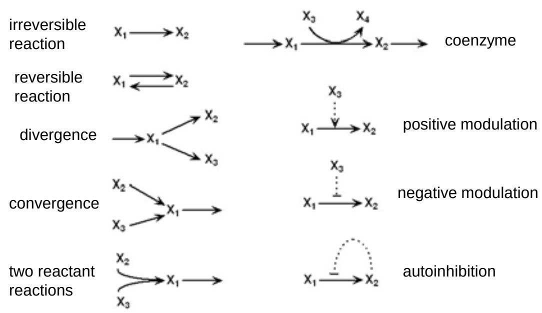
Han et al, Nature 443, 88 (2004)

Networks of chemical reactions –usual visualization



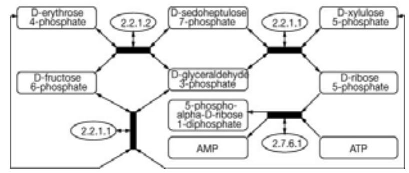
Enzymes shown in blue, co-enzymes (small molecules necessary for enzyme activity) in red. Double arrows mean reversible reactions. Reactants, products in black, box indicates that node appears in several locations.

Representation of chemical reactions+ regulation



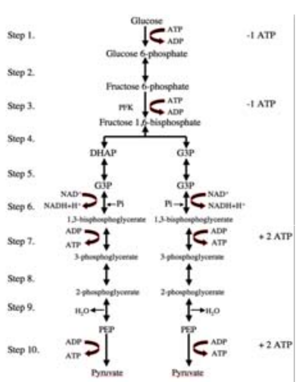
E. O. Voit, Computational Analysis of Biochemical Systems

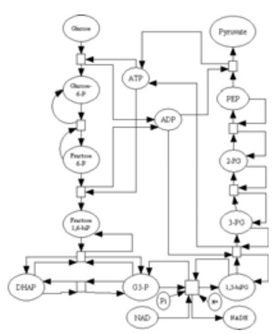
Tri-partite representation of metabolic network



- Node types:
  - Metabolites (substrates or products), open rectangles
  - No distinction between metabolites and coenzymes
  - Metabolite-enzyme complexes, black rectangles
  - Enzymes, open ovals
- Edges:
  - Substrate to complex or complex to product
  - Symmetrical edges between enzyme and complex

Ex. A traditional representation of the glycolysis pathway is given on the left. Draw a graph of the pathway.





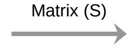
What improvements can be done to this graph?

Reaction Stoichiometry

Reaction Pathway

A + B → C + D (1)
A + D → E (2)
B + C → F (3)

Stoichiometric Matrix (S)



Reactions; rows: Reactants (Substrates/Metabolites)

	1	2	3
A	-1	-1	0
B	-1	0	-1
C	1	0	-1
D	1	-1	0
E	0	1	0
F	0	0	1

S_{ij} = Number of molecules of substrate i participating in reaction j

$S_{ij} < 0$ if substrate i is a reactant in reaction j

$S_{ij} > 0$ if substrate i is a product in reaction j

$i = 1,2,\ldots,N$ = # of substrates = # rows

$j = 1,2,\ldots,M$ = # of reactions = # columns

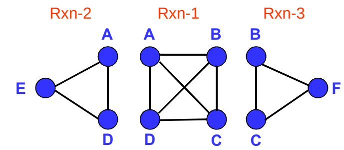
Ex. Represent these reactions by a bi-partite graph.

$A + B \rightarrow C + D$ (1)
$A + D \rightarrow E$ (2)
$B + C \rightarrow F$ (3)

Network Representation – Substrate Graph

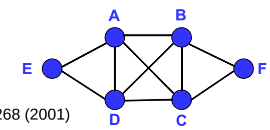
Reaction Pathway

$A + B \rightarrow C + D$ (1)
$A + D \rightarrow E$ (2)
$B + C \rightarrow F$ (3)



- One type of node: Substrate Node
- Un-directed edges
- Each reaction represented as a clique

Substrate Graph



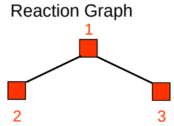
A. Wagner & D. Fell, Proc. Roy. Soc. 268 (2001)

Network Representation – Reaction Graph

Reaction Pathway

$A + B \rightarrow C + D$ (1)
$A + D \rightarrow E$ (2)
$B + C \rightarrow F$ (3)

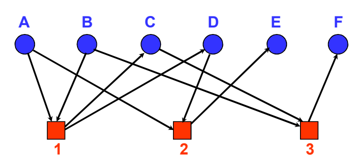
Reaction Graph



- One type of node: Reaction Node
- Un-directed edges
- An edge between two reactions if they share at least one substrate in common

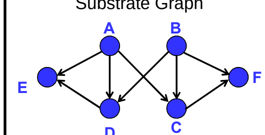
Three alternate network representations for the same reaction pathway !

