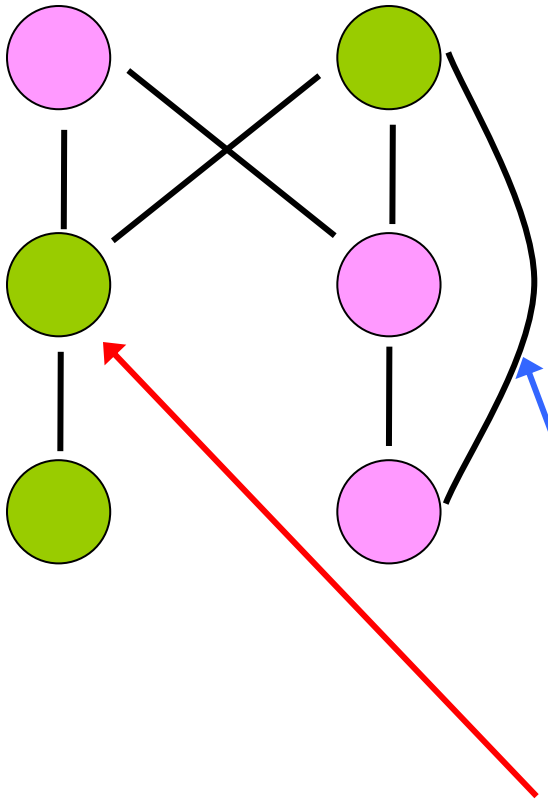
Biological Networks

Gavin Conant
163B ASRC
conantg@missouri.edu
882-2931

Types of Network

- Regulatory
- Protein-interaction
- Metabolic
- Signaling
- Co-expressing

General principle

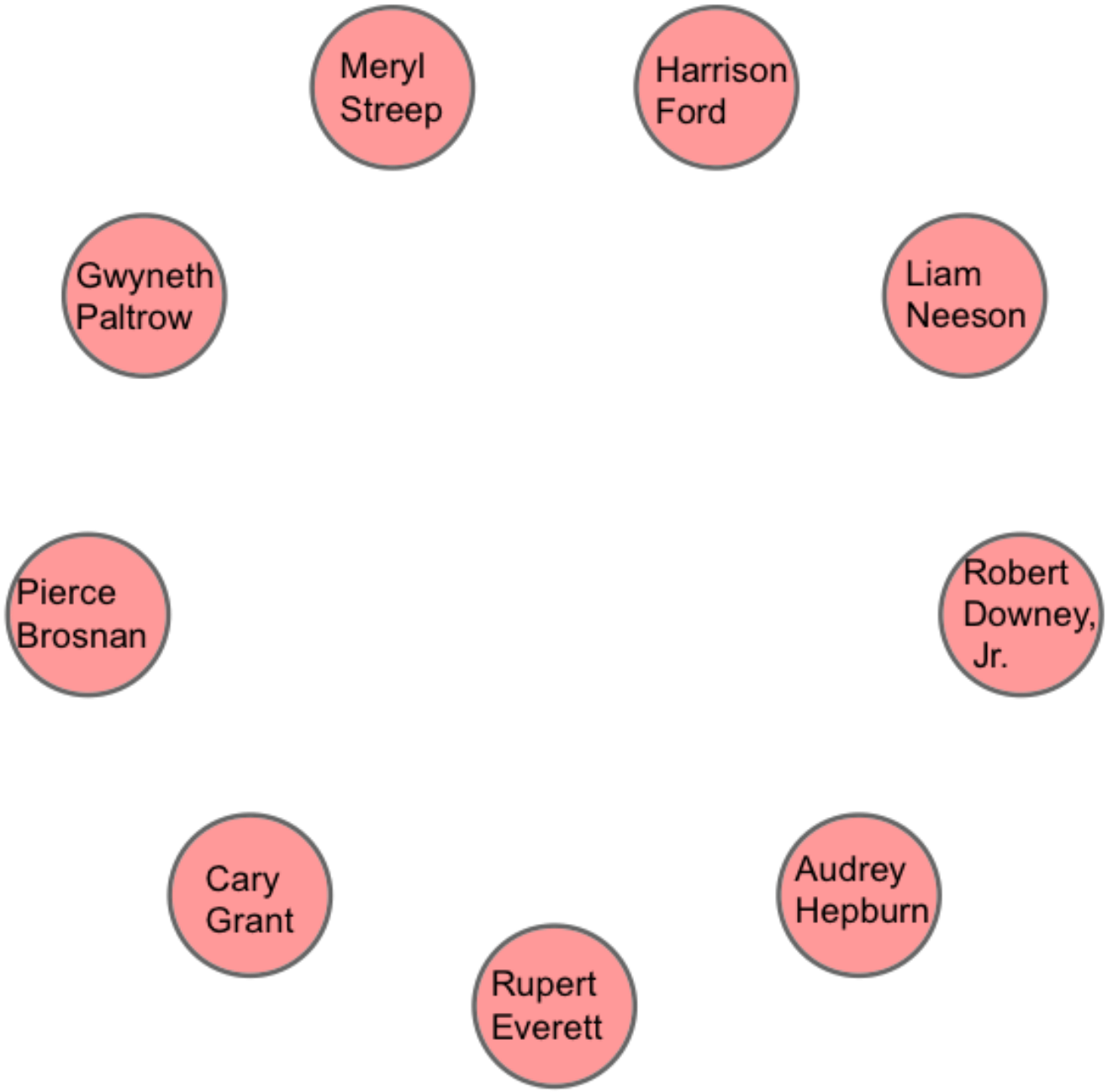


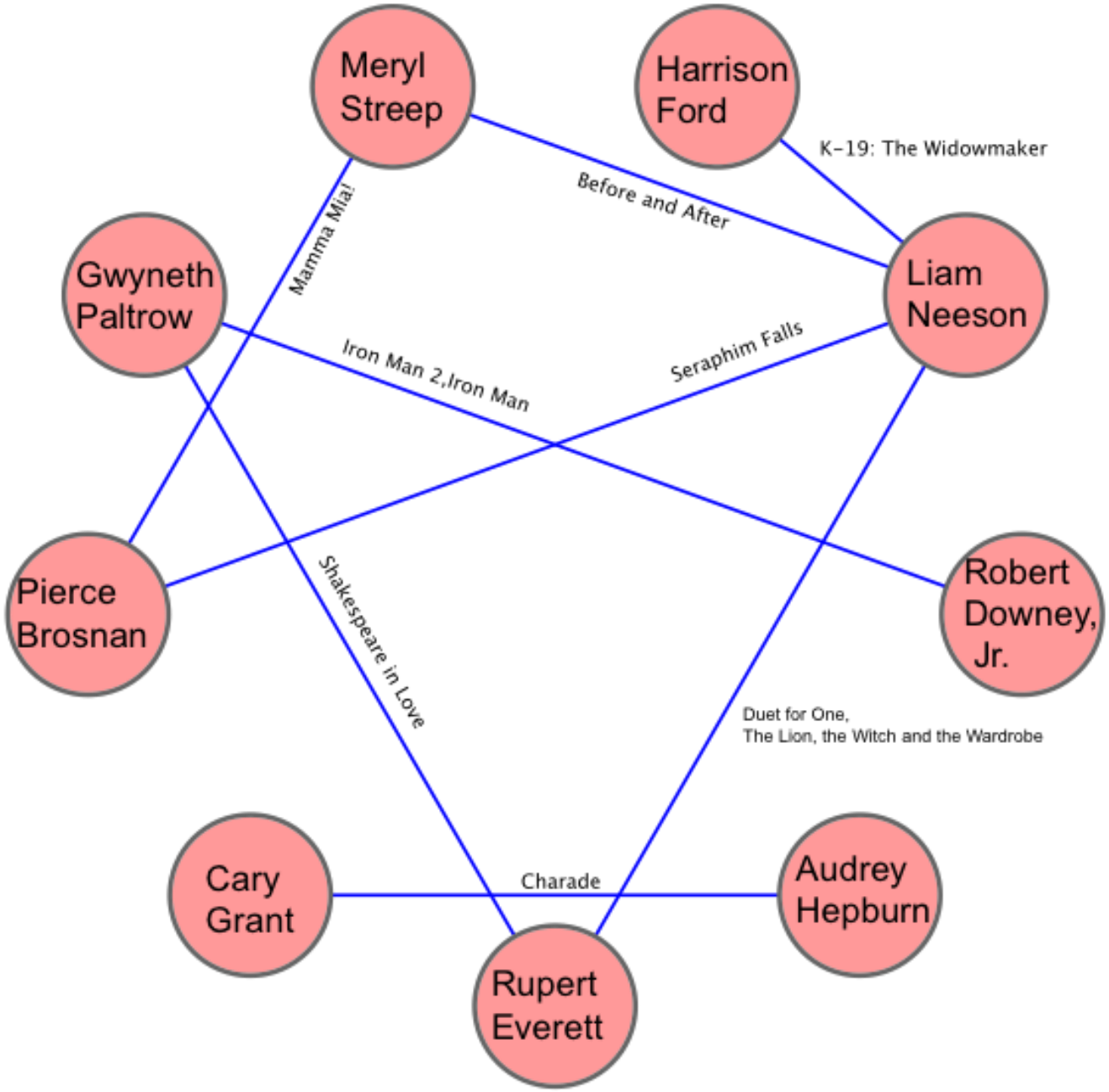
Relationship between genes

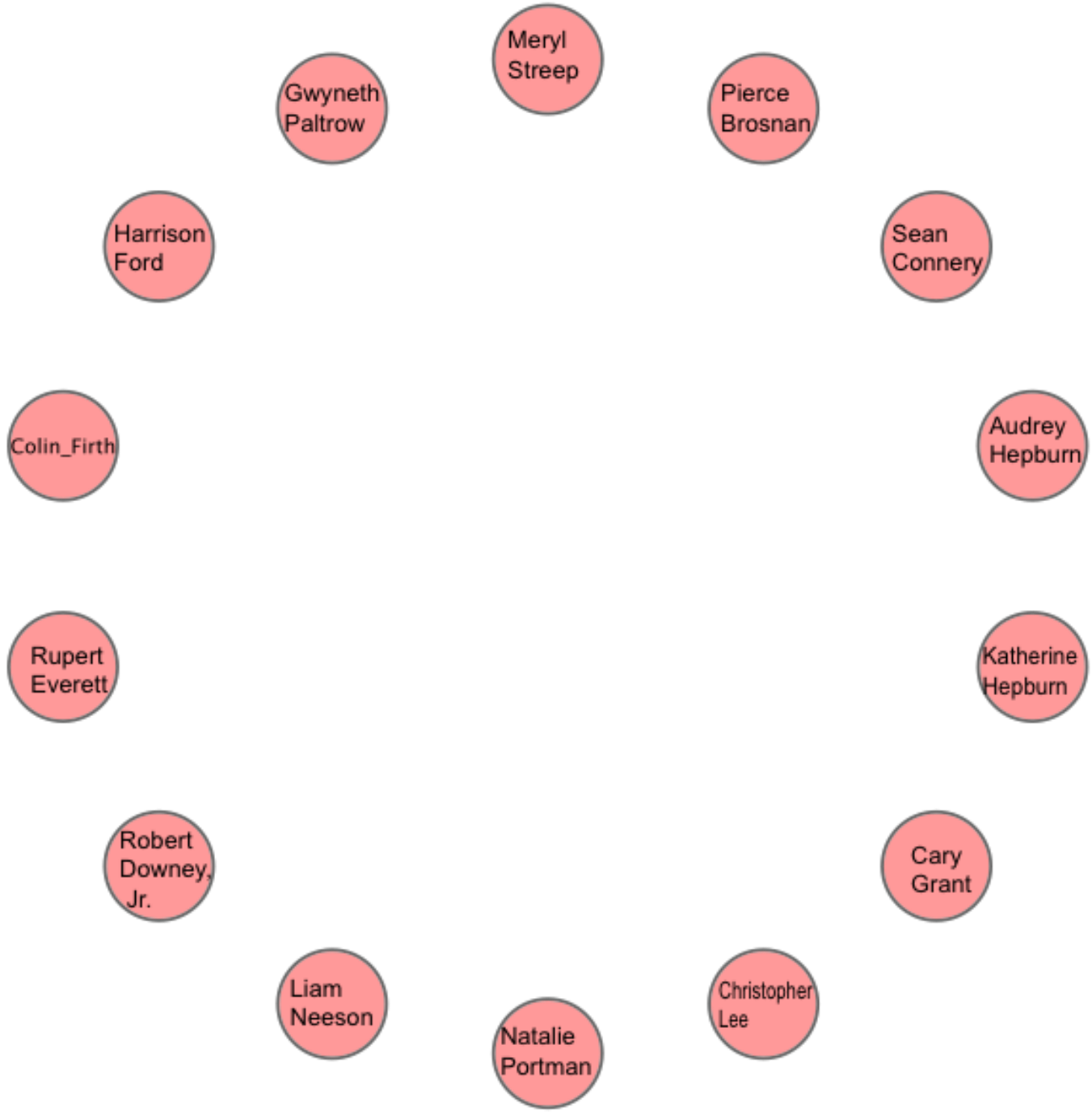
Gene/protein/enzyme

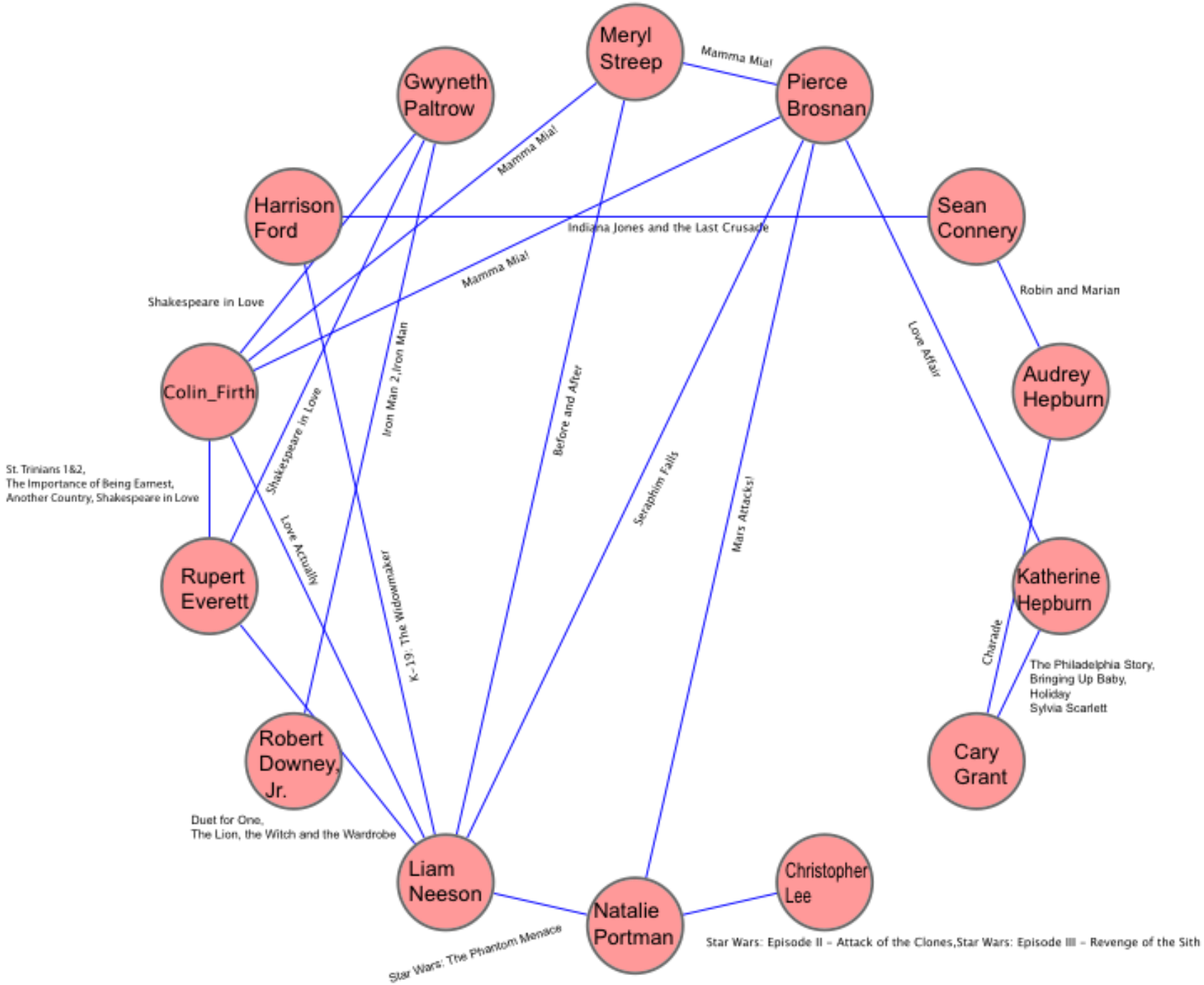
Example

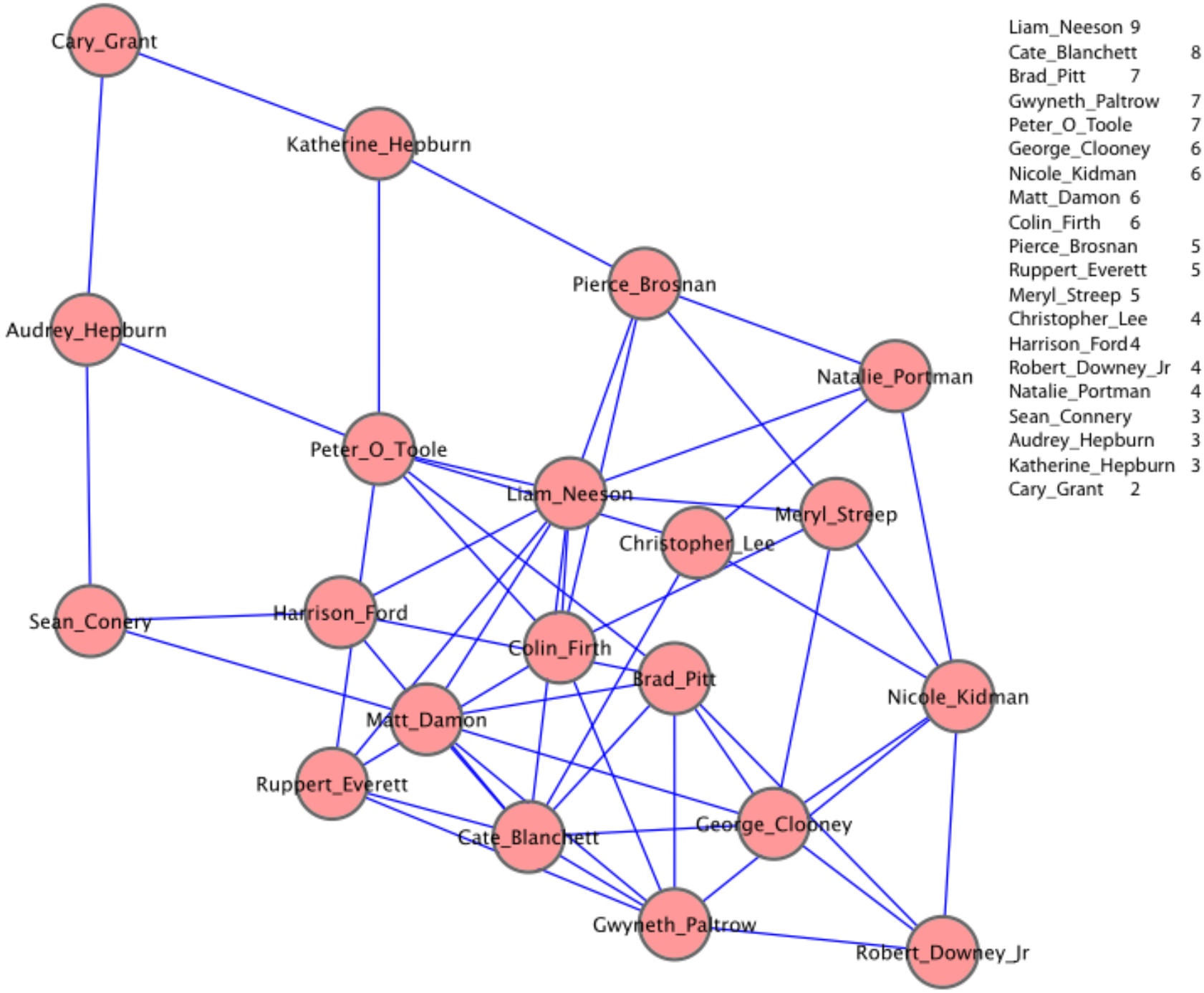
- We can create a network where actors and actresses are the nodes
