

Network biology minicourse (part 4)
Algorithmic challenges in genomics

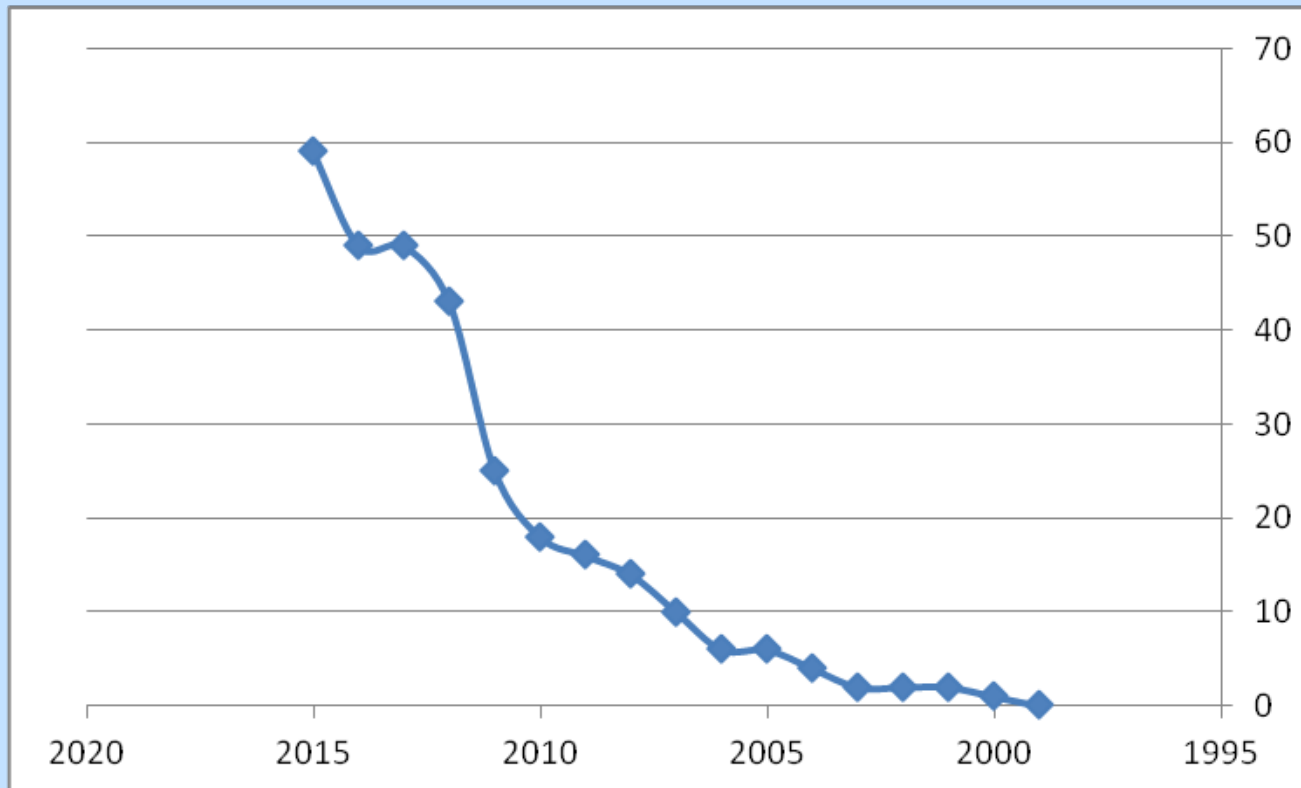
Network alignment and querying

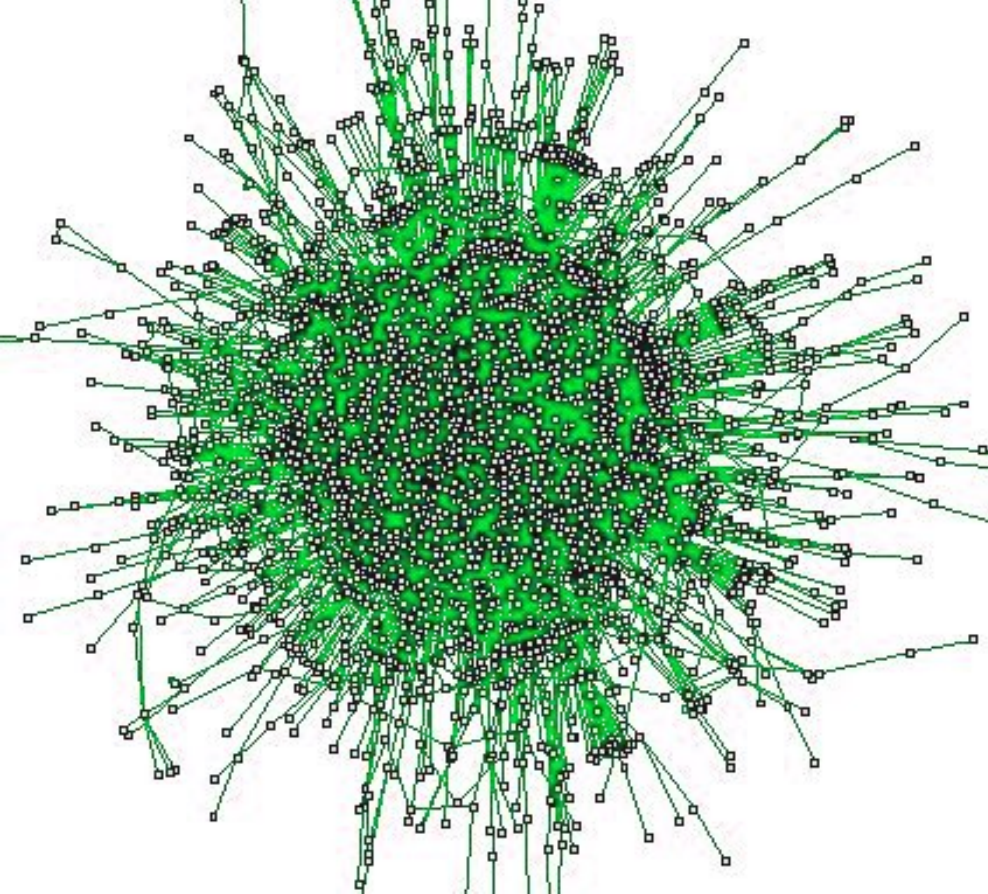
Roded Sharan

School of Computer Science, Tel Aviv University

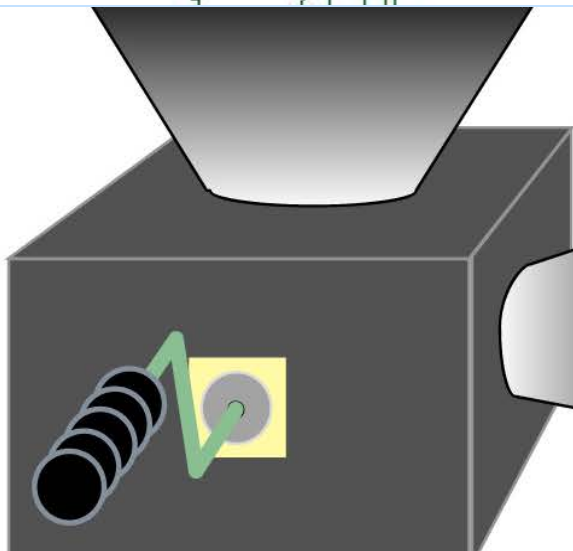
Multiple Species PPI Data

- Rapid growth in number of species measured.





