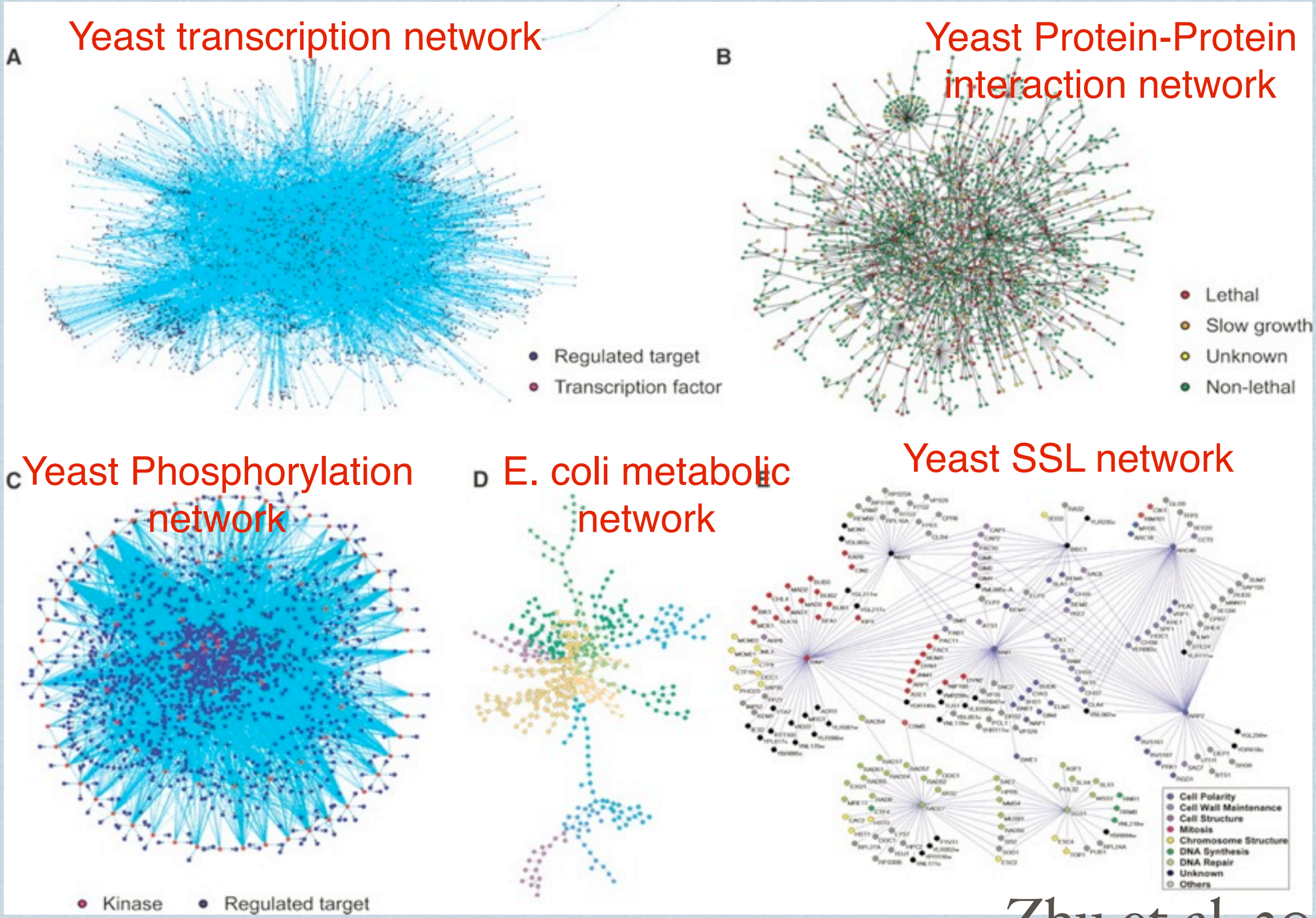
Network Alignment

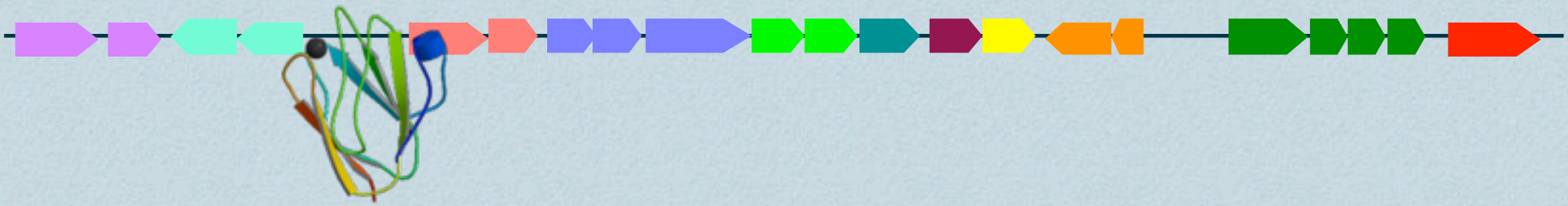
858L

Types of Biological Networks



Zhu et al, 2007.

Transcription network, aka regulatory network:



Transcription Factors =
proteins that bind to DNA
to activate or repress the
nearby, downstream genes.

regulates

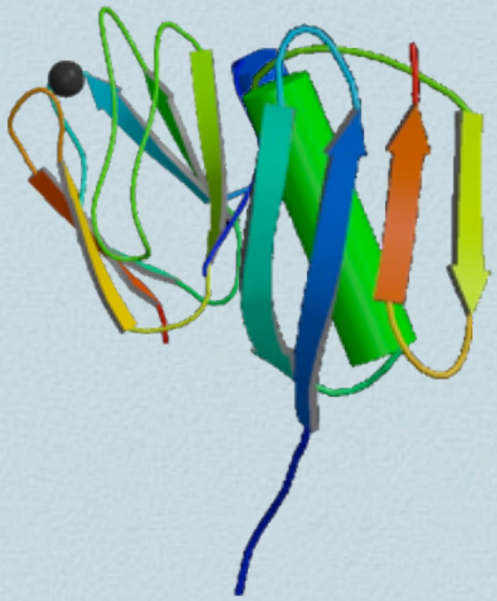
gene

gene

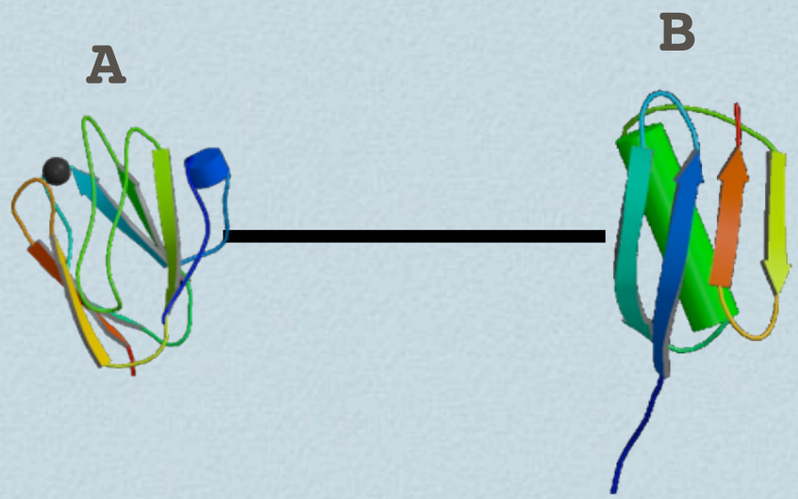
the regulated gene might
also be a transcription factor

leads to a
directed graph

Protein-Protein interaction network:



Proteins physically interact



Edge {A,B}

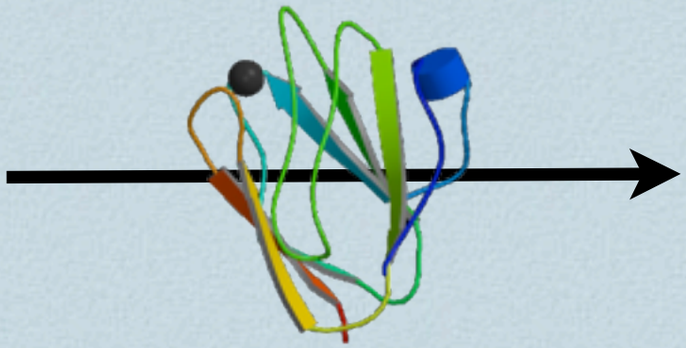
Assumption of binary interactions is imperfect.