- Two actors are joined if they co-stared in a film
- Kevin Bacon game

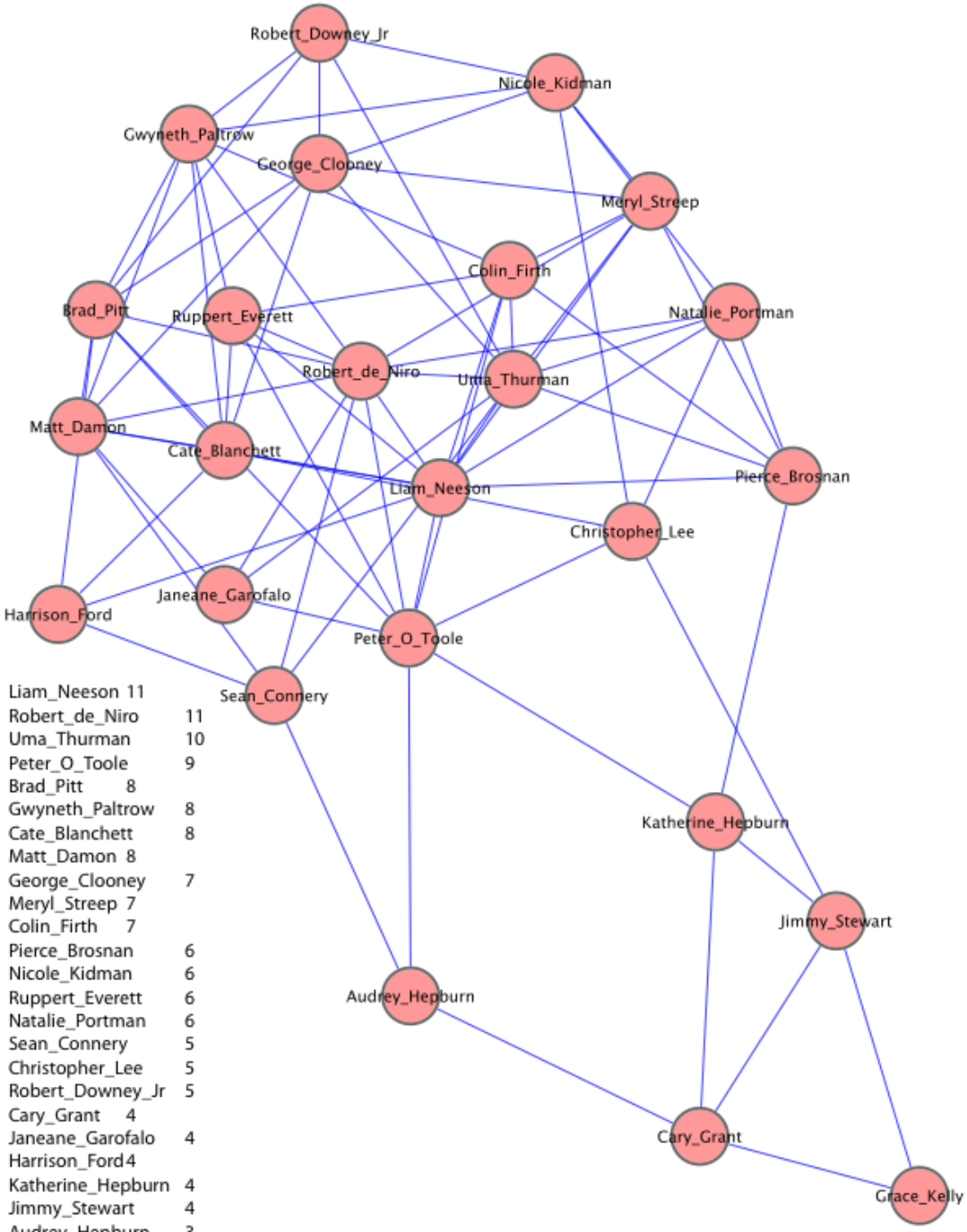

Meryl Streep
Harrison Ford
Gwyneth Paltrow
Liam Neeson
Pierce Brosnan
Robert Downey, Jr.
Cary Grant
Audrey Hepburn
Rupert Everett


Meryl Streep
Harrison Ford
K-19: The Widowmaker
Before and After
Mamma Mia!
Gwyneth Paltrow
Liam Neeson
Iron Man 2,Iron Man
Seraphim Falls
Pierce Brosnan
Robert Downey, Jr.
Shakespeare in Love
Duet for One,
The Lion, the Witch and the Wardrobe
Cary Grant
Charade
Audrey Hepburn
Rupert Everett