Distilling Modules



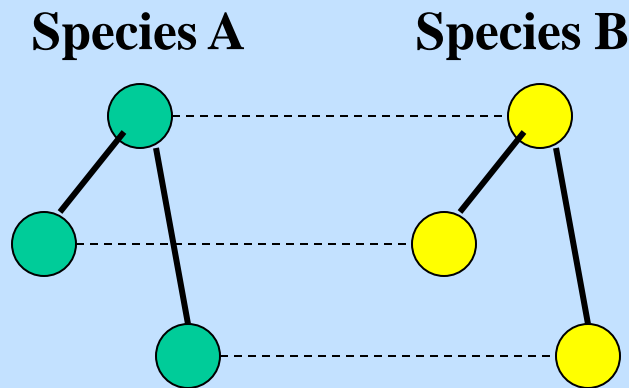
Problem:

Data is partial and noisy.

Being Comparative

Paradigm: Evolutionary conservation implies functional significance.

Conservation: similarity in sequence and interaction topology.



Main challenges

Local network alignment: detect *conserved* subnetworks across two (or more) networks.

Global network alignment: find 1-1 mapping between networks.

Network querying: given a query subnetwork in species A, find similar instances in the network of species B.

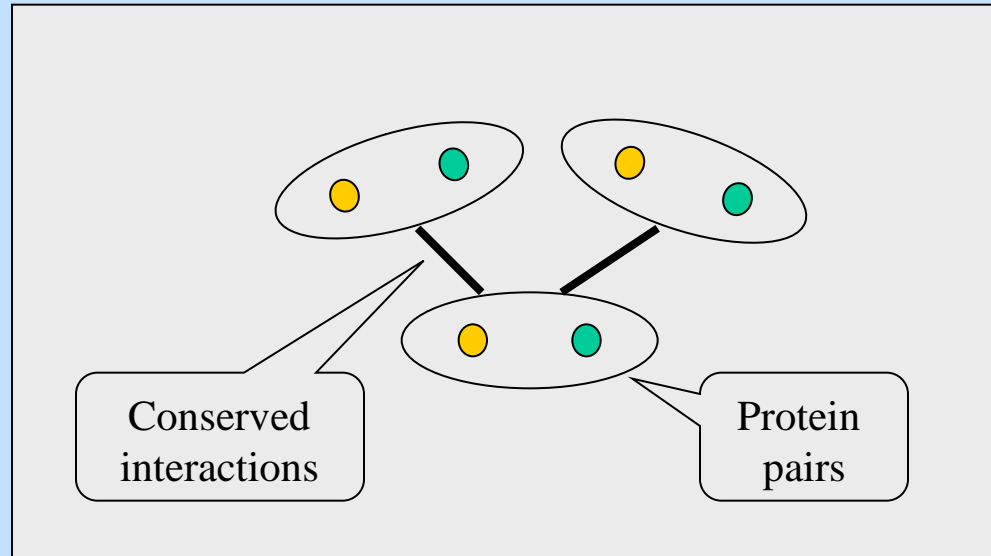
Local Pairwise Alignment

Problem definition

Given two networks (of two species), find pairs of subnetworks (one from each species) that are significantly similar.

- Similarity is measured both on vertices (sequence similarity) and edges (topological similarity).
- Under certain formulations reduces to subgraph isomorphism (NP hard).

Network Alignment



Alignment graph:

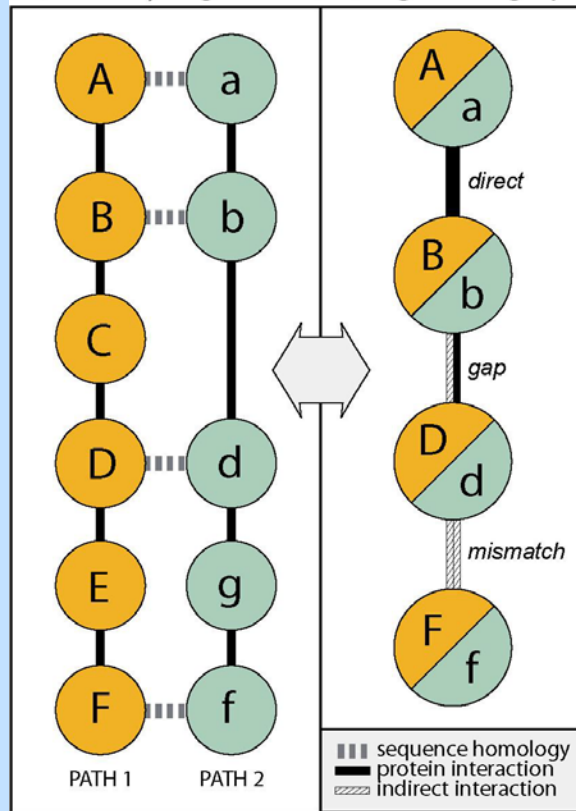
Nodes: pairs of sequence-similar proteins, one per species.