Sometimes several proteins must bind simultaneously for there to be any “interaction” (could be modeled as a hyperedge)

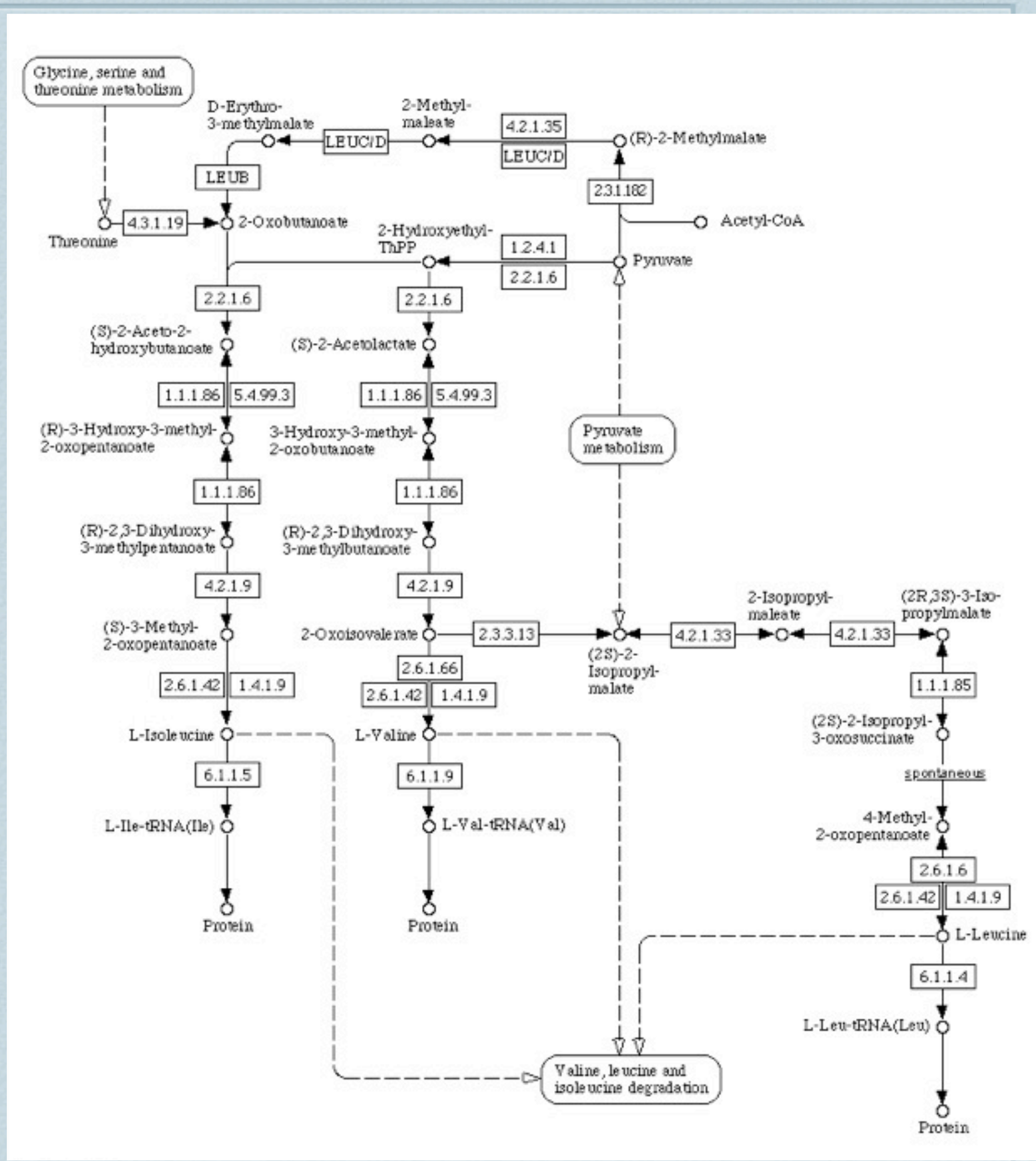
Metabolic network

Proteins are enzymes.

They label the *edges* and their substrates are the nodes.



Valine, Leucine, and Isoleucine biosynthesis (from KEGG)

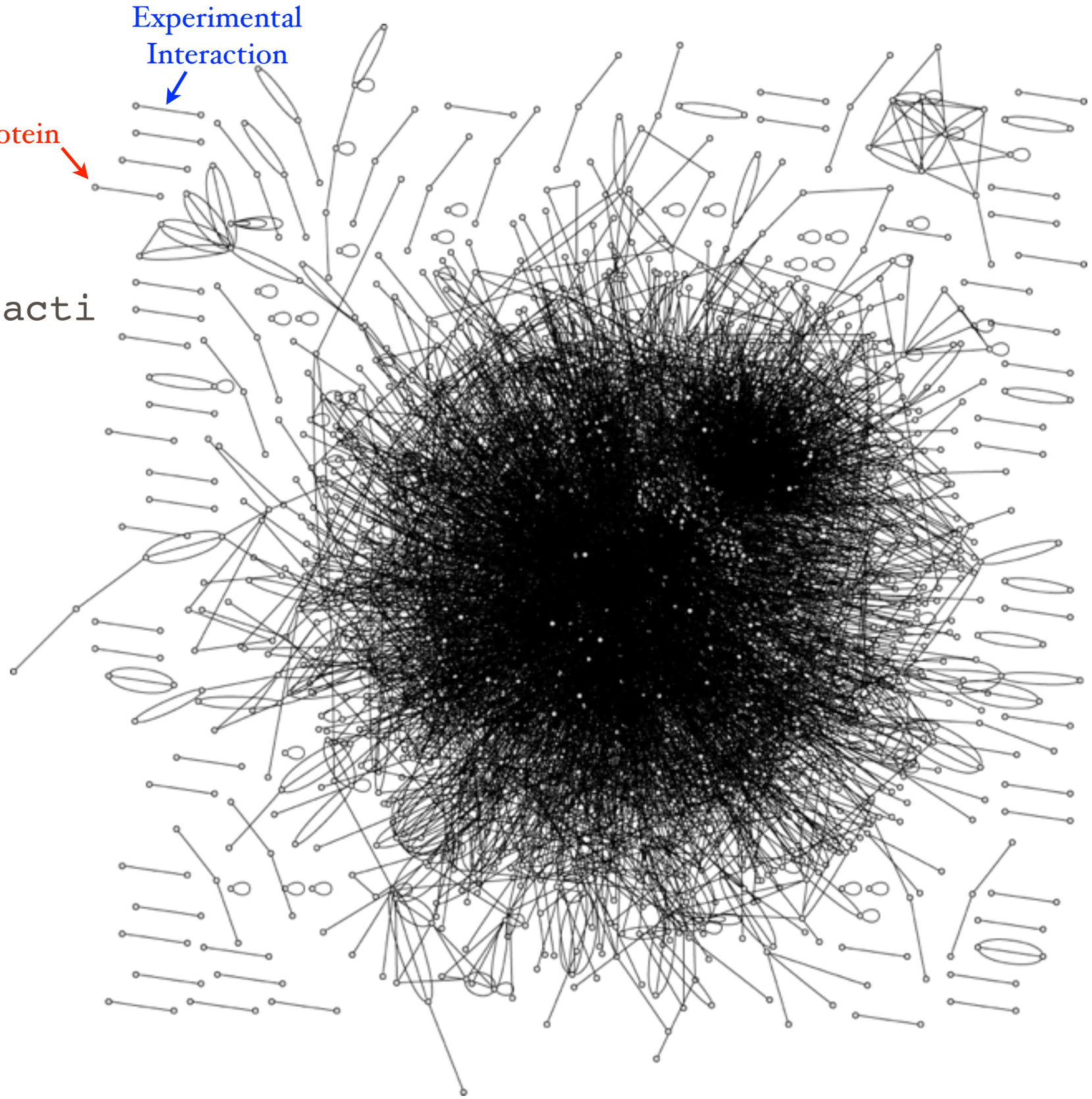


Yeast (aka *Saccharomyces cerevisiae*)interacti on network (Two-hybrid)

