Detecting topological patterns in protein networks

Sergei Maslov

Brookhaven National Laboratory

What defines a complex system?

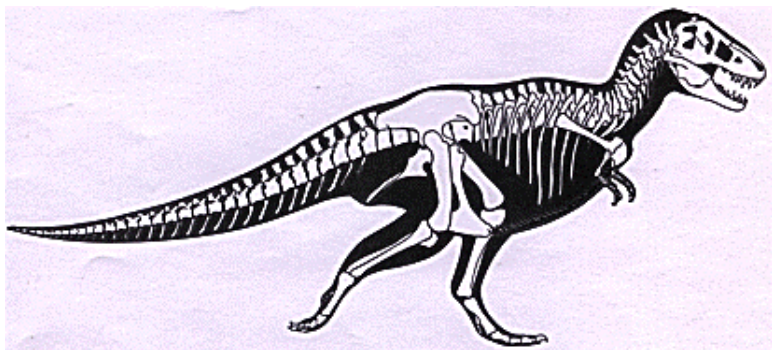
- Complex system has many interacting components (10^{11} neurons, 10^{4} types of proteins, 10^{6} routers, 10^{9} web pages)
- All components are different from each other
- Systems traditionally studied by physics also have many interacting components (10^{23} electrons in a superconductor)
- But they are all the same!

Networks in complex systems

- The simplest question about a complex system: who interacts with whom?
- The answer can be visualized as a network
- Network is the backbone of a complex system

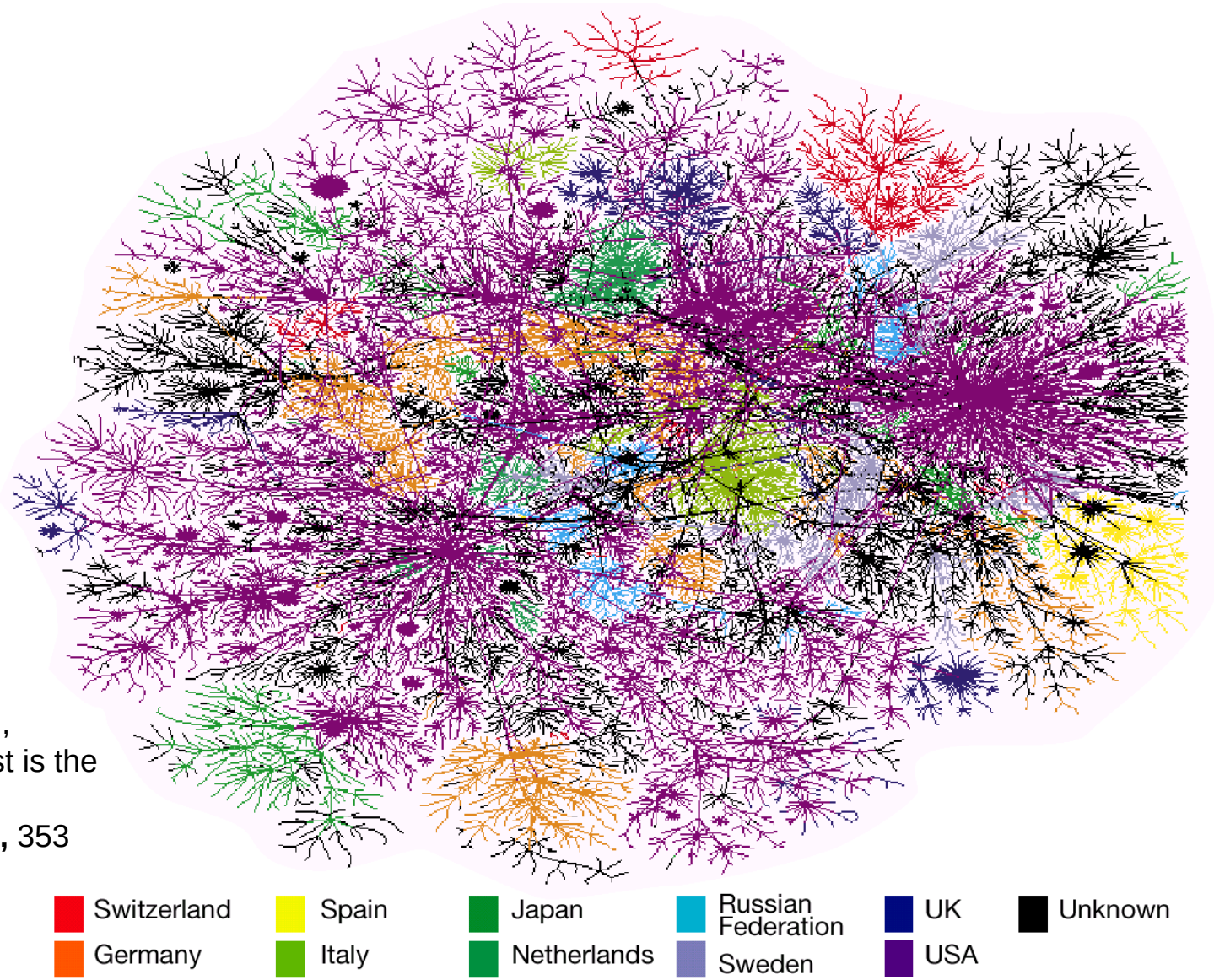
Why study the topology of complex networks?

- Lots of easily available data: that's where the state of the art information is (at least in biology)
- Large networks may contain information about basic design principles and/or evolutionary history of the complex system
- This is similar to paleontology: learning about an animal from its backbone



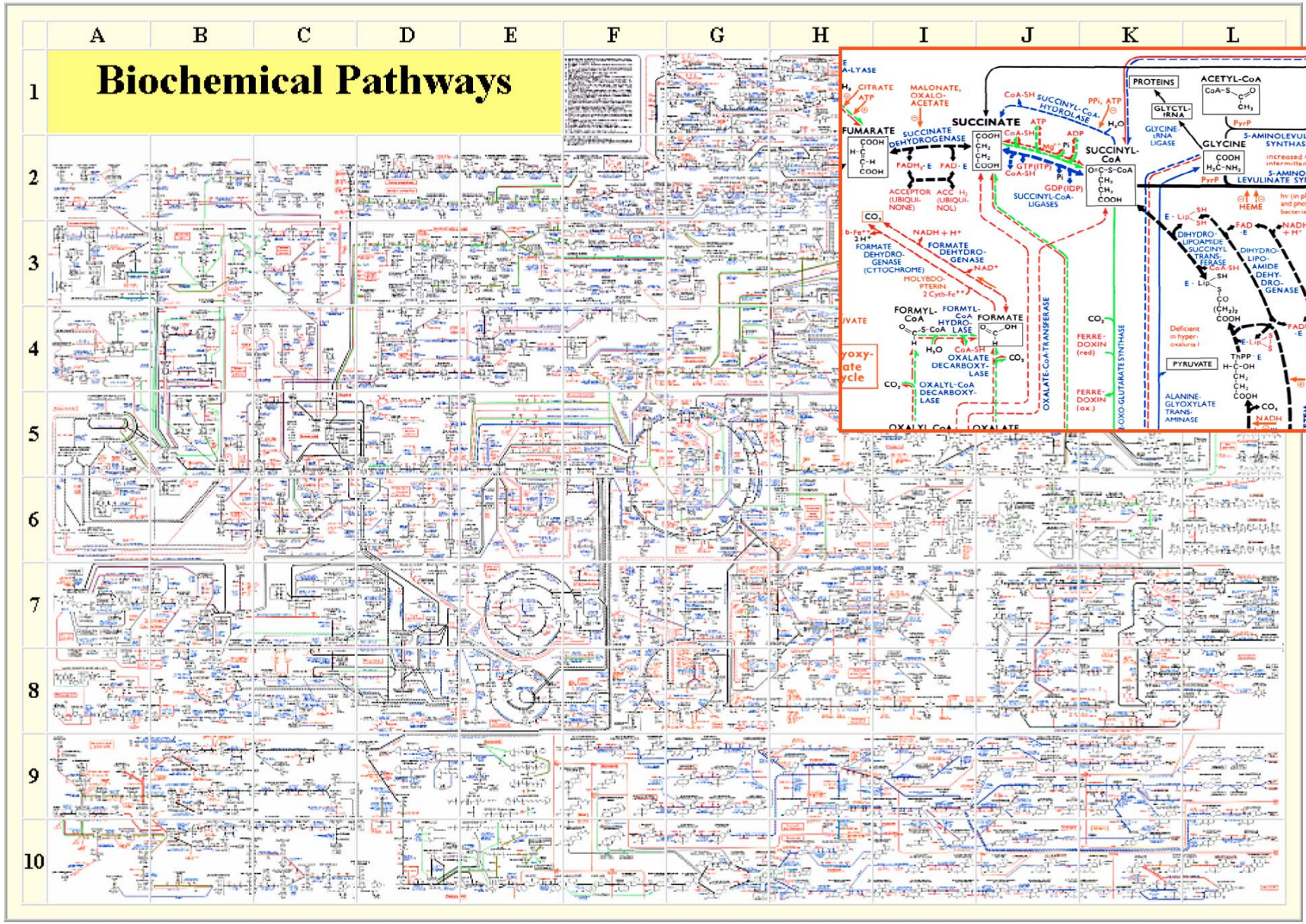
Complex networks are the right description when things are interconnected

Internet



From Y. Tu, "How robust is the Internet?" , *Nature* **406,** 353 (2000)

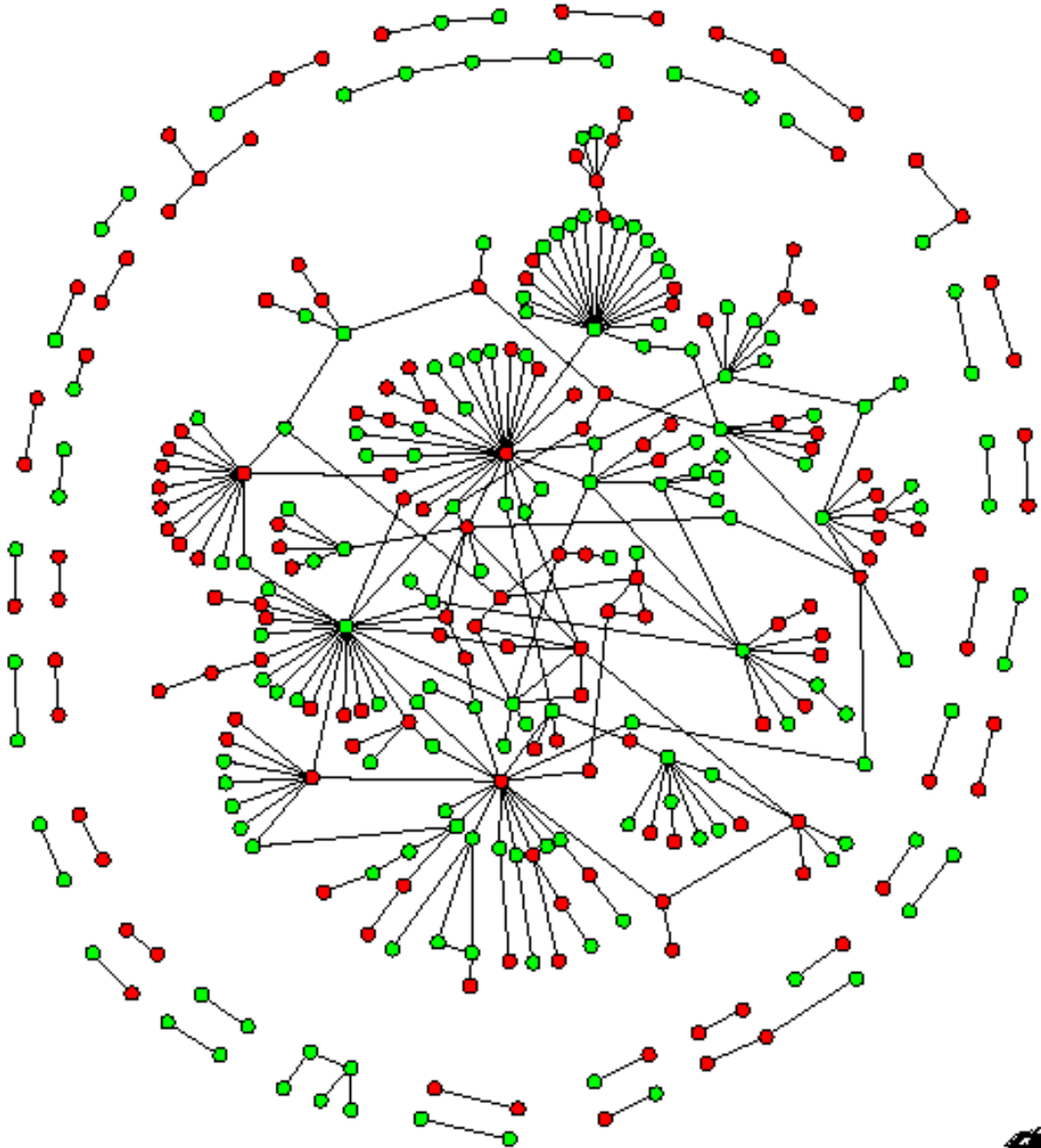
Biochemical Pathways

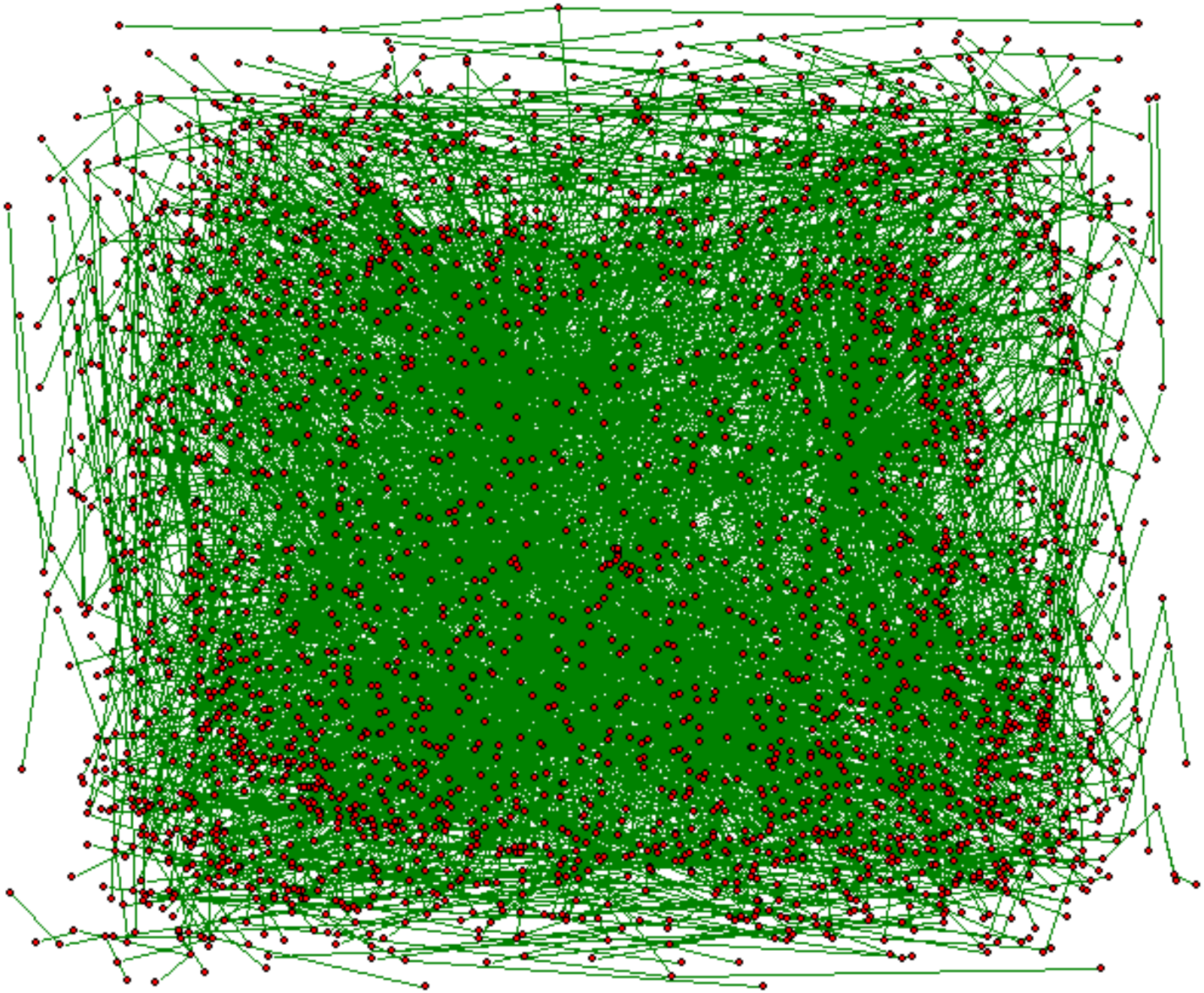


Hierarchy of bio-networks

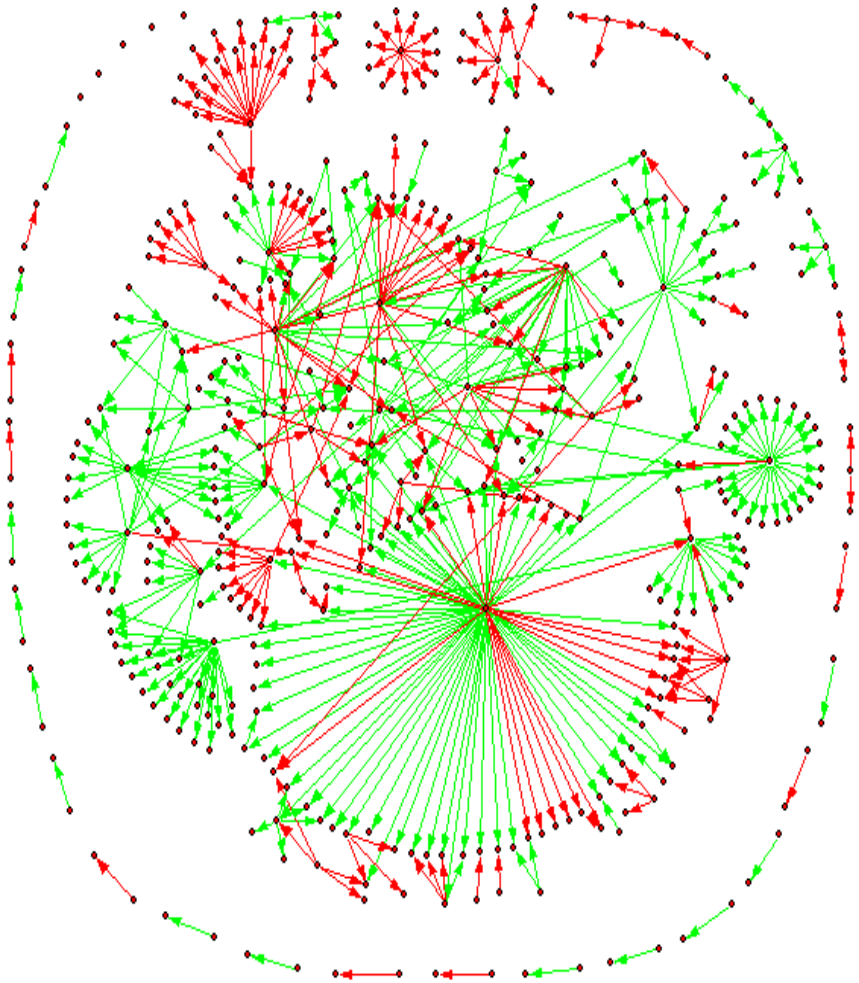
- Metabolic network: production of necessary chemical compounds
- Binding network: enzymes bind to their substrates in a metabolic network and to other proteins to form complexes
- Regulatory network: turns on and off particular groups of proteins in response to signals
- HIGHER LEVELS: cell-to cell communication (e.g. neurons in a brain), food webs, social networks, etc.