Edges: conserved interactions.

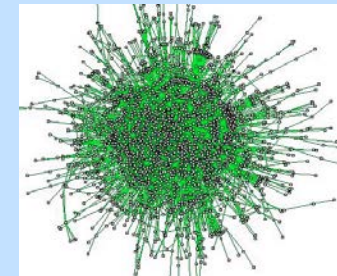
- Facilitates search for conserved subnetworks.
- First introduced by Ogata et al.'00 and Kelley et al.'03.

PathBLAST (Kelley et al.'03)

[a] Pathway alignment [b] Alignment graph



H. pylori
~1500 PPIs

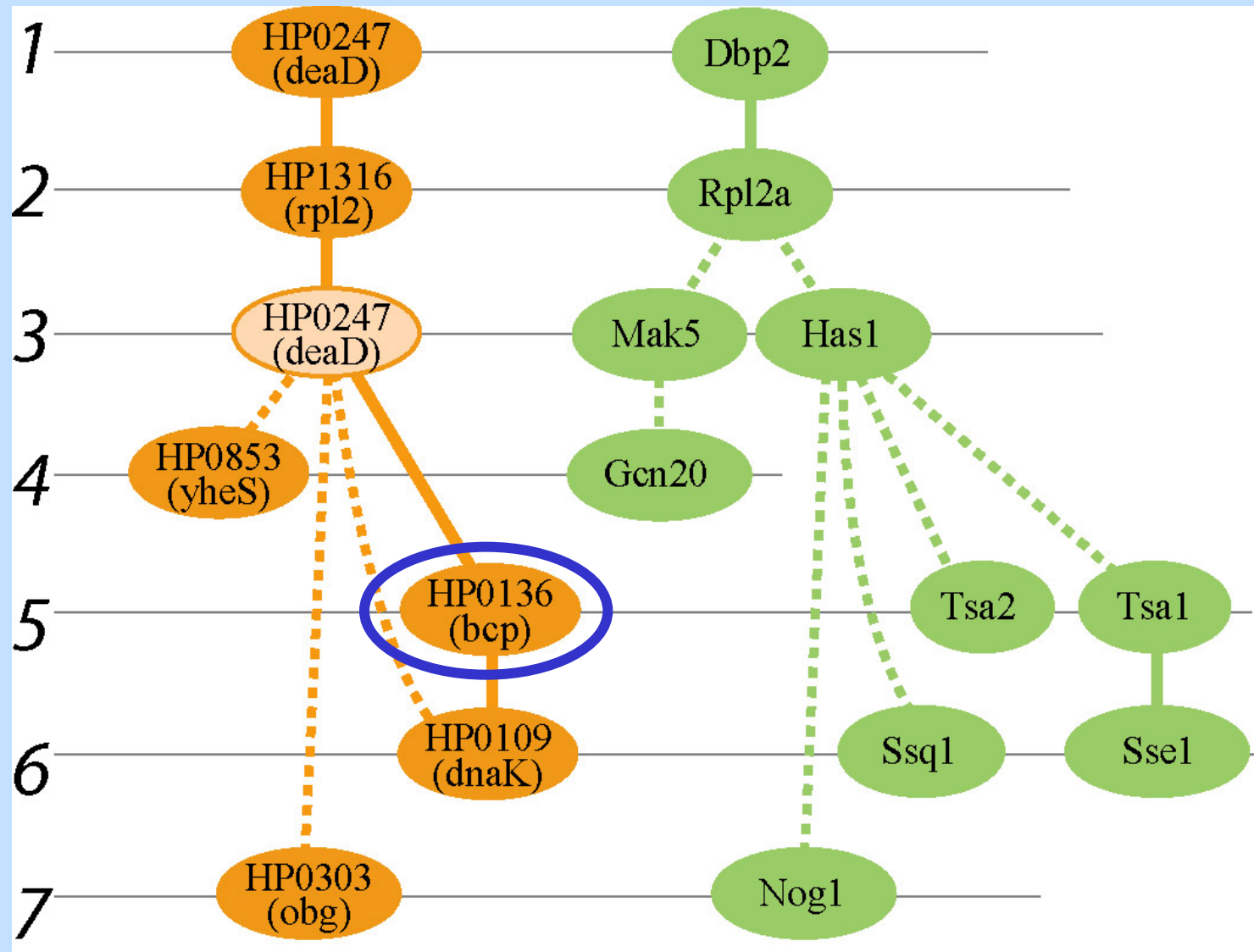


Yeast
~15000 PPIs

	Vertices (homologs)	Edges			
		Total	Direct	Gap	Mismatch