GRID (http://thebiogrid.org)

8,742 edges

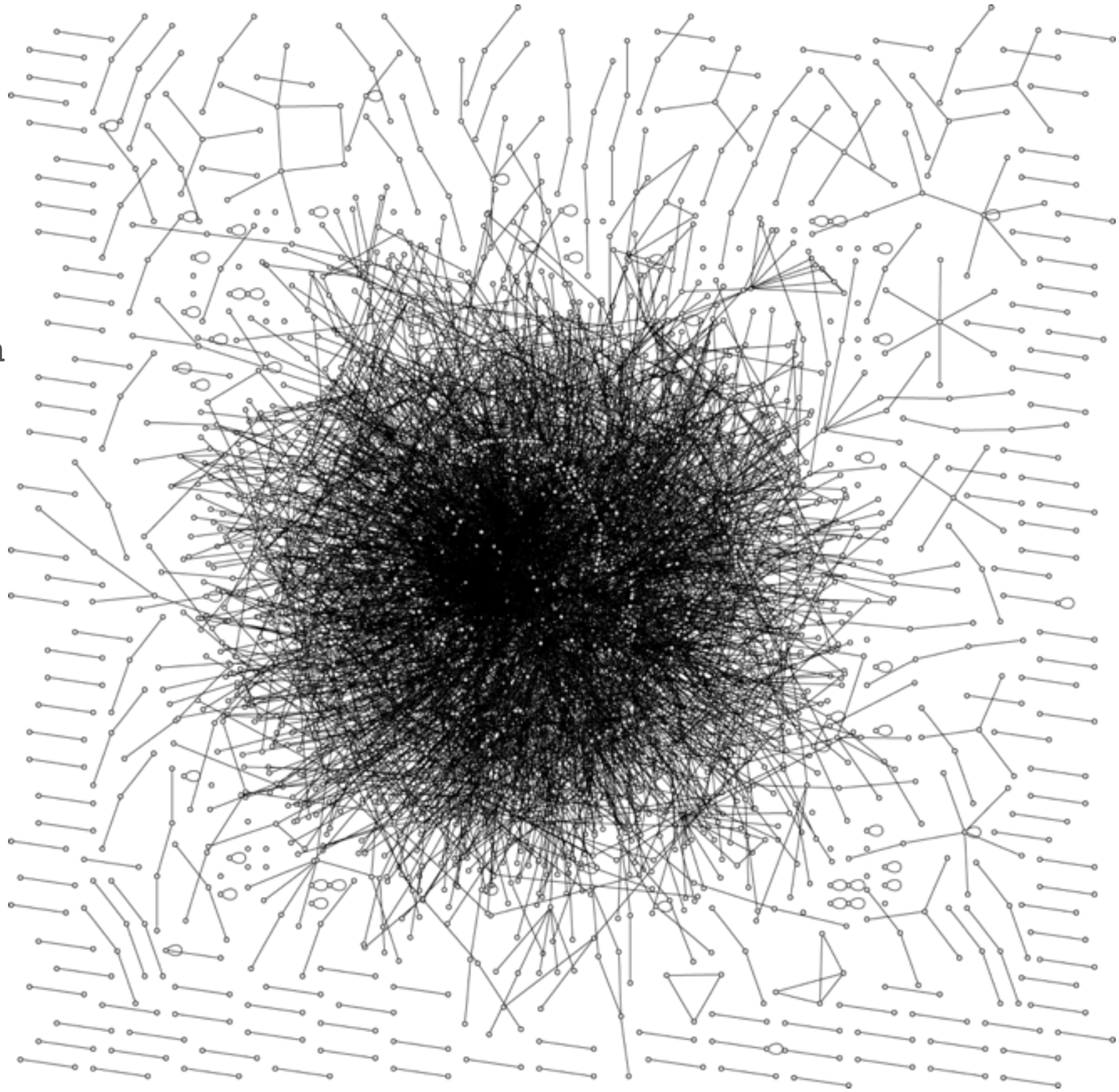
3113 nodes (= proteins)
(out of ~6,000 genes)



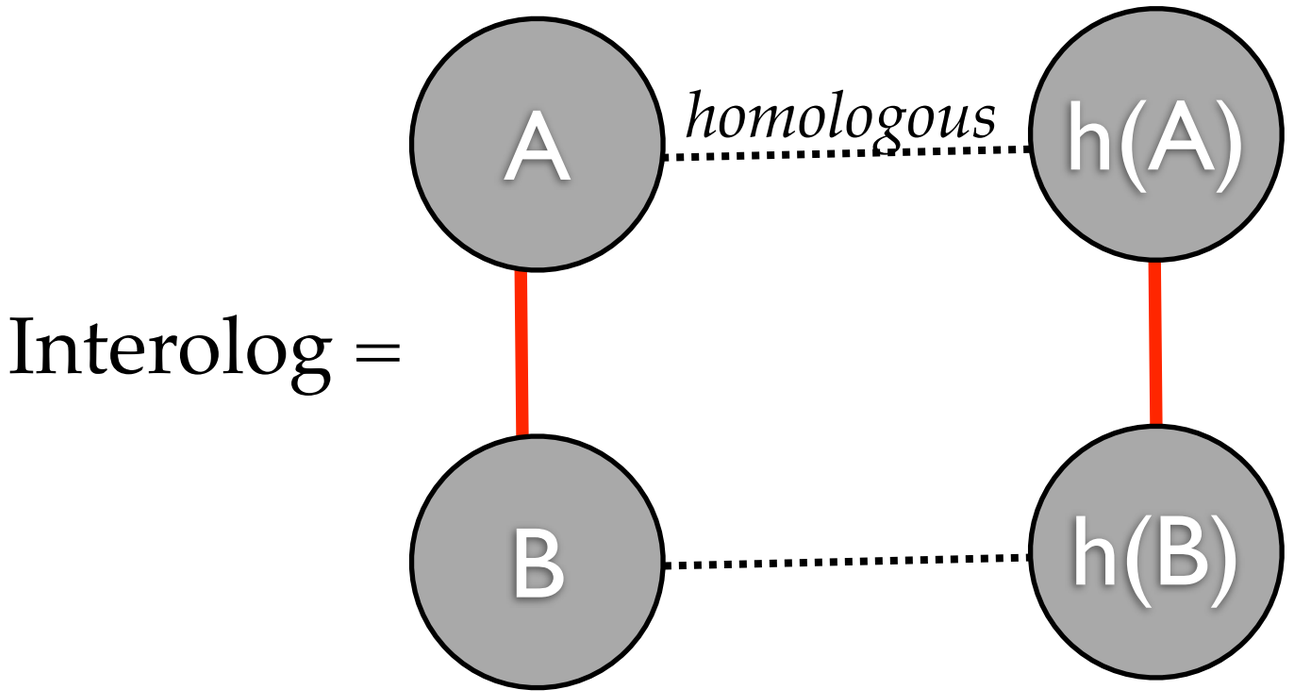
Human interaction network (Two-hybrid)

6,434 edges

4,083 nodes (out of ~25,000 genes)



Terms & Questions



Are there conserved pathways?
What is the minimum set of pathways required for life?
Can we compare networks to develop an evolutionary distance?