Protein binding network

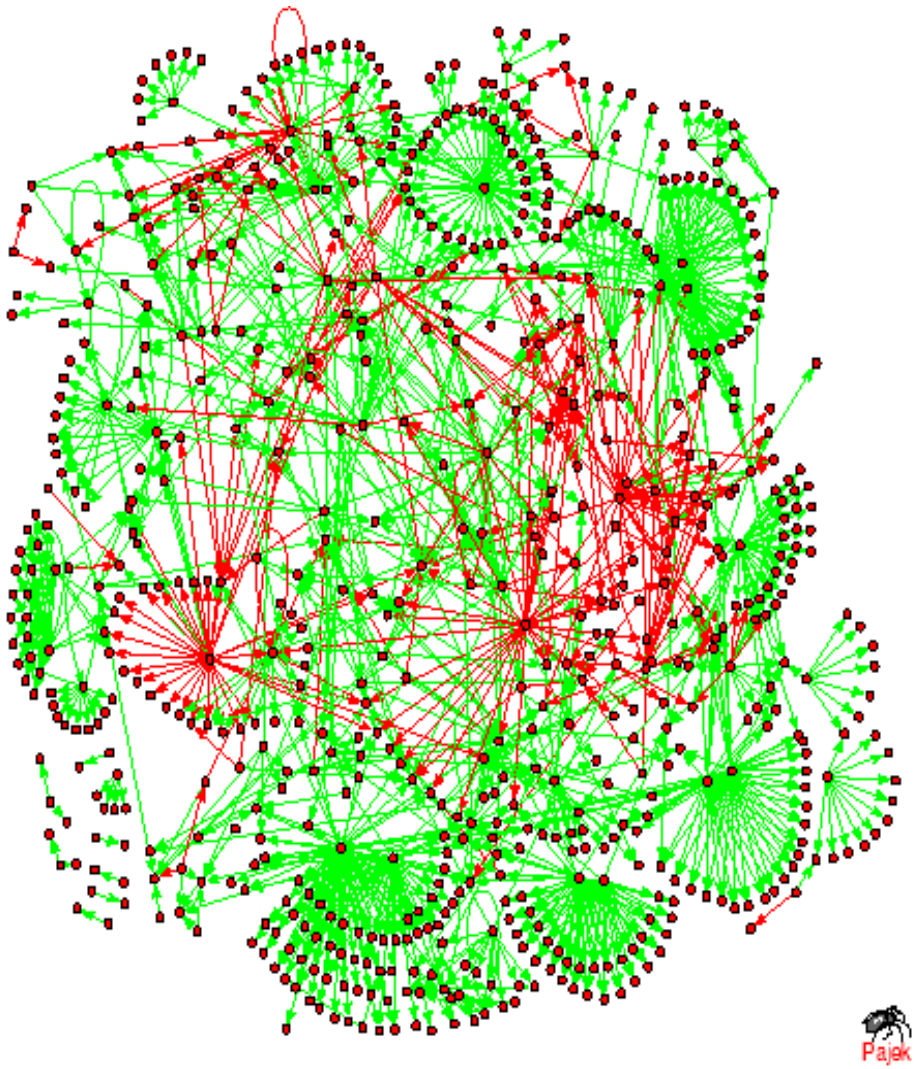




Transcription regulatory networks



Prokaryotic bacterium:
E. coli

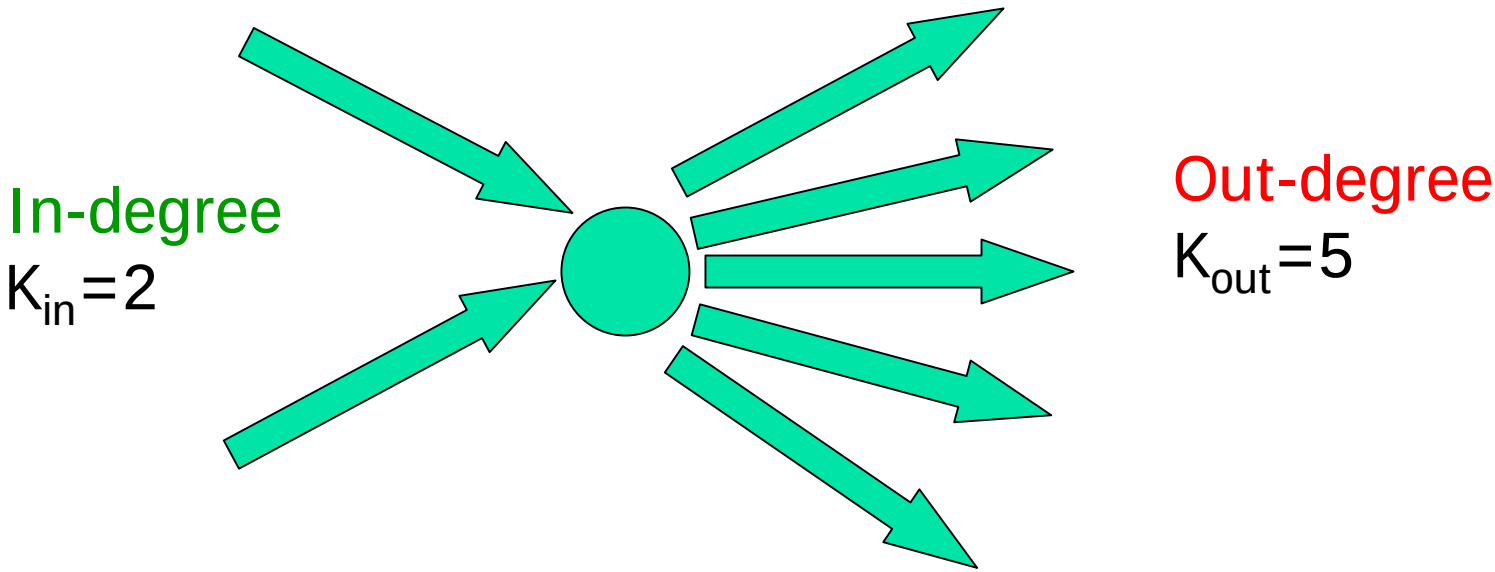


Single-celled eukaryote:
S. cerevisiae

General properties

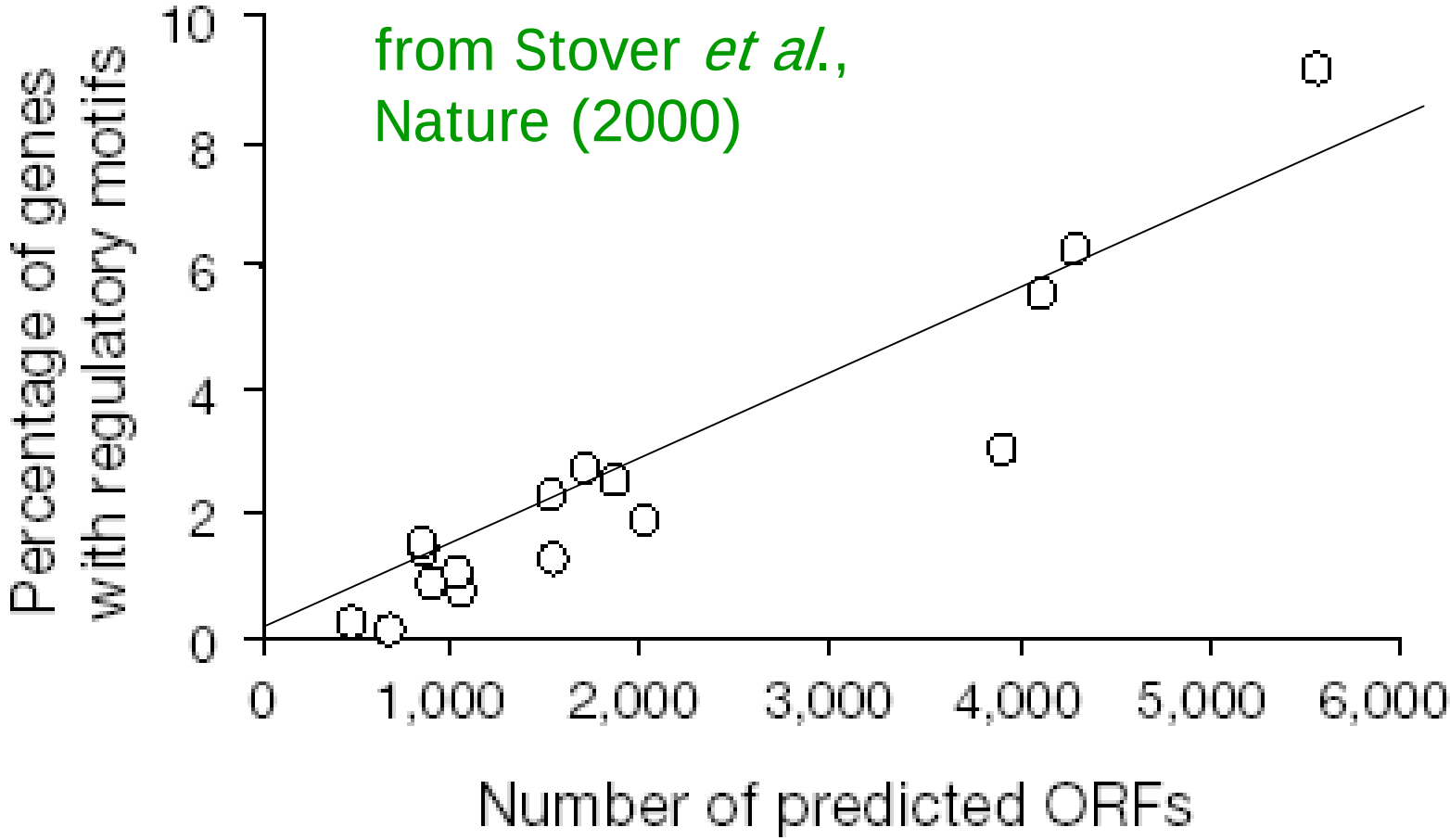
- Densely interconnected
- Not very modular: functional modules talk to each other
- Have many random features
- Few proteins (hubs) interact with a lot of neighbors: but most – with just one

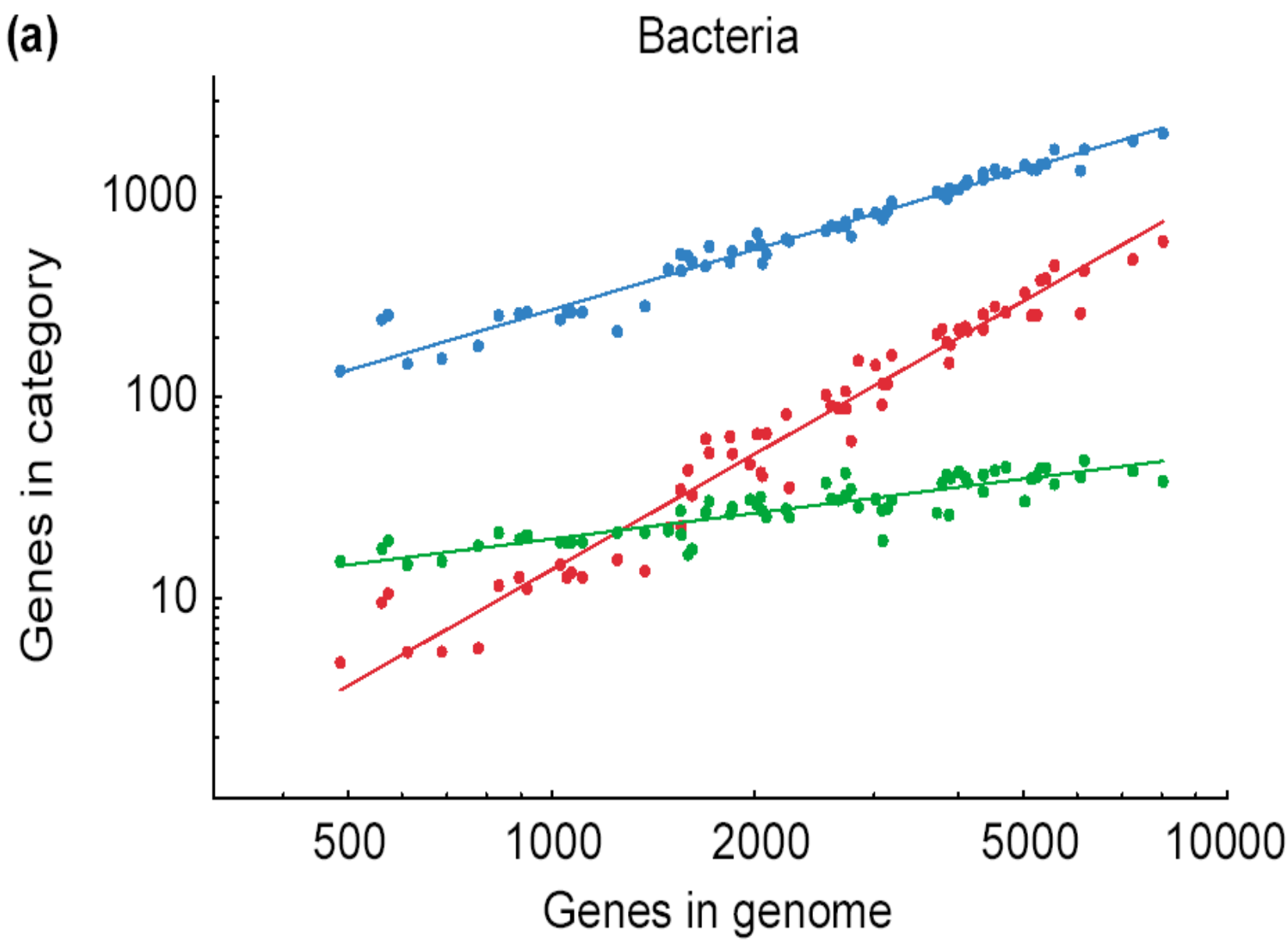
in- and out-degree of nodes



How many transcriptional regulators are out there?

Fraction of transcriptional regulators in bacteria





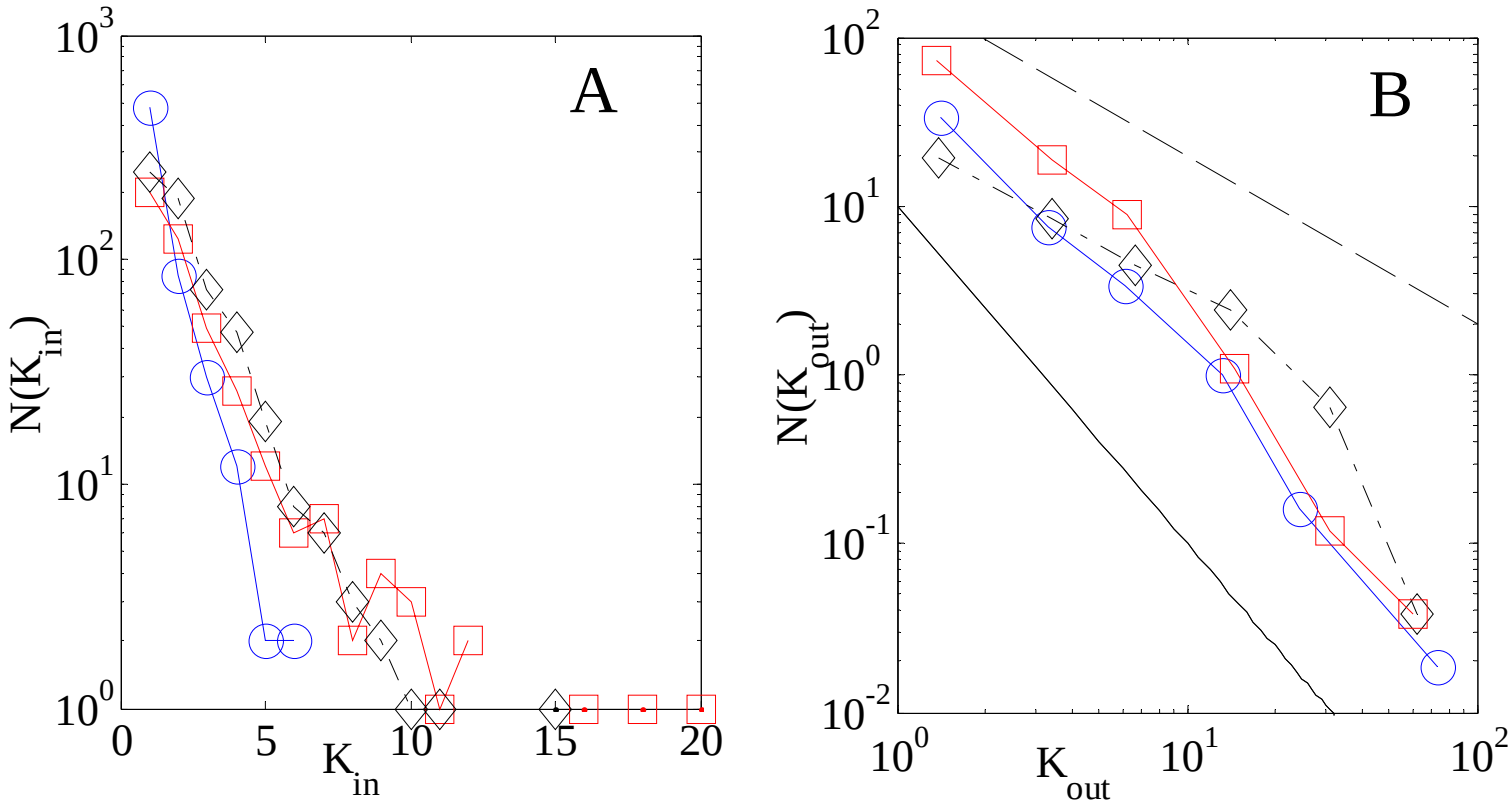
From E. van Nimwegen, Trends in Genetics, 2003

Complexity of regulation grows with complexity of organism

- $N_R \langle K_{out} \rangle = N \langle K_{in} \rangle$ = number of edges
- $N_R/N = \langle K_{in} \rangle / \langle K_{out} \rangle$ increases with N
- $\langle K_{in} \rangle$ grows with N
  - In bacteria $N_R \sim N^2$ (Stover, et al. 2000)
  - In eucaryots $N_R \sim N^{1.3}$ (van Nimwengen, 2002)
- Networks in more complex organisms are more interconnected then in simpler ones
- Life is not just a bunch of independent modules!

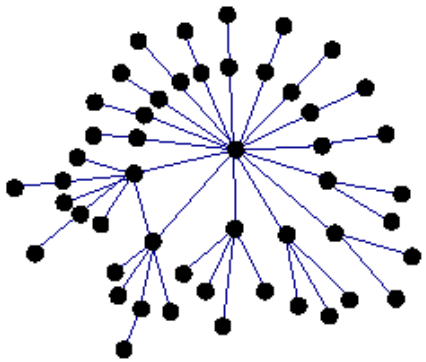
Complexity is manifested in K_{in} distribution

E. coli vs. S. cerevisiae vs. H. sapiens

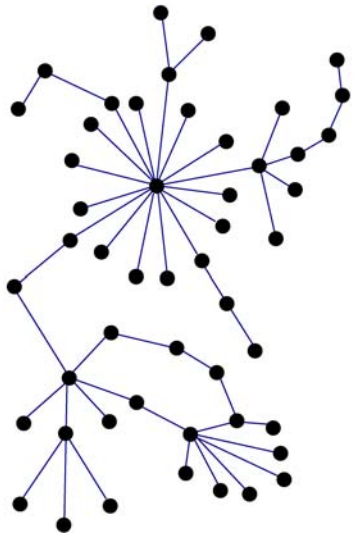


Beyond degree distributions: How is it all wired together?

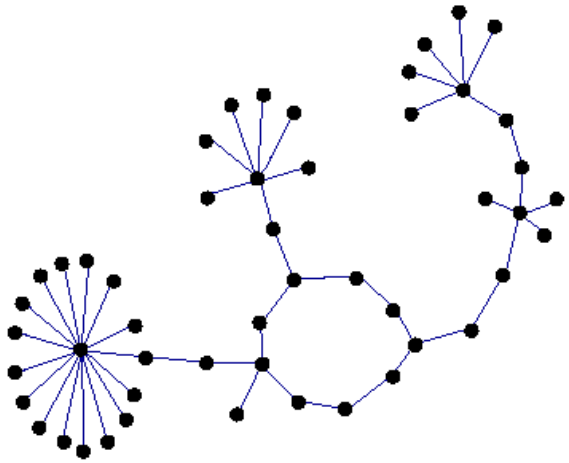
Central vs peripheral network architecture



central
(hierarchical)



random



peripheral
(anti-hierarchical)

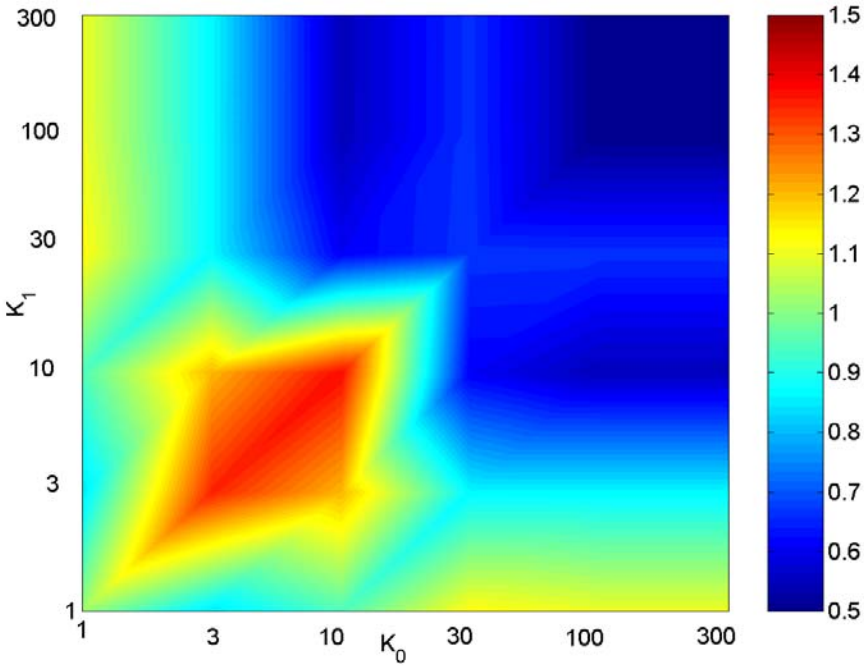
From A. Trusina, P. Minnhagen, SM, K. Sneppen, *Phys. Rev. Lett. (2004*)

