Comparison and alignment of networks

Sushmita Roy

sroy@biostat.wisc.edu

Computational Network Biology

Biostatistics & Medical Informatics 826

https://compnetbiocourse.discovery.wisc.edu

Nov 8th 2018

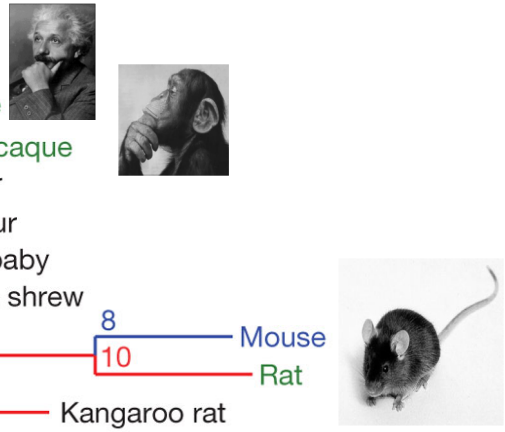
Plan for this section

- Aligning two networks to identify conserved linear paths or small complexes
  - PATHBLAST (Nov 8th)
- Global alignment of graphs using spectral methods
  - IsoRank (Nov 13th)
- Global alignment of graphs using matrix factorization
  - FUSE (Nov 15th)

Goals for today

- Introduction to the network alignment problem
- Classes of methods for network alignment
- Pairwise global network alignment
  - Identifying linear paths
  - Identifying complexes

How are these organisms related?



Toh et al, Nature, 2011

Organisms can be compared at multiple levels

- Comparison at the sequence level
  - Sequence alignment
  - Phylogenetic tree construction
- Comparison at the expression level

- Comparison at the network level

Some terminology

- Homology
  - Two sequences are said to be homologous if they are derived from a common ancestral sequence
- Orthology
  - Two proteins in two organisms are said to be orthologs of each other if they are related by a common ancestor
- Paralogy
  - Two proteins that are related by a duplication event within a species
- Match, mismatch, gaps
  - Terms used in sequence alignment
- BLAST
  - A software program used to align two molecular sequences that provides a statistical score (BLAST E-value) used to assess the quality of the alignment

How do sequences change between organisms?

- Substitutions

Sequence 1 T H I S S E Q U E N C E

mismatch

Sequence 2 T H A T S E Q U E N C E

- Deletions

Sequence 1 T H I S I S A S E Q U E N C E

Sequence 2 T H I S _ _ _ S E Q U E N C E

gap

- Insertions

Sequence 1 _ _ _ _ S E Q U E N C E

gap

Sequence 2 T H I S S E Q U E N C E

Alignment problem

- **Sequence alignment**
  - Given the genomic sequence of two species, find the differences and similarities of the sequence
  - Aims to find a correspondence between the positions of two sequences while minimizing the number of substitutions and gaps
- **Network alignment**
  - Given molecular interaction networks from different organisms find the differences and similarities between them at the subnetwork level
  - Aims to find a correspondence between the positions of two networks while minimizing the number of substitutions and gaps on the network

Sequence Alignment Examples

Sequence 1 T H I S S E Q U E N C E

T H + + S E Q U E N C E

Sequence 2 T H A T S E Q U E N C E

Sequence 1 T H I S - - - S E Q U E N C E

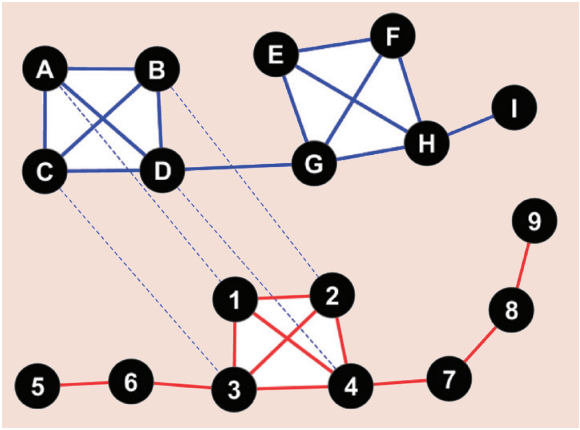
T H + + S E Q U E N C E

Sequence 2 T H A T I S A S E Q U E N C E

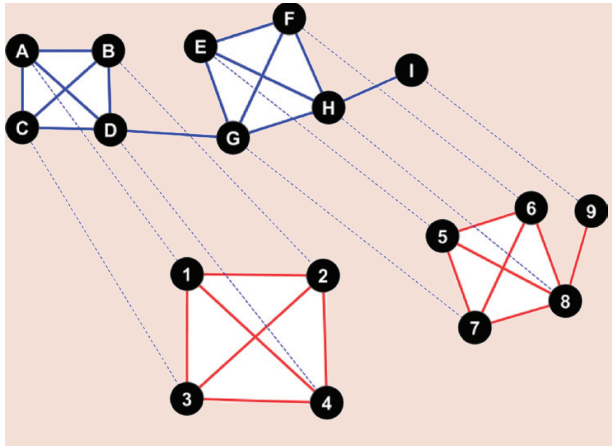
Why is network alignment important?

- Important from an evolutionary perspective
  - Are interactions of proteins with similar sequence conserved?
  - How do networks evolve?
  - Is there a minimal set of interactions common to all species?
- Refine existing interaction networks

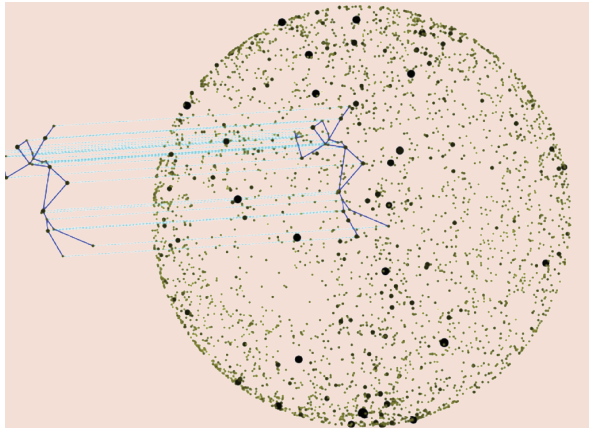
Different network alignment problems



Local Network alignment: Find locally similar subnetworks



Global network alignment: Align all nodes in one network to all nodes in the second network



Network query: Find instances of a small subnetwork in a larger network

Nir Atias and Roded Sharan, May 2012, ACM Communications

Different types of network alignment problems

Problem type	Description	Methods
Local alignment	Align small parts of the network	PathBLAST, LocalAli
Global alignment	Align the entire network	FUSE, IsoRank
Pairwise alignment	Align two networks	PathBLAST
Multiple network alignment	Align more than two networks	FUSE, IsoRank, LocalAli
NetworkQuery	Search for a small network in a larger network	NetGrep, QNet, QPath

Adapted from Nir Atias and Roded Sharan, May 2012, ACM Communications

What makes network alignment difficult?