Bi-partite Graph



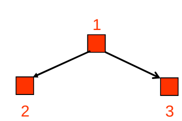
Derived

Directed Substrate Graph



Connect two substrates if there exists a 2-edge path in the bi-partite graph between them

Directed Reaction Graph



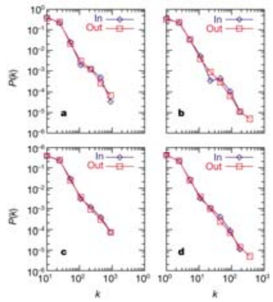
Connect two reactions if there exists at least one 2-edge path in the bi-partite graph between them

Key Properties of Metabolic Networks

- In- and out- degree distributions of substrate nodes in the bi-partite representation consistent with power-laws

$$P_{in}(k) \approx k^{-2.2}$$
$$P_{out}(k) \approx k^{-2.2}$$

- Existence of "hub" substrates such as ATP, ADP, NADP, NADPH... Most (but not all) are carrier metabolites.



a: *A. fulgidus* d: *C. elegans*

b: *E. coli* e: Average (43 organisms)

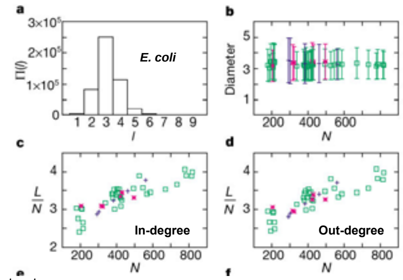
H. Jeong *et al.*, Nature 407, 651 (2000)

Distances in Metabolic Networks

Paths defined to connect substrates (reactants) to products, the average is calculated on the reachable pairs only.

Distance distribution

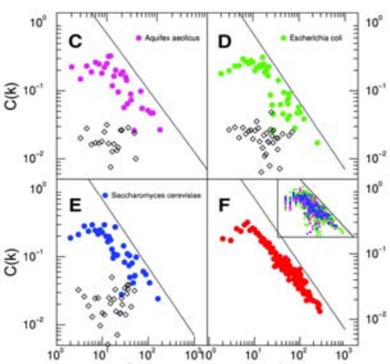
Average degree



Relatively small and constant network diameter across organisms

H. Jeong *et al.*, Nature 407, 651 (2000)

Clustering-degree relation in metabolic networks



Average clustering coefficient of nodes with degree k

Open symbols: a model with the same degree distribution

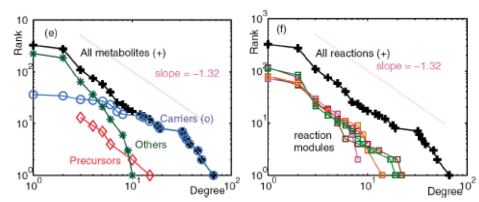
Straight line: $C(k) \sim k^{-1}$

Ravasz et al., Science 297, 1551 (2002)

Degree distributions in metabolite and reaction networks

Construct non-directed projections to metabolite and reaction networks

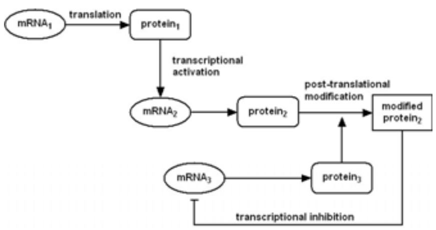
Rank vs. degree plot, similar to P(k>K).
The degree exponent γ=|slope|+1



Undirected substrate network   Undirected reaction network

Tanaka, Phys. Rev Lett. 94, 168101 (2005)

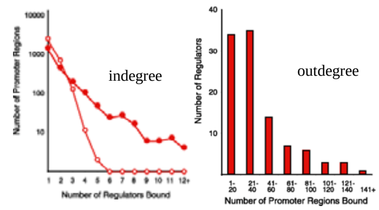
Gene regulatory networks

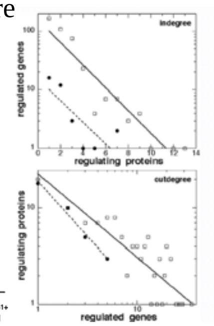


- nodes: mRNAs (ovals), proteins (boxes)
- edges: interaction or regulation
  - regulatory edges acting on edges – similar to catalysis
  - edges can be activating or inhibiting

Often-used simplification: merge all gene products into one.