Correlation profile

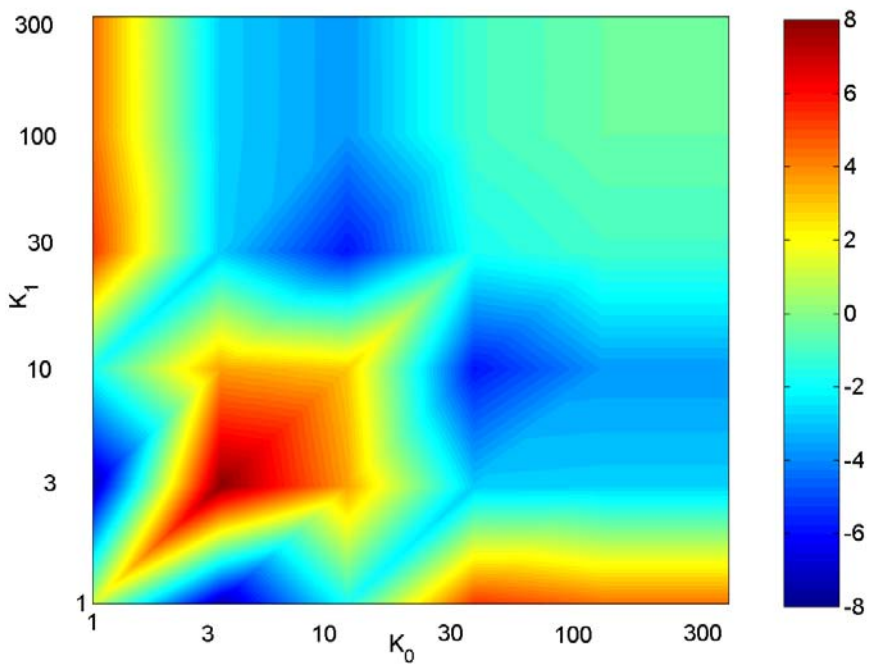
- Count $N(k_0,k_1)$ – the number of links between nodes with connectivities k_0 and k_1
- Compare it to $N_r(k_0,k_1)$ – the same property in a random network
- Qualitative features are very noise-tolerant with respect to both false positives and false negatives

Correlation profile of the protein interaction network

$R(k_0,k_1)=N(k_0,k_1)/N_r(k_0,k_1)$

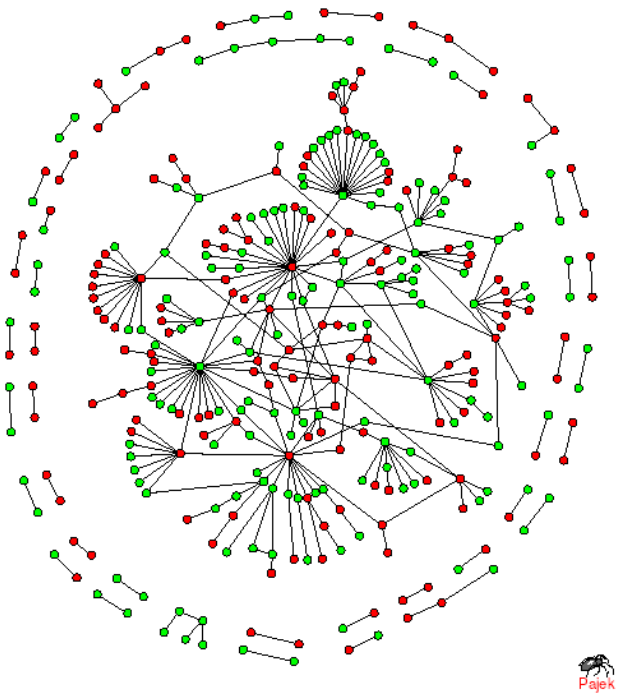


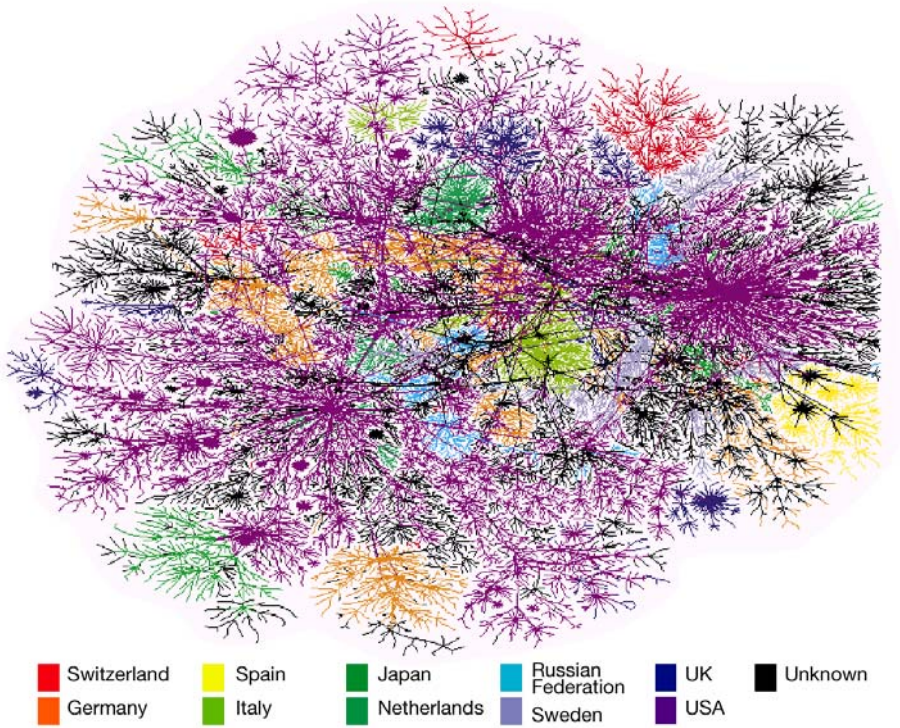
$Z(k_0,k_1) =(N(k_0,k_1)-N_r(k_0,k_1))/\Delta N_r(k_0,k_1)$



Similar profile is seen in the yeast regulatory network

Some scale-free networks may appear similar

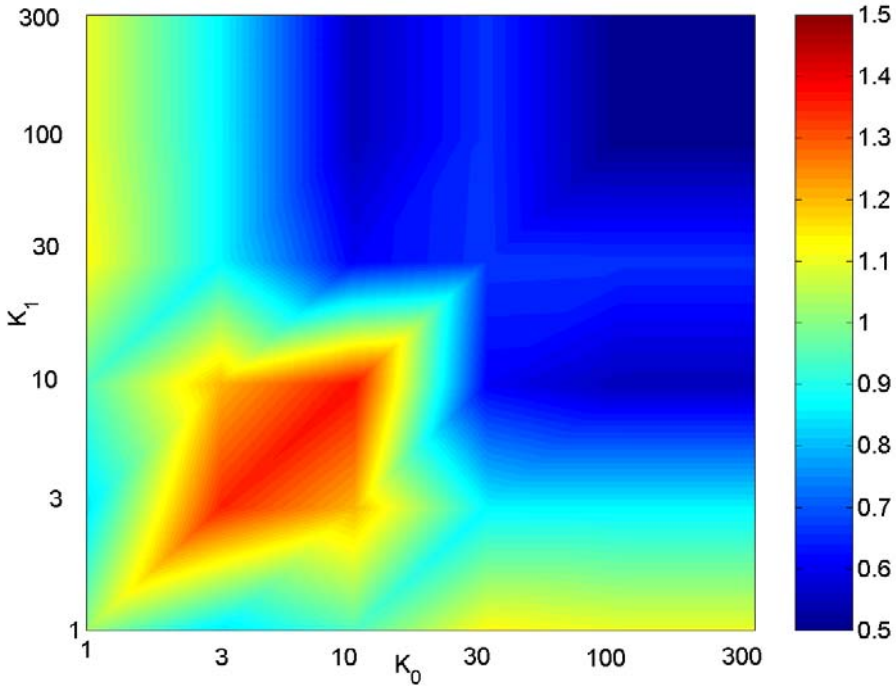




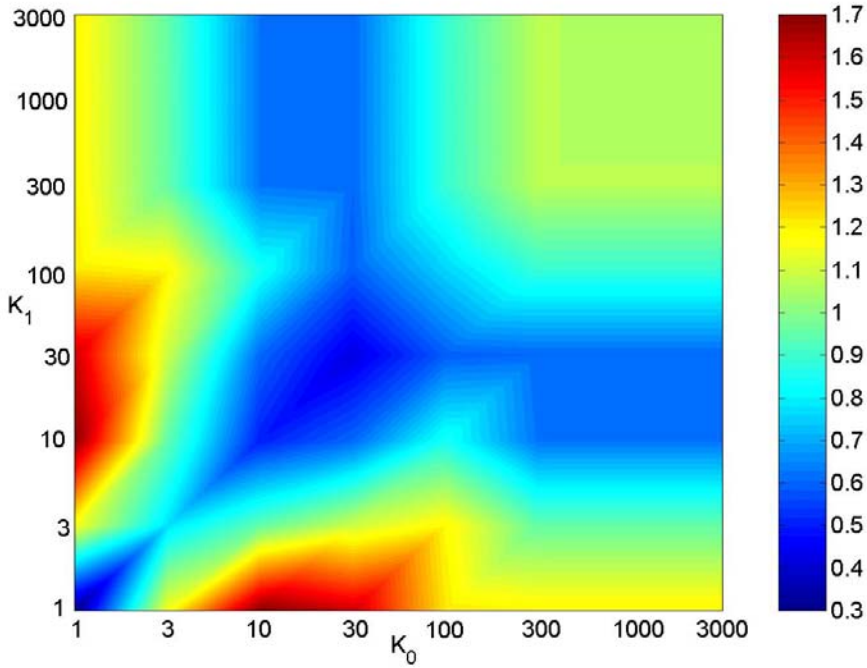
In both networks the degree distribution is scale-free $P(k) \sim k^{-\gamma}$ with $\gamma \sim 2.2\text{-}2.5$

But: correlation profiles give them unique identities

Protein interactions



Internet

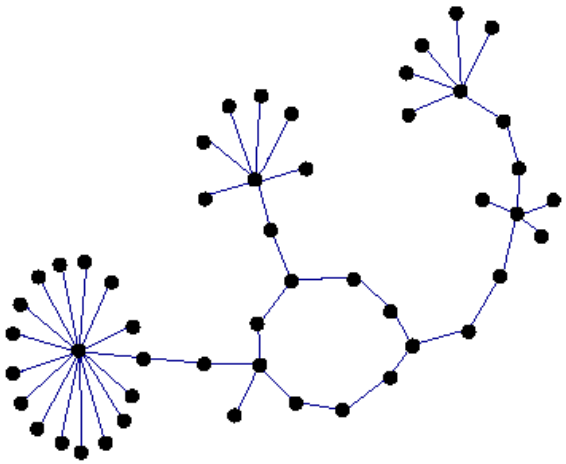


How to construct a proper random network?

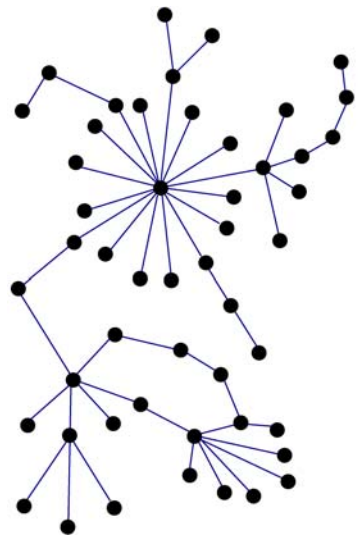
Null-model of a network

- Distribution of degrees is non-random: the degree of every node has to be conserved in a random network
- Other topological properties may be also conserved as well:
  - The extent of modularity (by function, sub-cellular localization, etc.)
  - Small motifs (e.g feed-forward loops)

Randomization

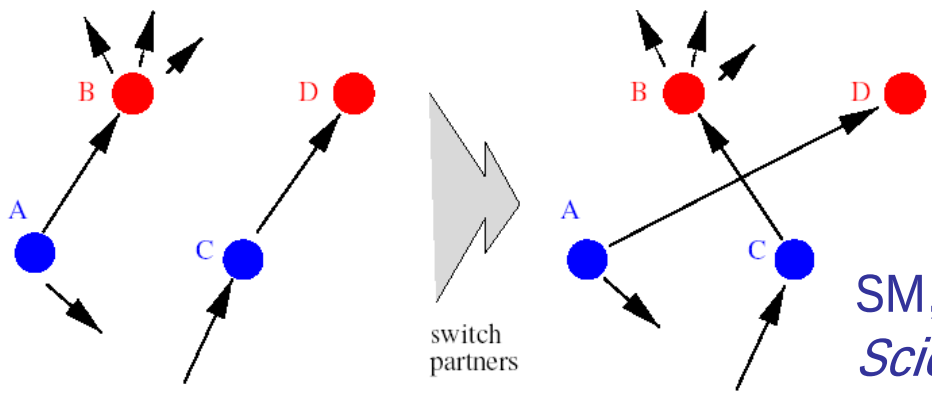


given complex network



random

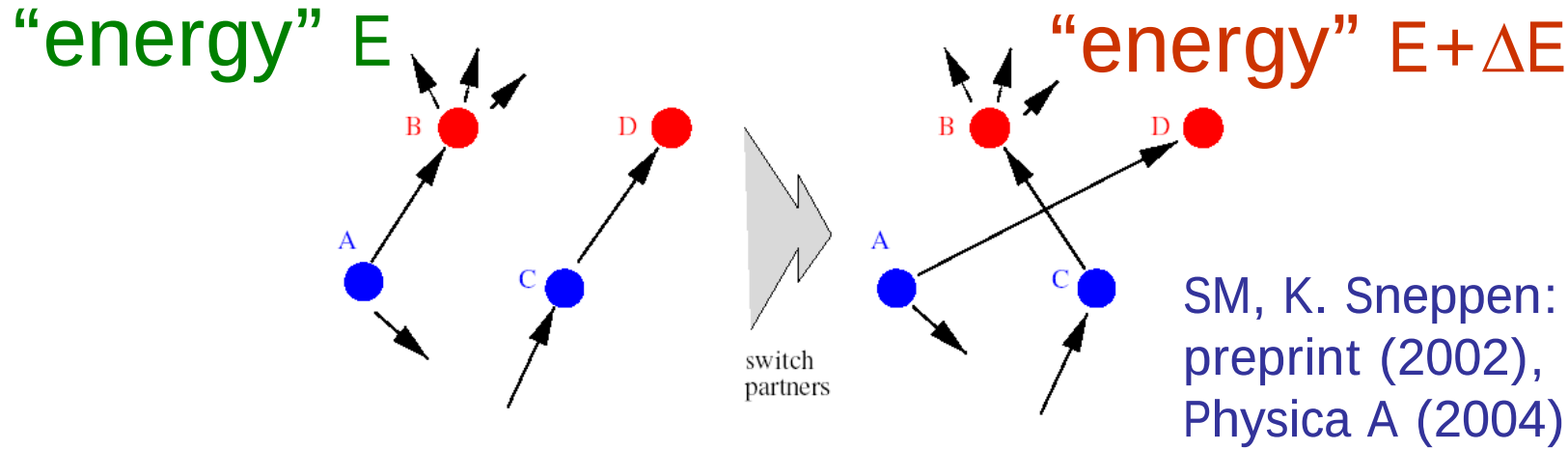
Edge swapping (rewiring) algorithm



SM, K. Sneppen, *Science* (2002)

- Randomly select and rewire two edges
- Repeat many times

Metropolis rewiring algorithm



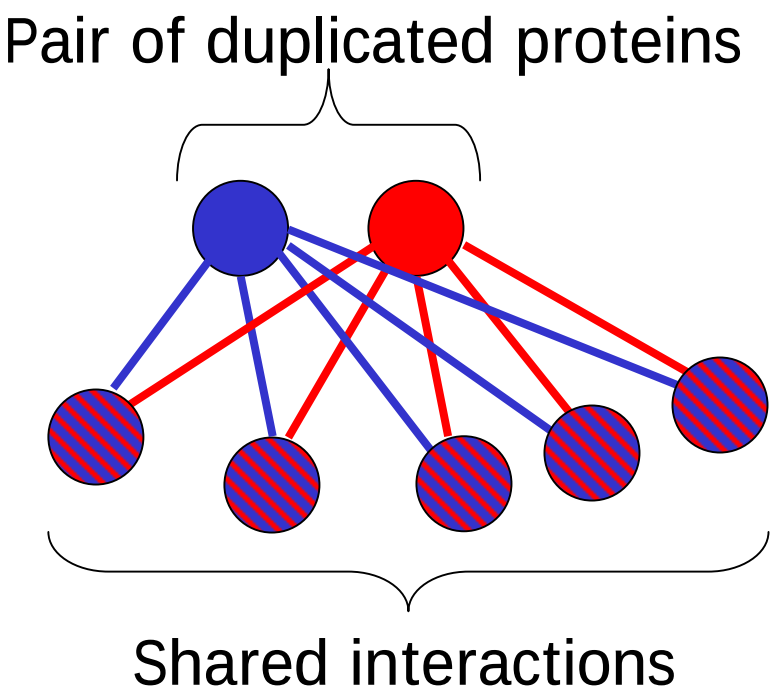
SM, K. Sneppen: preprint (2002), Physica A (2004)

- Randomly select two edges
- Calculate change ΔE in "energy function" $E=(N_{actual}-N_{desired})^2/N_{desired}$
- Rewire with probability $p=\exp(-\Delta E/T)$

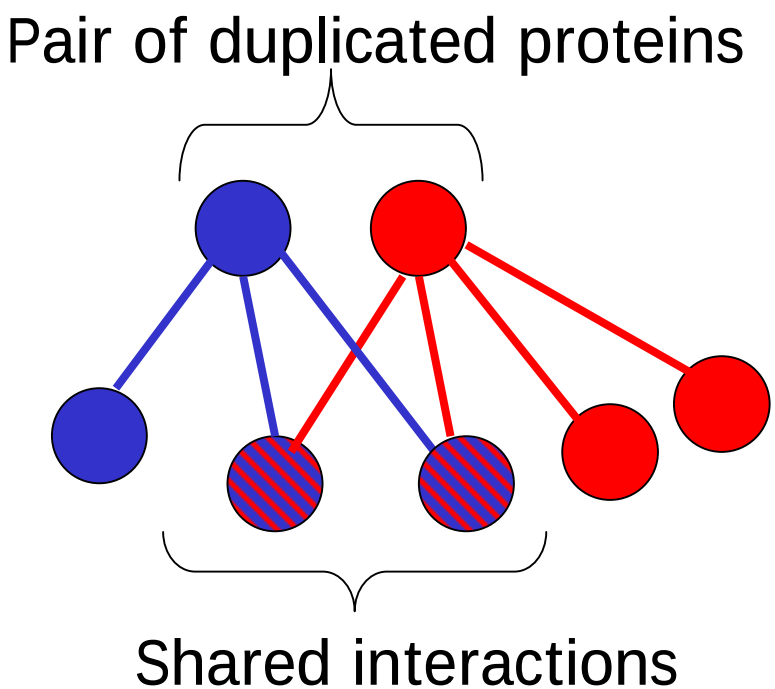
How do protein networks evolve?