Aligning Networks

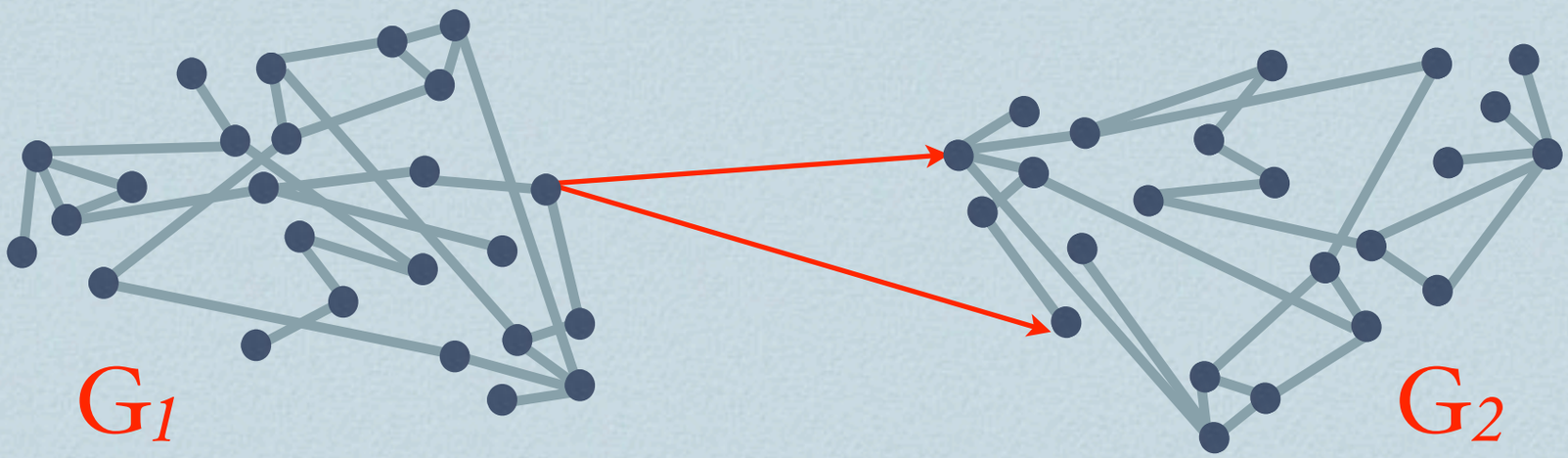
Combining Sequence and Network Topology

- Let $G_1 = (V_1, E_1)$, G_2, ... G_k be graphs, each giving noisy experimental estimations of interactions between proteins in organisms *1*,.., *k*.

- If $G_i = (V_i, E_i)$, we also have a function:

$$\mathrm{sim}(u,v) : V_i \times V_j \rightarrow \mathrm{R}$$

that gives the sequence similarity between *u* and *v*.



Conservation ⇒ Functional Importance

- If a structure has withstood millions of years of the randomizing process of mutations, then it likely has an important function.
- "Structure" = DNA sequence, protein sequence, protein shape, **network topology**.
- So: appearance of similar topology in two widely separated organisms indicates a real, fundamental set of interactions.
- Also, by comparing graphs we can transfer knowledge about one organism to another.

Local alignment:

1. Which nodes are dissimilar [low sim(u,v)] but have similar neighbors / neighborhoods? (e.g. Bandyopadhyay et al.)

   functional orthologs: proteins that play the same role, but may look very different.

2. Which edges are real and important, e.g. form a conserved pathway in the cell?

Global alignment:

Singh et al., 2007 propose:

Maximum common subgraph: Find the largest graph H that is isomorphic to subgraphs of two given graphs G_1 and G_2.

- **Graph Isomorphism:** Given graphs $G_1 = (V_1,E_1)$, $G_2 = (V_2,E_2)$ each with n nodes, decide whether there is a one-to-one and onto function

$$f : V_1 \rightarrow V_2$$

such that $(u,v) \in E_1 \Leftrightarrow (f(u), f(v)) \in E_2$

- **Subgraph Isomorphism:** Given graphs $G_1 = (V_1,E_1)$, $G_2 = (V_2,E_2)$, where G_1 has k nodes and G_2 has $n > k$ nodes, decide whether there is a one-to-one function

$$f : V_1 \rightarrow V_2$$

such that $(u,v) \in E_1 \Leftrightarrow (f(u), f(v)) \in E_2$

- **Graph Isomorphism:** Given graphs $G_1 = (V_1,E_1)$, $G_2 = (V_2,E_2)$ each with n nodes, decide whether there is a one-to-one and onto function

$$f : V_1 \rightarrow V_2$$

such that $(u,v) \in E_1 \Leftrightarrow (f(u), f(v)) \in E_2$

Not known to be NP-hard.

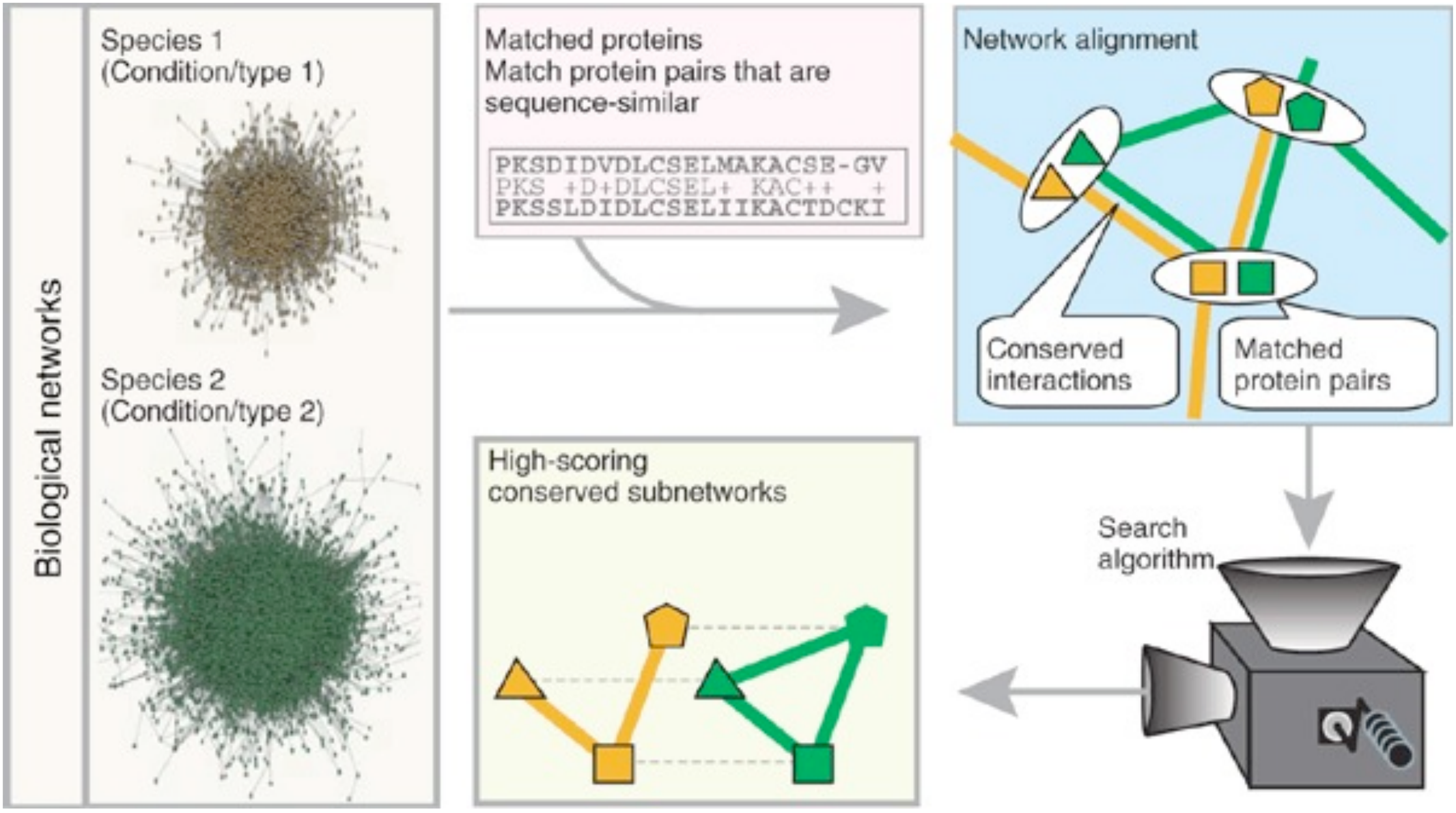
- **Subgraph Isomorphism:** Given graphs $G_1 = (V_1,E_1)$, $G_2 = (V_2,E_2)$, where G_1 has k nodes and G_2 has $n > k$ nodes, decide whether there is a one-to-one function

$$f : V_1 \rightarrow V_2$$

such that $(u,v) \in E_1 \Leftrightarrow (f(u), f(v)) \in E_2$

NP-complete.