Yeast vs. <i>H. pylori</i> ($E_{\text{cutoff}} = 10^{-2}$)	829	2,036	7	260	1,769
Random: mean \pm SD		509.0 \pm 128.0	2.5 \pm 1.9	68.8 \pm 23.8	437.7 \pm 110.3

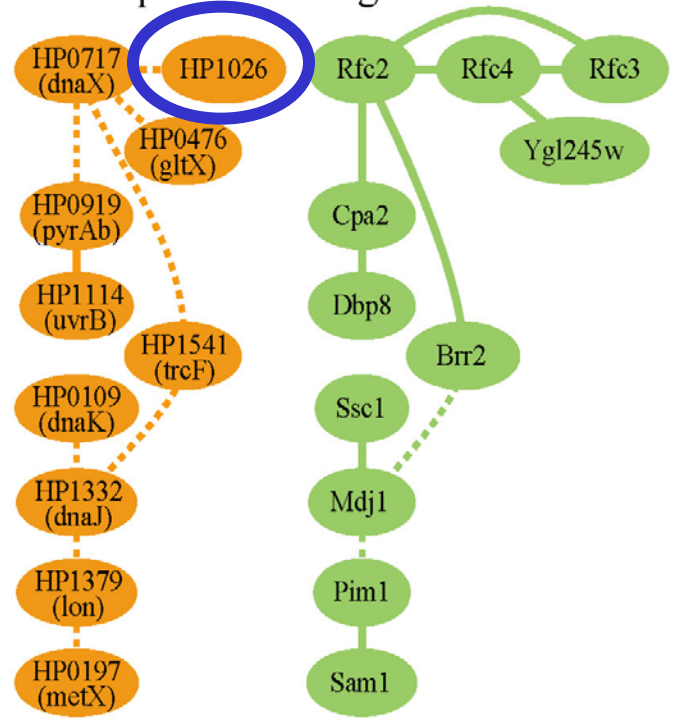
← Bacteria → ← Yeast →

↑ Protein sequence similarity ↓

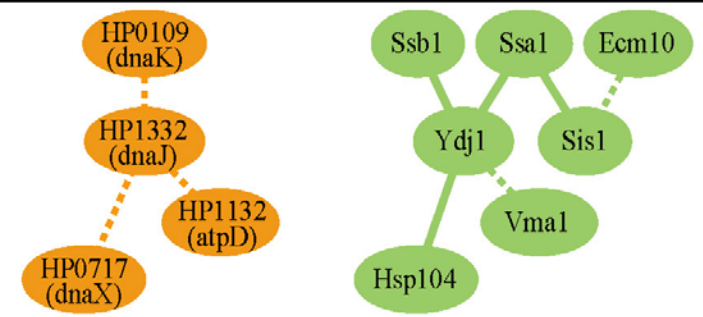


Best match of bcp is Dot5, which does not interact with the pathway's proteins.