Meryl Streep
Gwyneth Paltrow
Pierce Brosnan
Harrison Ford
Sean Connery
Colin_Firth
Audrey Hepburn
Rupert Everett
Katherine Hepburn
Robert Downey, Jr.
Cary Grant
Liam Neeson
Christopher Lee
Natalie Portman


Meryl Streep
Mamma Mia!
Pierce Brosnan
Gwyneth Paltrow
Mamma Mia!
Harrison Ford
Indiana Jones and the Last Crusade
Sean Connery
Mamma Mia!
Shakespeare in Love
Robin and Marian
Colin_Firth
Iron Man 2,Iron Man
Before and After
Love Affair
Audrey Hepburn
Shakespeare in Love
St. Trinians 1&2,
The Importance of Being Earnest,
Another Country, Shakespeare in Love
Seraphim Falls
Mars Attacks!
Love Actually
K-19: The Widowmaker
Rupert Everett
Katherine Hepburn
Charade
The Philadelphia Story,
Bringing Up Baby,
Holiday
Sylvia Scarlett
Robert Downey, Jr.
Cary Grant
Duet for One,
The Lion, the Witch and the Wardrobe
Liam Neeson
Natalie Portman
Christopher Lee
Star Wars: The Phantom Menace
Star Wars: Episode II – Attack of the Clones,Star Wars: Episode III – Revenge of the Sith


Cary_Grant
Katherine_Hepburn
Audrey_Hepburn
Pierce_Brosnan
Natalie_Portman
Peter_O_Toole
Liam_Neeson
Meryl_Streep
Christopher_Lee
Sean_Conery
Harrison_Ford
Colin_Firth
Brad_Pitt
Nicole_Kidman
Matt_Damon
Ruppert_Everett
George_Clooney
Cate_Blanchett
Gwyneth_Paltrow
Robert_Downey_Jr
Liam_Neeson 9
Cate_Blanchett 8
Brad_Pitt 7
Gwyneth_Paltrow 7
Peter_O_Toole 7
George_Clooney 6
Nicole_Kidman 6
Matt_Damon 6
Colin_Firth 6
Pierce_Brosnan 5
Ruppert_Everett 5
Meryl_Streep 5
Christopher_Lee 4
Harrison_Ford 4
Robert_Downey_Jr 4
Natalie_Portman 4
Sean_Connery 3
Audrey_Hepburn 3
Katherine_Hepburn 3
Cary_Grant 2



Liam_Neeson	11
Robert_de_Niro	11
Uma_Thurman	10
Peter_O_Toole	9
Brad_Pitt	8
Gwyneth_Paltrow	8
Cate_Blanchett	8
Matt_Damon	8
George_Clooney	7
Meryl_Streep	7
Colin_Firth	7
Pierce_Brosnan	6
Nicole_Kidman	6
Ruppert_Everett	6
Natalie_Portman	6
Sean_Connery	5
Christopher_Lee	5
Robert_Downey_Jr	5
Cary_Grant	4
Janeane_Garofalo	4
Harrison_Ford	4
Katherine_Hepburn	4
Jimmy_Stewart	4
Audrey_Hepburn	3
Grace_Kelly	2

Example

- We can create a network where actors and actresses are the nodes
- Two actors are joined if they co-stared in a film
- Kevin Bacon game
  - On average, two actors can be linked by 3.65 films
  - Christopher Lee is actually more highly connected than Kevin Bacon

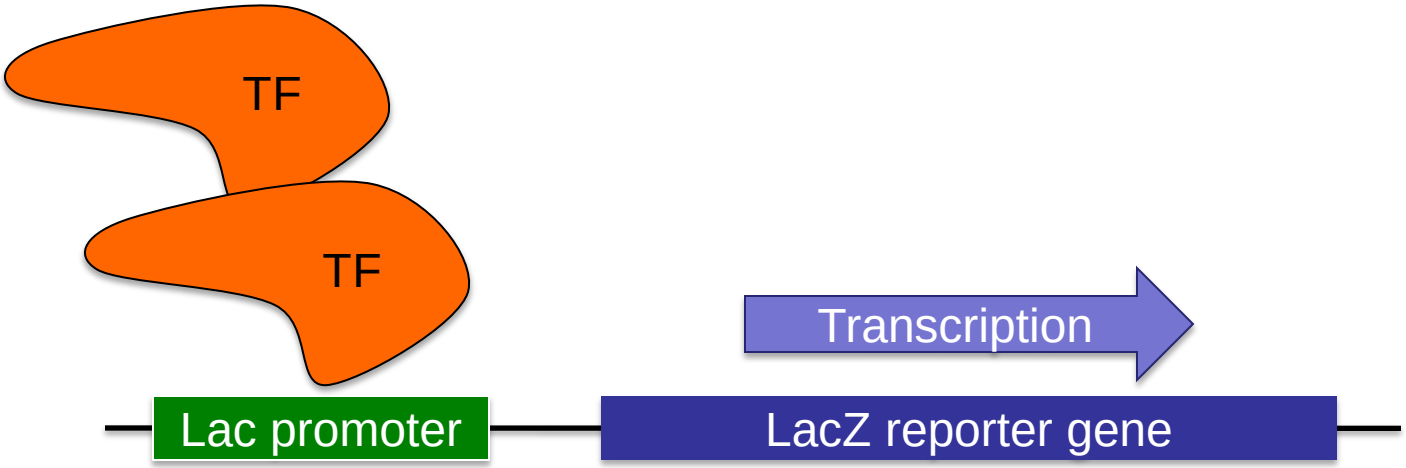
Watts, D. J., and S. H. Strogatz. 1998. Collective dynamics of 'small-world' networks. Nature **393**:440-442.

Example—Protein interaction networks

- Many or most proteins bind to other proteins
  - Generally not covalent
  - But can still be
    - Long term (multi-peptide enzyme complexes)
    - Short term (signal transduction)
  - Generically treat this as *protein interactions*

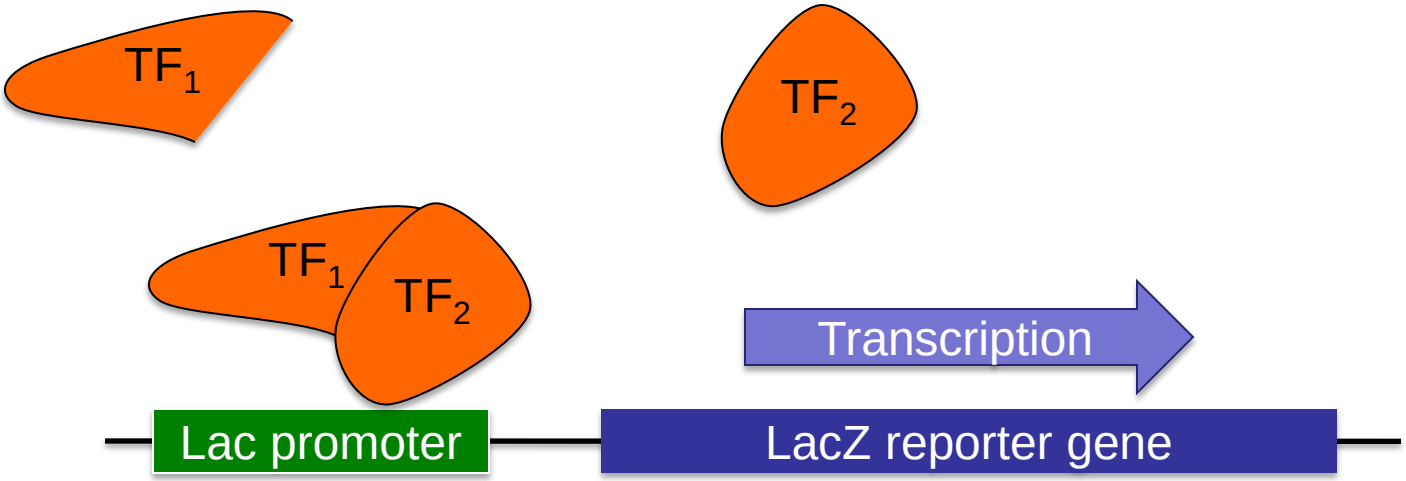
Methods to identify protein interactions

- Lots of small-scale experiments
- We will discuss the “genome-scale” methods
- Earliest: yeast two-hybrid
  - Uses yeast as a tool—can be done for proteins from any organism
  - But has been done at genome scale in yeast ☺



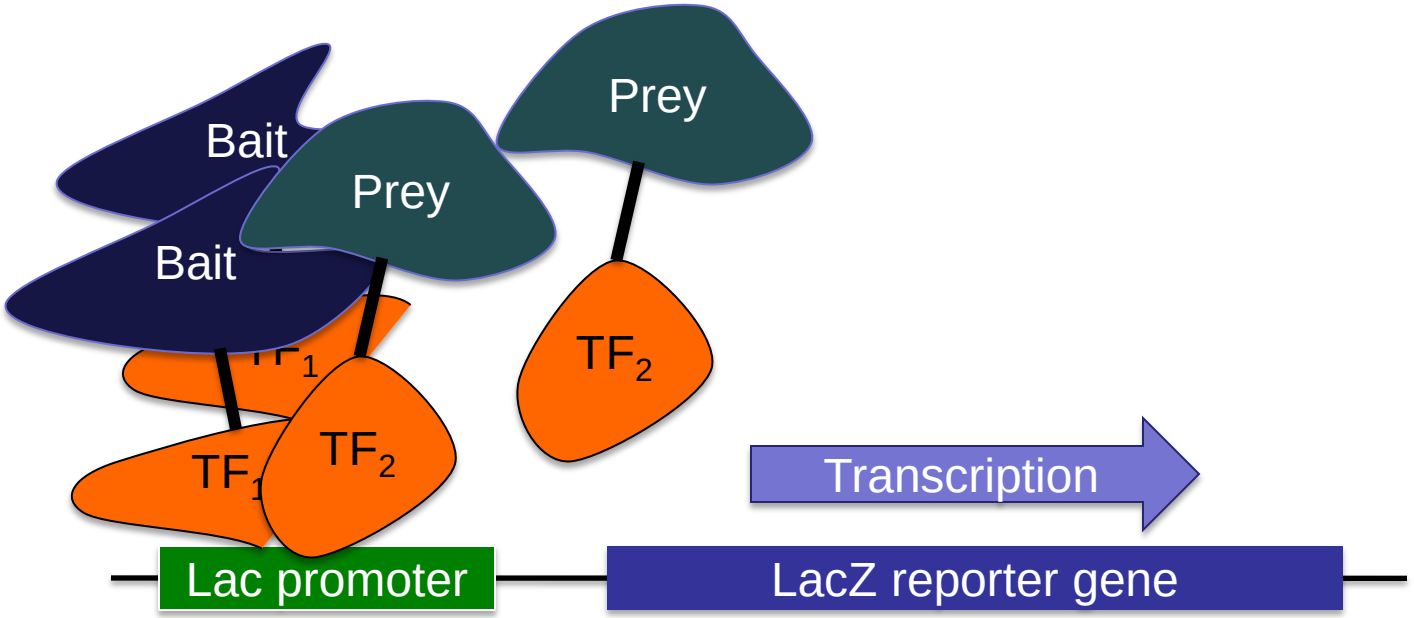
Introduce a plasmid with a reporter gene: turns colonies blue if LacZ is on and a galactose-analog (“X-gal”) is present

- Originally, the transcription factor induced transcription of the gene and turned the colonies blue



The different domains of the TF can function independently

- TF now functions if the two pieces are “close”
- Note however that the two pieces do not “stick together” on their own



Tether each piece of the TF to another protein

- We are testing if the bait and prey interact
- If they do, they will bring the two halves of the TF close enough to turn on LacZ
- If the colonies turn blue, the two proteins interaction

Yeast two-hybrid results

- > 60% of the proteins in the yeast genome have been tested for some interactions
- Question, if there are 5000 unique proteins in yeast, what is the possible number of interactions?
- Problems with two-hybrid method
  - We've altered the proteins—false negatives and false positives

Yeast two-hybrid results

- Problems with two-hybrid method (cont.)
  - Probably won't be able to sample all possible interactions
  - Doesn't account for time of protein expression or location
    - We could infer that two proteins interact when they are in fact never at the same place at the same time

Mass-spec. methods

- Start by adding a tag to the “bait” protein of interest:

- Now grow cells with this construct: copies of the bait protein will be tagged

Interactions vs. complexes

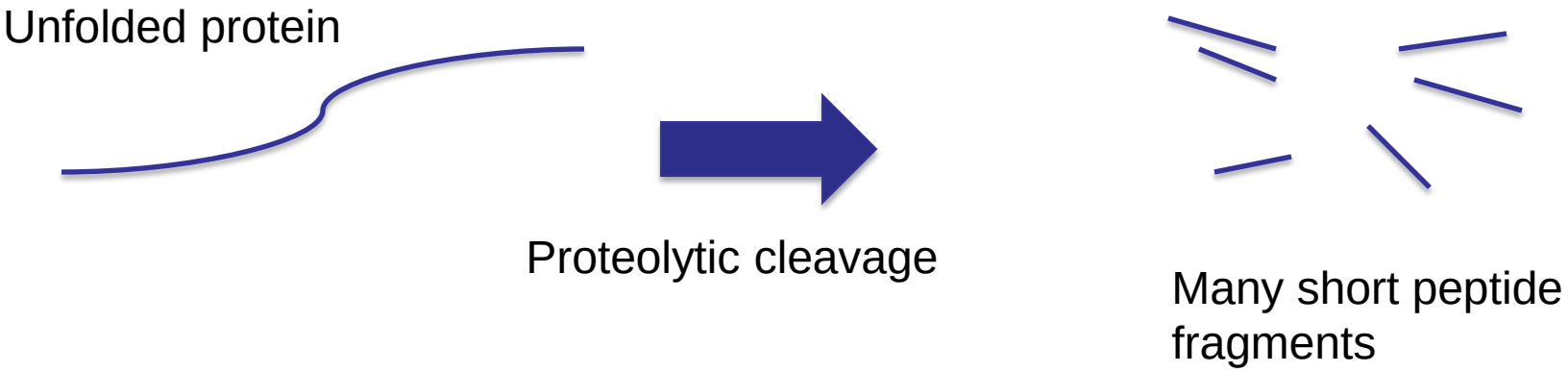
- Two-hybrid methods find pairwise protein interactions
- Here, we are looking at larger groups of proteins, aka complexes
- Complex is a somewhat vague term
  - Length of residue?
  - Functional?

Mass spec. continued

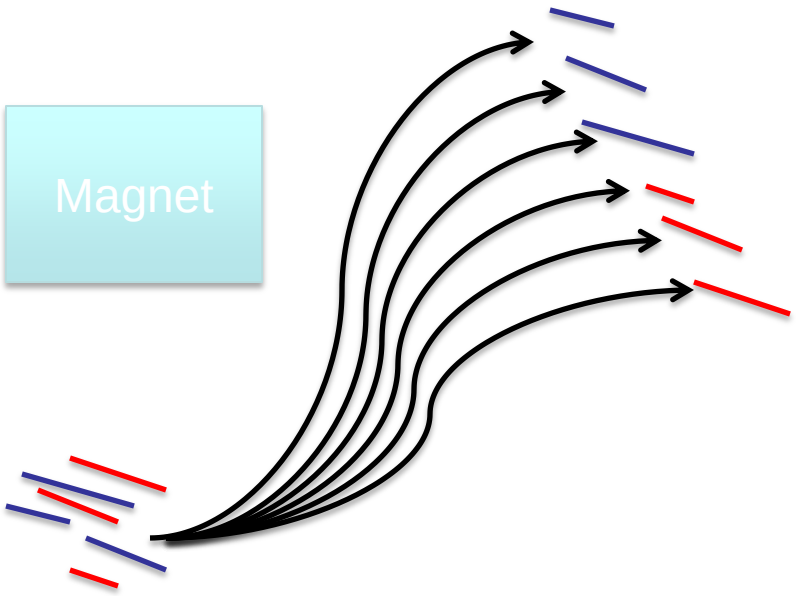
- Extract the cell proteins without disturbing the protein complexes
- From those proteins, extract any complexes with a member having the tag
  - Several steps
  - Uses an antibody to the tag for identification
- Result is complexes that have the bait protein as a member

Actual mass spec

- We now have a group of complexes which the bait protein is a member of
- What other proteins are present?



Separation of fragments



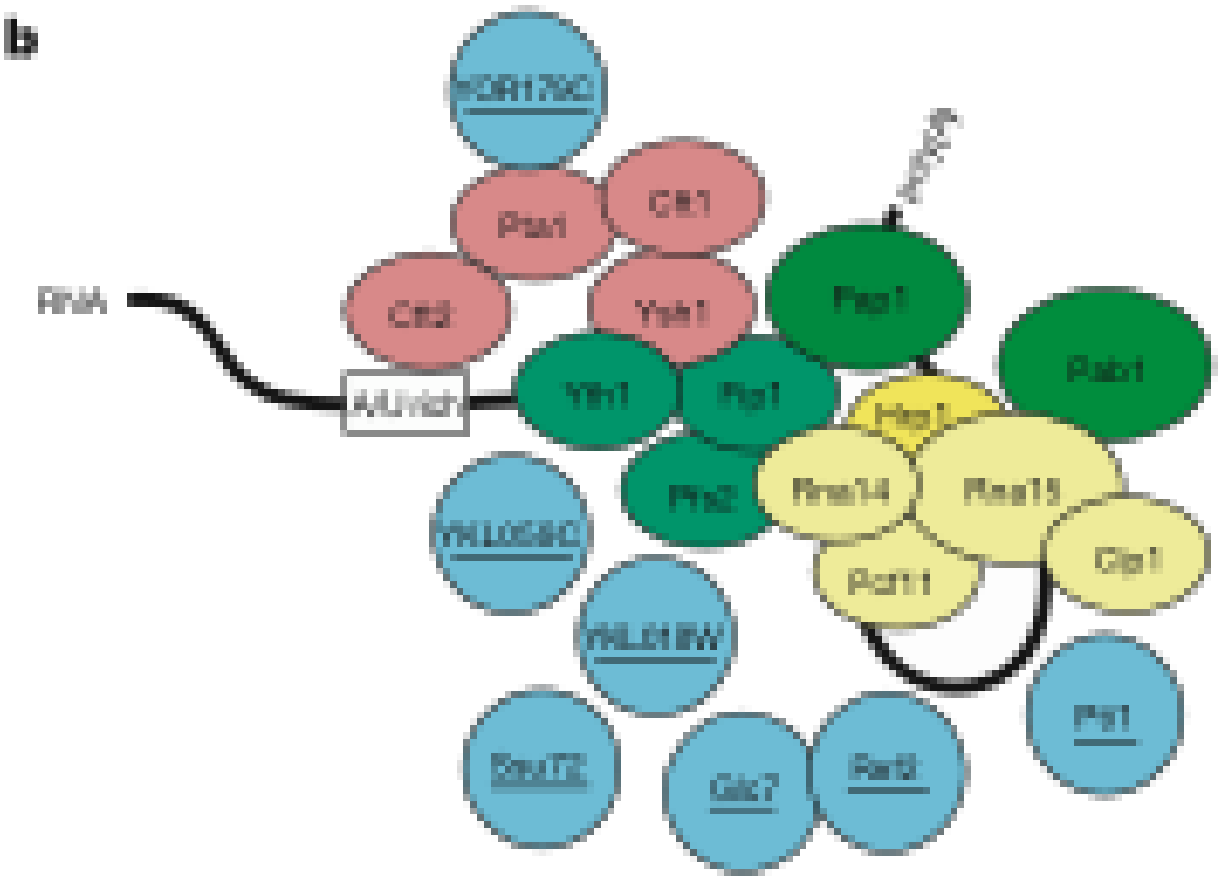
- Magnetic field separates fragments based on:
  - Charge
  - Mass
- Result is a list of ions with mass and change that are present

Identifying the ions

1. The mass spectrometer gives us a list of peptide fragment masses and charges
2. You search the genome for all possible peptide fragments and calculate their mass and charge
3. Match 1 to 2!
   - Obviously this is a computational challenge

Mass-spec protein complexes

- We put the bait-centered protein complexes into the mass-spec and identify the peptides present
- Look those up in the genome to identify the proteins present
- This gives us a list of the proteins in the complex for that bait protein



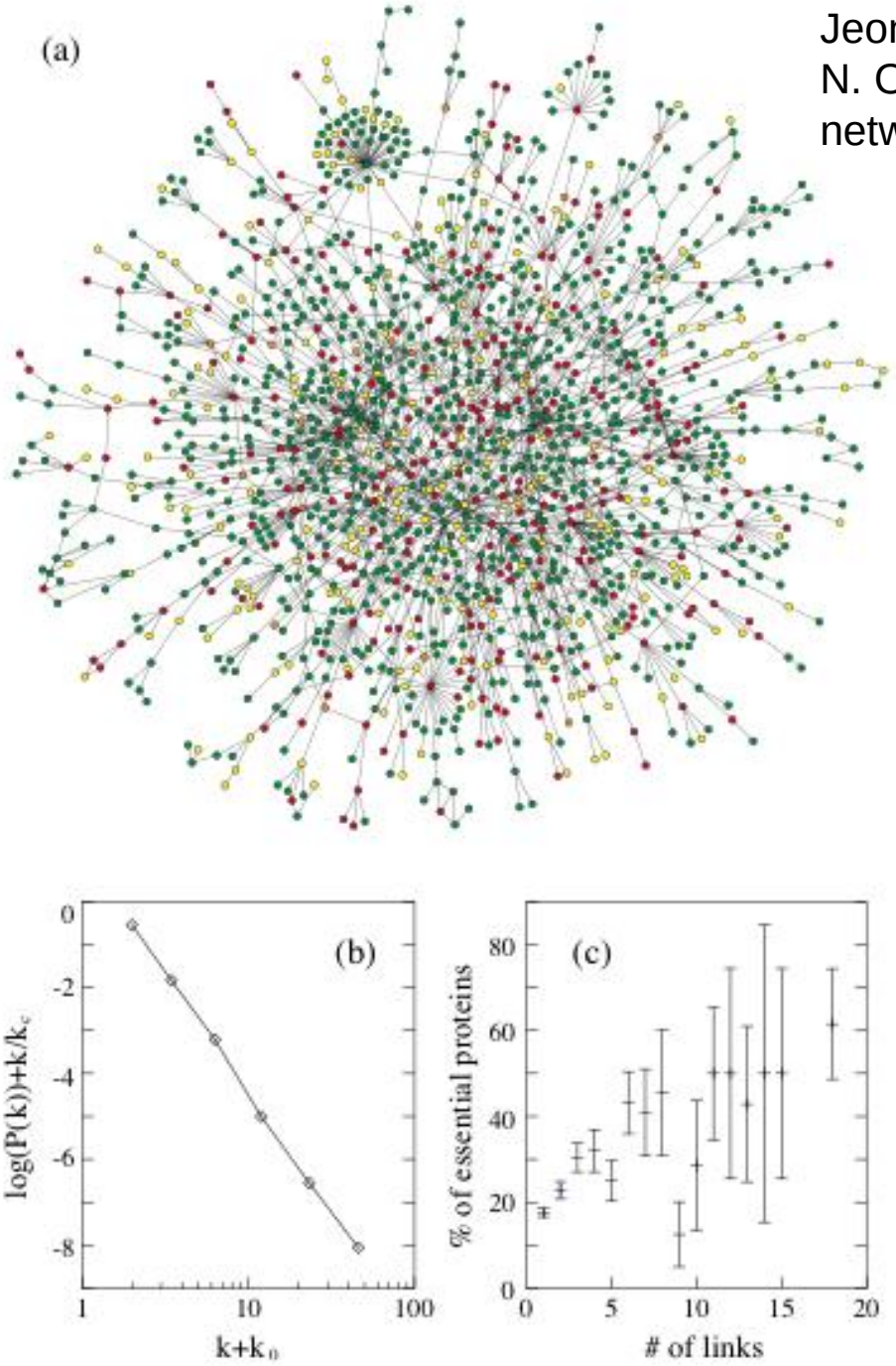
Gavin, A. C. et al., 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature **415**:141-147.

In yeast

- ~590 proteins used as bait
- ~ 232 complexes found among these
  - ~130 proteins used as baits did not return a complex
  - But complexes can also interact with each other

Building networks

- Protein interaction data has lots of uses
- One is to study biological networks
- Networks are an abstraction of the relationship among entities!



Jeong, H., S. P. Mason, A.-L. Barabási, and Z. N. Oltvai. 2001. Lethality and centrality in protein networks. Nature **411**:41-42.