- The set of genes/proteins between species are not the same
- The correspondence between genes of one species and the genes of another species is not one-to-one
  - Although many algorithms assume one-to-one mapping
- Underlying networks might be noisy and/or incomplete

Goals for today

- Introduction to the network alignment problem
- Classes of methods for network alignment
- Pairwise global network alignment
  - Identifying linear paths
  - Identifying complexes

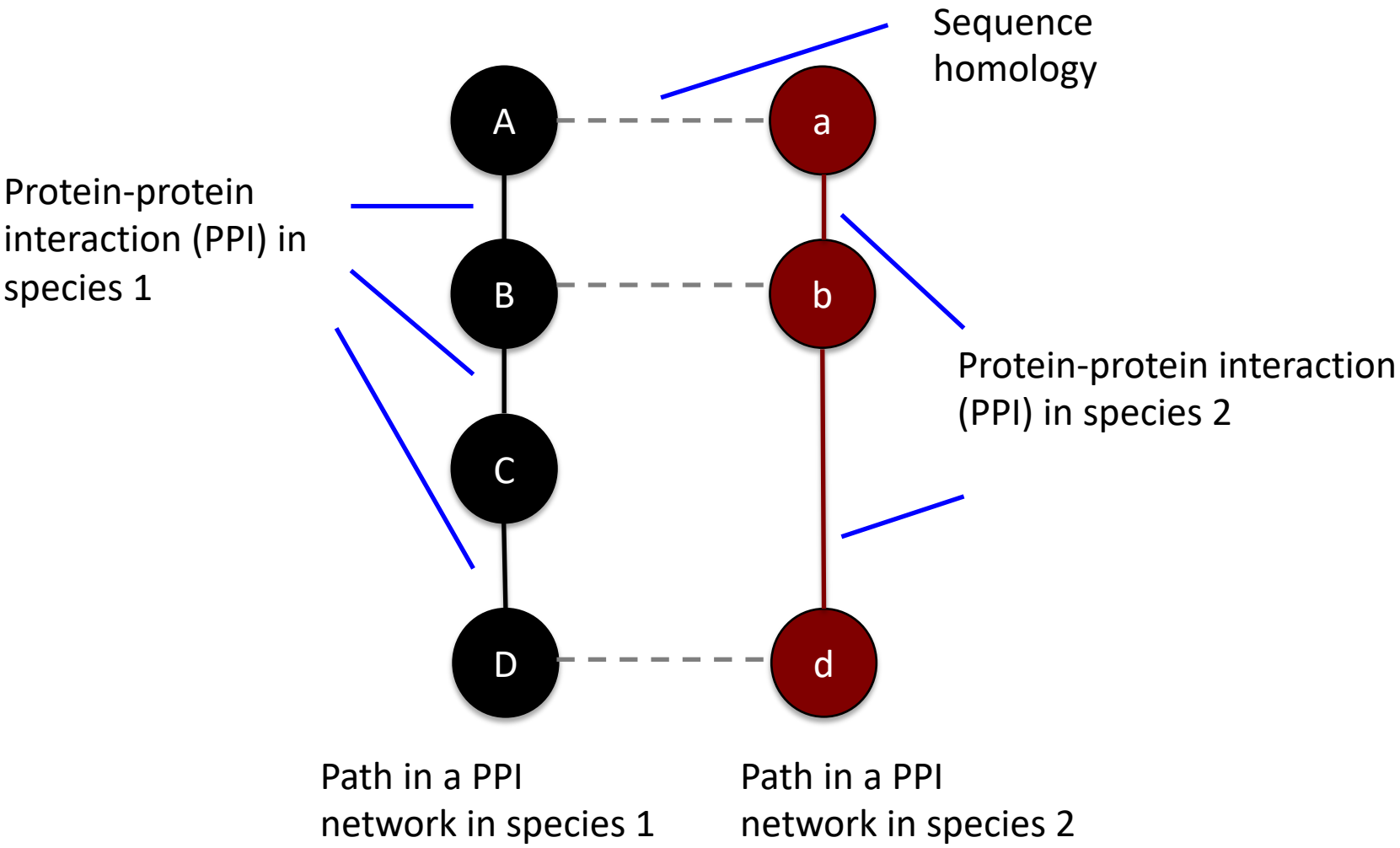
Pairwise network alignment approach

- Conserved pathways within bacteria and yeast as revealed by global protein network alignment
    - B. P. Kelley, R. Sharan, R. M. Karp, T. Sittler, D. E. Root, B. R. Stockwell, and T. Ideker, PNAS 2003
- Identification of Protein Complexes by Comparative Analysis of Yeast and Bacterial Protein Interaction Data
    - R. Sharan, T. Ideker, B. Kelley, R. Shamir, and R. M. Karp, Journal of Computational Biology, 2004

Defining a pairwise alignment problem

- Given
  - Two graphs
    - $G_1=(V_1,E_1)$
    - $G_2=(V_2,E_2)$
      - V_1 and V_2 correspond to vertex set
      - E_1 and E_2 correspond to the interaction set
  - A possibly incomplete many-to-many mapping between V_1 and V_2
- Do
  - Identify regions that are similar across the networks being compared
- Typically done by
  - Generating a network alignment graph
  - Defining a scoring function that assesses both node and topological similarity

Network alignment example



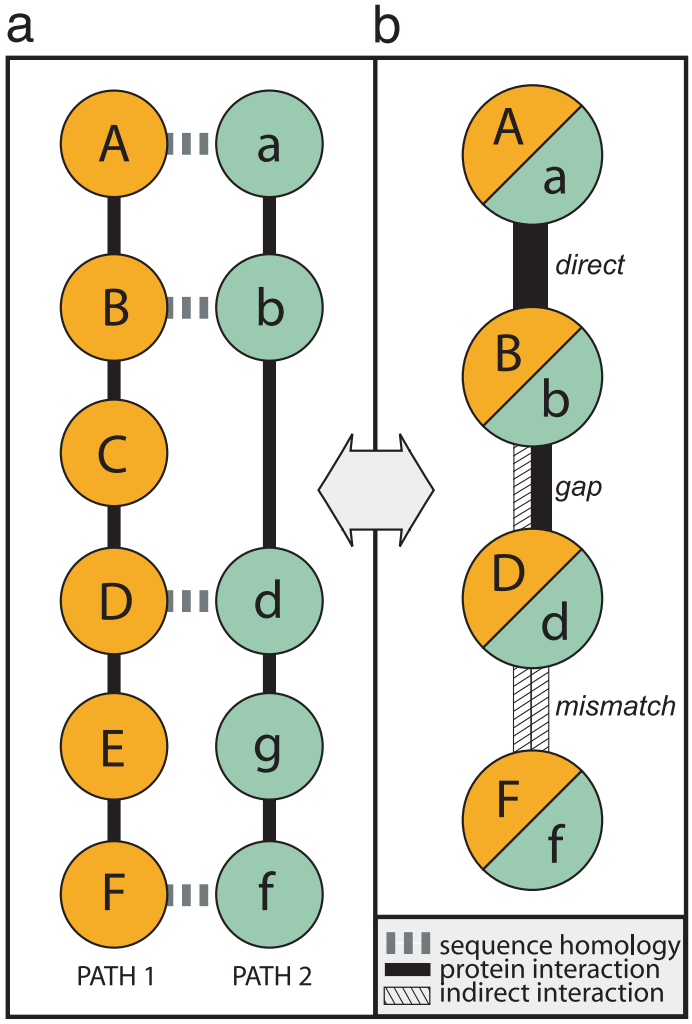
Gaps and mismatches in network alignment

- Mismatch
  - Occurs when aligned proteins in the network alignment do not share sequence homology
  - This boils down to pair (A,B) in one species connected to pair (a,b) in another species by a distance of 2.