Gene duplication

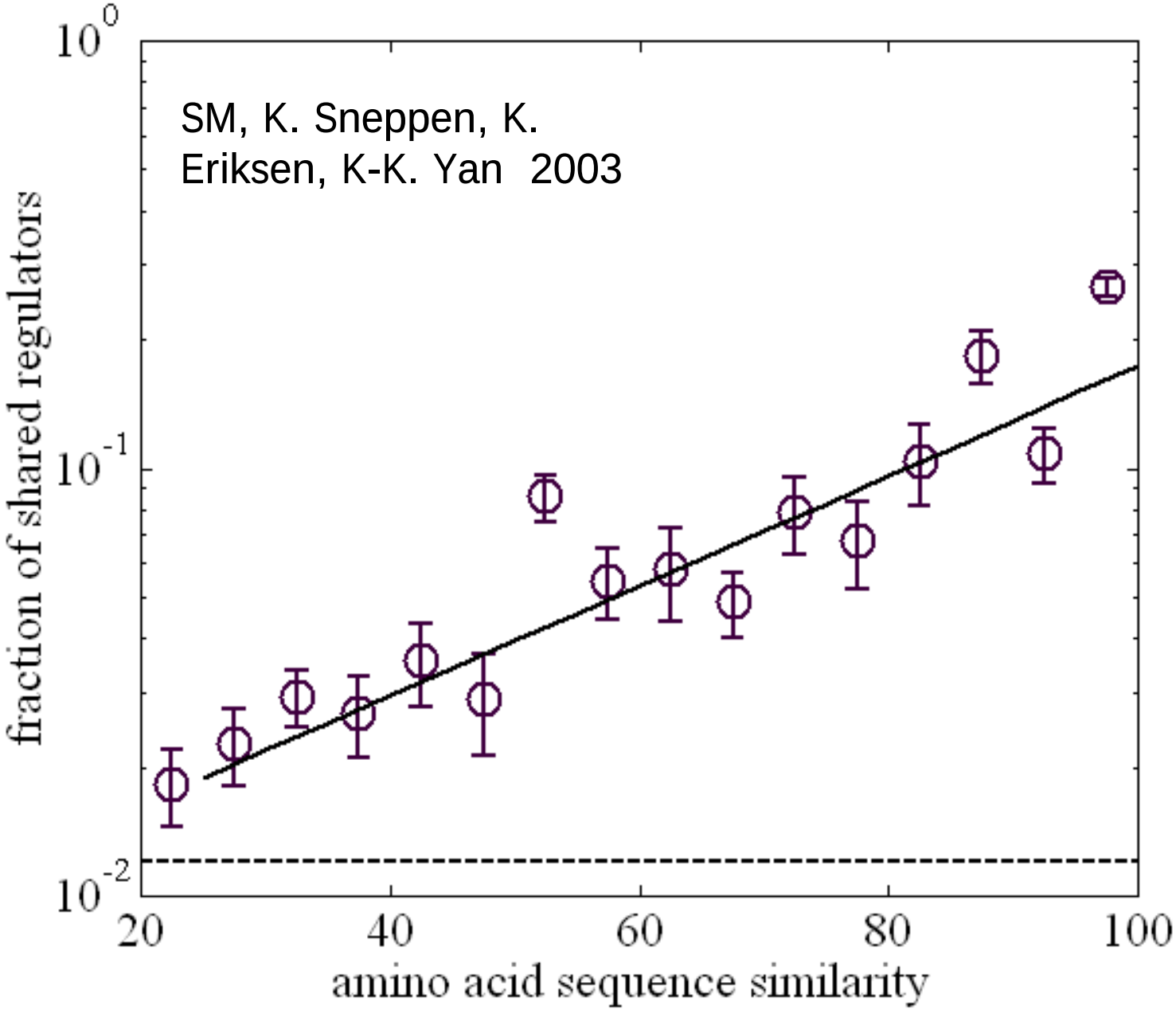
Right after duplication

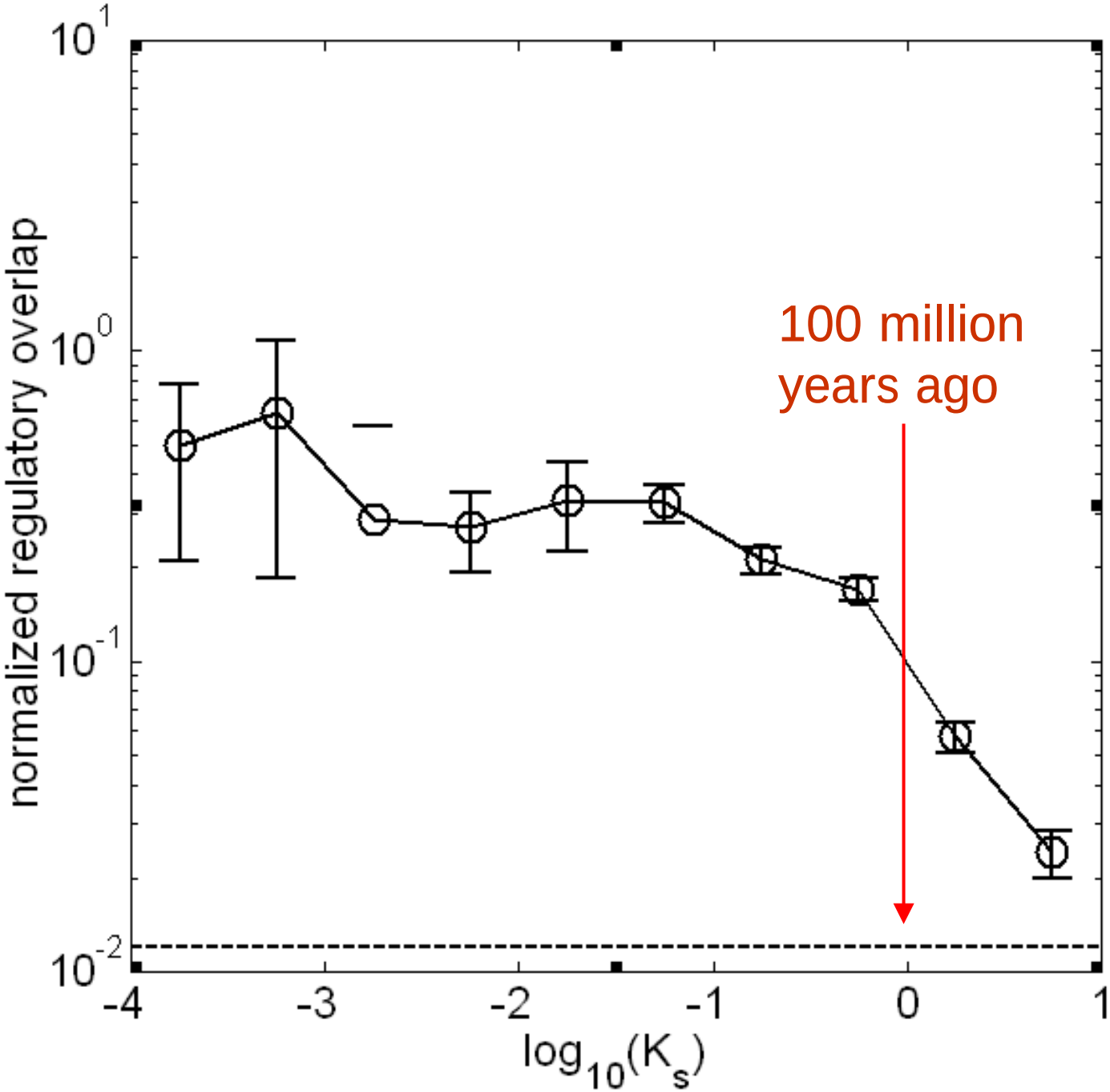


After some time



Yeast regulatory network




100 million
years ago
normalized regulatory overlap
10^{1}
10^{0}
10^{-1}
10^{-2}
-4
-3
-2
-1
0
1
$\log_{10}(K_s)$

Network properties of self-binding proteins

AKA homodimers

There are just TOO MANY homodimers

	N_{dimer}	$\langle k \rangle$
yeast	179	6.6 ± 0.2
worm	89	3.3 ± 0.1
fly	160	5.9 ± 0.1
human	1045	5.7 ± 0.1

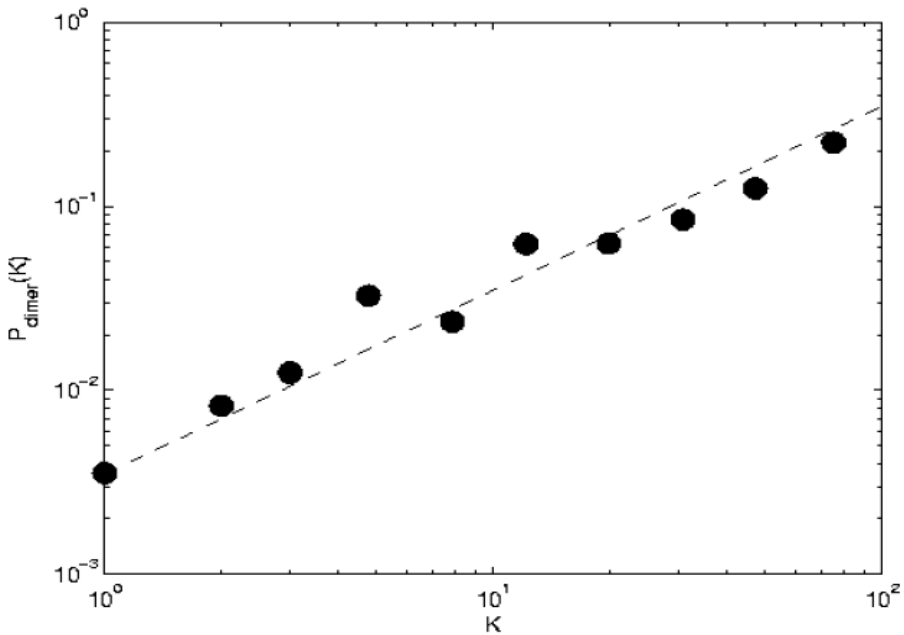
- Null-model
- $P_{self} \sim \langle k \rangle / N$
- $N_{dimer} = N \bullet P_{self} = \langle k \rangle$
- Not surprising as homodimers have many functional roles

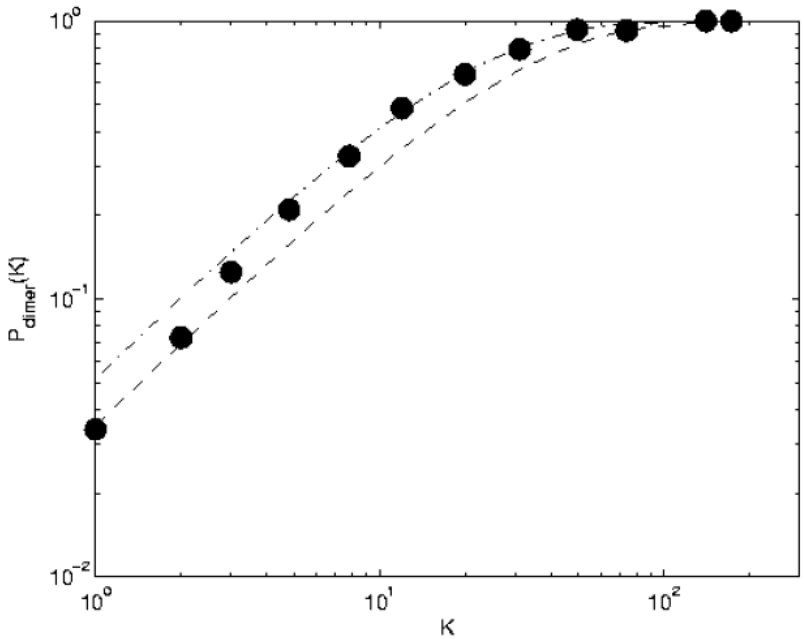
Network properties around homodimers

	$\langle k \rangle$	$\langle k \rangle_{\text{dimer}}$
yeast	6.6 ± 0.2	12.4 ± 1.2
worm	3.3 ± 0.1	13.1 ± 2.2
fly	5.9 ± 0.1	14.2 ± 1.2
human	5.7 ± 0.1	14.0 ± 0.6

Likelihood to self-interact vs. K

$$P_{dimer}(k) = 1 - (1 - p_{self})^k$$





Fly: two-hybrid data

P_{self}~0.003, P_{others}~0.0002

Human: database data

P_{self}~0.05, P_{others}~0.0002

What we think it means?

- In random networks $p_{dimer}(K) \sim K^2$ not $\sim K$ like our empirical observation
- K is proportional to the "stickiness" of the protein which in its turn scales with
  - the area of hydrophobic residues on the surface
  - # copies/cell
  - its' popularity (in datasets taken from databases)
  - etc.
- Real interacting pair consists of an "active" and "passive" protein and binding probability scales only with the "stickiness" of the active protein
- "Stickiness" fully accounts for higher than average connectivity of homodimers

Summary

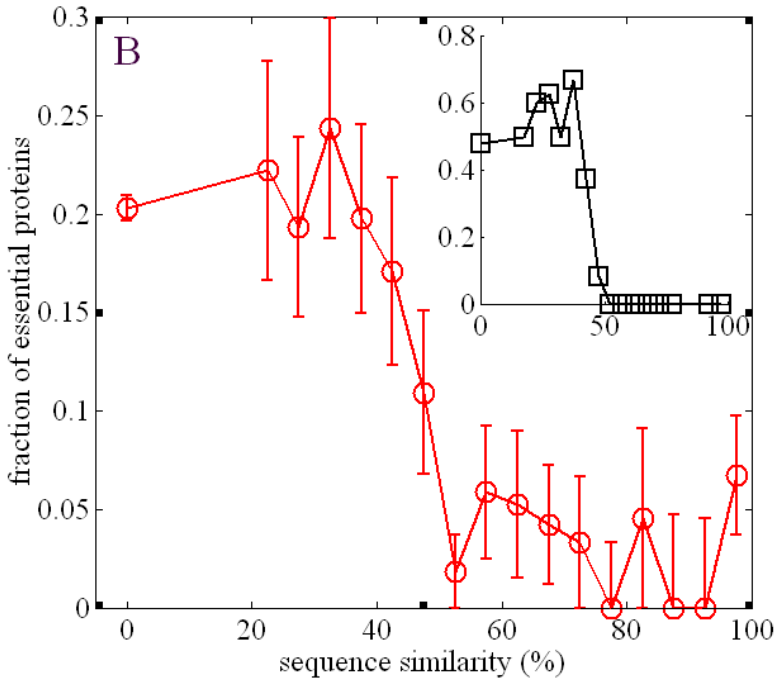
- Living cells contain many complex protein networks
- Networks in more complex organisms are more interconnected
- Most have hubs – highly connected proteins
- Hubs often avoid each other (networks are anti-hierarchical)
- Networks evolve by gene duplications
- There are many self-interacting proteins. Probability to self-interact linearly scales with the degree K.

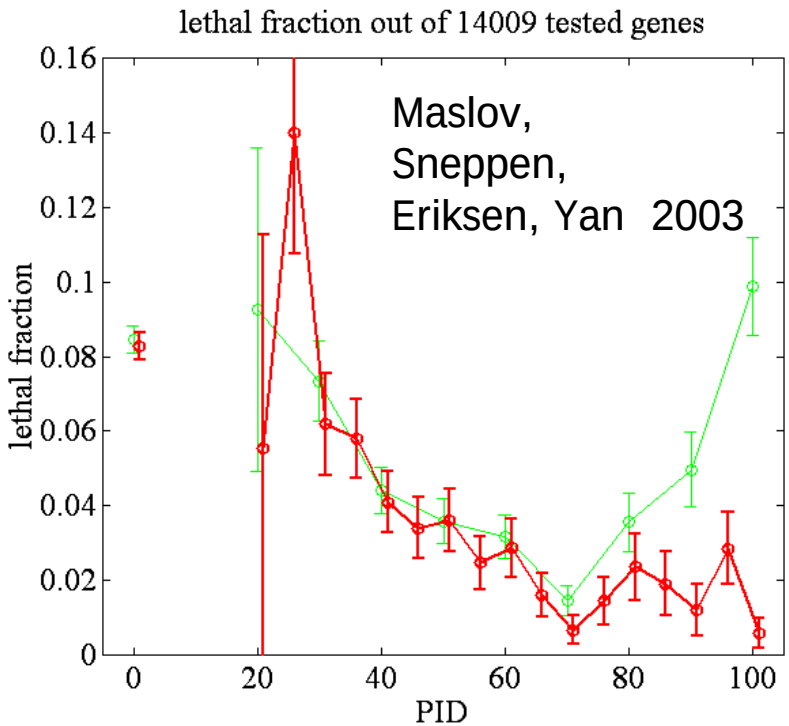
Collaborators:

- Kim Sneppen – U. of Copenhagen
- Kasper Eriksen – U. of Lund
- Koon-Kiu Yan – Stony Brook
- Ilya Mazo, Jaroslav Ispolatov, Anton Yuryev – Ariadne Genomics