Out-degree distribution long - tailed, in-degree distribution more limited





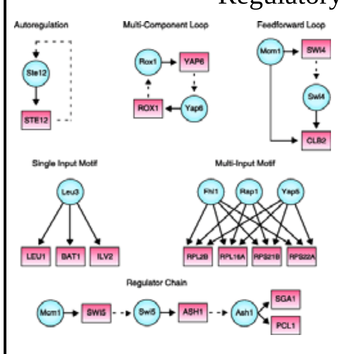
Guelzim et al, Nature Genetics 31, 60 (2002)
Lee et al, Science 298, 799 (2002)

S. cerevisiae

Other features of transcriptional regulatory networks

- No strongly connected component in E. coli and yeast - a unidirectional regulation mode.
- The subgraphs found by following the paths that start from non-transcriptionally regulated genes have relatively little overlap - distinct environmental signals tend to initiate distinct transcriptional responses.
- The source – sink distances are small in networks, e.g. the longest regulatory chain has only four (in E. coli) and five (in S. cerevisiae) edges.

Regulatory motifs



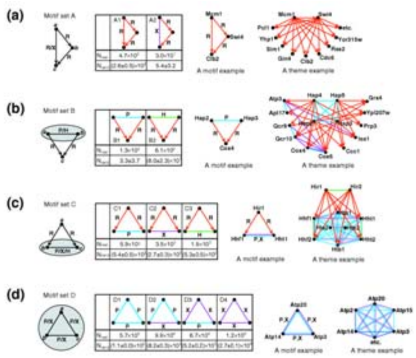
Regulators (TFs), blue circles
Genes, red rectangles
Dashed edges mean translation

Feedforward loop: convergent direct and indirect regulation; noise filter
Single input motif: one TF regulates several genes; temporal program
Multi-input motif: combinatorial regulation

Lee et al, Science 298, 799 (2002)

Regulatory themes

R: Transc. reg
P: Prot. interaction
H: Seq. homology
X: Correlated expression



Feed-forward

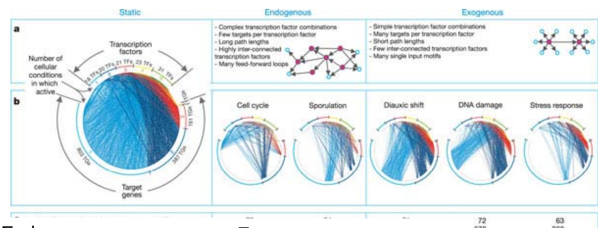
Co-pointing

Co-regulation

Protein complex

Zhang et al, J. Biol 4, 6 (2005)

Condition-dependent transcription sub-networks



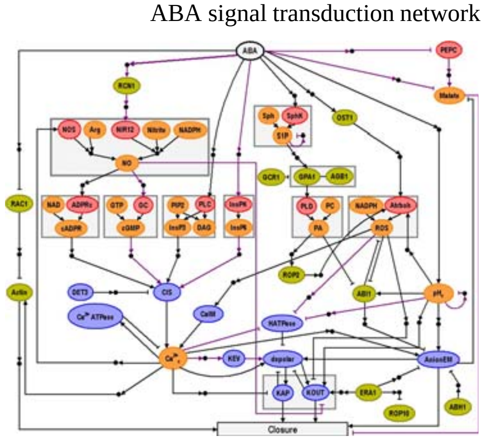
Endogenous	Exogenous
•Complex TF combination	•Simple TF combination
•Few targets per TF	•Many targets per TF
•Long path length	•Short path length
•Inter-connected TF	•Few inter-connected TF
•Many FFL	•Single input motifs

Luscombe et al, Nature 431, 308 (2004)

Ex. Draw a network corresponding to this verbal description of a signaling pathway.