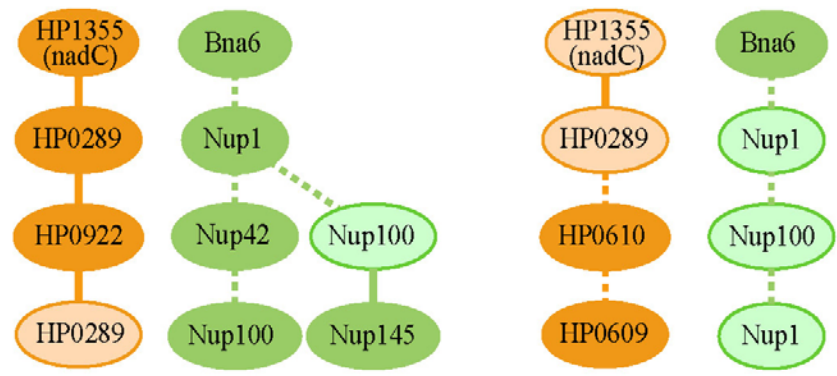
DNA replication /protein folding



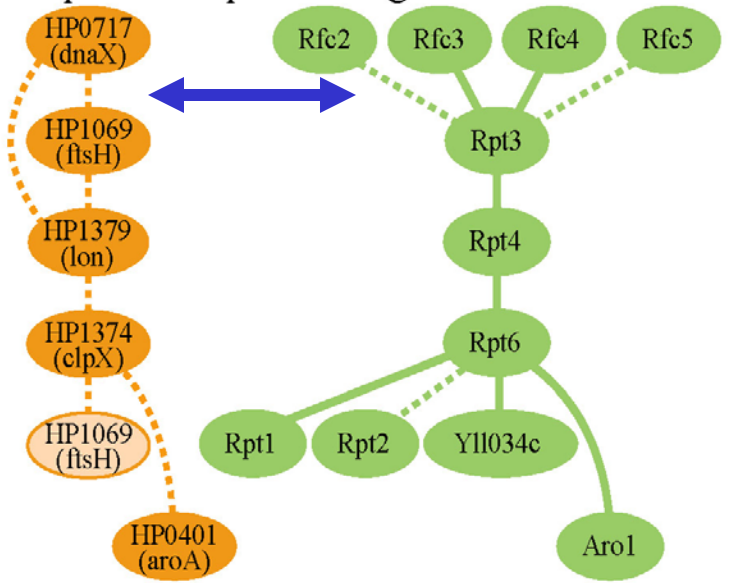
Heat shock and chaperone proteins



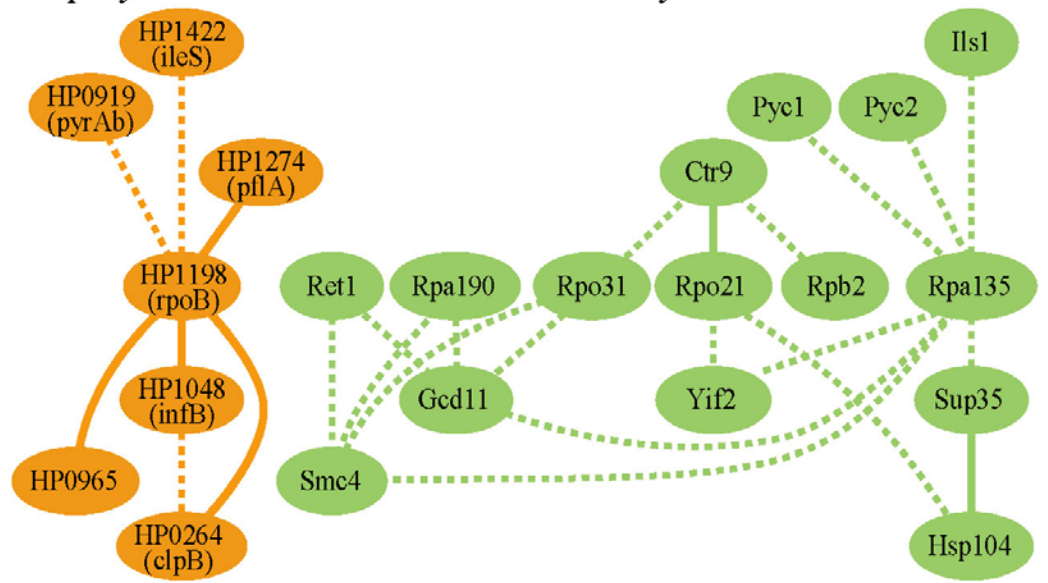
Cytoplasmic and nuclear membrane transport