THE END

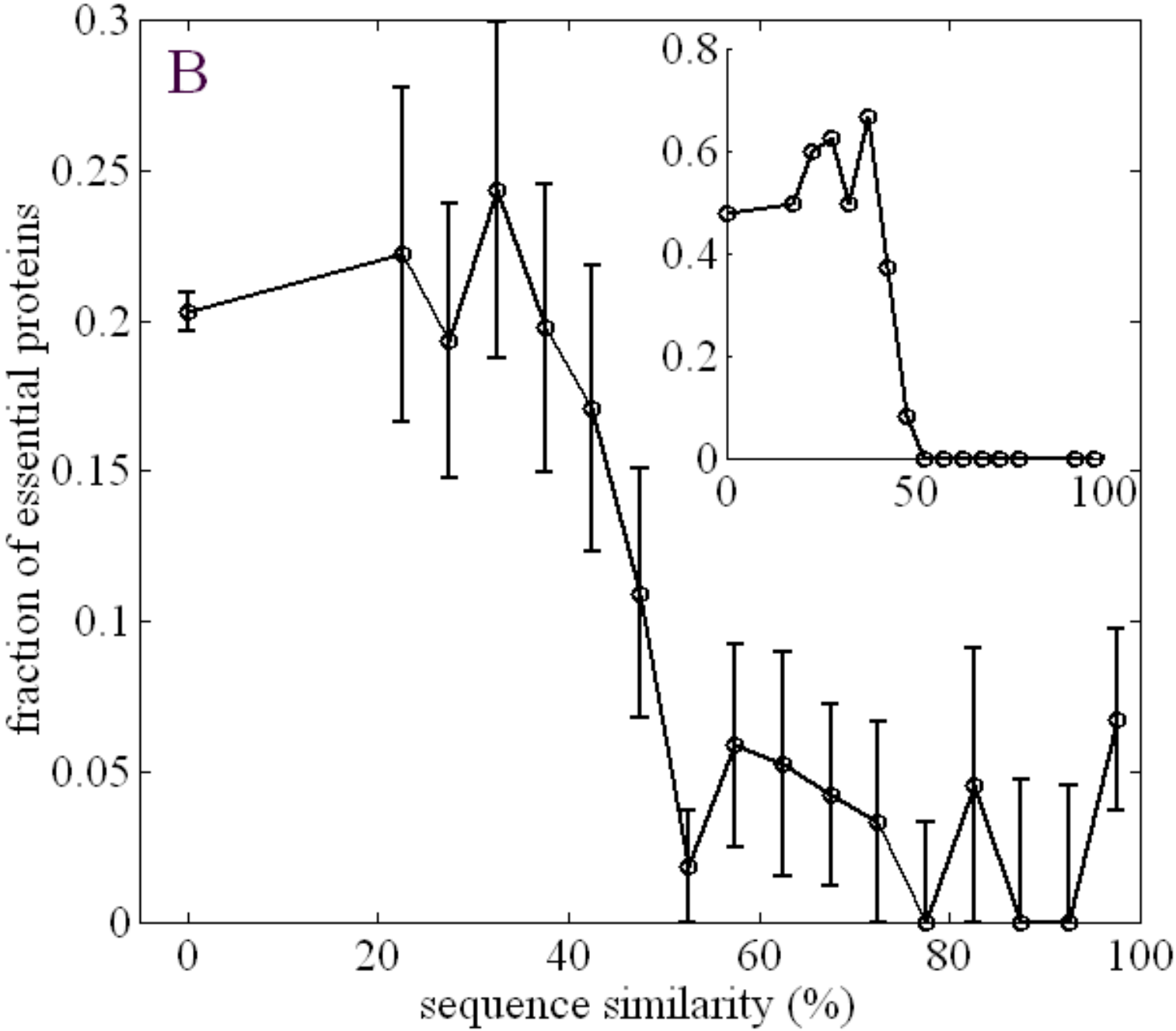
Protective effect of duplicates

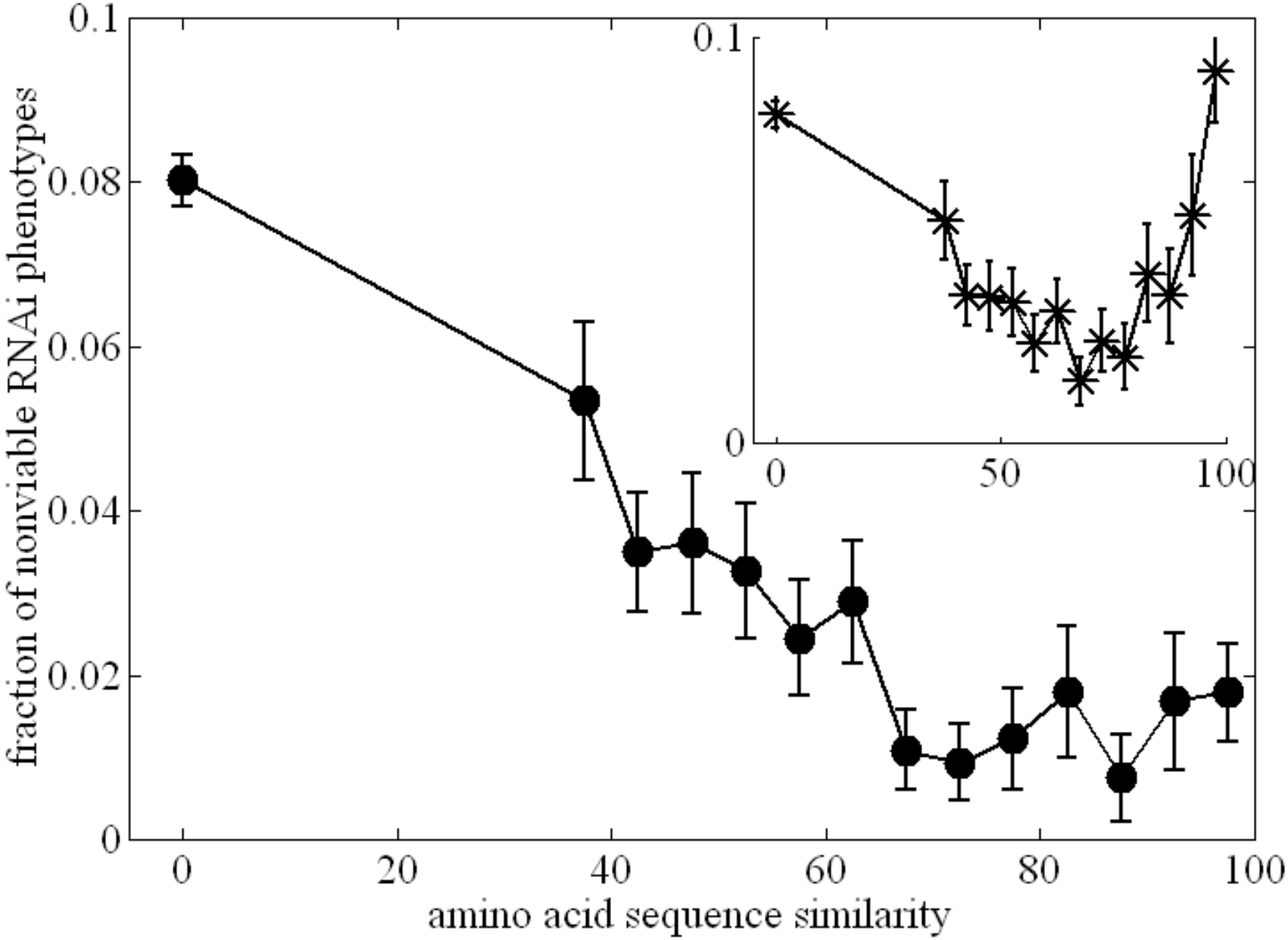


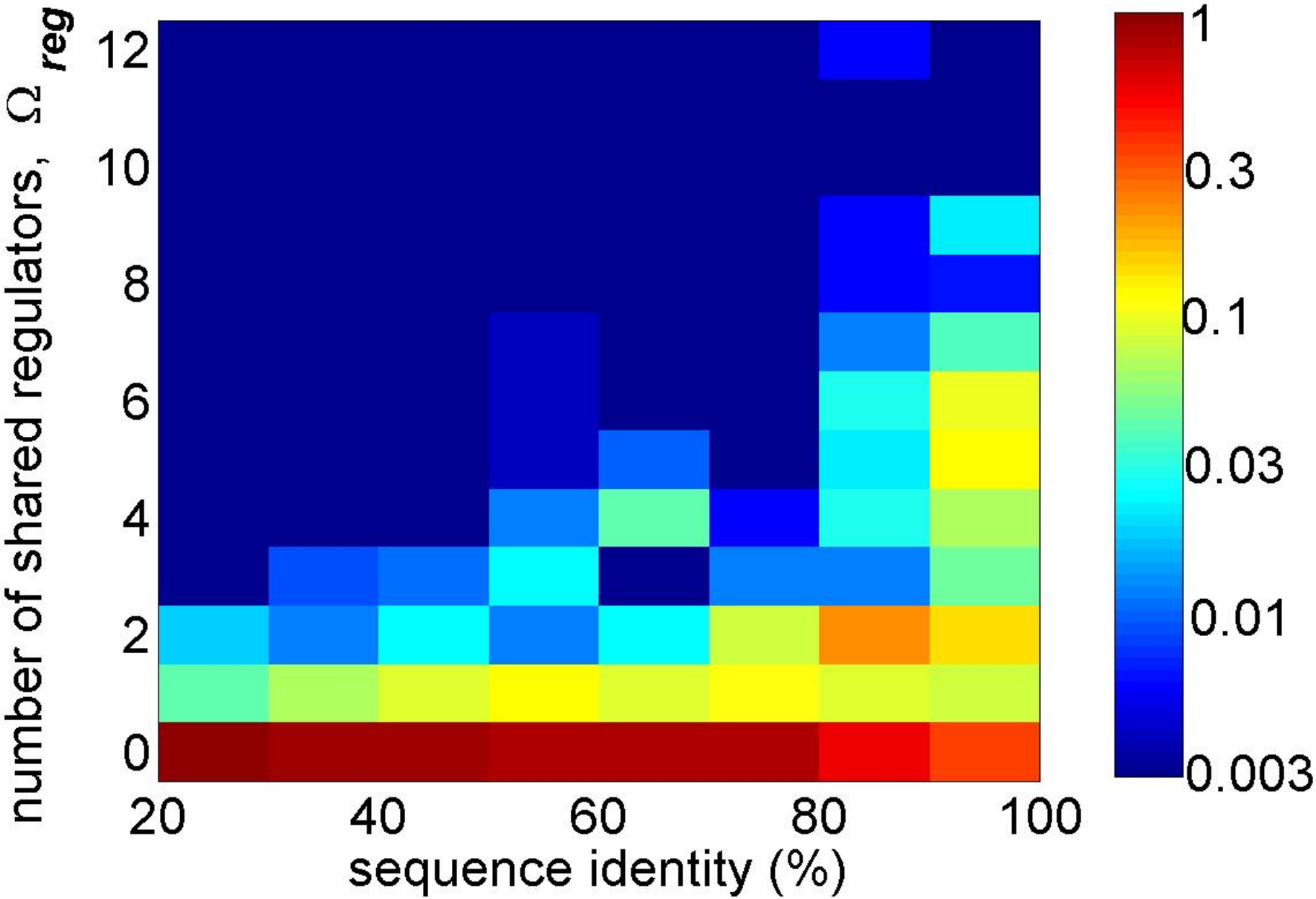


Yeast

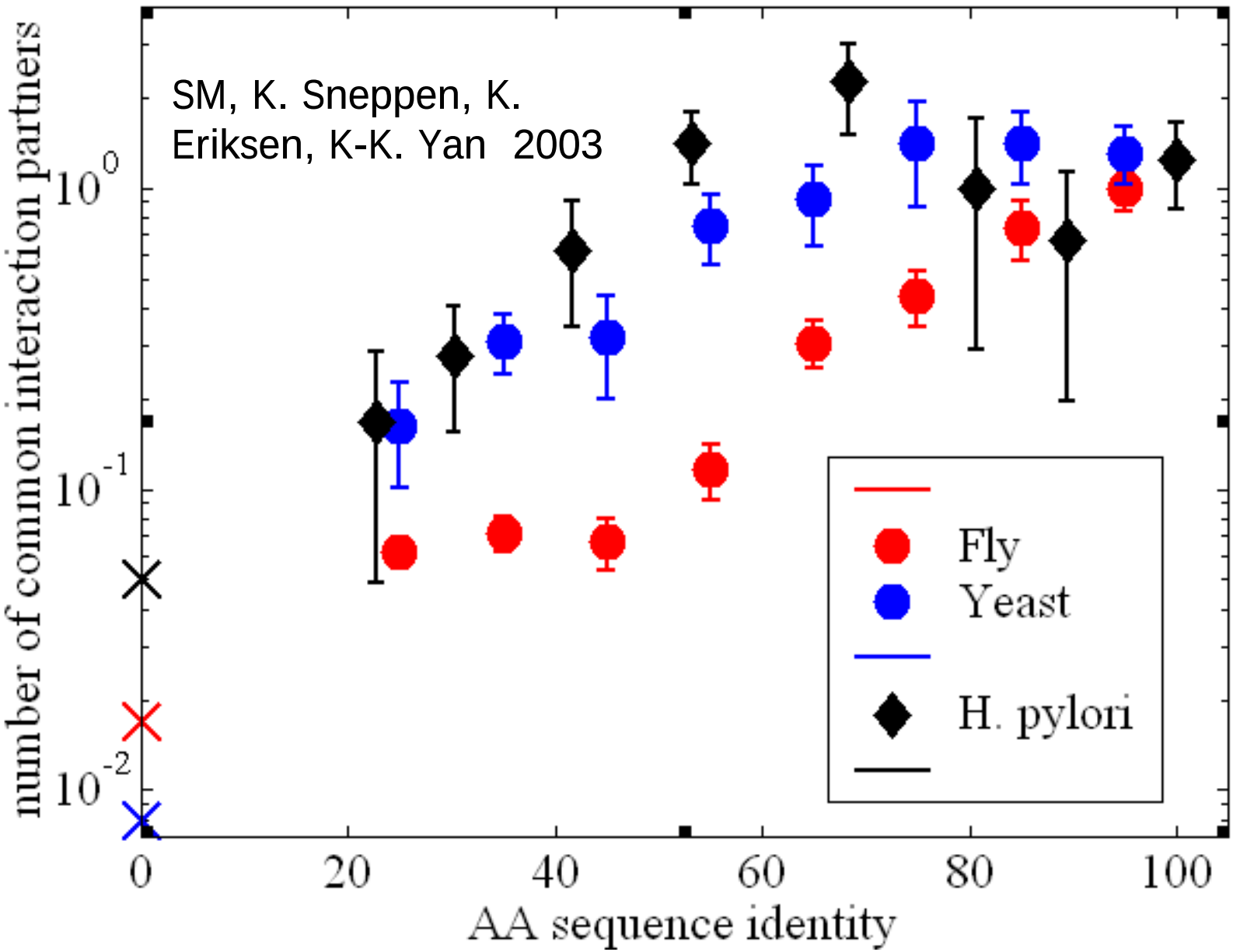
Worm


B
fraction of essential proteins
sequence similarity (%)
0
0.05
0.1
0.15
0.2
0.25
0.3
0
20
40
60
80
100
0.8
0.6
0.4
0.2
0
0
50
100


fraction of nonviable RNAi phenotypes
amino acid sequence similarity
0.1
0.08
0.06
0.04
0.02
0
0
20
40
60
80
100
0.1
0
0
50
100


number of shared regulators, Ω_{reg}
12
10
8
6
4
2
0
20
40
60
80
100
sequence identity (%)
1
0.3
0.1
0.03
0.01
0.003

Protein interaction networks



What shapes the topology of protein networks?

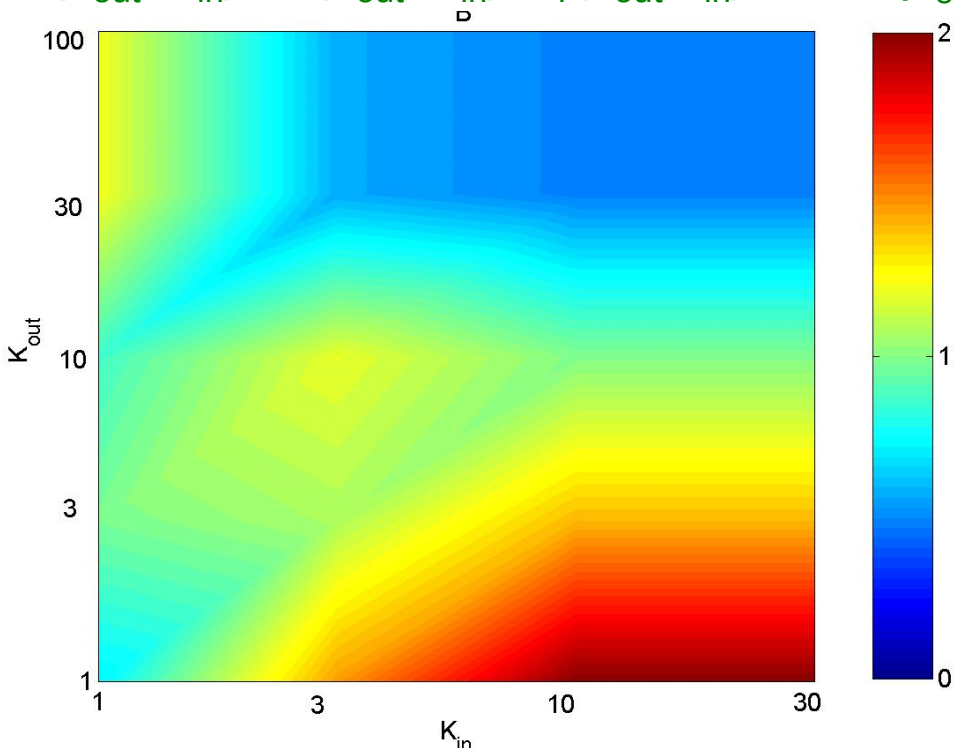
- Duplication-divergence models CAN account for their basic topological features
- BUT: functional organization dominates:
  - Most pairs with many shared interactions are NOT homologs
  - Hubs are not caused by mult. duplications
- Functional organization is more important than duplication-divergence

Genome-wide protein networks

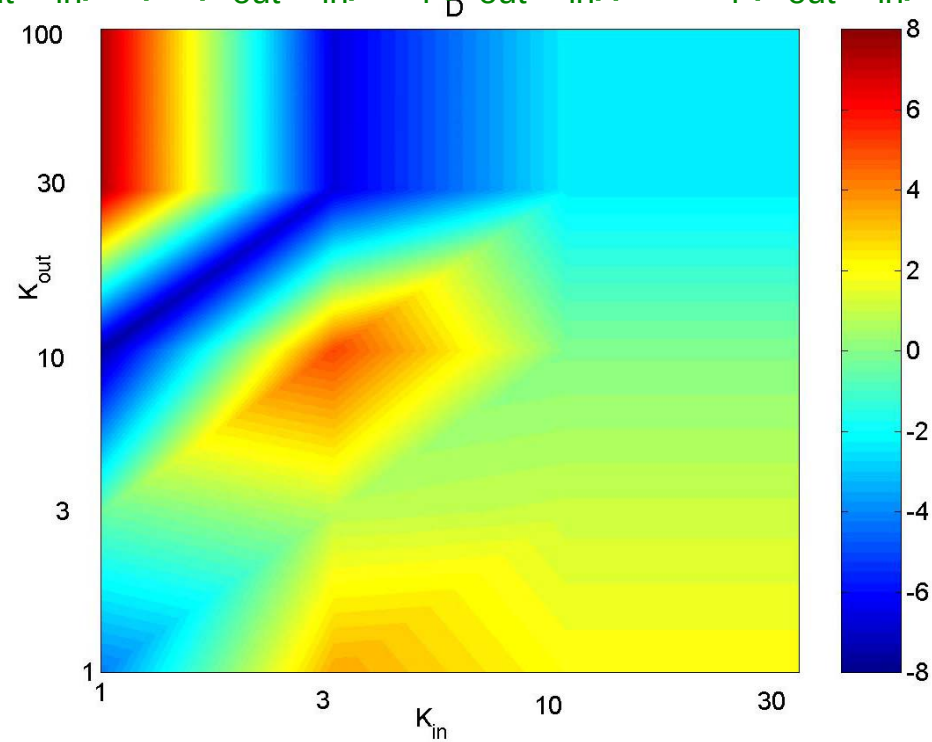
- Nodes - proteins
- Edges – interactions between proteins
  - Bindings (physical interactions)
  - Regulations (transcriptional, protein modifications, etc.)
  - Etc, etc, etc.

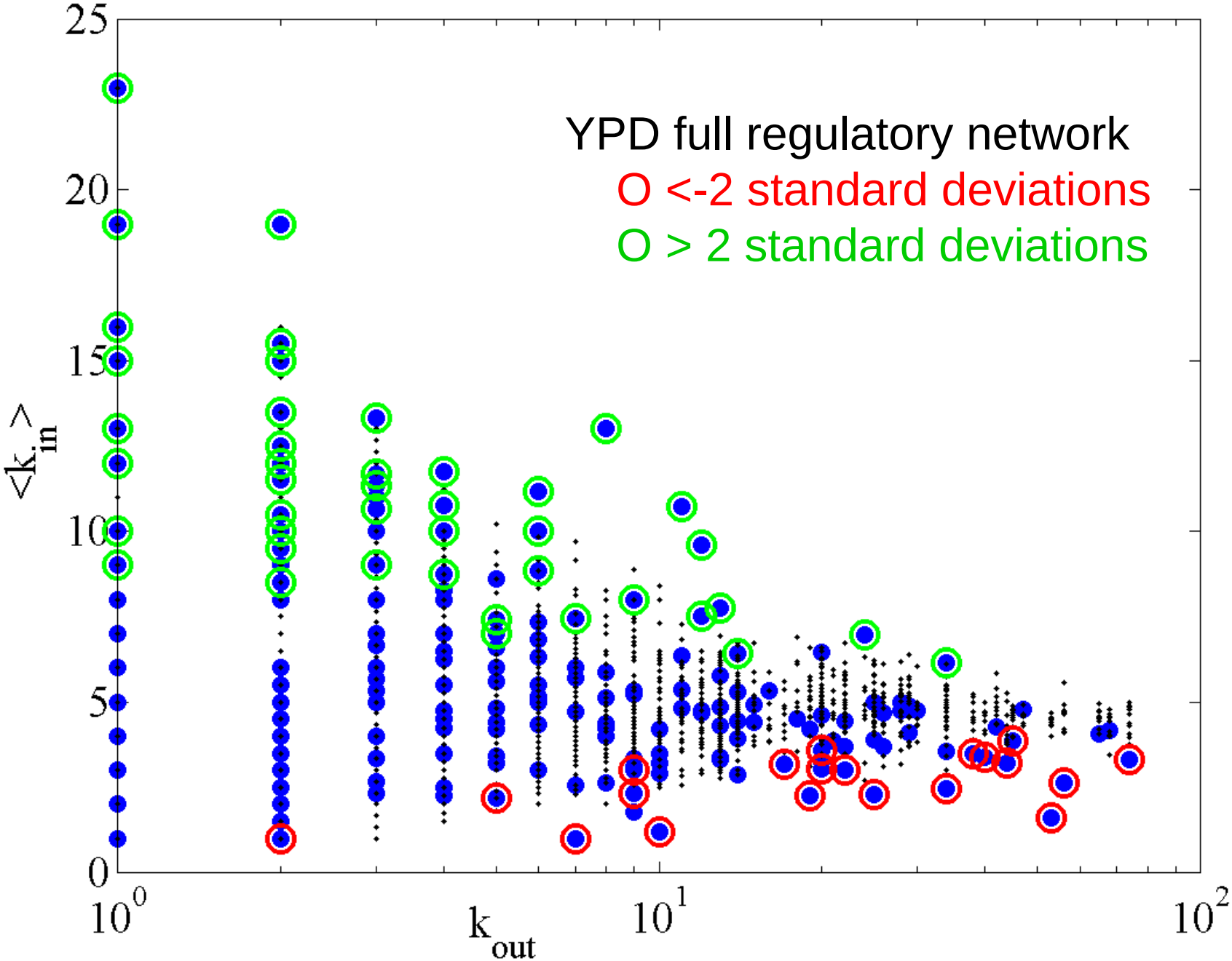
Correlation profile of the yeast regulatory network

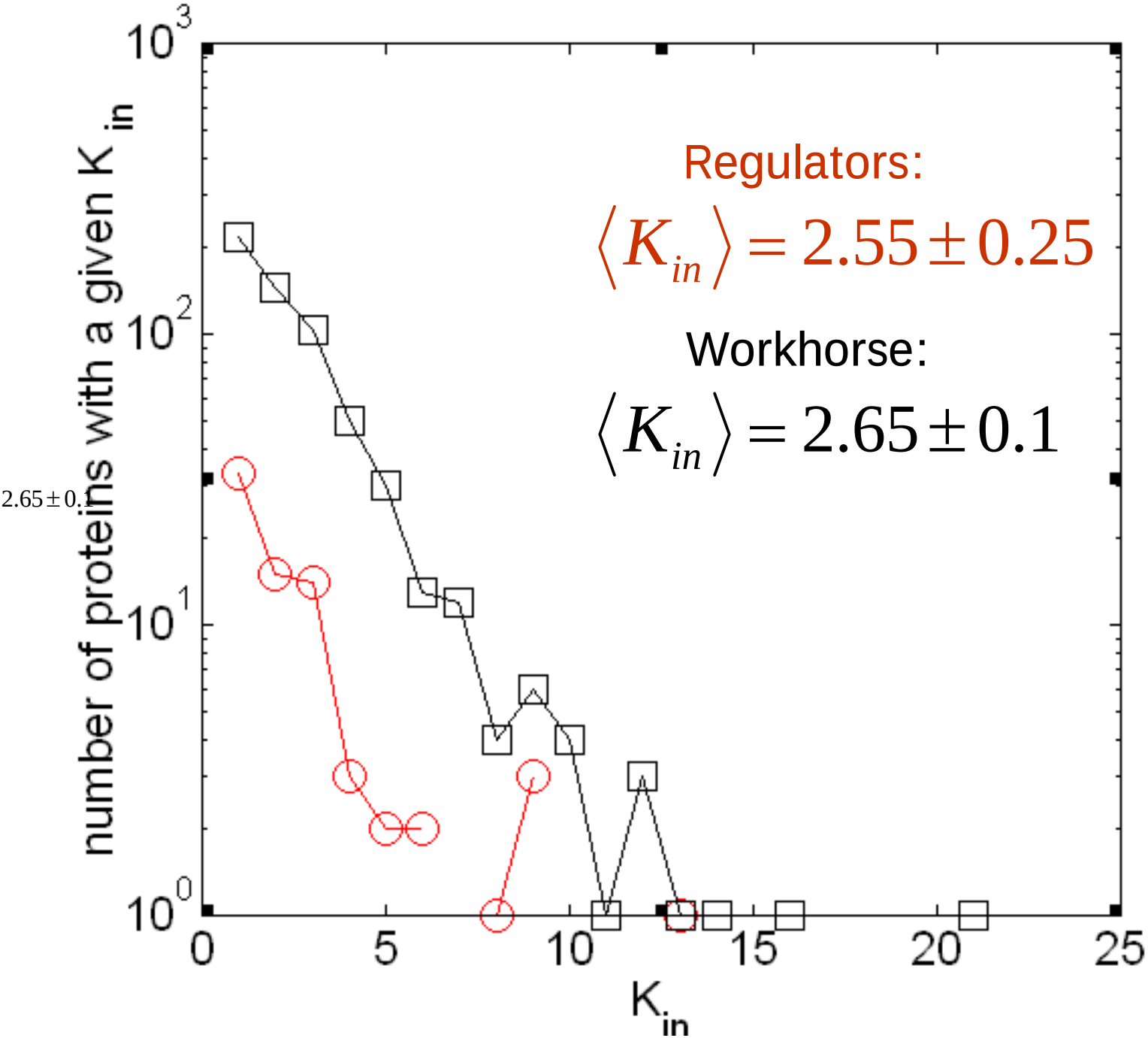
$R(k_{out}, k_{in})=N(k_{out}, k_{in})/N_r(k_{out},k_{in})$