- A protein ligand FASL binds to the receptor FAS. The interaction activates intracellular protein FADD, which in turn activates the proteolysis of procaspase-8, giving active caspase-8. Caspase-8 leads to the proteolytic activation of caspase-3, activating programmed cell death.

ABA signal transduction network



Red: enzymes
Blue: transport
Orange: small molecules
Green: sign. transd proteins
Black points: unknown intermediary nodes

Li, Assmann, Albert, PLoS Biology 2006

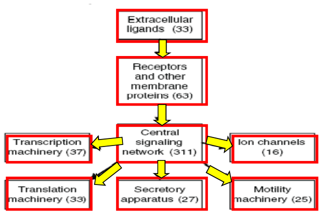
Signal transduction network of the hippocampal CA1 neuron

Data (binary interactions) collected form the experimental literature
System of interacting cellular components involved in phenotypic behavior

Edges can be directed or undirected (neutral)
Directed edges are activating or inhibitory

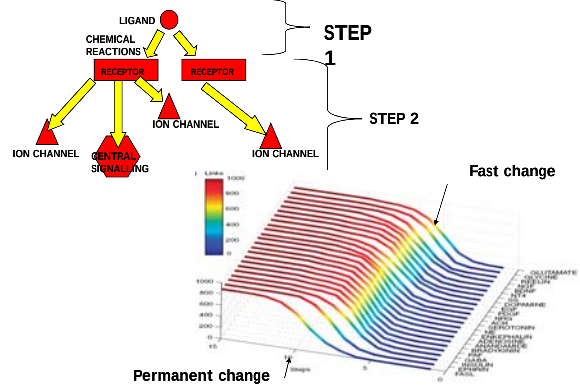
In and out degree distribution broad tailed

Highest degree nodes:
MAPK, CaMKII, PKA and PKC

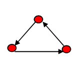
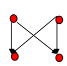


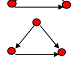
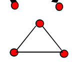
Ma'ayan et al, Science 309, 1078 (2005)

Signal propagation as links per step starting at a specific ligand



Motif abundance, homeostasis, and plasticity

Feedback loops  **bifans** 

Feed-forward loops  **scaffolds** 

Rapid-change ligands engage more motifs in fewer steps;

At early steps, more FFL than expected; at later steps, more +FBL than expected

Motif counts increase linearly with steps for all regulators – preferential paths to key effectors;

Positive and negative motifs are balanced for glutamate and BDNF - homeostasis;

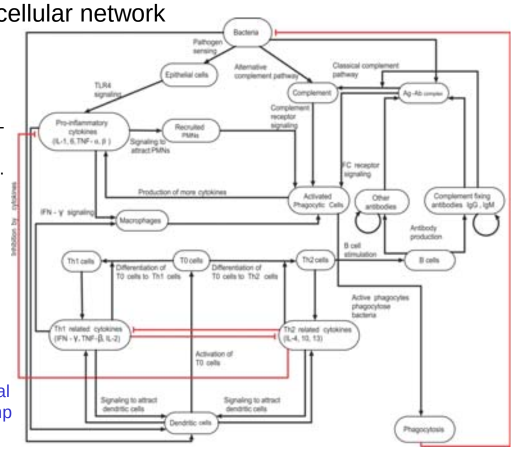
More positive than negative FBL and FFL in NE – long- term info storage

Intercellular network

Positive feedback - mediated by pro-infl. cytokines

Double negative feedback

Self-loops



Thakar et al PLoS Comp Bio 2007

Graph analysis uncovered common architectural features of cellular networks

(Weakly) Connected,
short path length,
heterogeneous (approximately power law degree distribution),
conserved interaction motifs

Can you think of reasons and/or consequences of these features in addition to what we already talked about?

Importance of a dynamical understanding

Only subsets of the genome-wide interaction networks are active in a given external condition

Han et al. 2004 – dynamical modularity of protein interaction networks
Luscombe et al. 2004 – endogeneus and exogeneus transcriptional subnetworks

Network topology needs to be complemented by a description of network dynamics – states of the nodes and changes in the state
First step - pseudo-dynamics: propagation of reactions in chemical(interaction) space, starting from a source (signal)

Complete dynamical description is only feasible on smaller networks (modules):
Signal transduction in bacterial chemotaxis, NF-kB signaling module, the yeast cell cycle, Drosophila embryonic segmentation