Network Measurements

- Simple: *Degree* distribution
  - How many edges?
  - Expressed as a *distribution*

Network Measurements

- Simple: *Degree* distribution
  - How many edges?
  - Expressed as a *distribution*
- Common degree distributions:
  - Normal: Mean number of edges is μ
  - Exponential
    - $P(x) \approx e^{x}$
- Power-law
  - $P(x) \approx x^{a}$

Comparing exponential and "scale-free" networks

- Exponential
  - $P(x) \approx e^x$
  - Most nodes have similar numbers of edges
  - Sensitive to node loss
- Scale-free
  - $P(x) \approx x^a$
  - Few nodes with many edges (aka hubs)
  - Less sensitive to random node loss

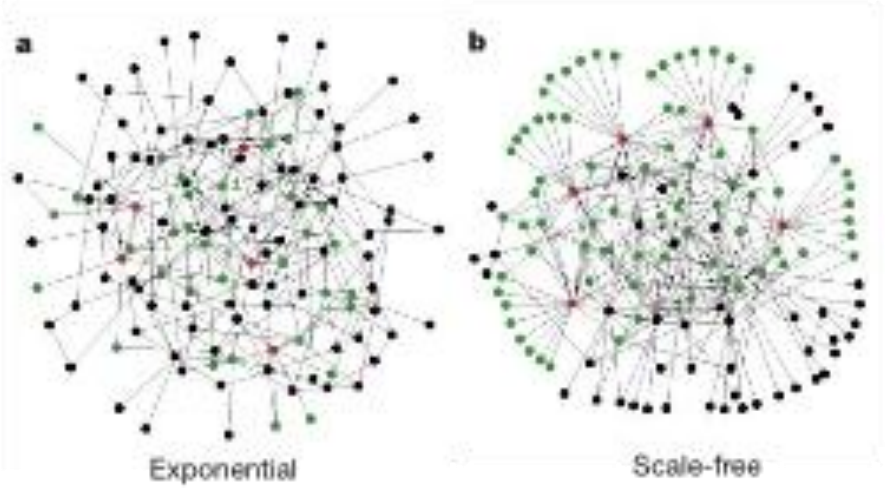


Figure 1 Visual illustration of the difference between an exponential and a scale-free network. **a**, The exponential network is homogeneous: most nodes have approximately the same number of links. **b**, The scale-free network is inhomogeneous: the majority of the nodes have one or two links but a few nodes have a large number of links, guaranteeing that the system is fully connected. Red, the five nodes with the highest number of links; green, their first neighbours. Although in the exponential network only 27% of the nodes are reached by the five most connected nodes, in the scale-free network more than 60% are reached, demonstrating the importance of the connected nodes in the scale-free network. Both networks contain 130 nodes and 215 links ($\langle k \rangle = 3.3$). The network visualization was done using the Pajek program for large network analysis: (http://vlado.fmf.uni-lj.si/pub/networks/pajek/pajekman.htm).

Albert, R., H. Jeong, and A. L. Barabasi. 2000. *Nature* **406**:378-382.

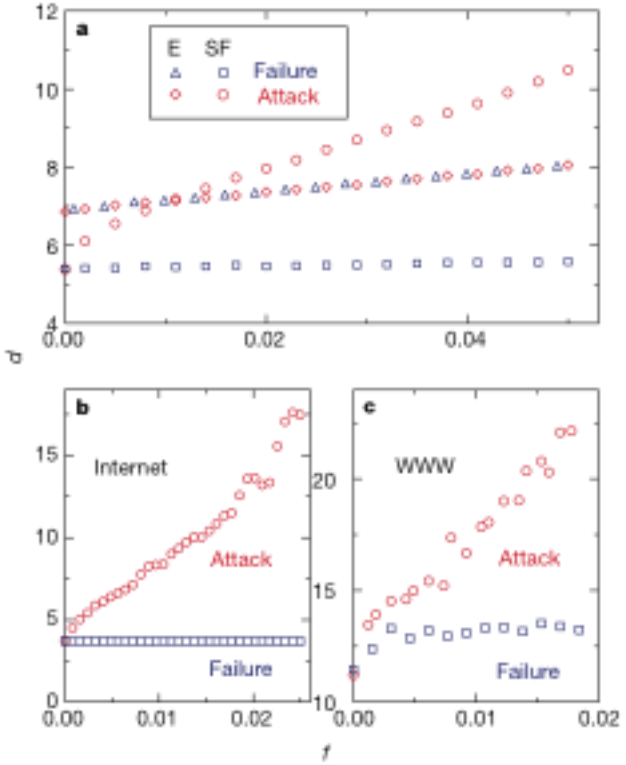


Figure 2 Changes in the diameter d of the network as a function of the fraction f of the removed nodes. **a**, Comparison between the exponential (E) and scale-free (SF) network models, each containing $N = 10{,}000$ nodes and 20,000 links (that is, $\langle k \rangle = 4$). The blue symbols correspond to the diameter of the exponential (triangles) and the scale-free (squares) networks when a fraction f of the nodes are removed randomly (error tolerance). Red symbols show the response of the exponential (diamonds) and the scale-free (circles) networks to attacks, when the most connected nodes are removed. We determined the f dependence of the diameter for different system sizes ($N = 1{,}000$; 5,000; 20,000) and found that the obtained curves, apart from a logarithmic size correction, overlap with those shown in **a**, indicating that the results are independent of the size of the system. We note that the diameter of the unperturbed ($f = 0$) scale-free network is smaller than that of the exponential network, indicating that scale-free networks use the links available to them more efficiently, generating a more interconnected web. **b**, The changes in the diameter of the Internet under random failures (squares) or attacks (circles). We used the topological map of the Internet, containing 6,209 nodes and 12,200 links ($\langle k \rangle = 3.4$), collected by the National Laboratory for Applied Network Research (http://moat.nlanr.net/Routing/rawdata/). **c**, Error (squares) and attack (circles) survivability of the World-Wide Web, measured on a sample containing 325,729 nodes and 1,498,353 links[3], such that $\langle k \rangle = 4.59$.

Other statistics

- Mean path length
- Longest path (aka diameter)
- Clustering coefficient
  - Number of connections between your “neighbors” over the possible number of connections
- Number of components

Features of protein interaction networks

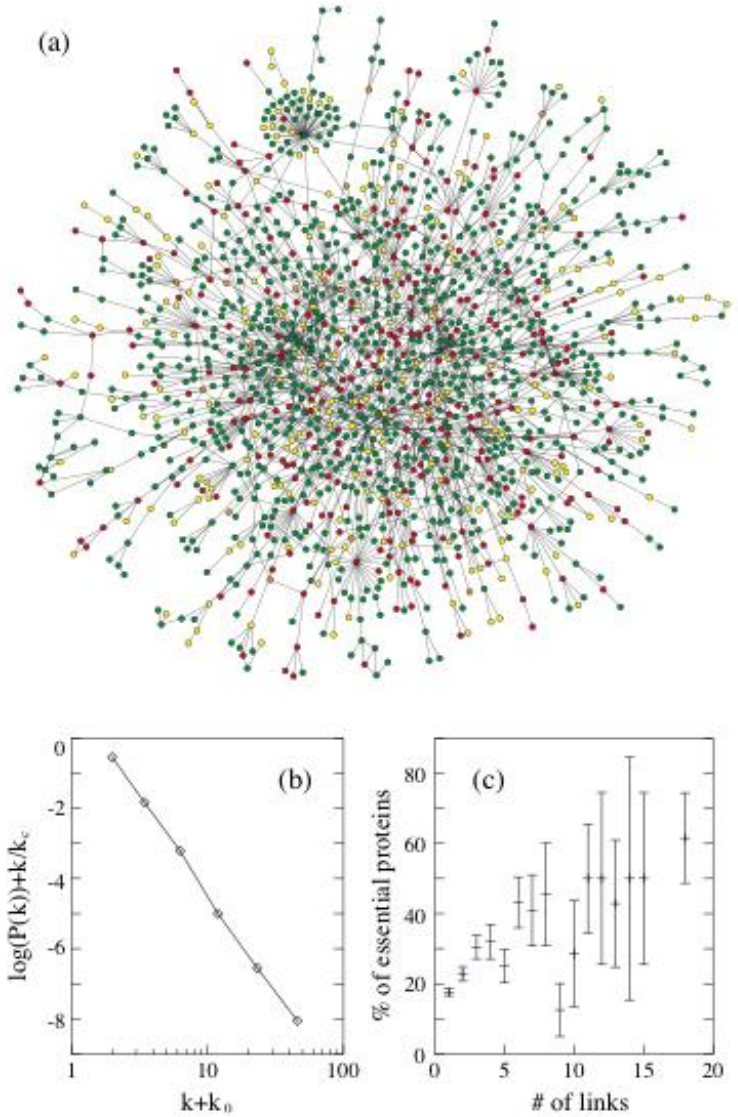
- “Small world”
- There are a few proteins with very many interactions (hubs)
- Proteins are “cliquish”:
  - If you and I interact, and I interact with Susan
  - It is more likely you and Susan will interact
- Most proteins “talk” to every other protein in very few steps

Uses of interaction networks

- Predict disease-related genes
  - The density of protein interactions that a protein shares with other proteins can be used to predict whether it is likely to influence a particular disease
- Follow cell-signaling?

Predictions from interaction networks

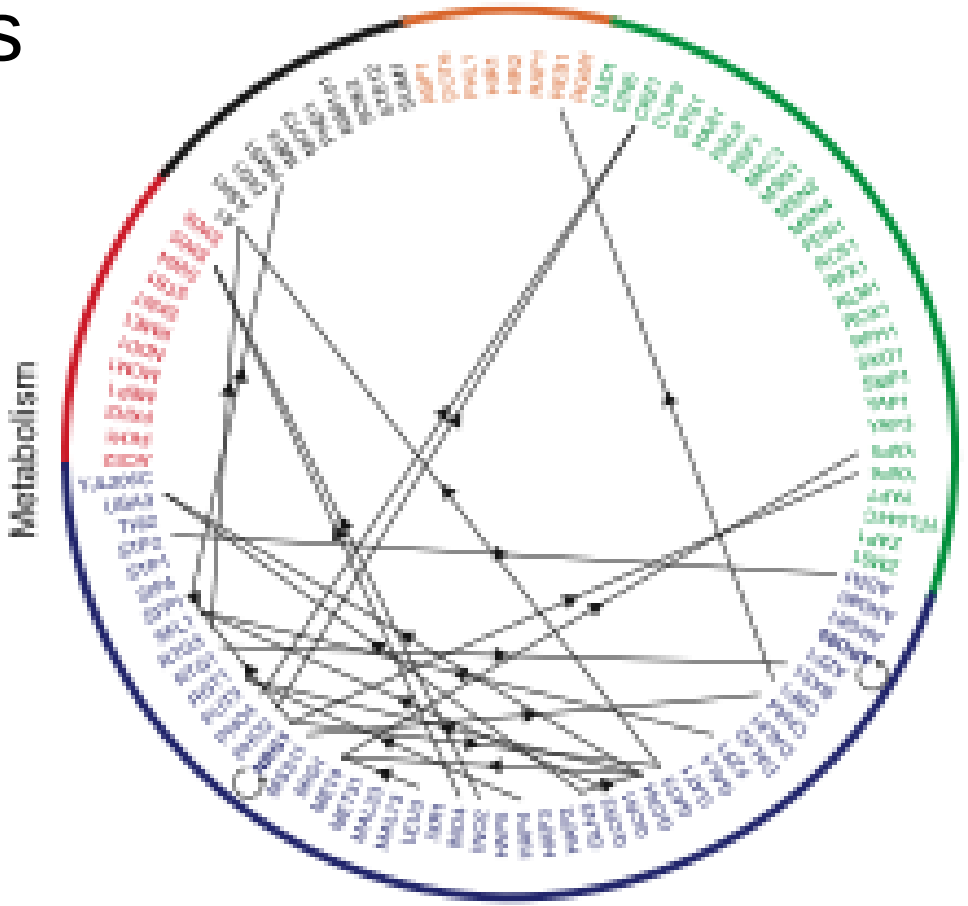
- The scale-free character of biological networks suggests that hub nodes may be more evolutionarily important
- Jeong et al., tested this by asking if genes for yeast protein interaction hubs were more likely to kill the cell when “knocked out” than other genes
  - Jeong, H., S. P. Mason, A.-L. Barabási, and Z. N. Oltvai. 2001.*Nature* **411**:41-42.



Characteristics of the yeast proteome. **a**, Map of protein–protein interactions. The largest cluster, which contains ~78% of all proteins, is shown. The colour of a node signifies the phenotypic effect of removing the corresponding protein (red, lethal; green, non-lethal; orange, slow growth; yellow, unknown). **b**, Connectivity distribution $P(k)$ of interacting yeast proteins, giving the probability that a given protein interacts with k other proteins. The exponential cut-off[5] indicates that the number of proteins with more than 20 interactions is slightly less than expected for pure scale-free networks. In the absence of data on the link directions, all interactions have been considered as bidirectional. The parameter controlling the short-length scale correction has value $k_0 = 1$. **c**, The fraction of essential proteins with exactly k links versus their connectivity, k, in the yeast proteome. The list of 1,572 mutants with known phenotypic profile was obtained from the Proteome database[13]. Detailed statistical analysis, including $r = 0.75$ for Pearson's linear correlation coefficient, demonstrates a positive correlation between lethality and connectivity. For additional details, see http://www.nd.edu/~networks/cell.