$Z(k_{out},k_{in})=(N(k_{out},k_{in})-N_r(k_{out},k_{in}))/ \Delta N_r(k_{out},k_{in})$




YPD full regulatory network
O <-2 standard deviations
O > 2 standard deviations
25
20
15
10
5
0
$<k_{in}>$
10^0
10^1
10^2
k_{out}


Regulators:
$\langle K_{in} \rangle = 2.55 \pm 0.25$
Workhorse:
$\langle K_{in} \rangle = 2.65 \pm 0.1$
number of proteins with a given K_{in}
K_{in}
10^0
10^1
10^2
10^3
0
5
10
15
20
25

2.65 ± 0.1

Amplification ratios

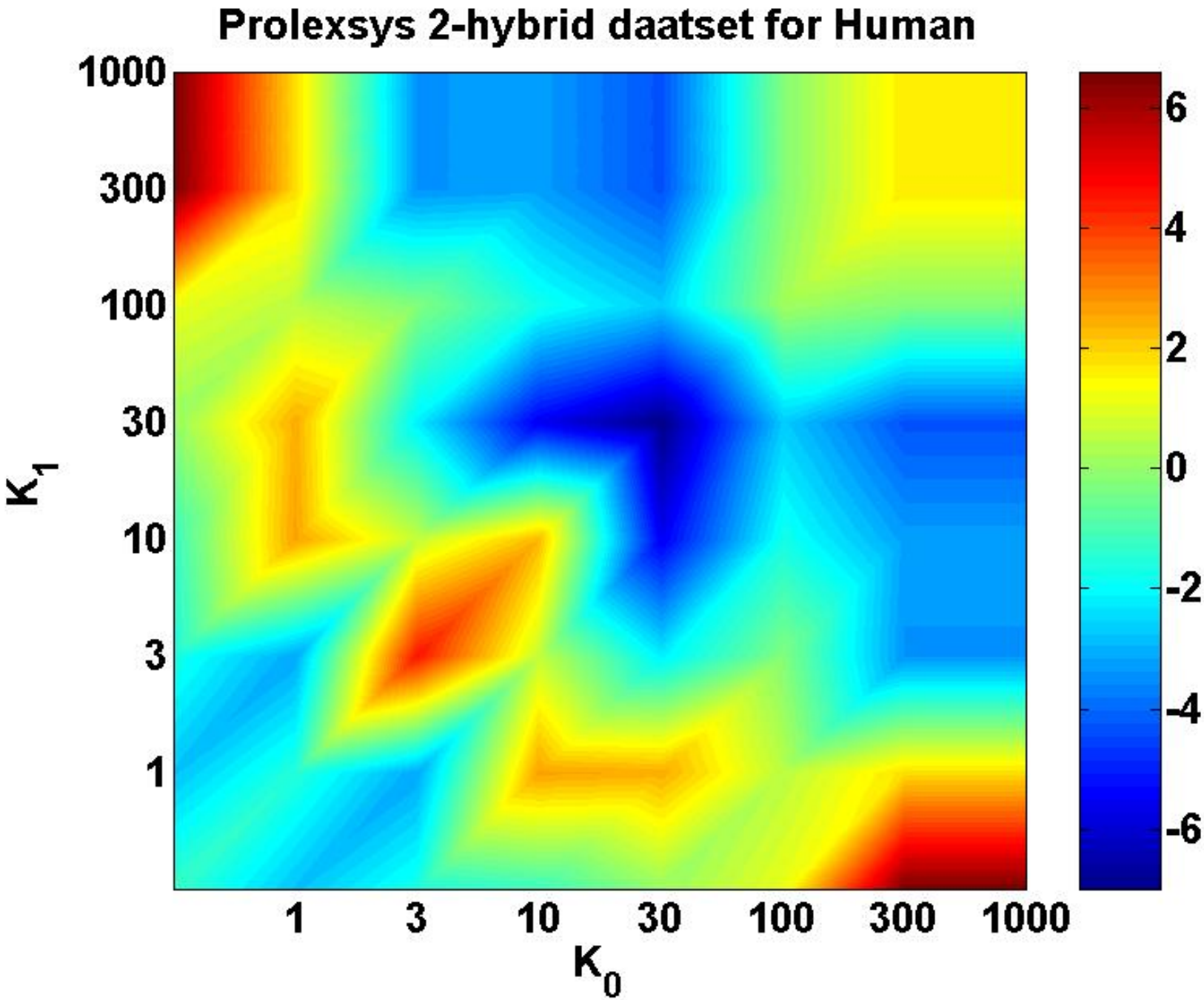
$$\mathcal{A}^{(dir)} = \frac{\langle K_{in} K_{out} \rangle}{\langle K_{in} \rangle} \qquad \mathcal{A}^{(undir)} = \frac{\langle K(K-1) \rangle}{\langle K \rangle}$$

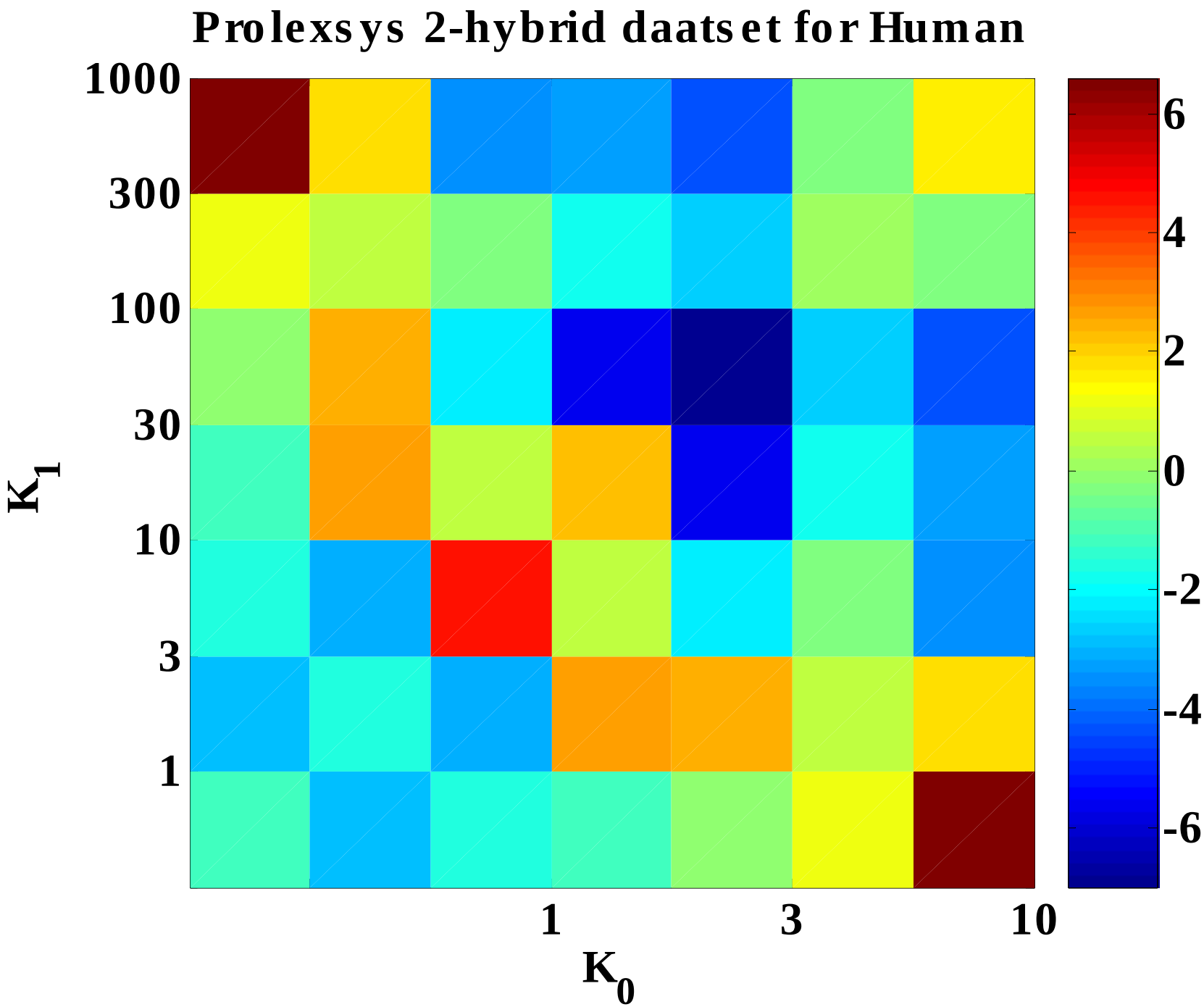
- $A^{(dir)}$: 1.08 - E. Coli, 0.58 - Yeast
- $A^{(undir)}$: 10.5 - E. Coli, 13.4 – Yeast
- $A^{(PPI)}$: ? - E. Coli, 26.3 - Yeast

Two-hybrid experiment

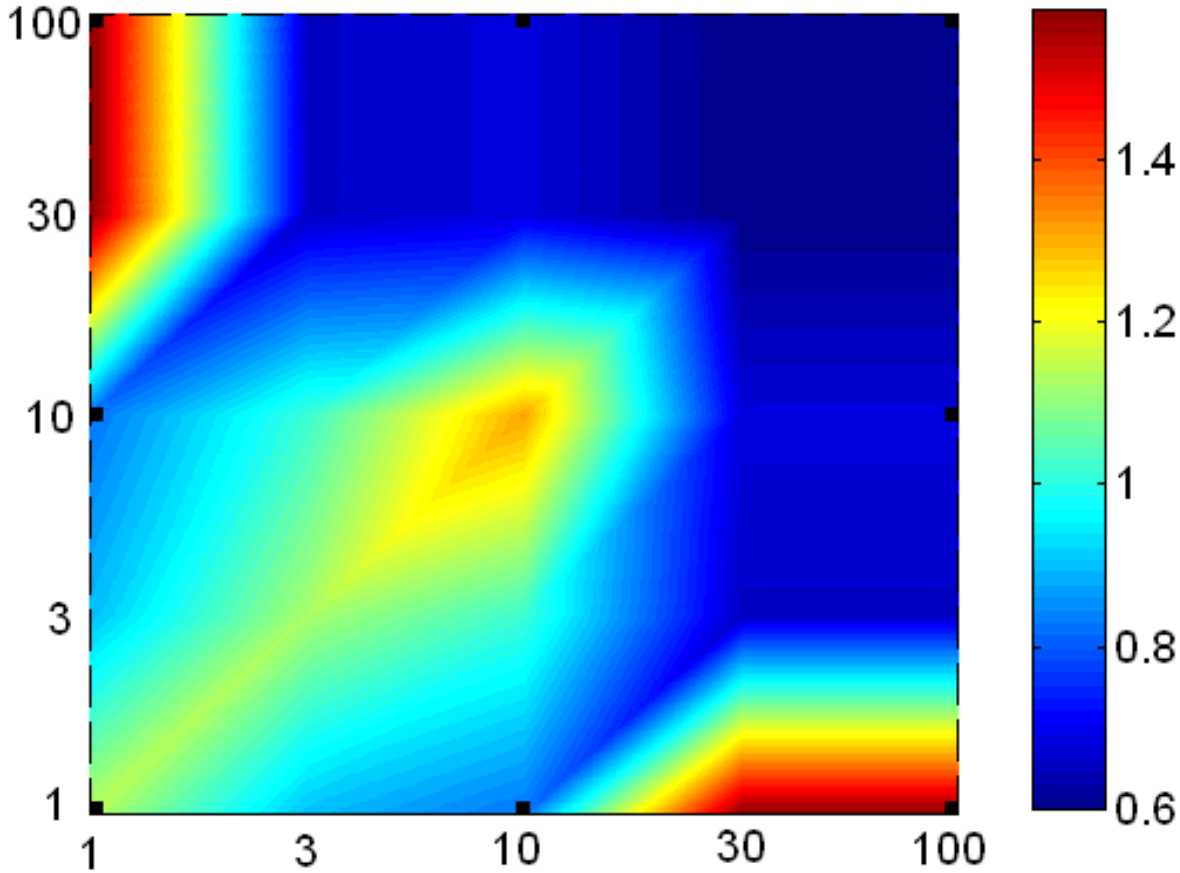
- To test if A interacts with B create two hybrids A* (with Gal4p DNA-binding domain) and B* (with Gal4p activation domain)

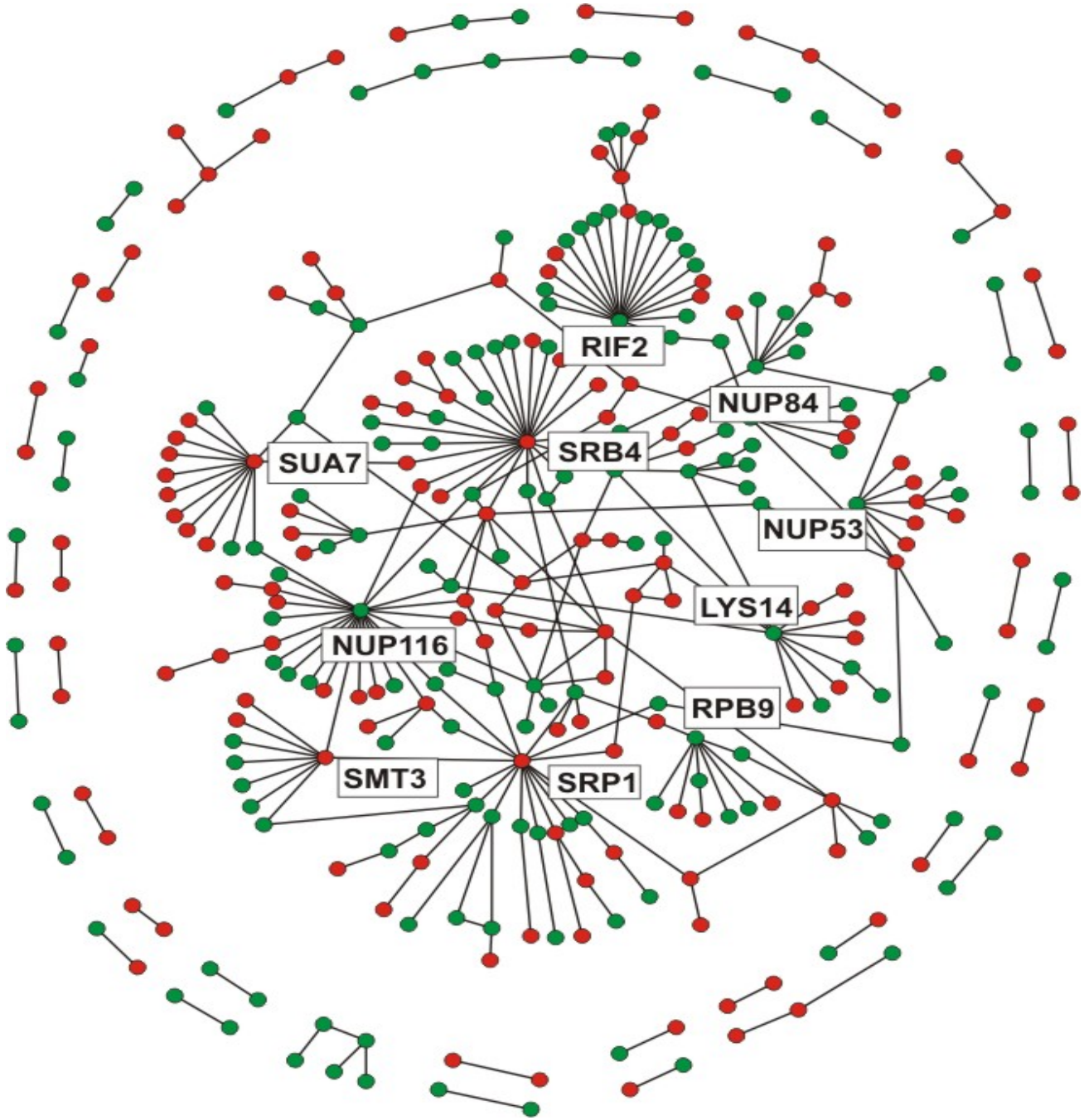
- High-throughput: all pairs among 6300 yeast proteins are tested: Uetz, *et al.* Nature (2000), Ito, *et al.*, PNAS (2001).

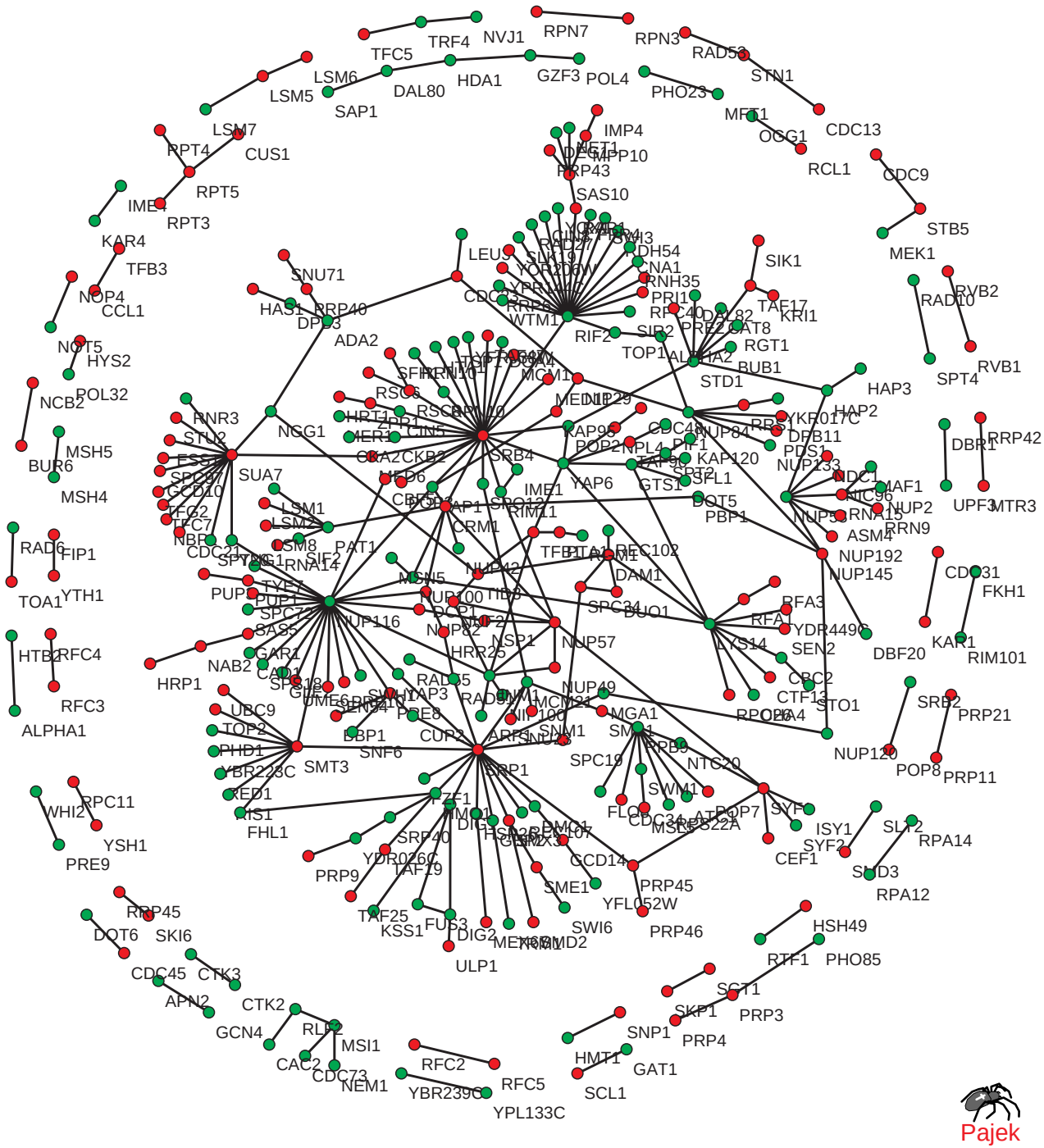

Prolexsys 2-hybrid daatset for Human
1000
300
100
30
10
3
1
K1
1
3
10
30
100
300
1000
K0
6
4
2
0
-2
-4
-6


Prolexsys 2-hybrid daatset for Human
1000
300
100
30
10
3
1
K1
1
3
10
K0
6
4
2
0
-2
-4
-6