PathBLAST:



(Kelley et al, 2003)

PathBLAST Alignment Graph

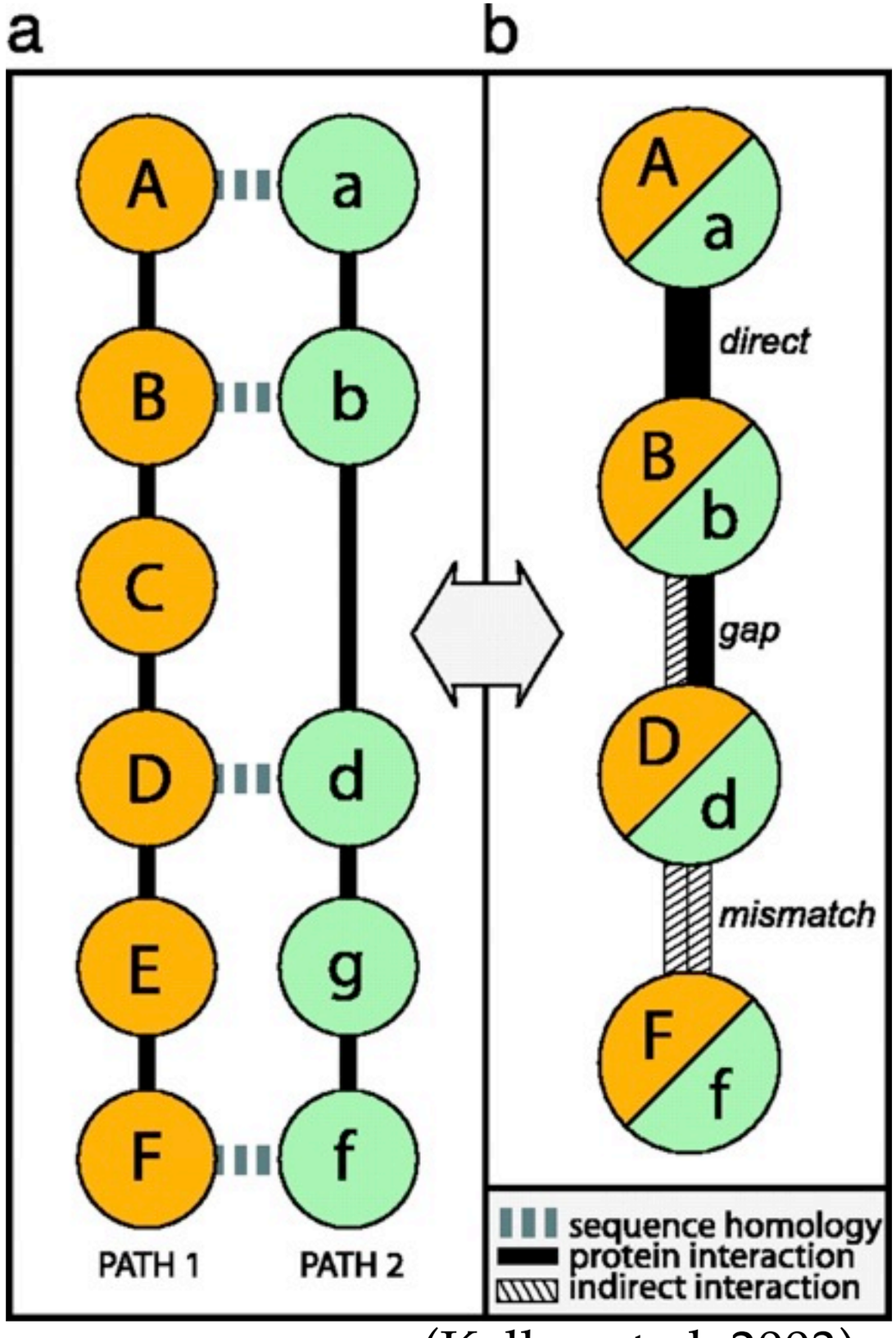
Nodes correspond to homologous pairs (**A, a**) where **A** is from one species, and **a** is from the other.

Edges come in 3 types:

- **Direct**. **A-B** and **a-b** interactions are present.

- **Gap**. Edge **A-B** is present, and **a** & **b** are separated by 2 hops.

- **Mismatch**. Both (**A** & **B**) and (**a** & **b**) are both separated by 2 hops.



(Kelley et al, 2003)

PathBLAST Scoring Function

$p(E_v \mid H)$ estimated from E_v distributions in COG:

$$p(v) = p(H \mid E_v) = \frac{p(E_v \mid H)\, p(H)}{p(E_v)}$$

p(v) = probability that proteins in v are really homologs.

$p(v) = \Pr[\text{Homology} \mid E_v]$

	Pr[interaction]
1	0.1
2	0.3
≥ 3	0.9

$$q(e) = \prod_{i \in e} \Pr[i]$$

q(e) = product of interaction edges "contained" within the alignment edge

sum over logs = product over scores

$$S(P) := \sum_{v \in P} \log \frac{p(v)}{p_{\text{random}}} + \sum_{e \in P} \log \frac{q(e)}{q_{\text{random}}}$$

P is a path in the alignment graph.

p_{random} and q_{random} are the average values of $p(v)$ and $q(e)$ in the graph.

PathBLAST Search Procedure

If G is directed, acyclic (DAG) then its easy to find a high-scoring path via dynamic programming. S(v,L) = max-scoring path of length L that ends at v:

$$S(v,L) = \arg \max_{u \in \mathrm{pred}(v)} \left[S(u, L-1) + \log \frac{p(v)}{p_{\mathrm{random}}} + \log \frac{q(u \to v)}{q_{\mathrm{random}}} \right]$$

Because G is not directed, acyclic they randomly create a large number of DAGs by removing edges as follows:

1. Randomly rank vertices.
2. Direct edges from low to high rank.

Run dynamic program on the random DAGs and take the highest scoring path.

2/L! chance that a path will be preserved.

So repeat 5L! times.

H. pylori & S. cerevisiae

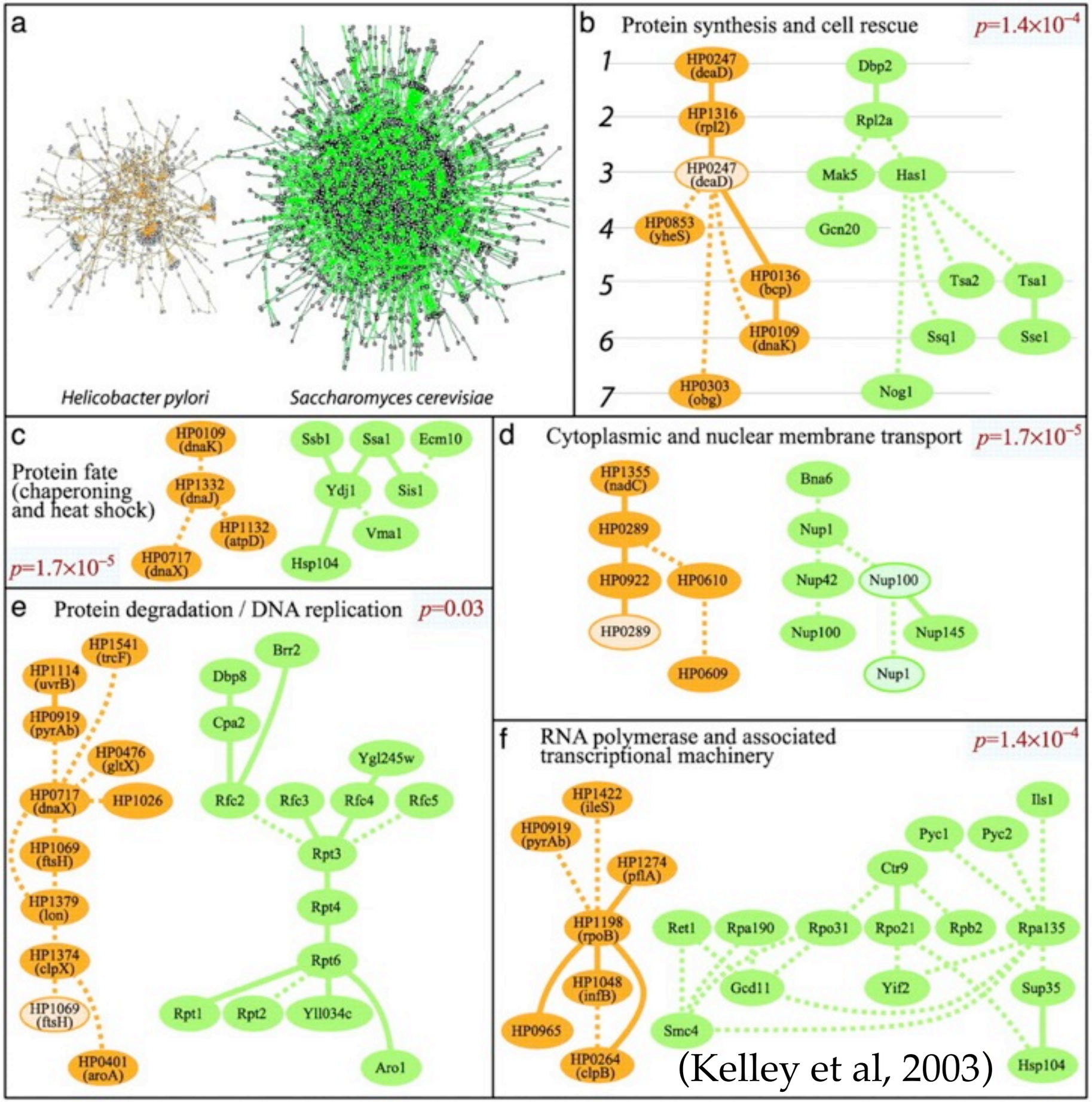
Find several (50) high-scoring paths

Then, remove those edges & vertices and repeat.