- Gap
  - Occurs when a protein interaction in one path skips over a protein in the other network

Network alignment graph

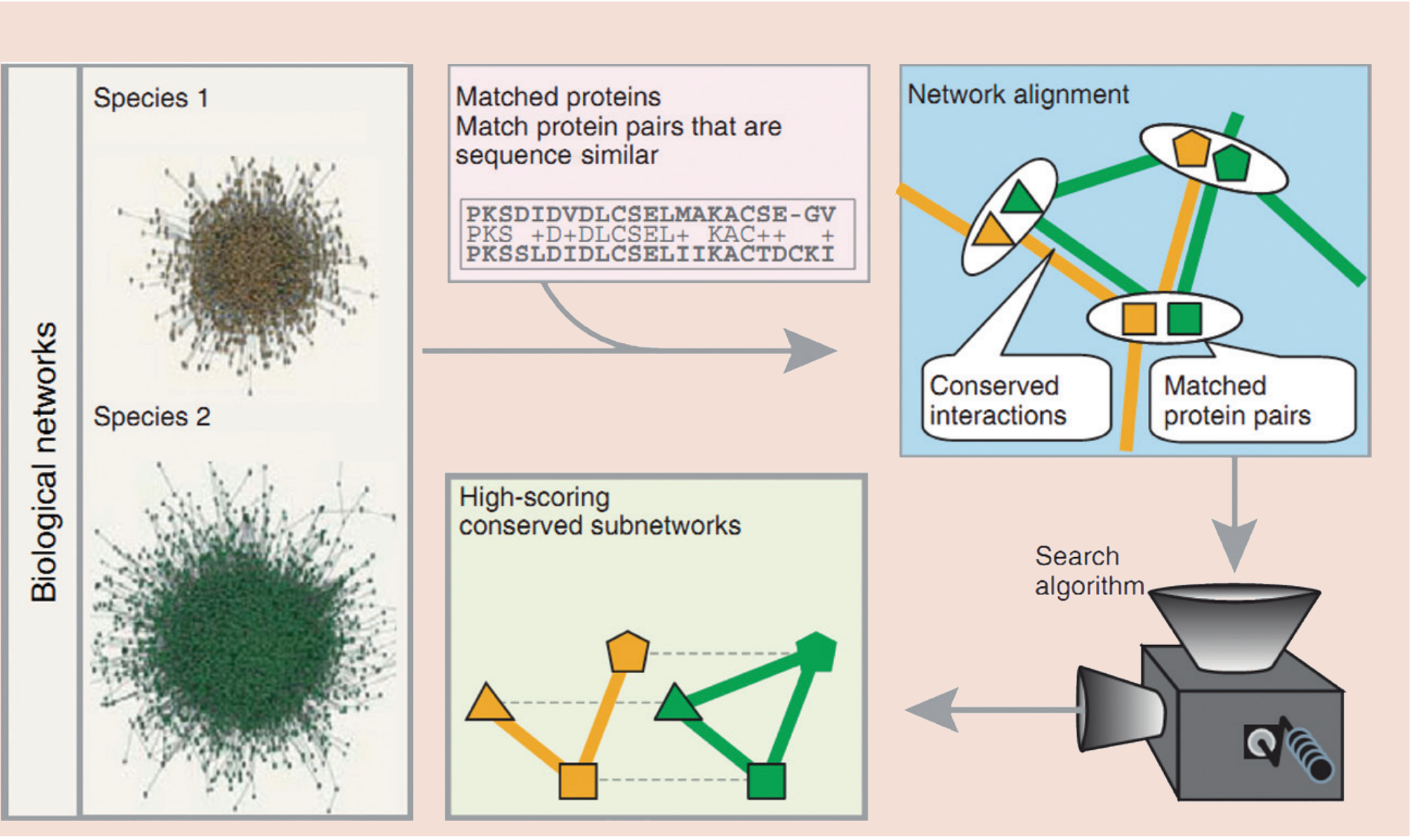
The network alignment problem of finding conserved paths boils down to finding a high scoring path in a new network called the “network alignment” graph.



Each vertex in the alignment graph is represented by a pair of vertices and a corresponding numeric score (BLAST E-value) that specifies the sequence similarity

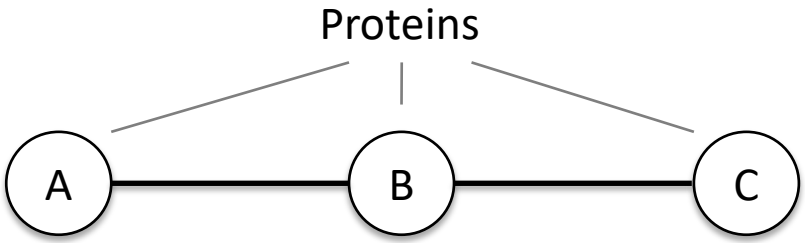
Each edge corresponds to to “interactions” in the original graph

Sketch of a network alignment approach



PathBLAST algorithm for network alignment

- PathBLAST algorithm aims to find conserved linear pathways
  - A pathway is defined as a sequence of protein-protein interactions forming a connected path
- PathBLAST algorithm has two steps:
  1. Combine two protein-protein interaction networks to create the network alignment graph
     - A path through this graph is a conserved pathway
  2. Search the network alignment graph for high-scoring paths



A pathway here is a linear path

Scoring a path P

- Let $S(P)$ denote the score of a path P
- PathBLAST defines S(P) in a decomposable manner as follows

$$S(P) = \sum_{v \in P} \log_{10} \frac{p(v)}{p_{\text{random}}} + \sum_{e \in P} \log_{10} \frac{q(e)}{q_{\text{random}}},$$

- $p(v)$ is the probability of true homology between two proteins represented by the pair v given the BLAST score associated with v
- $q(e)$ is the probability that the interaction e is real
- p_{random} and q_{random} are expected values of $p(v)$ and $q(e)$ over all vertices and edges in the alignment graph

Estimating the node and edge probabilities

- $p(v)$ is obtained from sequence alignment scores E_v of the two proteins making up v

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

- H represents the event of true homology (obtained from a curated set of homologs)
- $q(e)$ is obtained from underlying protein-protein interactions e represents

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

Number of studies	Pr(i)	# yeast interactions
1	0.1	9966
2	0.3	1597
≥ 3	0.9	1591

e can have 2 (direct), 3 (gap) or 4 (mismatch) interactions

Strategy to find high scoring path

- Find the highest scoring pathway alignment of a fixed pre-specified length L
- Based on a dynamic programming algorithm
- The highest scoring path of length l =2..L ending in vertex v will have score:

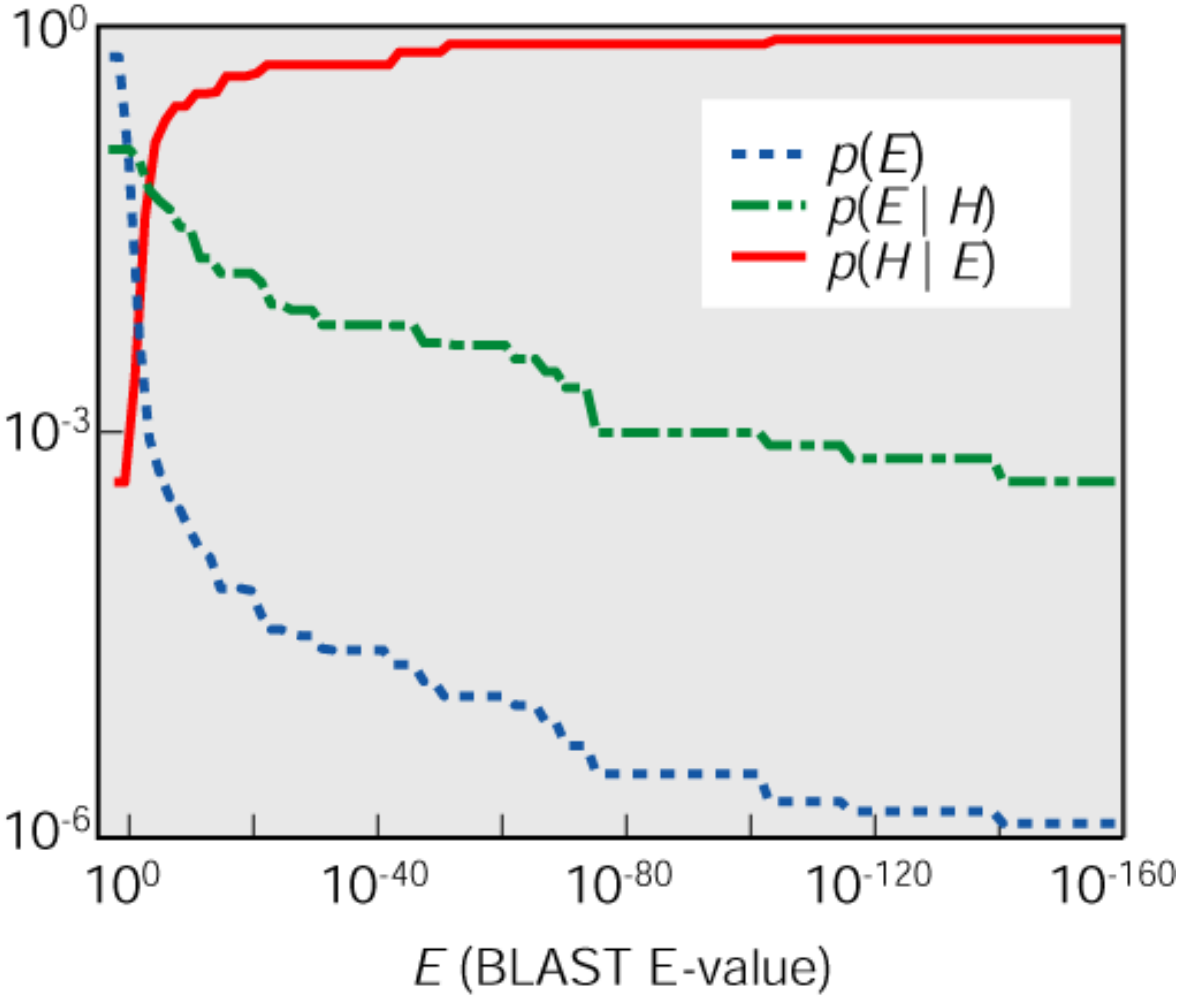
$$S(v,l) = \arg\max_{u \in parents(v)} \left[S(u,l-1) + \log \frac{p(v)}{p_{random}} + \log \frac{q(e_{u \to v})}{q_{random}} \right]$$

- With base case

$$S(v,1) = \log \frac{p(v)}{p_{random}}$$

- Thus the score of length l path is computed from score of path of length l-1

Estimating the probability of true homology



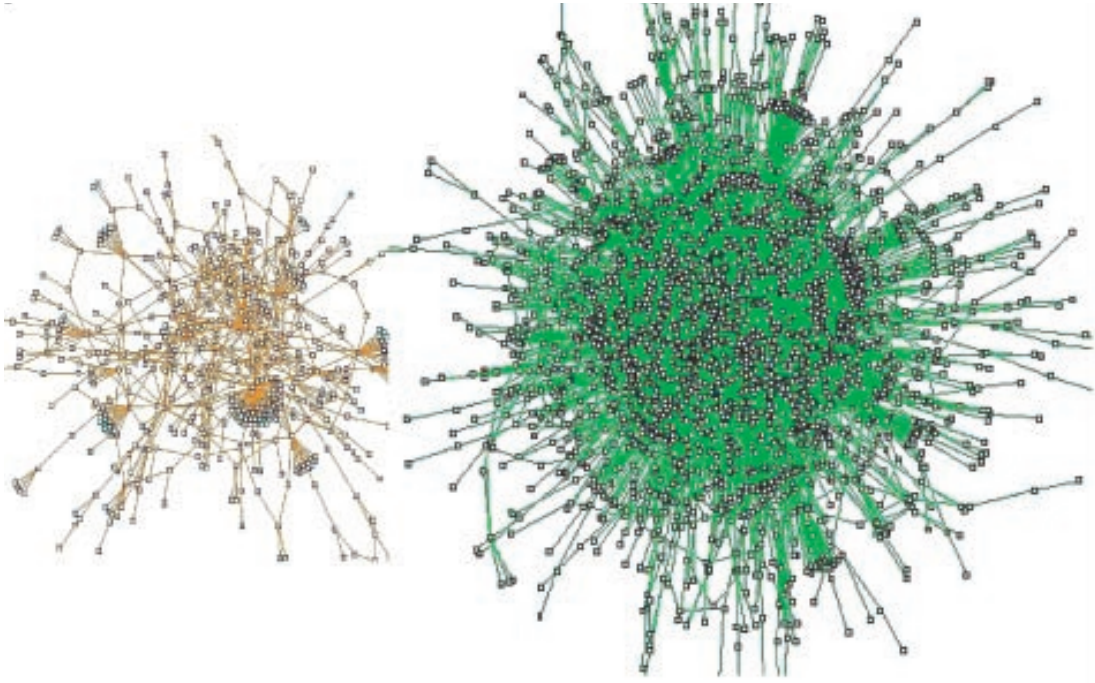
Results

- Perform alignment of Yeast (*S. cerevisiae*) vs Bacteria (*H. pylori*) protein-protein interaction networks

- Perform alignment of Yeast vs Yeast
  - Find paralogous (duplicate) pathways within the same species
- Query the large interaction network to find instances of a smaller network

Network dataset description

- Yeast network
  - 14,489 interactions and 4,688 proteins
- Bacteria network
  - 1,465 interactions among 732 proteins