DIP core correlation profile



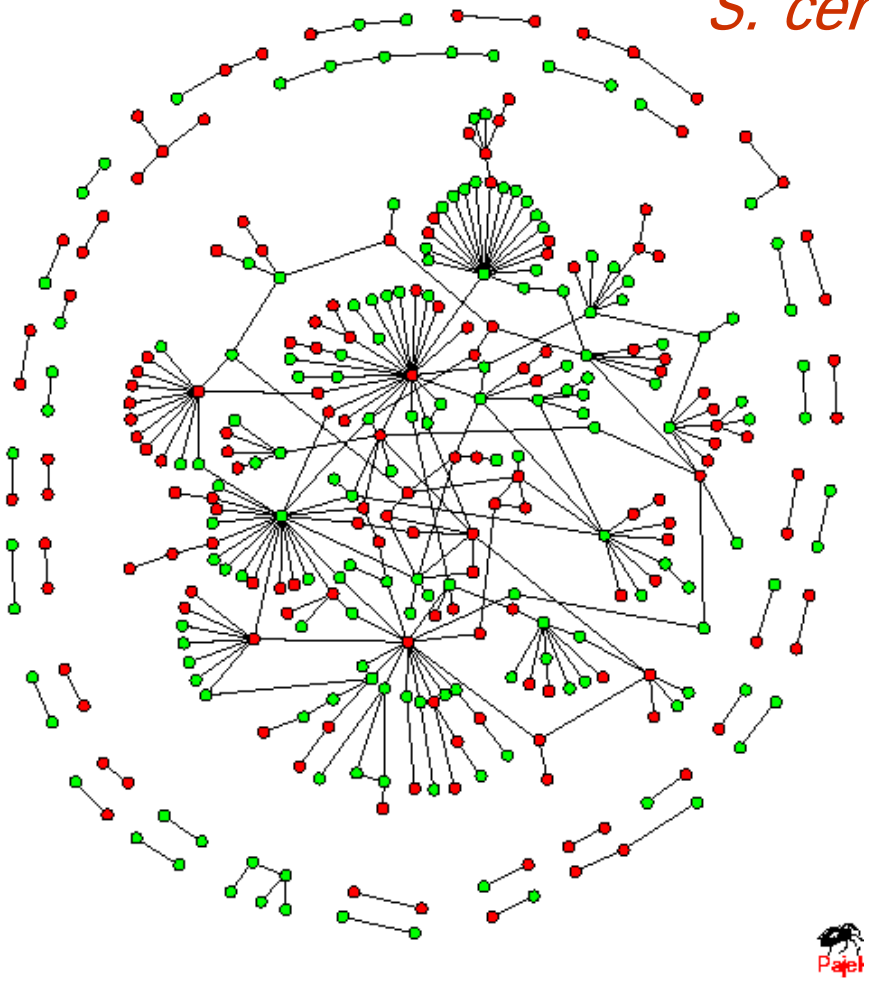

RIF2
NUP84
SRB4
SUA7
NUP53
LYS14
NUP116
RPB9
SMT3
SRP1

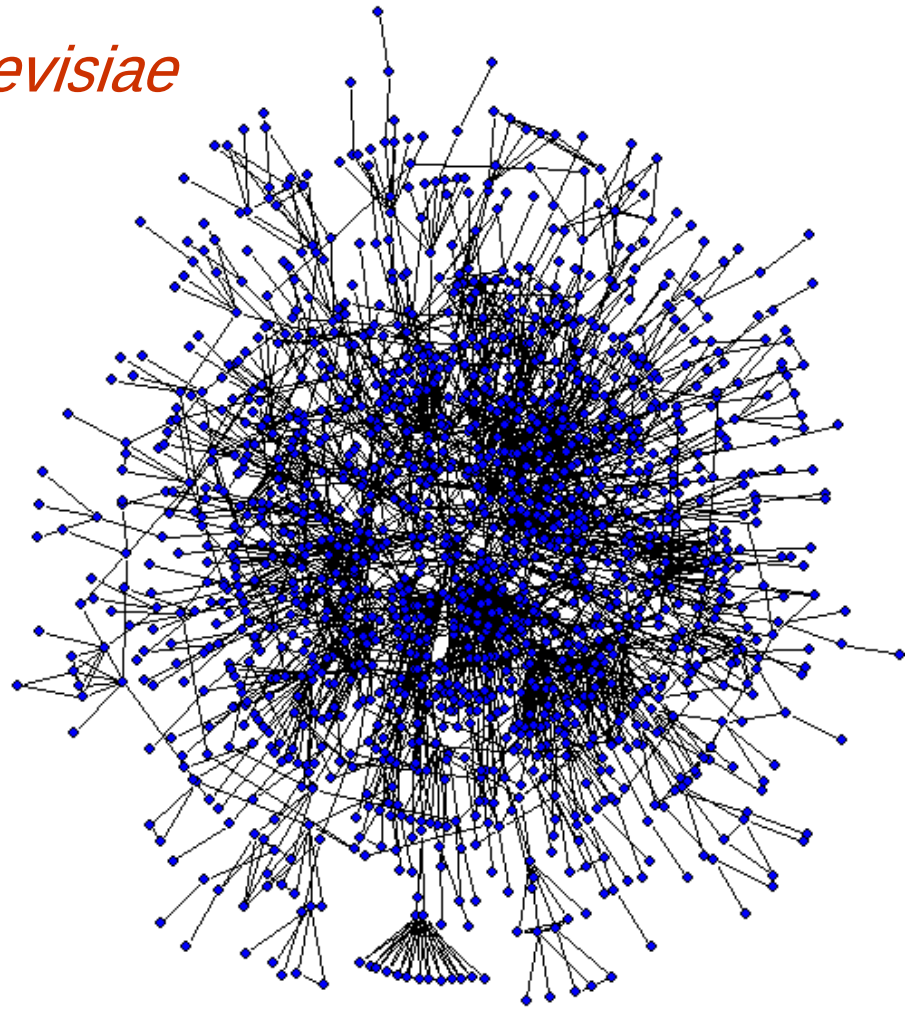


TFC5
TRF4
NVJ1
RPN7
RPN3
RAD53
STN1
LSM6
LSM5
DAL80
HDA1
GZF3
POL4
PHO23
SAP1
MFT1
OGG1
CDC13
LSM7
RPT4
CUS1
IMP4
MPP10
RCL1
CDC9
RPT5
IME4
RPT3
SAS10
STB5
KAR4
TFB3
MEK1
SIK1
SNU71
NOP4
CCL1
HAS1
RRP40
DPL3
ADA2
WTM1
RIF2
TAF17
KRI1
RAD10
RVB2
NOT5
HYS2
TOP1
RGT1
BUB1
RVB1
SPT4
NCB2
POL32
STD1
HAP3
HAP2
RNR3
NGG1
MSH5
BUR6
MSH4
SUA7
SRB4
YAP6
KAP120
DBR1
RRP42
UPF3
MTR3
RAD6
FIP1
TOA1
YTH1
PAT1
CRM1
DOT5
PBP1
NUP192
NUP145
CDC31
FKH1
KAR1
RIM101
DBF20
HTB2
RFC4
RFC3
ALPHA1
HRP1
NAB2
NUP116
NUP57
SEN2
CBC2
STO1
SRB2
PRP21
UBC9
TOP2
PHD1
SMT3
SNF6
CUP2
SPC19
NTC20
NUP120
POP8
PRP11
RPC11
WHI2
YSH1
PRE9
FHL1
SRP40
PRP9
TAF19
GCD14
SME1
PRP45
YFL052W
PRP46
ISY1
SYF2
CEF1
SLT2
RPA14
SHD3
RPA12
RRP45
DOT6
SKI6
TAF25
KSS1
FUS3
DIG2
SWI6
ULP1
HSH49
RTF1
PHO85
CDC45
CTK3
APN2
CTK2
GCN4
MSI1
CAC2
CDC73
NEM1
RFC2
YBR239C
RFC5
YPL133C
SNP1
PRP4
HMT1
GAT1
SCL1
SGT1
SKP1
PRP3
Pajek

Protein binding networks

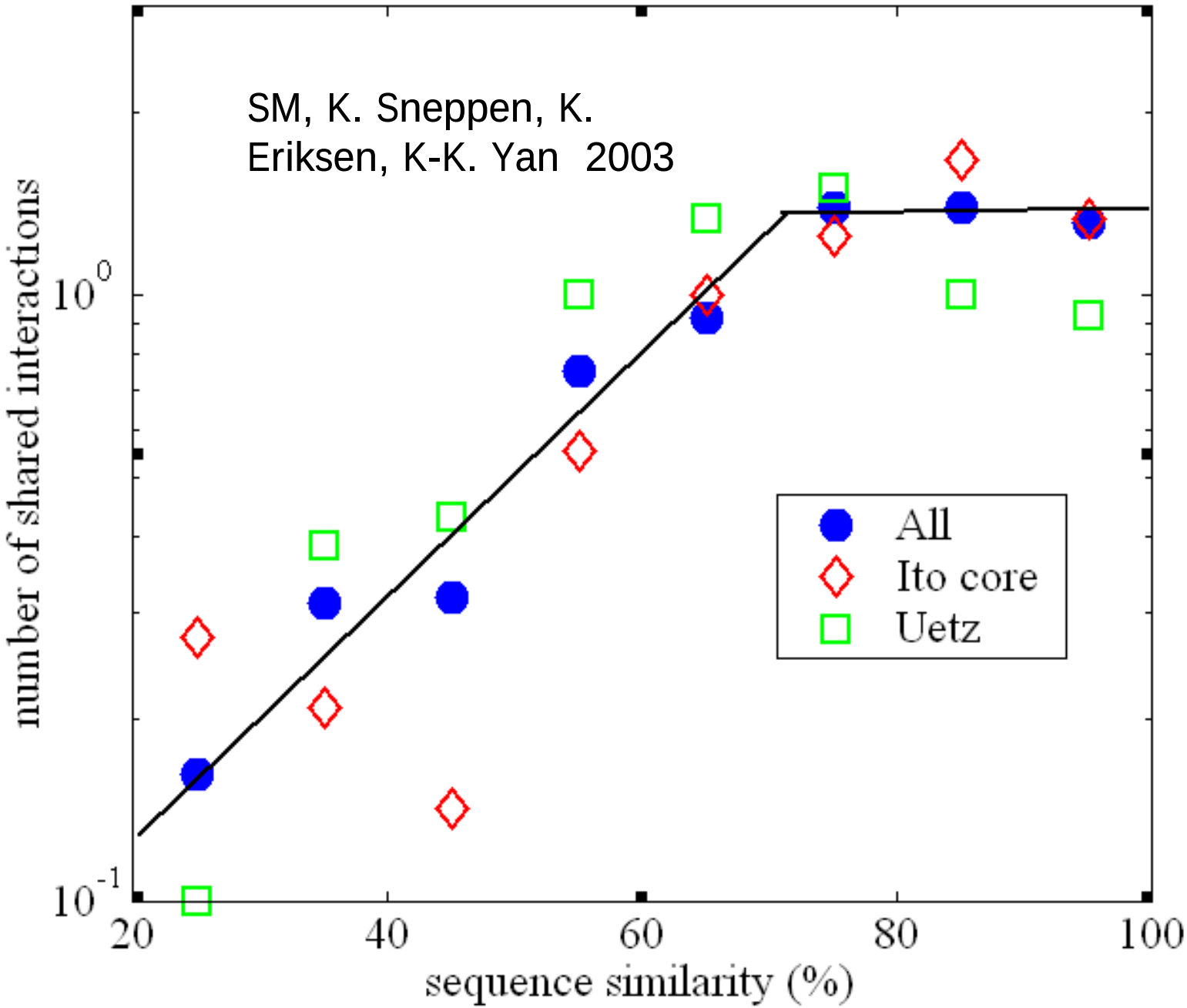
S. cerevisiae



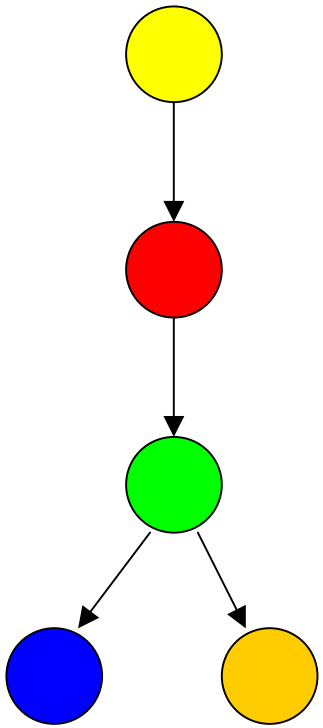


Two-hybrid nuclear

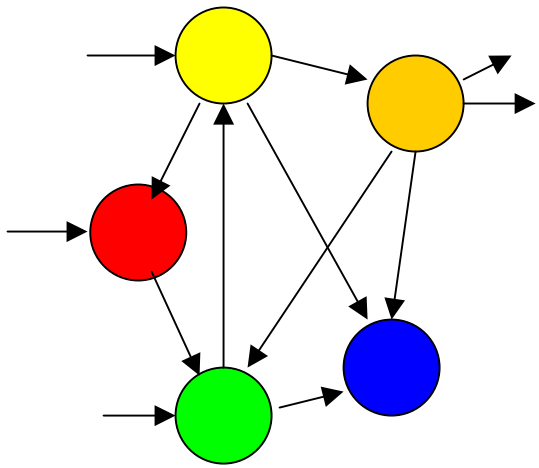
Database (DIP) core set


SM, K. Sneppen, K. Eriksen, K-K. Yan 2003
number of shared interactions
10^0
10^{-1}
20
40
60
80
100
sequence similarity (%)
All
Ito core
Uetz

Pathway vs. network paradigm



Pathway



Network

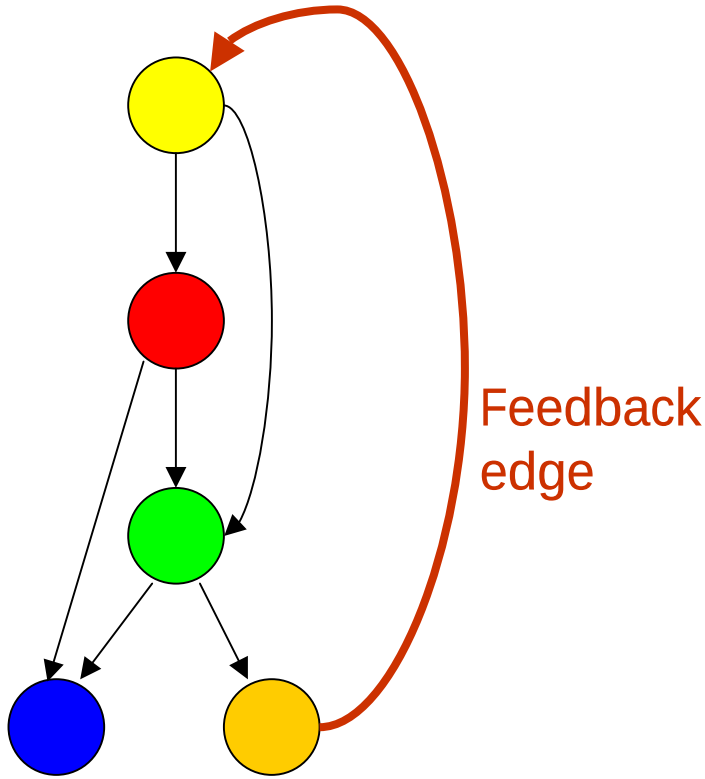
How to deal with feedback loops?

Propagation of signals often involves feedback loops

- Feedback could be used as a checkpoint that the signal reached its destination:
  - The main direction of information flow A→B→C
  - “Weak” feedback signal C→A
- Some closed loops like A→B→C→A are not used to send feedback but to regulate periodic processes:
  - Cell cycle
  - Circadian rhythms

Finding the feedback links

Feedback links go against the main direction of the information flow

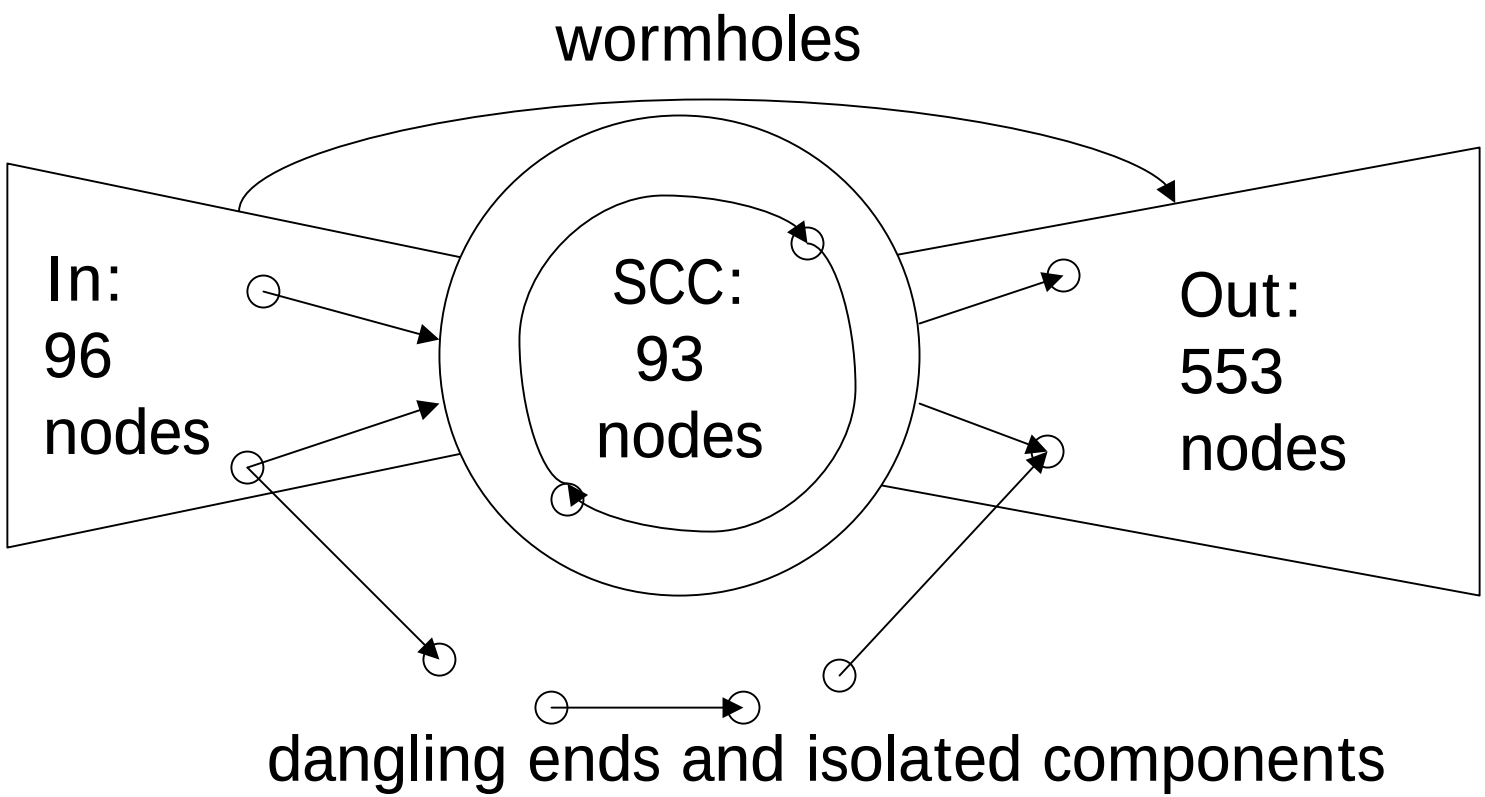


Algorithm

- Goal: remove the smallest number of edges to make the network feedback-free
- Solution:
  - Randomly assign hierarchy weights H_i to all nodes. Links $i \rightarrow j$ such that $H_i < H_j$ are candidate feedback links
  - Try to switch H_m and H_n on randomly selected nodes m,n. Accept if this switch reduces the number of feedback links. If increases – accept with a small probability (a la Metropolis)
  - The final network obtained by removing feedback links is feedback-free