Overlay the identified paths.

Revealed 5 conserved pathways.

Contains proteins from both: DNA polymerase and Proteosome => evidence that they interact



(Kelley et al, 2003)

H. pylori & S. cerevisiae

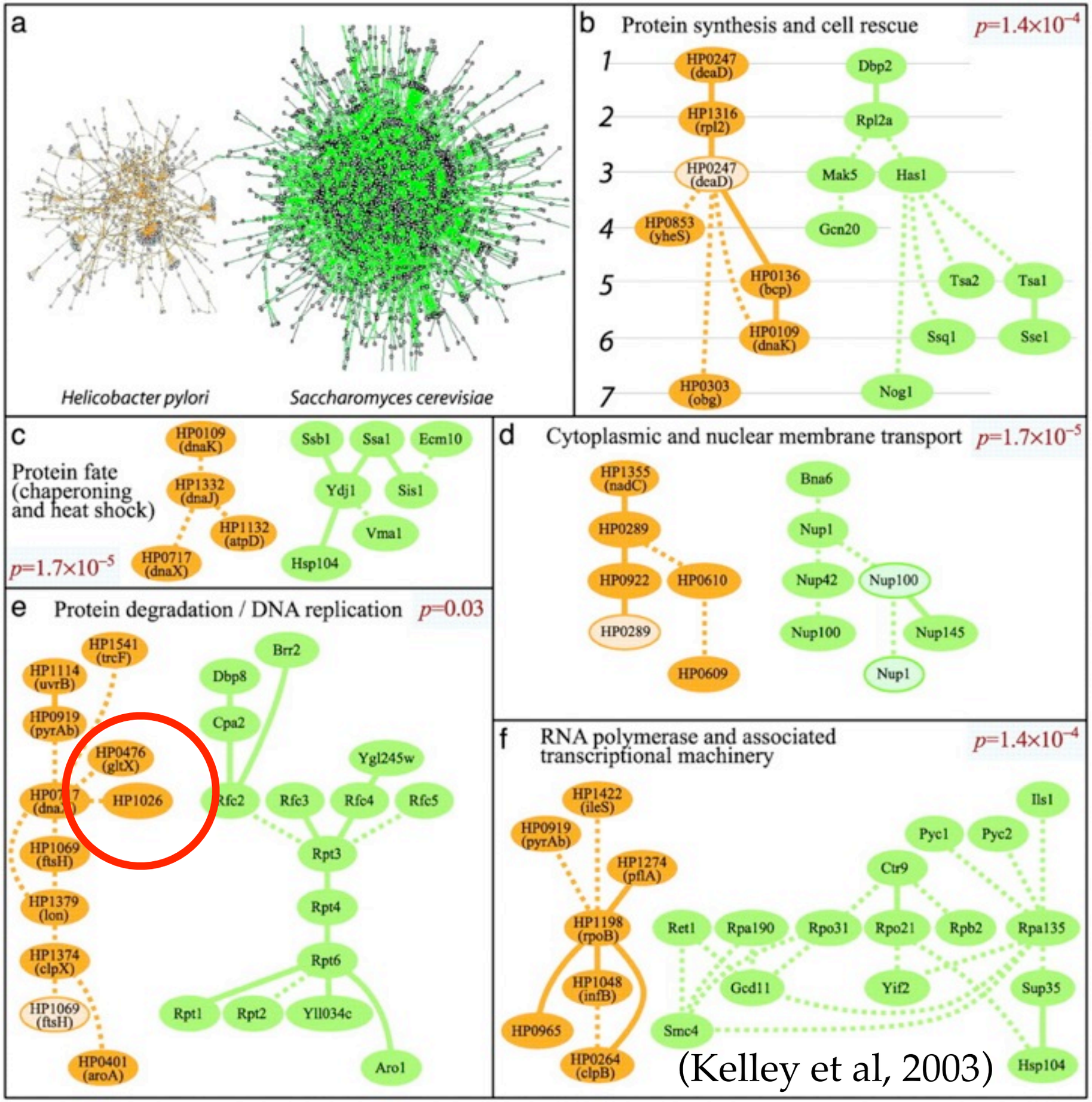
Find several (50) high-scoring paths

Then, remove those edges & vertices and repeat.

Overlay the identified paths.

Revealed 5 conserved pathways.

Contains proteins from both: DNA polymerase and Proteosome => evidence that they interact



(Kelley et al, 2003)

H. pylori & S. cerevisiae

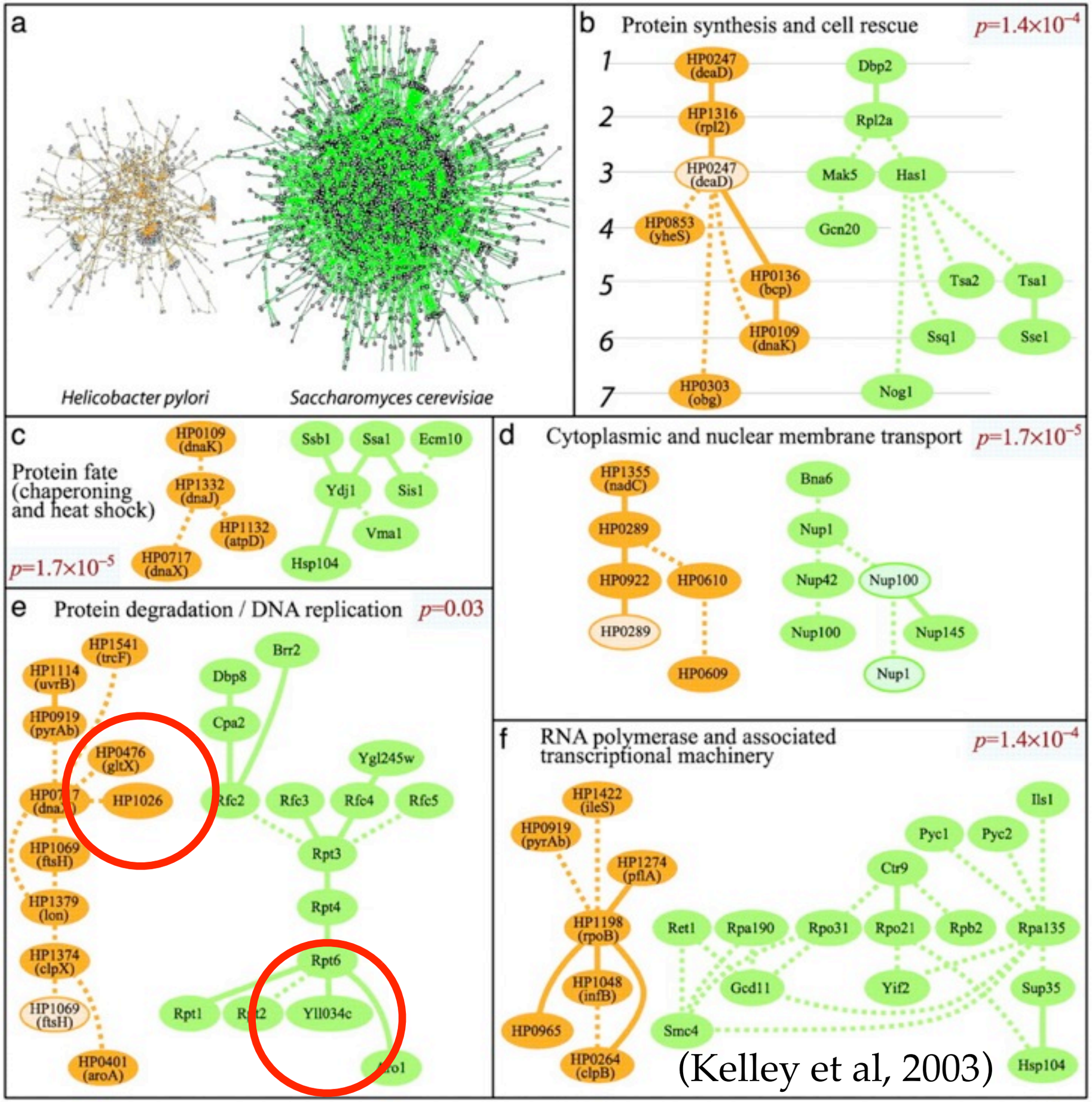
Find several (50) high-scoring paths

Then, remove those edges & vertices and repeat.