Helicobacter pylori

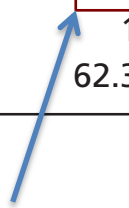
Bacteria

Saccharomyces cerevisiae

Yeast

A few statistics of the networks to be aligned

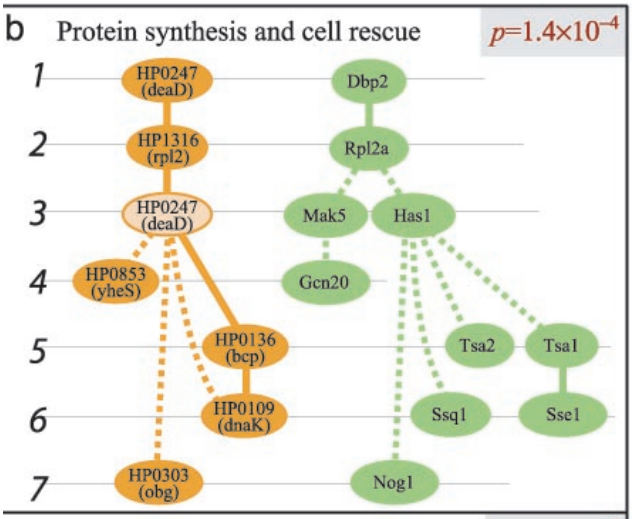
	Vertices (homologs)	Edges			
		Total	Direct	Gap	Mismatch
Yeast vs. *H. pylori* ($E_{cutoff} = 10^{-2}$)	829	2,036	7	260	1,769
Random: mean ± SD		509.0 ± 128.0	2.5 ± 1.9	68.8 ± 23.8	437.7 ± 110.3
Yeast vs. yeast ($E_{cutoff} = 10^{-10}$)	5,593	1,389	1,389	N/A	N/A
Random: mean ± SD		62.3 ± 29.4	62.3 ± 29.4	N/A	N/A

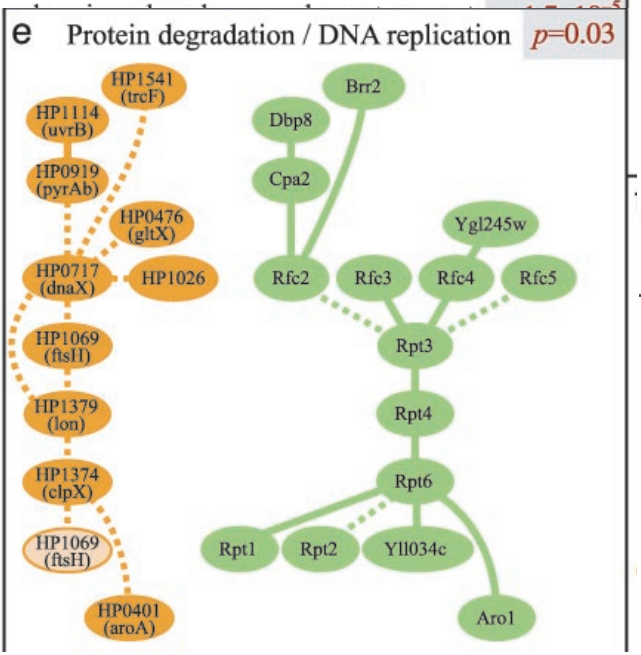


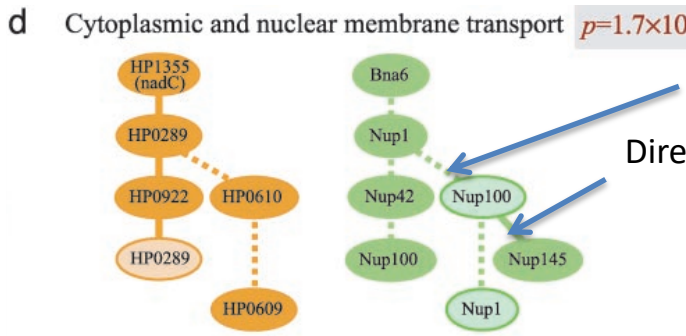
Very few direct interactions!

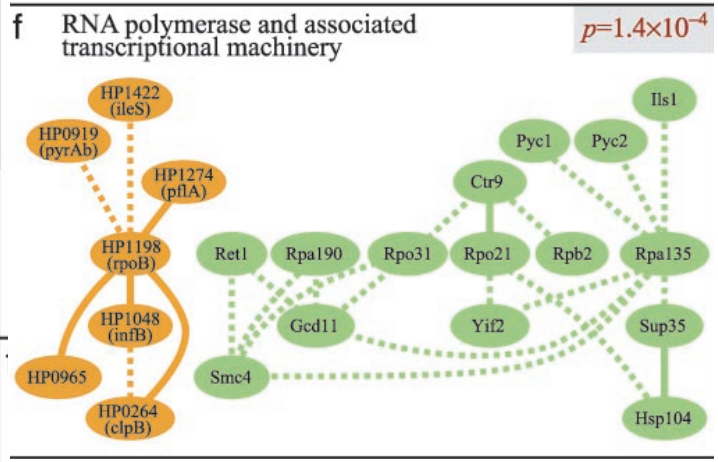
Randomization: Permute the protein names

Aligning yeast to bacteria









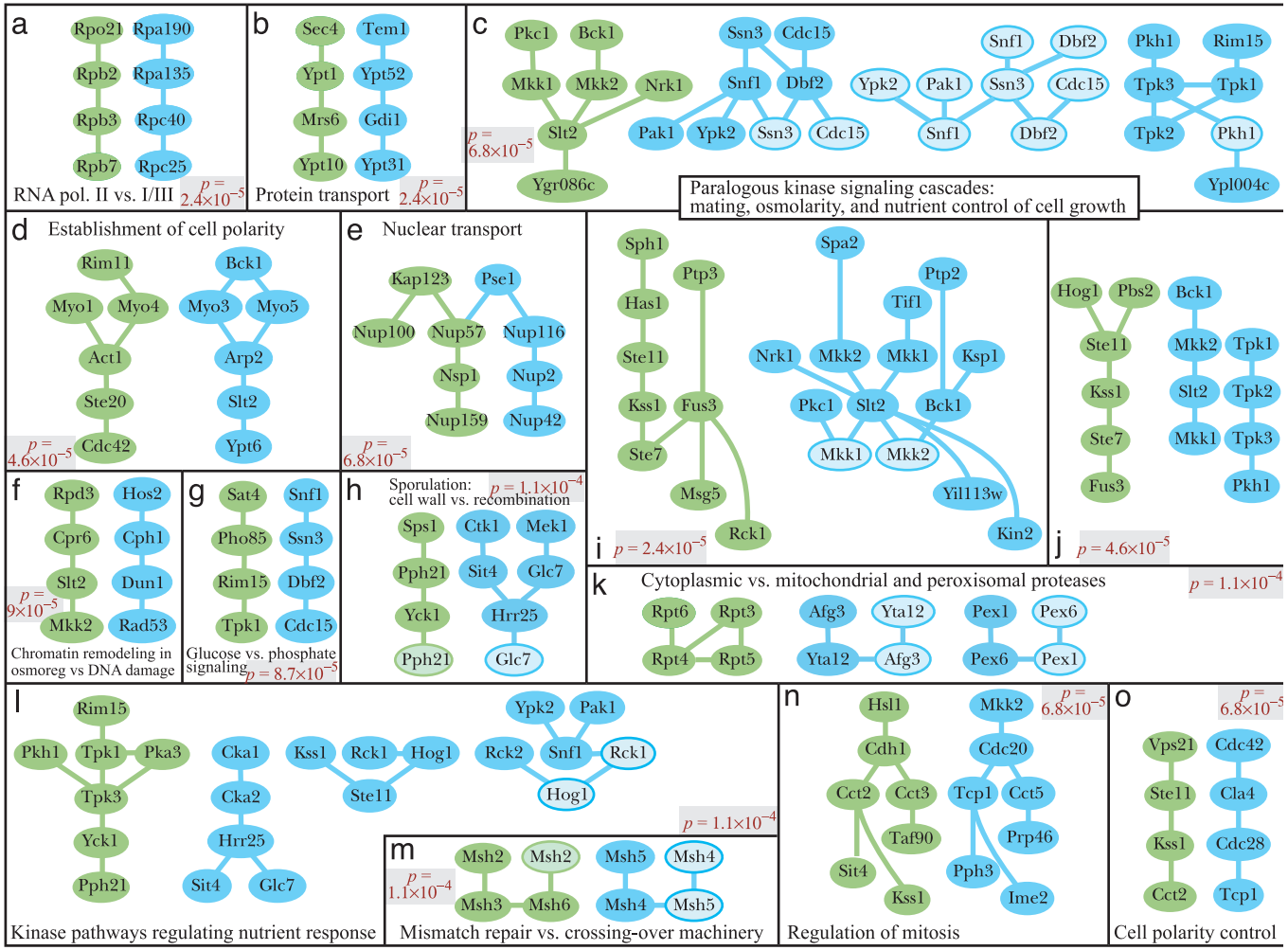
Indirect (gap/mismatch)