Other kinds of networks

- Regulatory networks
- Metabolic Networks
- Neural networks
- Gene expression networks



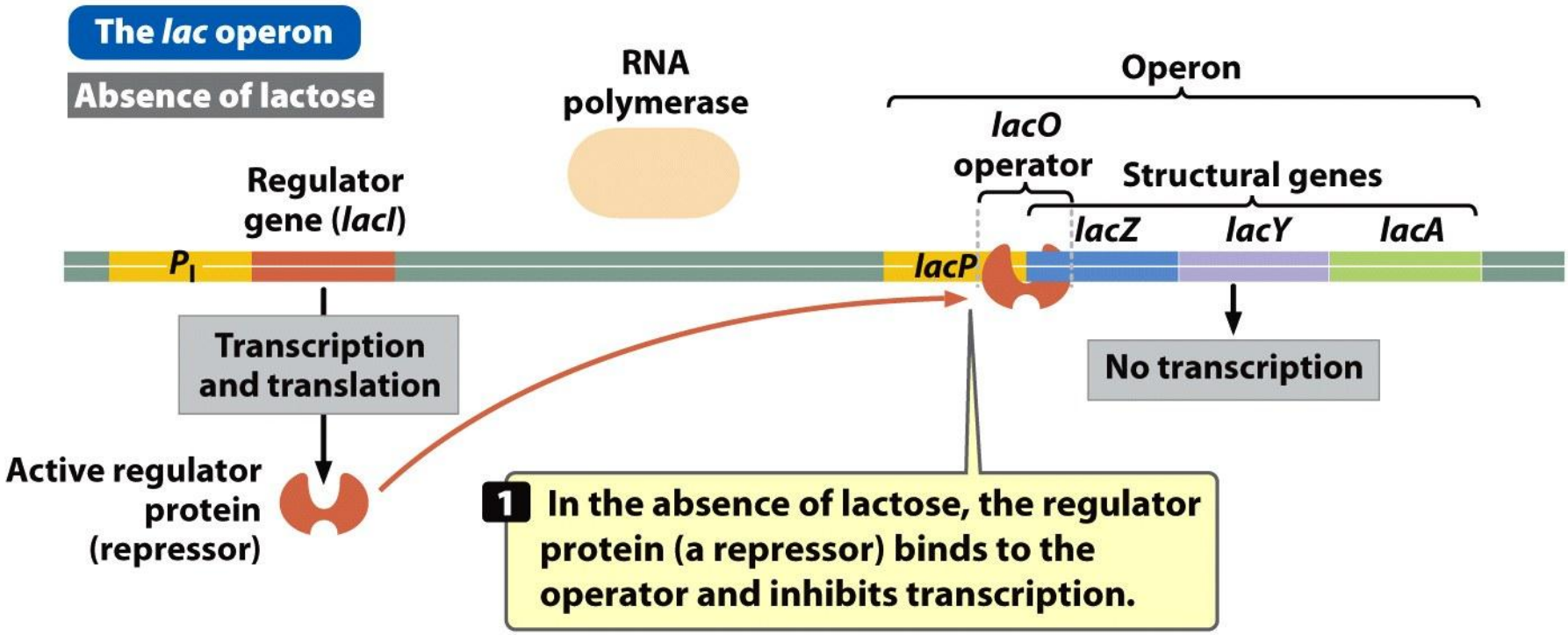


Figure 16-8a
Genetics: A Conceptual Approach, Third Edition

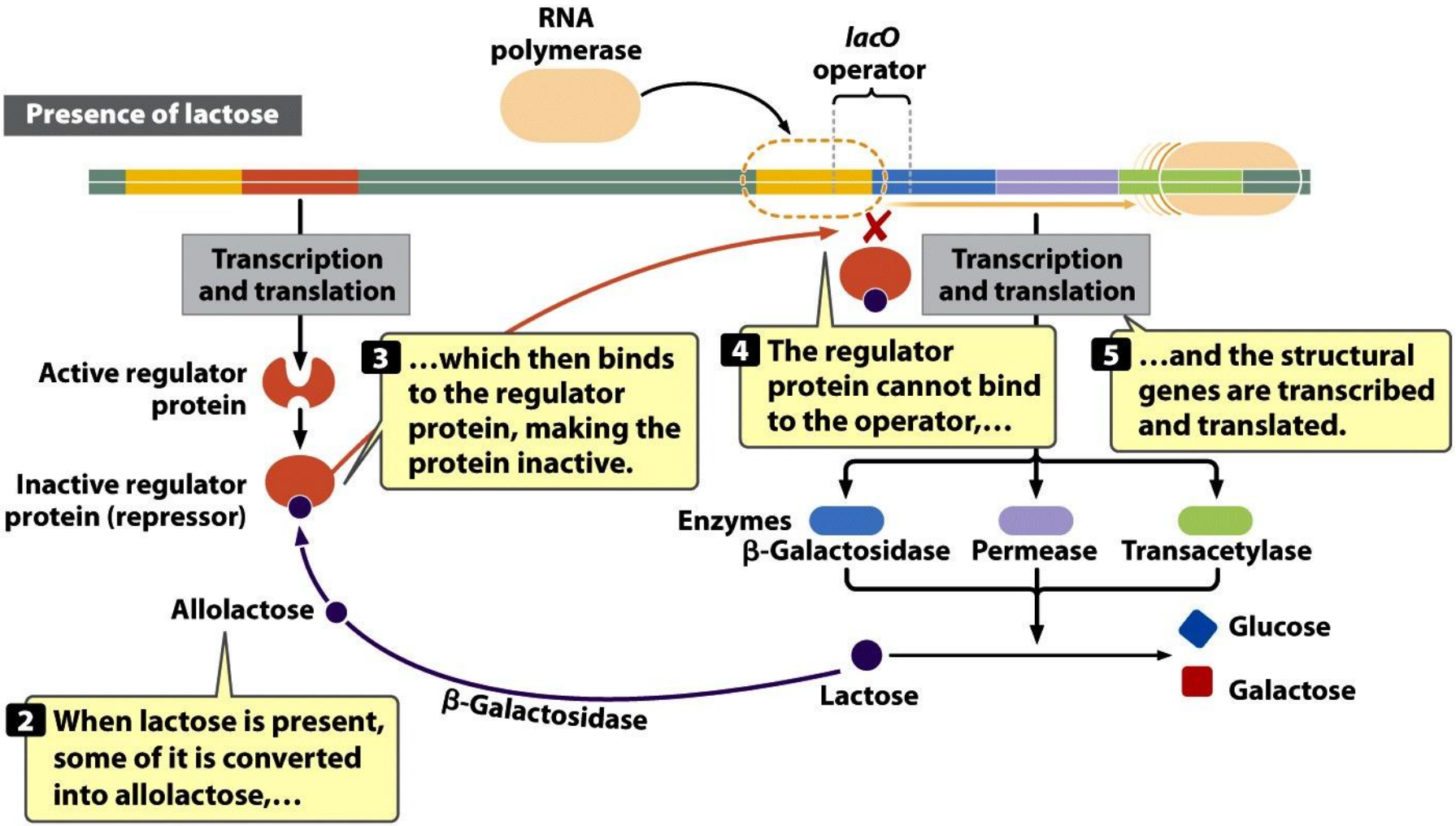


Figure 16-8b
Genetics: A Conceptual Approach, Third Edition

Why study gene expression?

- Identify genes associated with pathways or diseases
- Diagnostics of disease
- Reconstruct the *gene regulatory network*

Inferring the yeast regulatory network

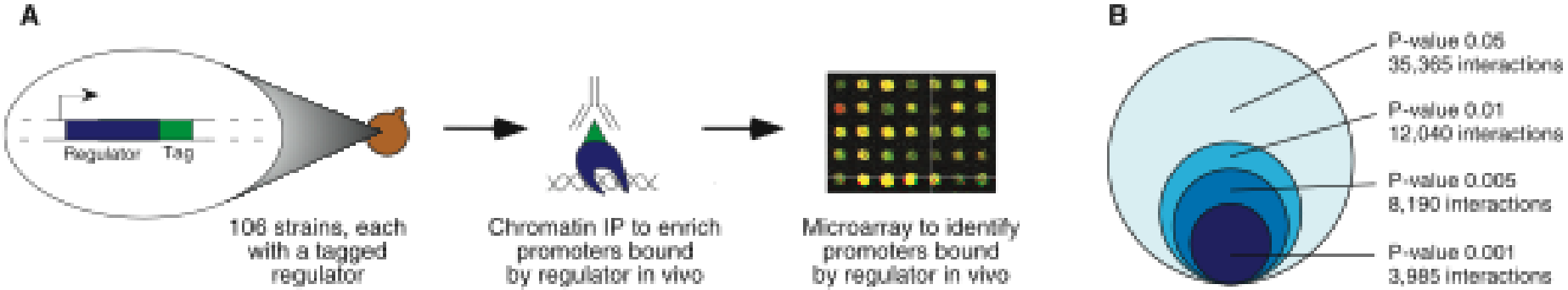
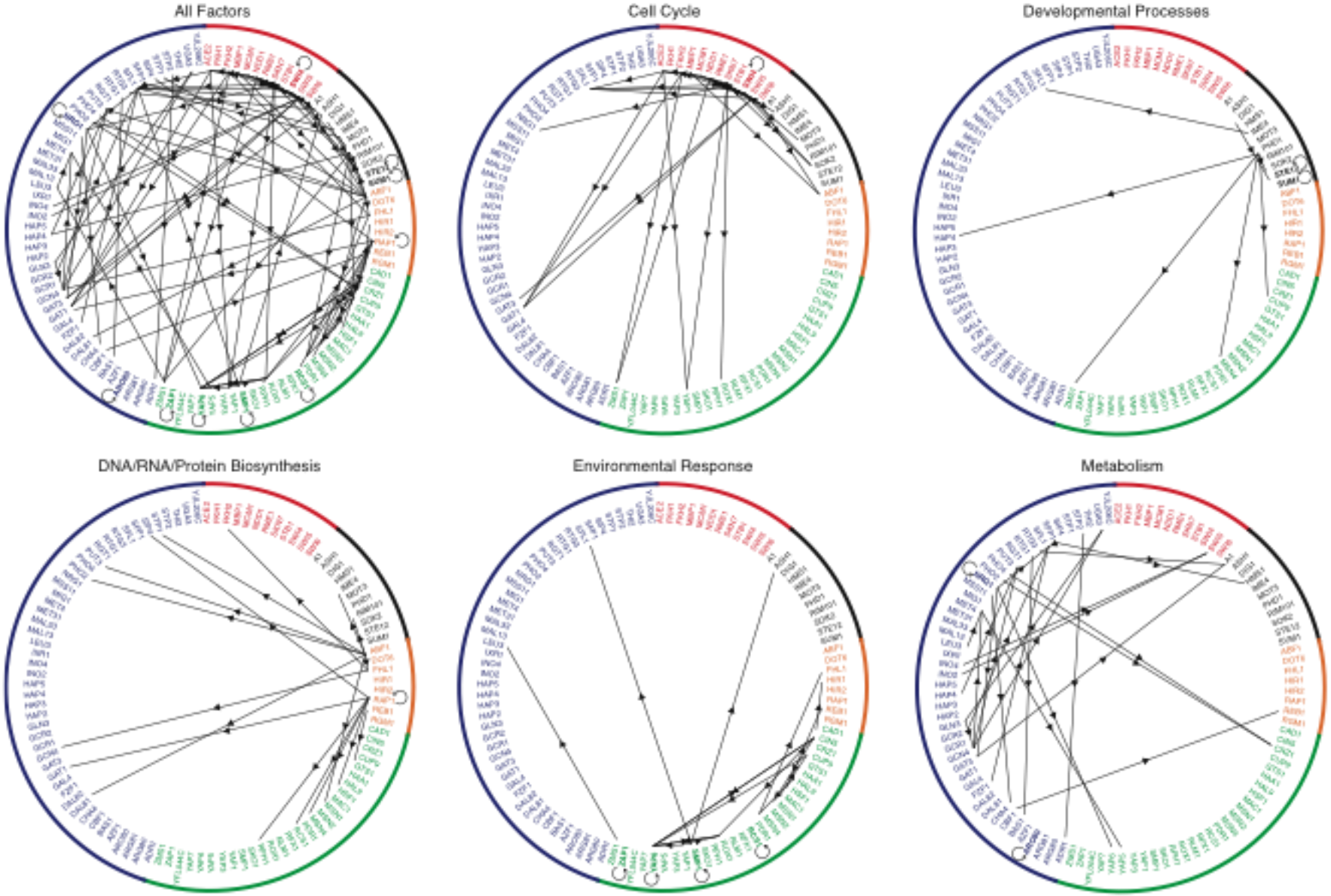


Fig. 1. Systematic genome-wide location analysis for yeast transcription regulators. (**A**) Methodology. Yeast transcriptional regulators were tagged by introducing the coding sequence for a c-myc epitope tag into the normal genomic locus for each regulator. Of the yeast strains constructed in this fashion, 106 contained a single epitope-tagged regulator whose expression could be detected in rich growth conditions. Chromatin immunoprecipitation (ChIP) was performed on each of these 106 strains. Promoter regions enriched through the ChIP procedure were identified by hybridization to microarrays containing a genome-wide set of yeast promoter regions. (**B**) Effect of P value threshold. The sum of all regulator-promoter region interactions is displayed as a function of varying P value thresholds applied to the entire location data set for the 106 regulators. More stringent P values reduce the number of interactions reported but decrease the likelihood of false-positive results.