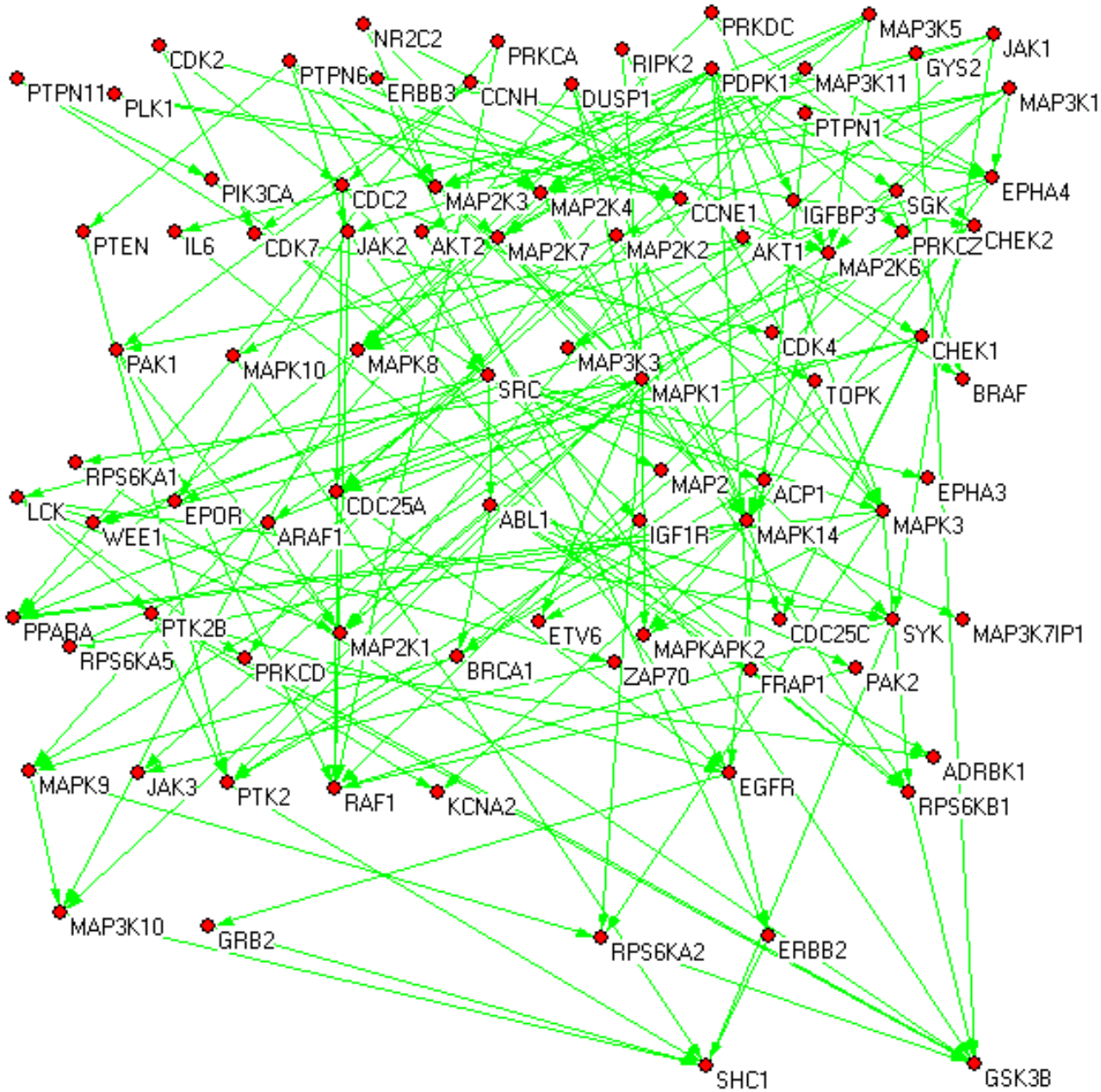
Bow-tie diagram

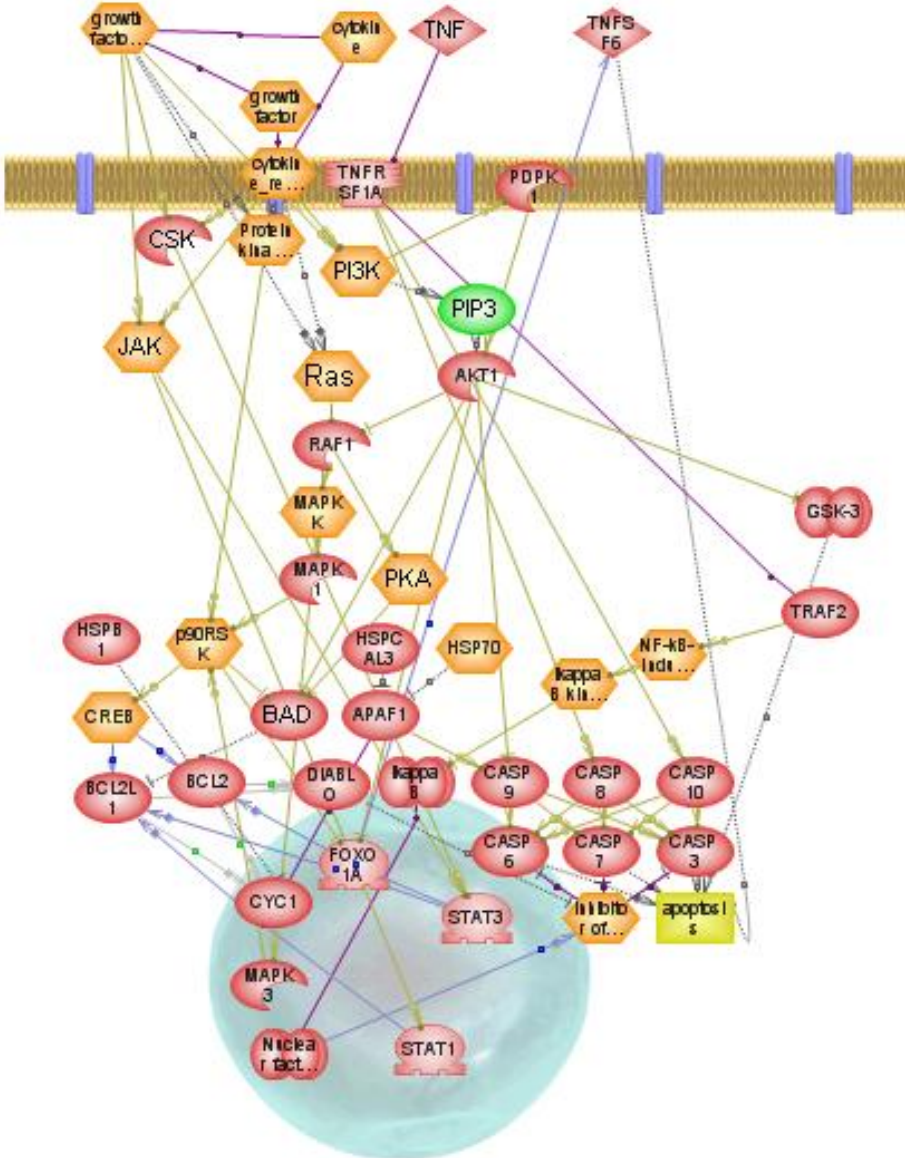
Human protein modification network: 1100 nodes



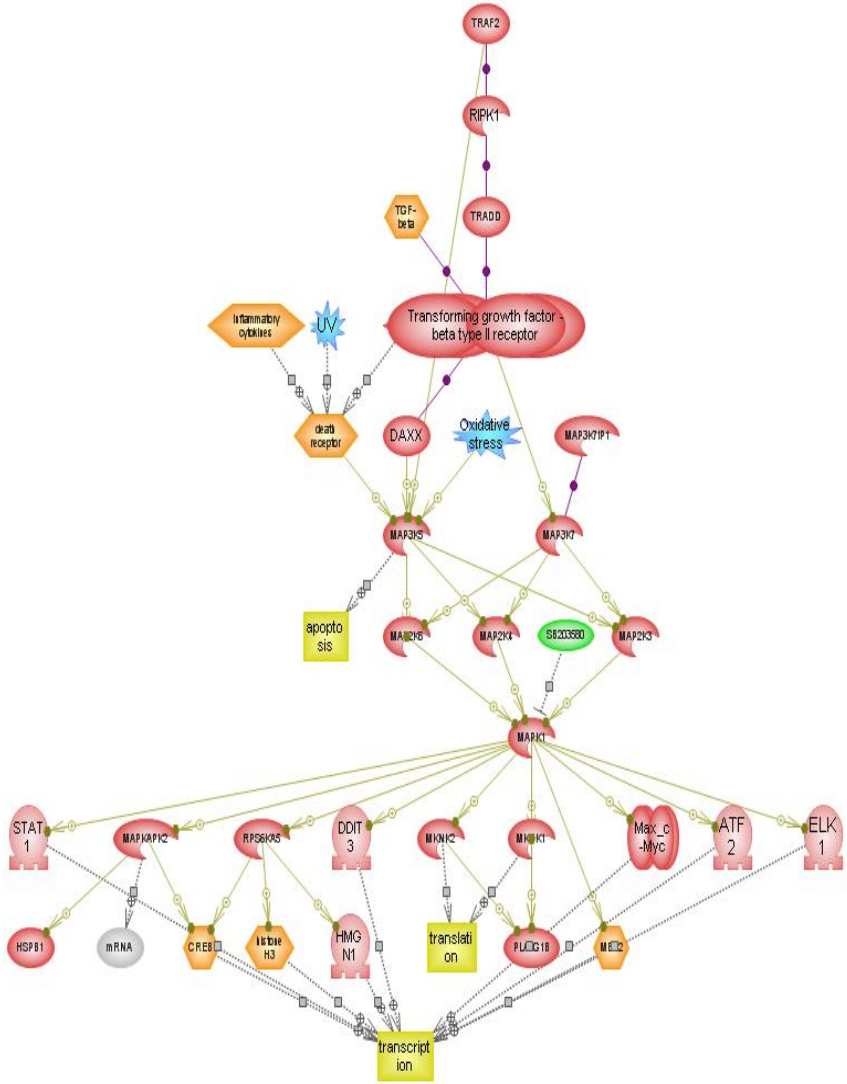
Feedback loops in human protein-modification network

- Removal of only 48 out of 2200 edges makes human protein modification network feedback-free
- Most links are “reproducible”: rerunning experiment gives almost the same answer
- Run algorithm 10 times. Each time record all upstream-downstream relations. Look for reproducible relationships: the correspond to main direction of the information flow


PRKDC
MAP3K5
JAK1
NR2C2
CDK2
PRKCA
RIPK2
GYS2
PTPN6
PDPK1
MAP3K11
PTPN11
ERBB3
CCNH
DUSP1
MAP3K1
PLK1
PTPN1
PIK3CA
CDC2
MAP2K3
MAP2K4
CCNE1
IGFBP3
SGK
EPHA4
PTEN
IL6
CDK7
JAK2
AKT2
MAP2K7
MAP2K2
AKT1
PRKCZ
CHEK2
MAP2K6
CDK4
CHEK1
PAK1
MAPK10
MAPK8
MAP3K3
SRC
MAPK1
TOPK
BRAF
RPS6KA1
MAP2
ACP1
EPHA3
LCK
EPOR
CDC25A
ABL1
WEE1
ARAF1
IGF1R
MAPK14
MAPK3
PPARA
PTK2B
ETV6
CDC25C
SYK
MAP3K7IP1
MAP2K1
MAPKAPK2
RPS6KA5
PRKCD
BRCA1
ZAP70
FRAP1
PAK2
ADRBK1
MAPK9
JAK3
PTK2
RAF1
KCNA2
EGFR
RPS6KB1
MAP3K10
GRB2
RPS6KA2
ERBB2
SHC1
GSK3B

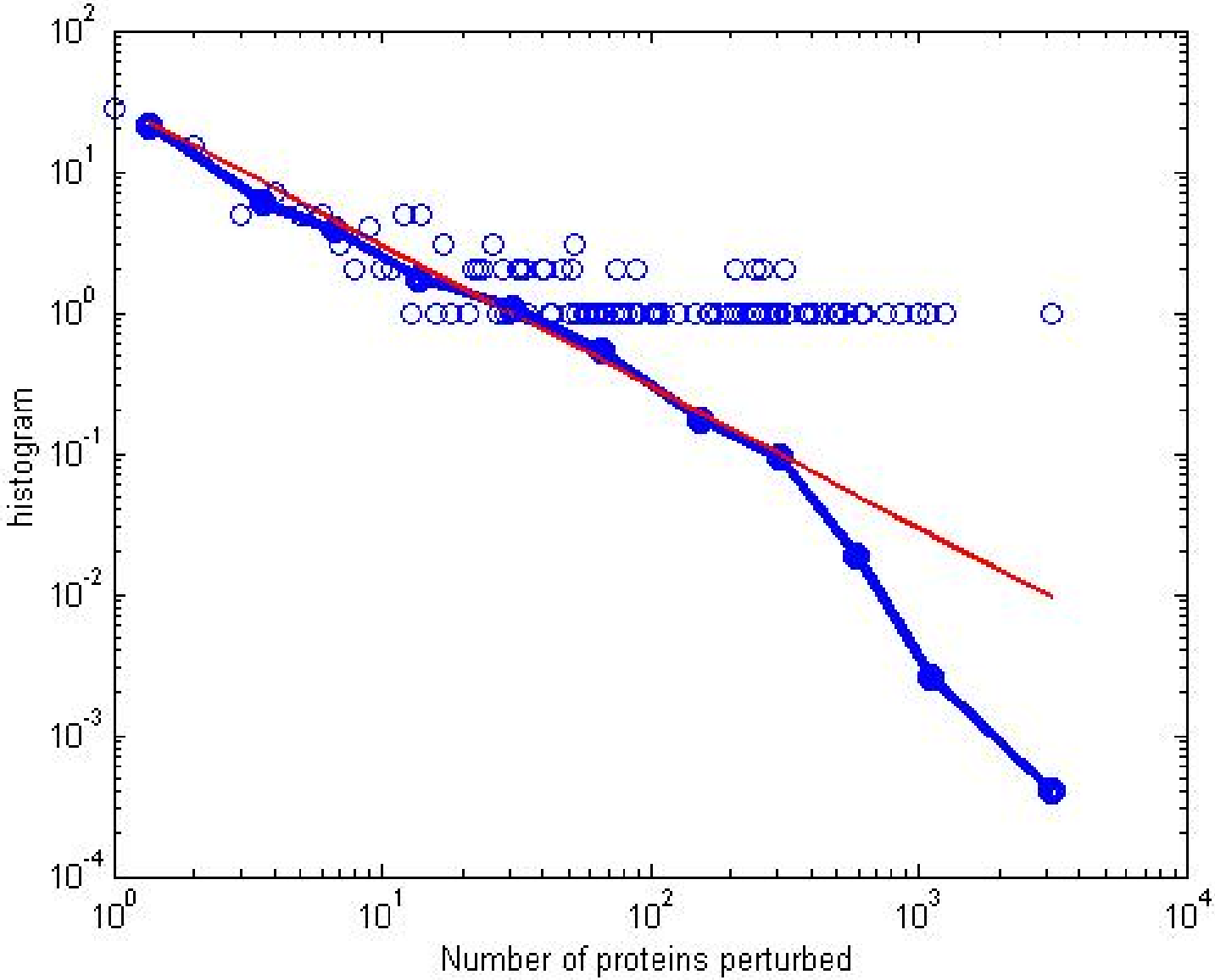


Inhibition of apoptosis

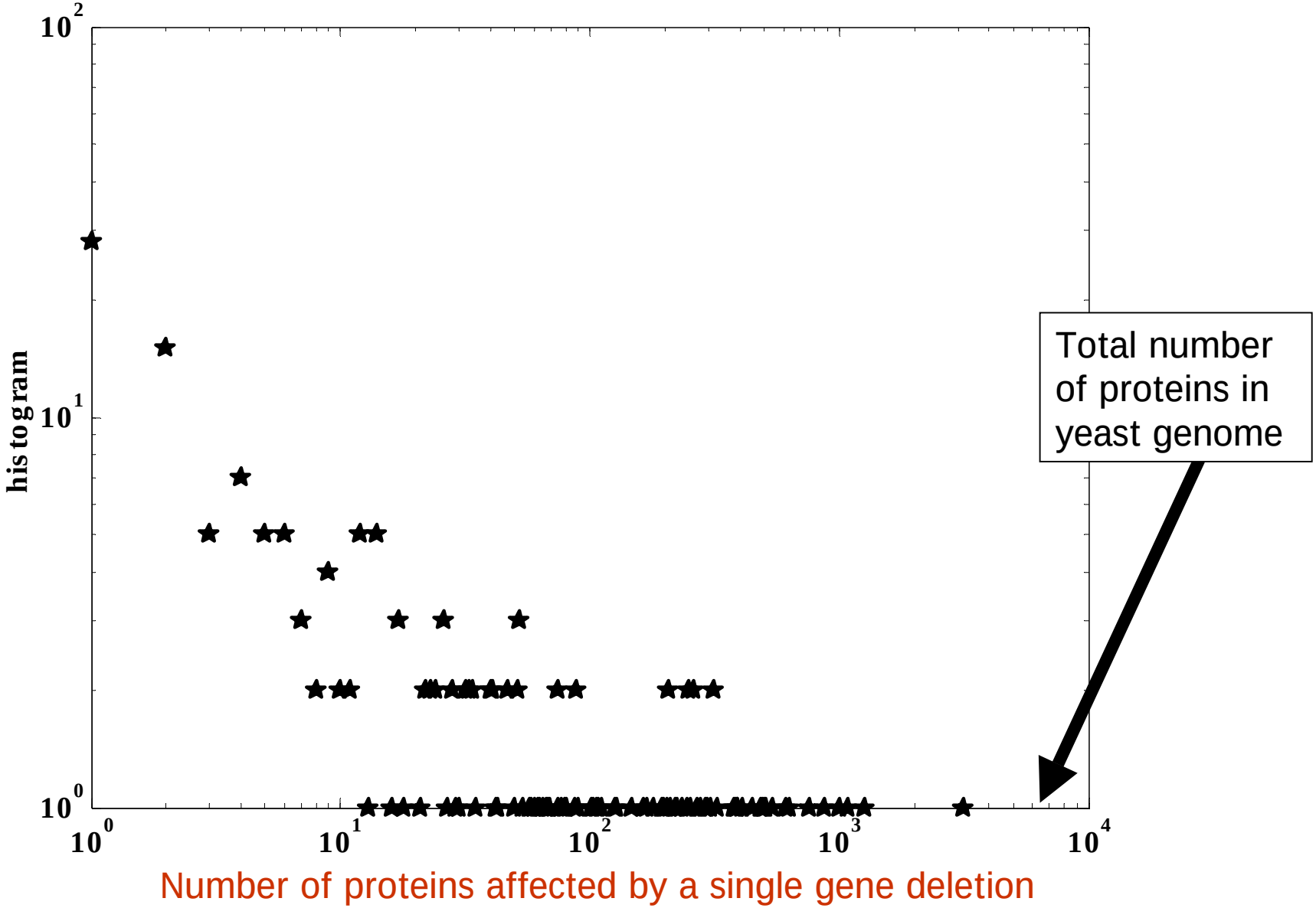


MAPK signaling

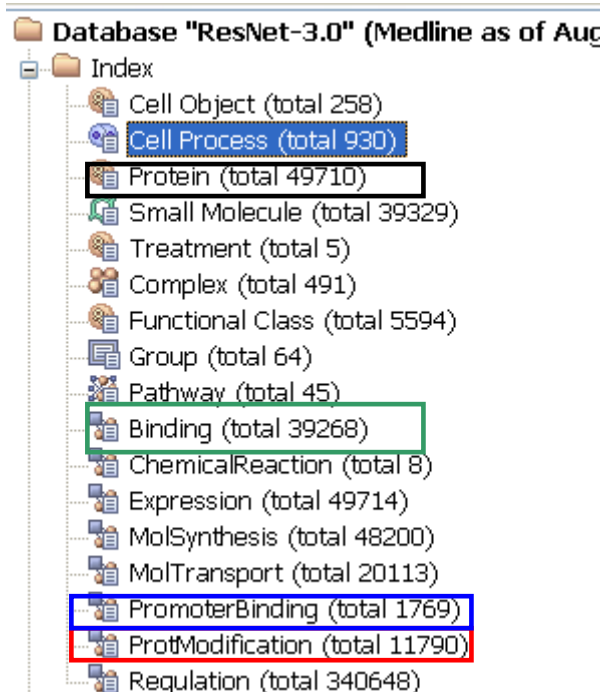
Gene disruptions in yeast

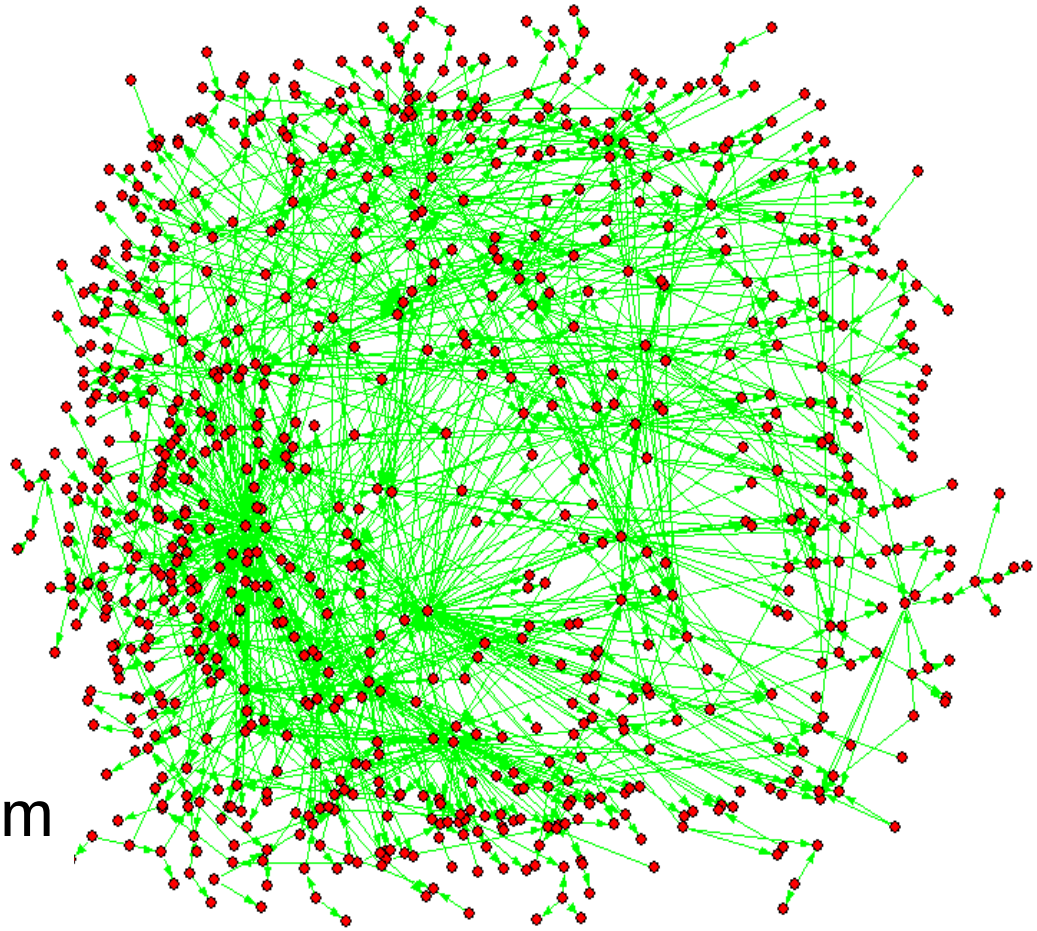


Repercussions of gene deletions



Homo sapiens





Total: 120,000 interacting protein pairs extracted from PubMed as of 8/2004

Data from Ariadne Genomics