Direct interactions

Highlights of results:

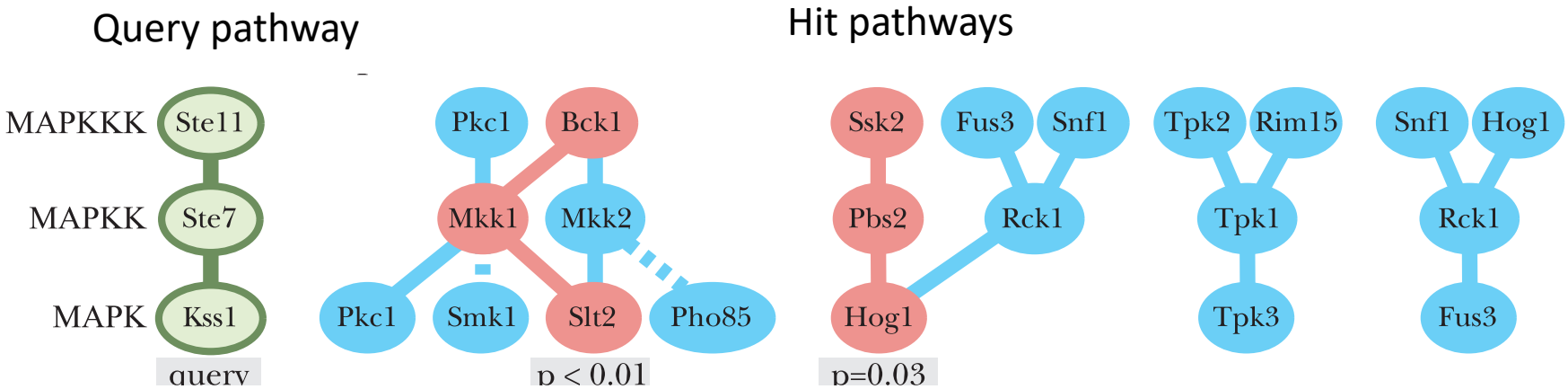
1. 150 high scoring paths of length 4 were identified
2. A total of 4.1% *H. pylori* and 1.2% *S. cerevisiae* proteins were in the high scoring alignments
3. Paths were enriched for diverse biological processes (protein synthesis, cell rescue, degradation)

Aligning yeast to yeast



1. A total of 300 high scoring pathways
2. This can help identify paralogous pathways
3. Identified several known distinct but similarly functioning complexes

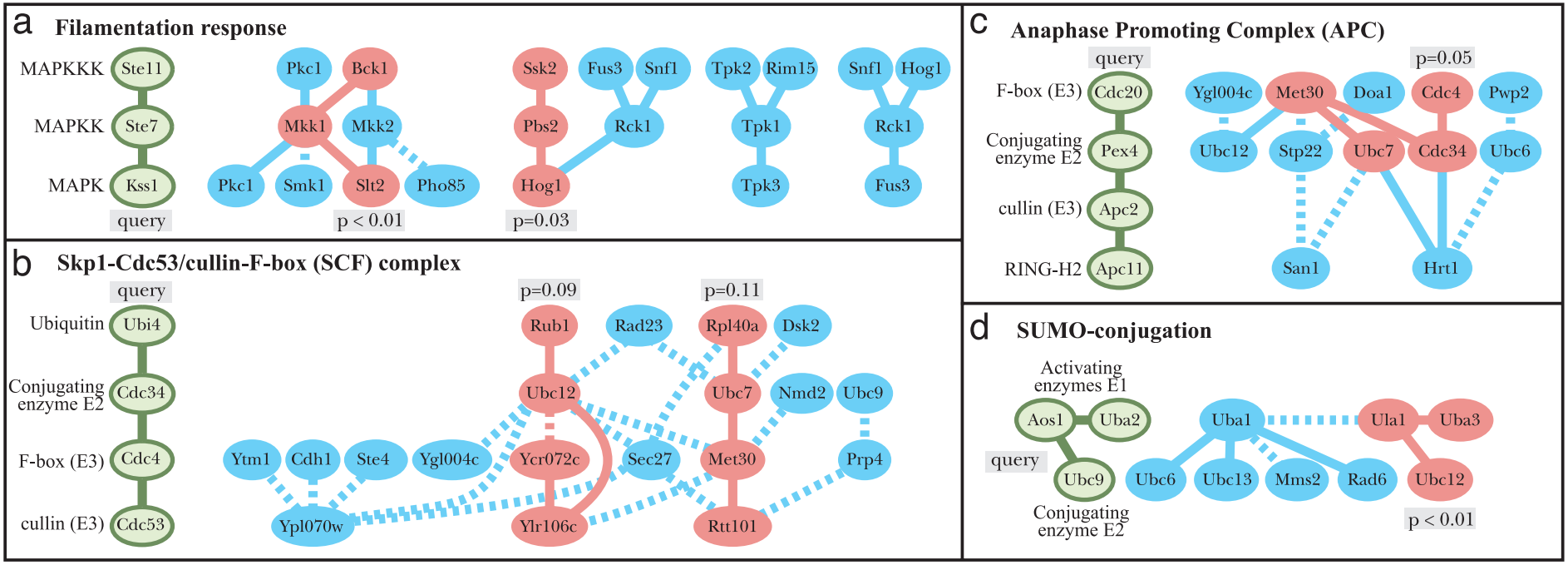
Network query: Align a query pathway to a larger network