Lee, T. I. *et al., Science* **298**, 799 (2002).


All Factors
Cell Cycle
Developmental Processes
DNA/RNA/Protein Biosynthesis
Environmental Response
Metabolism
Cell Cycle
Developmental Processes
DNA/RNA/Protein Biosynthesis
Environmental Response
Metabolism

Network “Motifs”

- We’ve looked at the “macro” structure of networks—what about their “micro” structure”
- Milo et al., defined the set of “motifs” for a network
  - Aka “subgraphs”
  - Milo, R., S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii., and U. Alon. 2002. *Science* **298**:824-827.

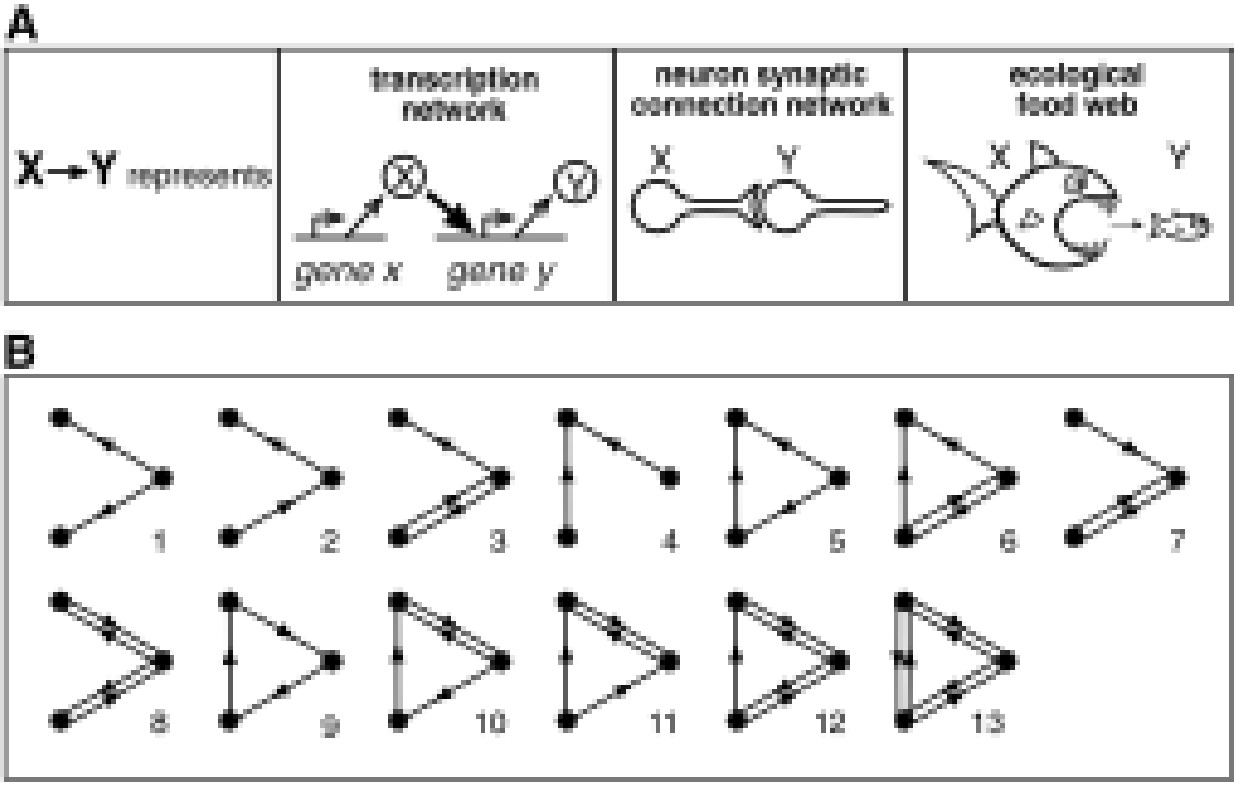


Fig. 1. (**A**) Examples of interactions represented by directed edges between nodes in some of the networks used for the present study. These networks go from the scale of biomolecules (transcription factor protein X binds regulatory DNA regions of a gene to regulate the production rate of protein Y), through cells (neuron X is synaptically connected to neuron Y), to organisms (X feeds on Y). (**B**) All 13 types of three-node connected subgraphs.

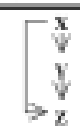
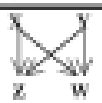
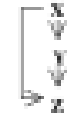
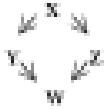
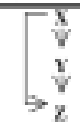
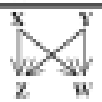
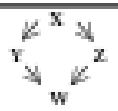
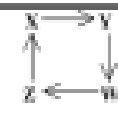
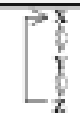
Network	Nodes	Edges	N_{real}	$N_{rand} \pm SD$	Z score	N_{real}	$N_{rand} \pm SD$	Z score	N_{real}	$N_{rand} \pm SD$	Z score
Gene regulation (transcription)					**Feed-forward loop**			**Bi-fan**			
E. coli	424	519	40	7 ± 3	10	203	47 ± 12	13			
*S. cerevisiae**	685	1,052	70	11 ± 4	14	1812	300 ± 40	41			
Neuron					**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	11,851	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	9000

Table 1. Network motifs found in biological and technological networks. The numbers of nodes and edges for each network are shown. For each motif, the numbers of appearances in the real network (N_{real}) and in the randomized networks ($N_{rand} \pm$ SD, all values rounded) (*17*, *18*) are shown. The *P* value of all motifs is $P < 0.01$, as determined by comparison to 1000 randomized networks (100 in the case of the World Wide Web). As a qualitative measure of statistical significance, the *Z* score = $(N_{real} - N_{rand})/SD$ is shown. NS, not significant. Shown are motifs that occur at least $U = 4$ times with completely different sets of nodes. The networks are as follows (*18*): transcription interactions between regulatory proteins and genes in the bacterium *E. coli* (*11*) and the yeast *S. cerevisiae* (*20*); synaptic connections between neurons in *C. elegans*, including neurons connected by at least five synapses (*24*); trophic interactions in ecological food webs (*22*), representing pelagic and benthic species (Little Rock Lake), birds, fishes, invertebrates (Ythan Estuary), primarily larger fishes (Chesapeake Bay), lizards (St. Martin Island), primarily invertebrates (Skipwith Pond), pelagic lake species (Bridge Brook Lake), and diverse desert taxa (Coachella Valley); electronic sequential logic circuits parsed from the ISCAS89 benchmark set (*7*, *25*), where nodes represent logic gates and flip-flops (presented are all five partial scans of forward-logic chips and three digital fractional multipliers in the benchmark set); and World Wide Web hyperlinks between Web pages in a single domain (*4*) (only three-node motifs are shown). e, multiplied by the power of 10 (e.g., 1.46e6 = 1.46×10^6).

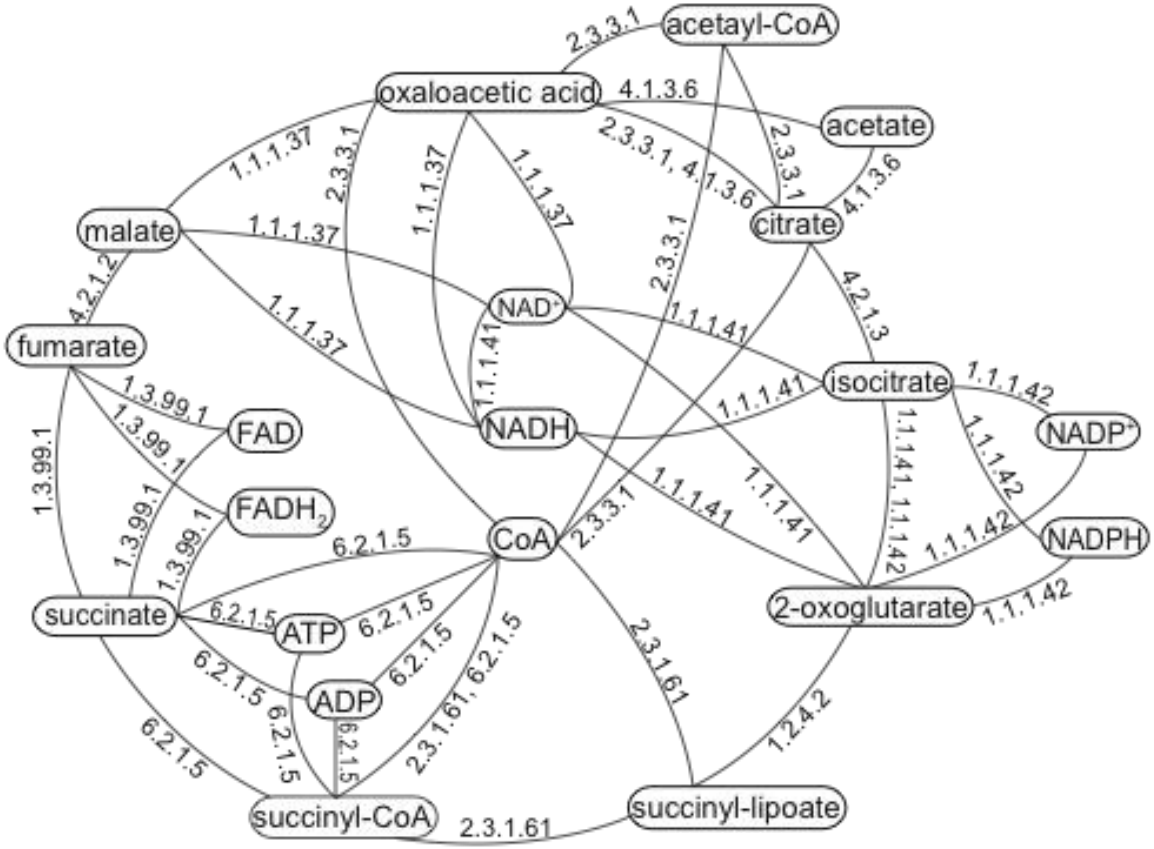
UNDERSTANDING REGULATORY NETWORKS

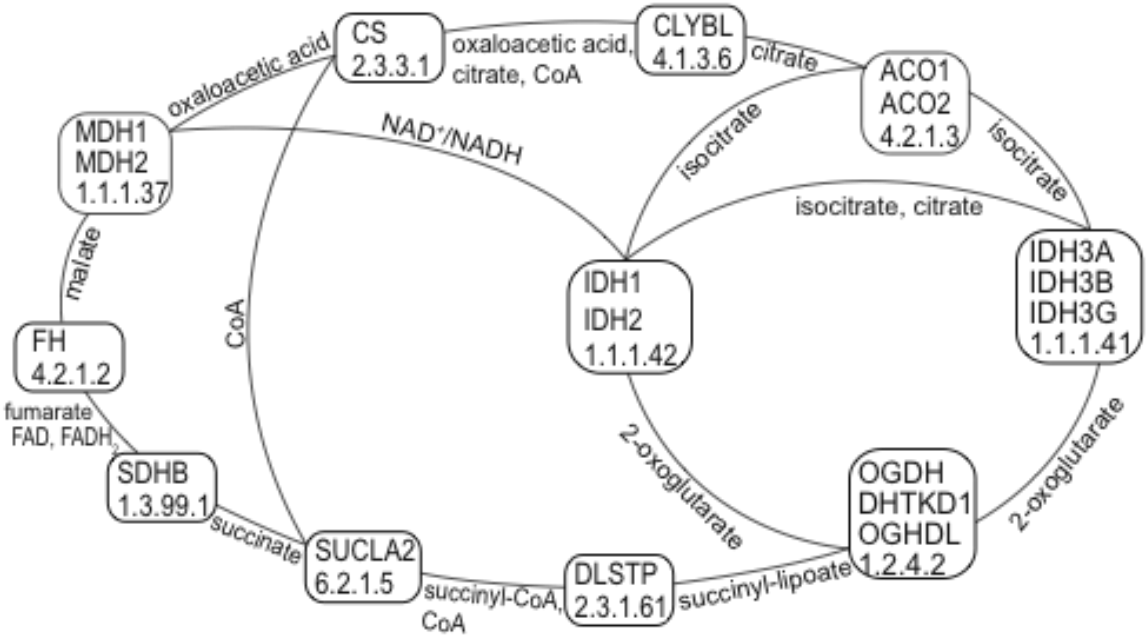
Genes have “switches”