Overlay the identified paths.

Revealed 5 conserved pathways.

Contains proteins from both: DNA polymerase and Proteosome => evidence that they interact



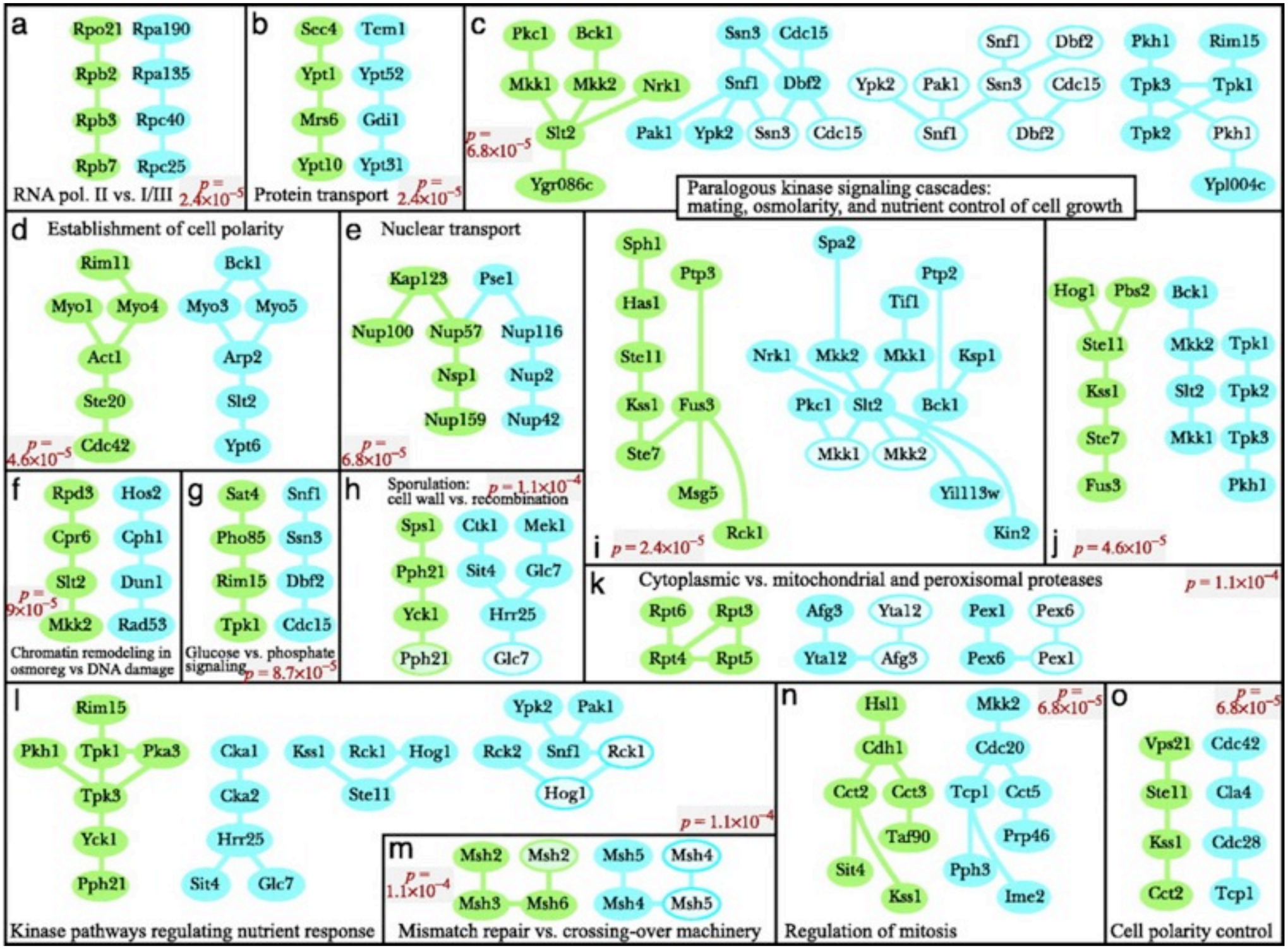
(Kelley et al, 2003)

Some Notes

- Goal: use a well-studied organism (yeast) to learn about a less-studied organism (H. pylori).
- There were only 7 directly shared edges between yeast & H. pylori. (you would expect 2.5 shared edges).
  - Gap & mismatch edges were essential!
- Within conserved pathways, proteins often were not paired with the protein with the most similar sequence.
  - 22% of the proteins in previous figure did not pair with their best sequence match
- Single pathways in bacteria often correspond to multiple pathways in yeast. (Yeast is suspected of having undergone multiple whole-genome duplications.)

Proteins were not allowed to pair with themselves or their network neighbors.

Yeast Paralogous Pathways



(Kelley et al, 2003)

Proteins were not allowed to pair with themselves or their network neighbors.