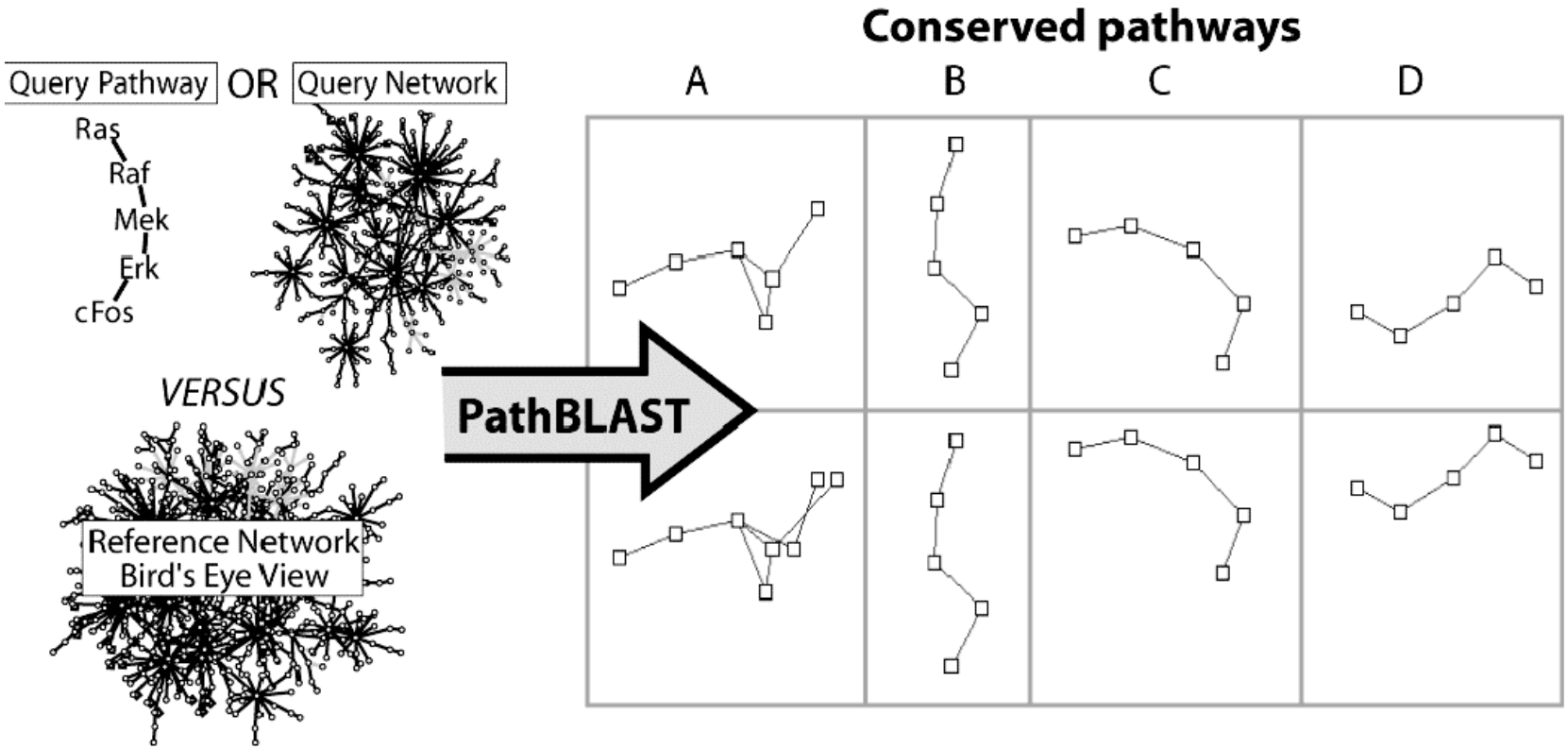
Highest scoring pathways in red

Results for the MAPK pathway as a query pathway.

Network query: Align a query pathway to a larger network



PathBLAST server



Kelly et al., 2004

http://www.pathblast.org

Take away points for PathBLAST

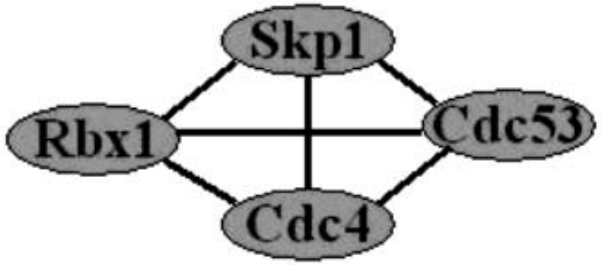
- A pairwise network alignment program that can identify conserved linear paths
- Applied to ask two main questions:
  - How similar are Yeast and Bacteria PPI networks?
    - Application to Yeast and Bacteria networks enabled using information from a well-studied organism to study a poorly studied organism
    - Pathway alignments often links two pathways that were not known to associate before
    - Proteins with high sequence similarity did not necessarily pair with each in the same pathway
  - Are there redundant pathways in yeast PPI networks?

Pairwise network alignment approach

- Conserved pathways within bacteria and yeast as revealed by global protein network alignment
    - B. P. Kelley, R. Sharan, R. M. Karp, T. Sittler, D. E. Root, B. R. Stockwell, and T. Ideker, PNAS 2003
- Identification of Protein Complexes by Comparative Analysis of Yeast and Bacterial Protein Interaction Data
    - R. Sharan, T. Ideker, B. Kelley, R. Shamir, and R. M. Karp, Journal of computational biology 2004

Identification of Protein Complexes between two species

- PathBLAST focused on finding chain like subnetworks
- Sharan et al, 2004 aimed to find protein complexes
  - more densely connected



- Key properties of this approach:
  - A more formal probabilistic model for finding dense subgraphs
  - Create an orthology graph (similar to the network alignment graph)
    - nodes correspond to pairs of putative orthologous proteins
    - a protein may appear in multiple nodes with different orthologs

Probabilistic model to find dense complexes in one species

- Let U denote a set of vertices corresponding to proteins
- Score its tendency to be a complex based on a likelihood score

$$L(U) = \log \frac{Pr(O_U|M_c)}{Pr(O_U|M_n)}$$

O_U : Observed interactions among U
M_C :The protein complex model specifying the interaction probabilities for a complex
M_n : The null model specifying the interaction probabilities if edges were randomly distributed

The protein complex model M_c

$$Pr(O_U|M_c) = \prod_{(u,v)\in U\times U} Pr(O_{uv}|M_c)$$ Assume independence of edges

$$= \prod_{(u,v)\in U\times U} (Pr(O_{uv}|T_{uv}, M_c)Pr(T_{uv}|M_c) + Pr(O_{uv}|F_{uv}, M_c)Pr(F_{uv}|M_c))$$

Sum over true edge status (hidden), T_{uv} and F_{uv}

$$= \prod_{(u,v)\in U\times U} (\beta Pr(O_{uv}|T_{uv}) + (1-\beta)Pr(O_{uv}|F_{uv}))$$

Set empirically (a prior)

Obtained from Deng et al (2013) that provides a reliability measure for observed interactions

The null model M_n

- The null model assumes that edges in a graph G are drawn randomly

$$Pr(O_U|M_n) = \prod_{(u,v)\in U\times U} (p_{uv} Pr(O_{uv}|T_{uv}) + (1 - p_{uv}) Pr(O_{uv}|F_{uv})) .$$

- To estimate the p_{uv}, the authors generated random graphs by doing a series of edge crosses
  - Pick two edges (a,b) and (c,d) and replace with (a,c) and (b,d)
  - p_{uv} is the proportion of random graphs in which (u,v) was present