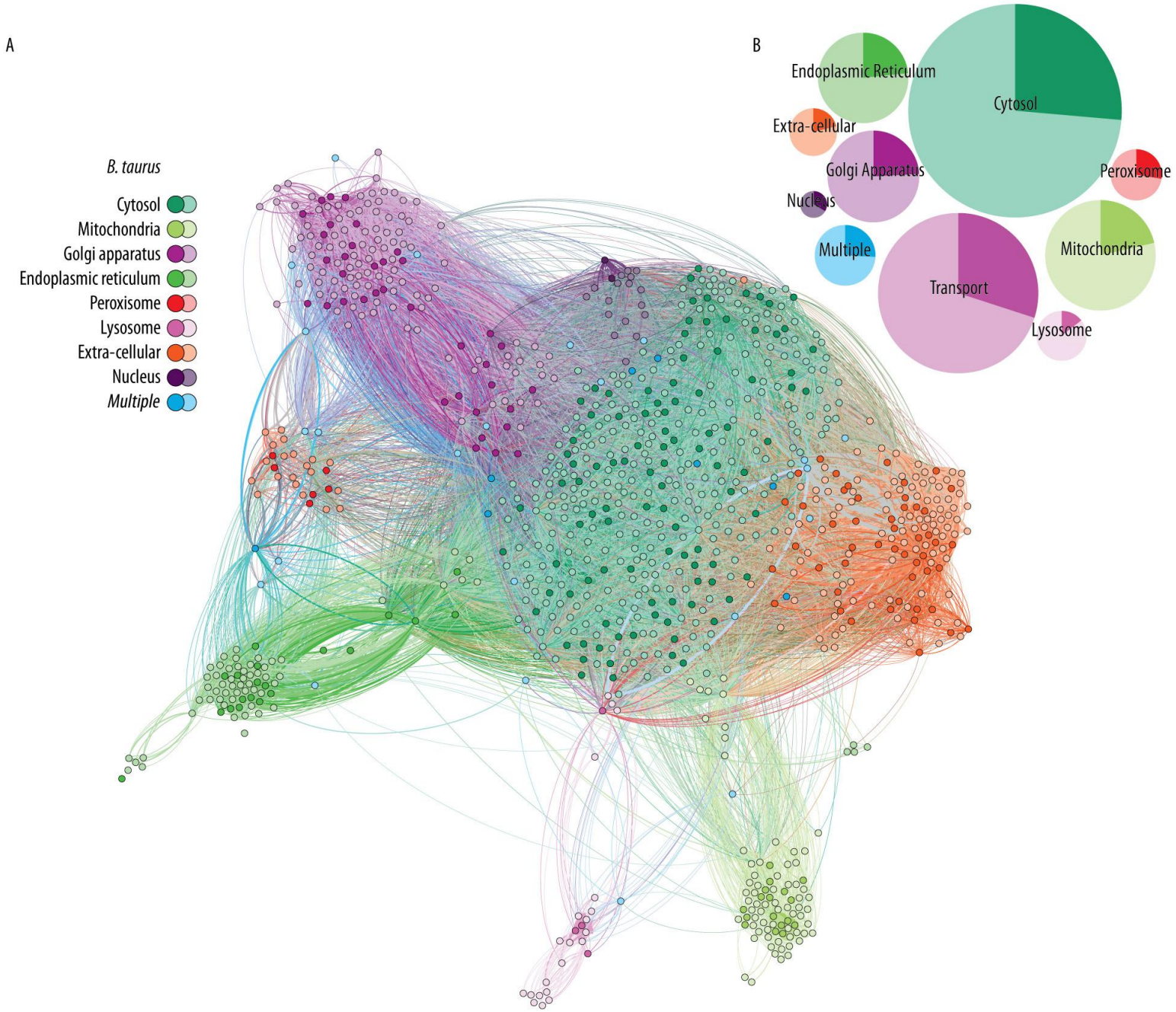
- Transcription is controlled at a number of levels
- Conceptually, there are proteins (transcription factors) that bind to specific non-coding DNA near transcribed genes
- These proteins can turn *on* transcription or turn *off* transcription
- The transcription factors themselves respond to various cellular signals

Metabolic networks

- Two types of entities
  - Metabolites: Compounds the cell degrades or synthesizes
  - Enzymes: Proteins that do this
- First work: Complete catalog of all enzymes in yeast and *E. coli*
  - Still don’t have effective methods to detect all metabolites


acetayl-CoA
2.3.3.1
oxaloacetic acid
4.1.3.6
acetate
1.1.1.37
2.3.3.1
1.1.1.37
1.1.1.37
2.3.3.1, 4.1.3.6
2.3.3.1
4.1.3.6
malate
1.1.1.37
citrate
4.2.1.2
NAD+
2.3.3.1
4.2.1.3
1.1.1.37
1.1.1.41
fumarate
isocitrate
1.1.1.42
1.3.99.1
1.1.1.41
NADP+
1.3.99.1
FAD
NADH
1.1.1.41
1.1.1.42
1.3.99.1
1.1.1.41, 1.1.1.42
FADH2
1.1.1.41
1.1.1.41
1.3.99.1
1.3.99.1
6.2.1.5
CoA
2.3.3.1
1.1.1.42
NADPH
6.2.1.5
6.2.1.5
6.2.1.5
succinate
ATP
6.2.1.5
2-oxoglutarate
1.1.1.42
6.2.1.5
6.2.1.5
6.2.1.5
2.3.1.61
ADP
6.2.1.5
2.3.1.61, 6.2.1.5
1.2.4.2
6.2.1.5
succinyl-CoA
2.3.1.61
succinyl-lipoate


CS
2.3.3.1
oxaloacetic acid
oxaloacetic acid,
citrate, CoA
CLYBL
4.1.3.6
citrate
ACO1
ACO2
4.2.1.3
isocitrate
isocitrate
isocitrate, citrate
MDH1
MDH2
1.1.1.37
NAD+/NADH
malate
FH
4.2.1.2
fumarate
FAD, FADH2
CoA
IDH1
IDH2
1.1.1.42
IDH3A
IDH3B
IDH3G
1.1.1.41
2-oxoglutarate
OGDH
DHTKD1
OGHDL
1.2.4.2
2-oxoglutarate
SDHB
1.3.99.1
succinate
SUCLA2
6.2.1.5
succinyl-CoA,
CoA
DLSTP
2.3.1.61
succinyl-lipoate


A
B. taurus
Cytosol
Mitochondria
Golgi apparatus
Endoplasmic reticulum
Peroxisome
Lysosome
Extra-cellular
Nucleus
Multiple
B
Endoplasmic Reticulum
Cytosol
Extra-cellular
Golgi Apparatus
Peroxisome
Nucleus
Multiple
Mitochondria
Transport
Lysosome

Human metabolic network

- Two versions
- ~1475 to ~2269 genes coding for enzymes
- 2478-3188 metabolites named
- 1052-3732 reactions

Duarte et al., *PNAS*. **104**:1777
Ma et al.,*Molecular Systems Biology* **3**:135.

Fig. 1. Human metabolic knowledge landscape. Colors represent the percentage of reactions within a pathway that have a confidence score of 3 (biochemical or genetic evidence), 2 (physiological data or evidence from a nonhuman mammalian cell), 1 (modeling evidence), or 0 (unevaluated). Metabolic pathways (primarily defined by the Kyoto Encyclopedia of Genes and Genomes LIGAND database) were classified into three categories based on their level of characterization as detailed in the text.

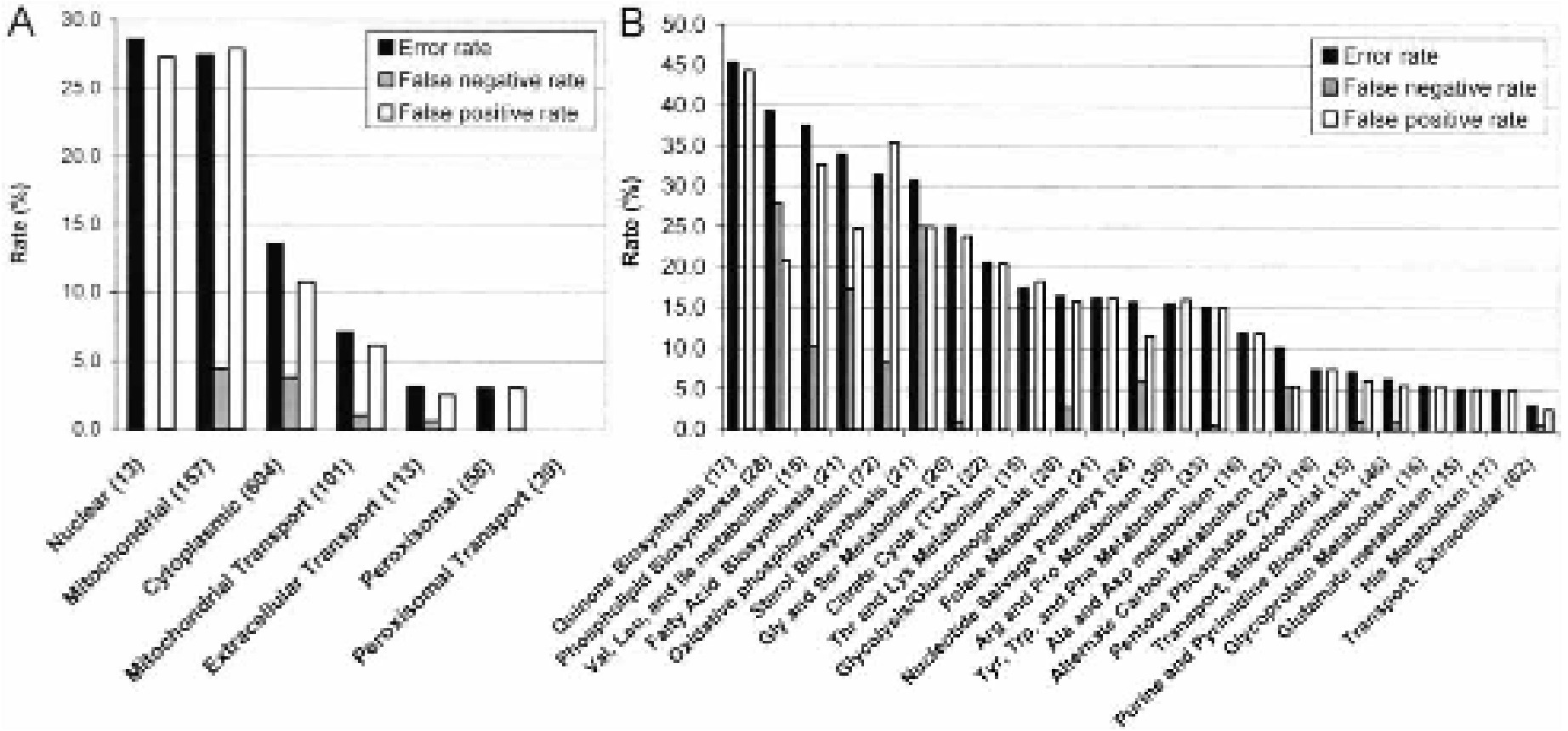


Figure 3 False prediction percentages for genes in particular cellular compartments (*A*) and particular metabolic subsystems (*B*). The overall error rate is the percentage of false predictions out of all of the predictions. The false-negative (FN) rate is the percentage of FN predictions out of all predictions in which the experimental data show normal growth. The false-positive (FP) rate is the percentage of FP predictions out of all predictions in which the experimental data show retarded growth. Genes that participate in transport functions between compartments are classified according to Table 2. Compartments with at least 10 genes and metabolic subsystems with at least 15 genes are included.