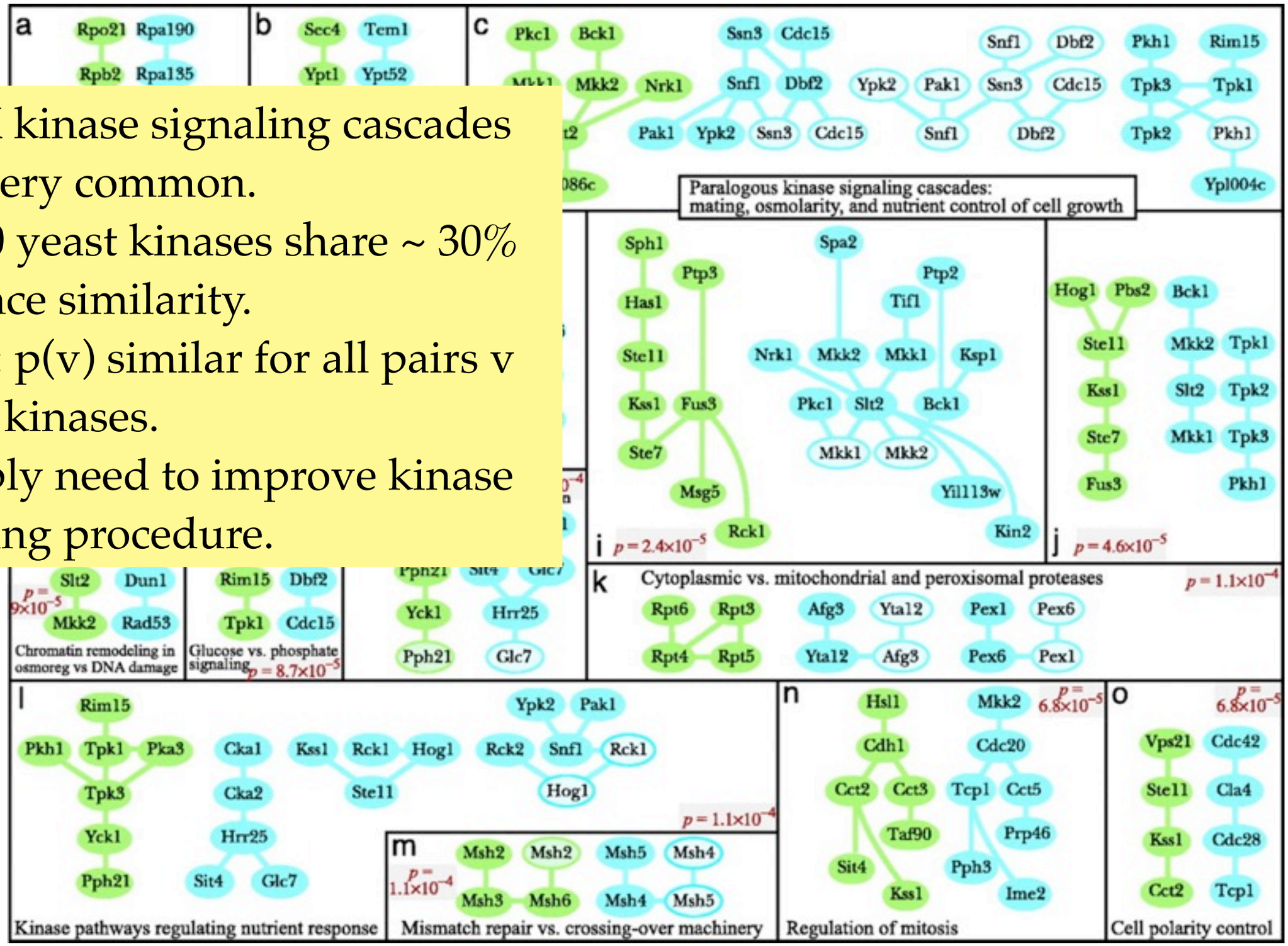
Yeast Paralogous Pathways

MAPK kinase signaling cascades were very common.

All 120 yeast kinases share ~ 30% sequence similarity.

Hence: p(v) similar for all pairs v of two kinases.

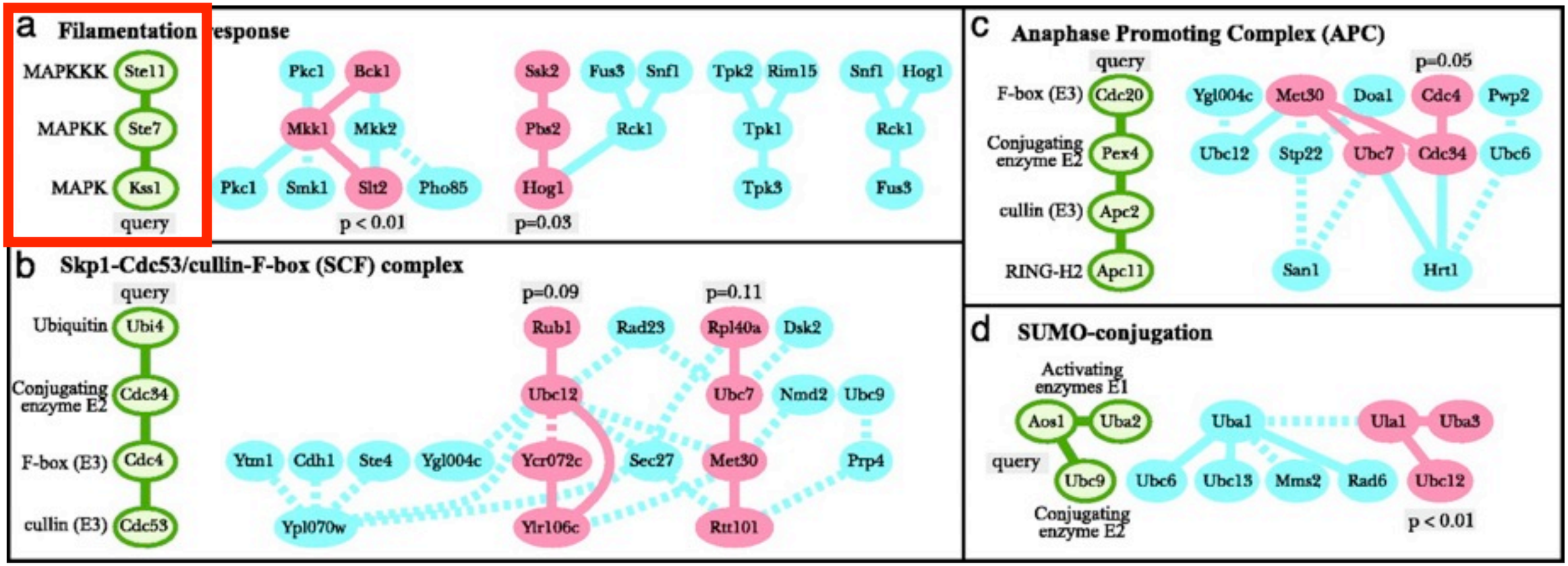
Probably need to improve kinase matching procedure.



(Kelley et al, 2003)

Searching

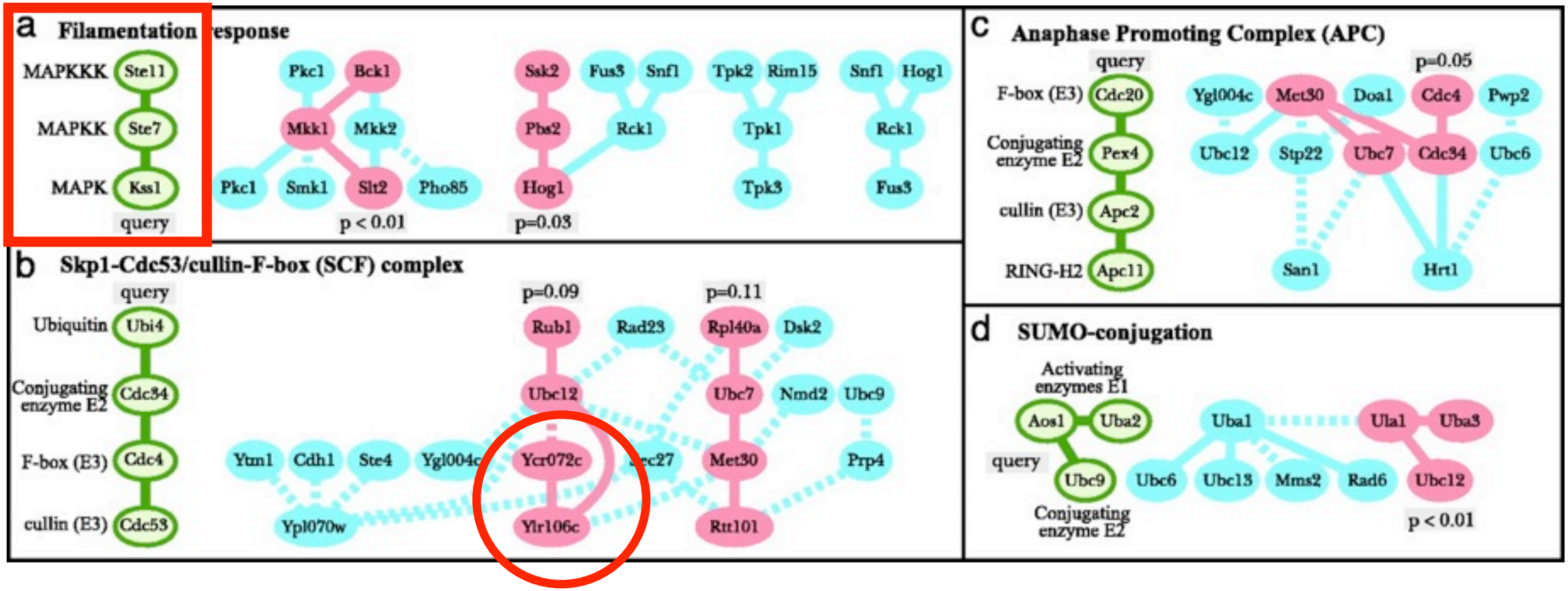
Can use local alignment to search: align a small query network to the large network.



(Kelly et al, 2003)

Searching

Can use local alignment to search: align a small query network to the large network.



(Kelly et al, 2003)

PathBLAST Summary

- Local graph alignment
- Takes into account sequence similarity & topological patterns
- Allows gaps and mismatches of length 1.
- Scoring function ~ probability of the path existing.
- Algorithm: fast, reasonable, but definitely a heuristic.
- Searching & local alignment are very related.