Extension to two species

- Let U^1 be the set of vertices in species 1 and V^2 be the set of vertices in species 2
- Let Θ denote the mapping from U^1 and V^2
- Score of a conserved complex is

$$S_\theta(U^1, V^2) = L(U^1, V^2) + \sum_{i=1}^{k_1} \sum_{v_j^2 \in \theta(u_i^1)} \log \frac{Pr(h_{u_i^1 v_j^2} | E_{u_i^1 v_j^2})}{Pr(h)}.$$

True homology

Prior probability of homology

Similarity score of pairs of proteins

where

$$L(U^1, V^2) = \log \frac{Pr(O_{U^1} | M_c^1)}{Pr(O_{U^1} | M_n^1)} + \log \frac{Pr(O_{V^2} | M_c^2)}{Pr(O_{V^2} | M_n^2)}.$$

Searching for complexes

- Define a complete edge and node weighted orthology graph (extends the network alignment graph of Kelly et al)
  - Weight of a node is proportional to the probability of the constituent nodes to be homologous
  - Weight of an edge is derived from the M_C and M_n models
  - Consider all pairs of edges between two interacting proteins
- An induced subgraph of an orthology graph corresponds to a subset of proteins from each species
  - NP hard in general

- Heuristic search
  - Find heavy seeds
  - Refine seeds exhaustively
  - Expand seed by local search

Heuristic search of heavy subgraphs

- A strong edge on the orthology graph is defined as an edge that has a positive score
- A seed is defined around each vertex v using all its neighbors u
- If seed has >10 nodes, remove nodes with minimal scores until seed has 10 nodes
- Consider all subsets of size 3 or more of the seed that contain v
- Expand seed with local search to increase the score of the subgraph or until a max size was reached (20)
  - Add a new node
  - Remove a node

Conserved complexes in yeast and bacteria

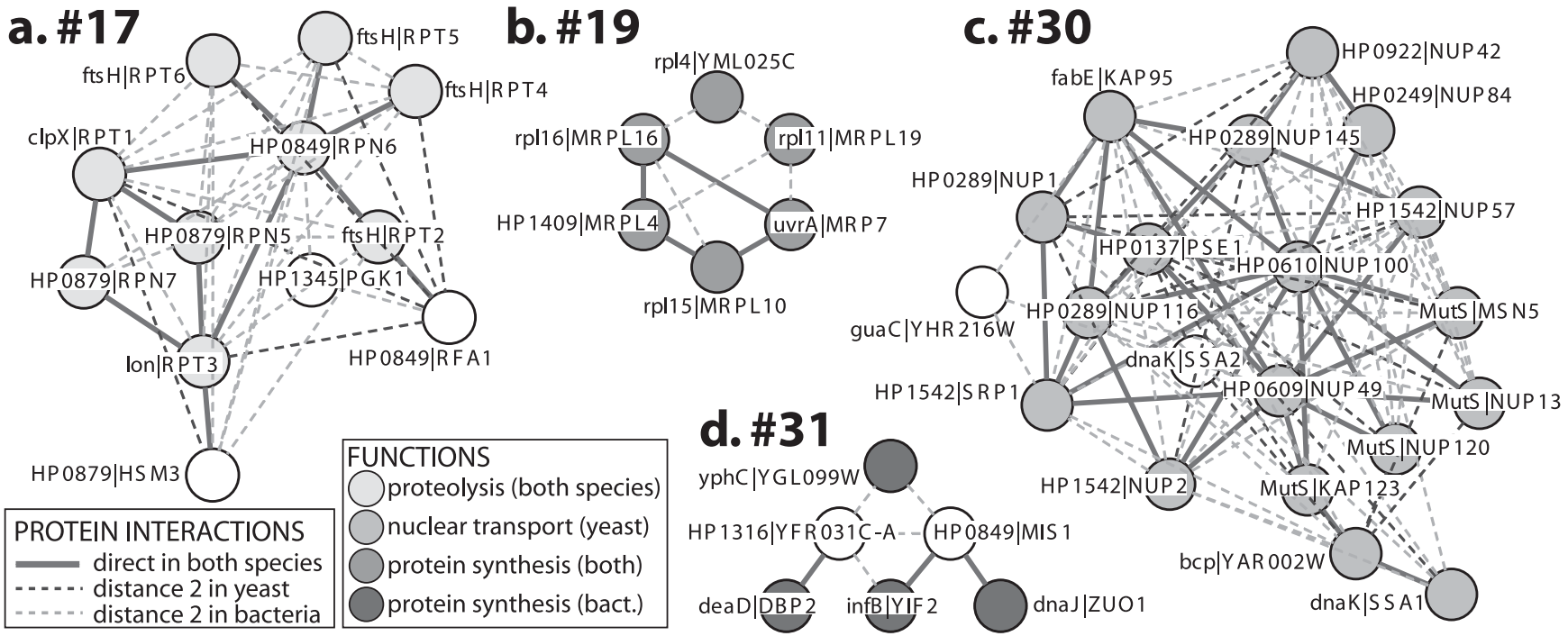


FIG. 2. Conserved protein complexes for proteolysis (**panel a**), protein synthesis (**panels b and d**), and nuclear transport (**panel c**). Conserved complexes are connected subgraphs within the bacteria-yeast orthology graph, whose nodes represent orthologous protein pairs and edges represent conserved protein interactions of three types: direct interactions in both species (solid edges); direct in bacteria but distance 2 in the yeast interaction graph (dark dashed edges); and distance 2 in the bacterial interaction graph but direct in yeast (light dashed edges). In the algorithm, both nodes and edges are assigned weights according to the probabilistic model. The number of each complex indicates the corresponding complex ID listed in Table 1.

Comparison to existing approaches

TABLE 2. PERFORMANCE COMPARISON OF THREE ALGORITHMS FOR COMPLEX DETECTION

Algorithm	*Jaccard*	*Sensitivity*	*Specificity*
This study	0.32	0.33	0.7
Kelley *et al*. (2003)	0.22	0.44	0.4
Yeast only	0.33	0.67	0.48

Sharan et al have higher specificity compared to existing approaches

Concluding remarks

- PathBLAST and its extension to complexes can be useful for
  - Transferring information from a well-annotated organism (*S. cerevisiae*) to shed insight into a poorly annotated organism (*H. pylori*)
  - Infer function of poorly annotated proteins
  - Accurately identify protein complexes
- Some questions remain
  - How to deal with multiple networks?
  - Are there better algorithms to search the orthology graph/network alignment graph

References

- N. Atias and R. Sharan, "Comparative analysis of protein networks: Hard problems, practical solutions," *Commun. ACM*, vol. 55, no. 5, pp. 88-97, May 2012. [Online]. Available: http://dx.doi.org/10.1145/2160718.2160738
- B. P. Kelley, R. Sharan, R. M. Karp, T. Sittler, D. E. Root, B. R. Stockwell, and T. Ideker, "Conserved pathways within bacteria and yeast as revealed by global protein network alignment," *Proceedings of the National Academy of Sciences*, vol. 100, no. 20, pp. 11 394-11 399, Sep. 2003. [Online]. Available: http://dx.doi.org/10.1073/pnas.1534710100
- R. Sharan, T. Ideker, B. Kelley, R. Shamir, and R. M. Karp, "Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data." *Journal of computational biology : a journal of computational molecular cell biology*, vol. 12, no. 6, pp. 835-846, Jul. 2005. [Online]. Available: http://dx.doi.org/10.1089/cmb.2005.12.835