DNA replication /protein degradation



RNA polymerase and associated machinery



Identifying conserved Complexes

- Generalize single-species scoring
- Given two protein subsets, one in each species, with a many-to-many correspondence between them, wish:
 1. Each subset induces a dense subgraph.
 2. Matched protein pairs are sequence-similar.

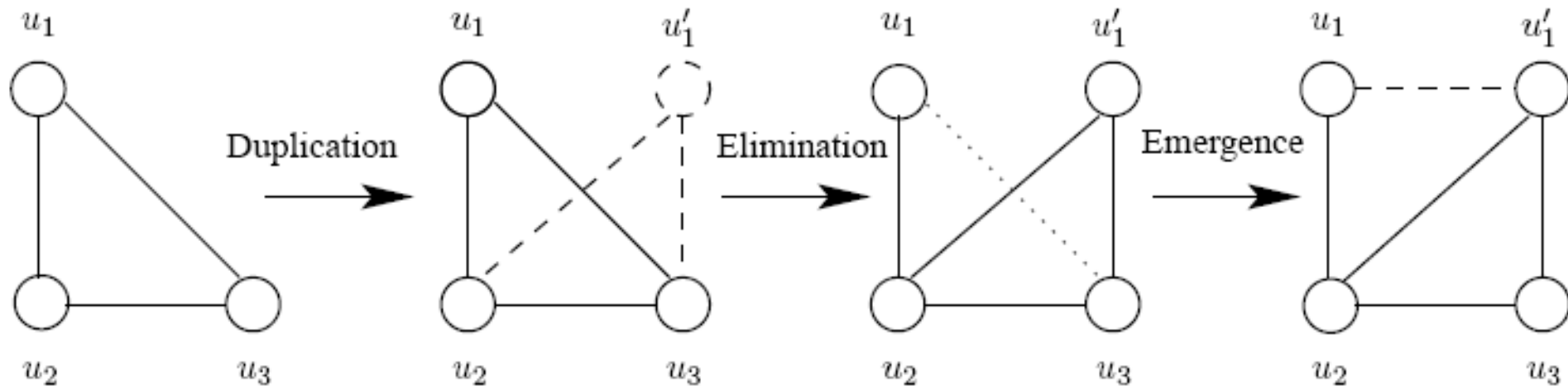
$$L(C, C') = L(C) \cdot L(C') \cdot \prod_{u,v \text{ matched}} \frac{\Pr(S_{u,v} \mid \text{homologs})}{\Pr(S_{u,v} \mid \text{random})}$$

$$\text{Recall: } L(C) = \prod_{(u,v) \in E'} \frac{p}{p(u,v)} \prod_{(u,v) \notin E'} \frac{1-p}{1-p(u,v)}$$

Evolutionary-based Scoring

A Word on PPI Evolution

- PPI networks are shaped by duplication and indel events.
- Indel events arise due to mutations that change protein surface and are much more frequent.



Scoring (MaWish)

- The score of two aligned protein subsets is based on the **match**, **mismatch** and **duplication** events they induce.
- Each event is associated with a parameter (heuristically set) which determines its relative weight.

$$\sigma(\mathcal{A}) = \sum_{M \in \mathcal{M}} \mu(M) + \sum_{N \in \mathcal{N}} \nu(N) + \sum_{D \in \mathcal{D}} \delta(D).$$

Match reward
 $mS(u,v)S(u',v')$

Indel penalty
 $-nS(u,v)S(u',v')$

**Duplication
reward/penalty**
 $d(S(u,u')-s)$

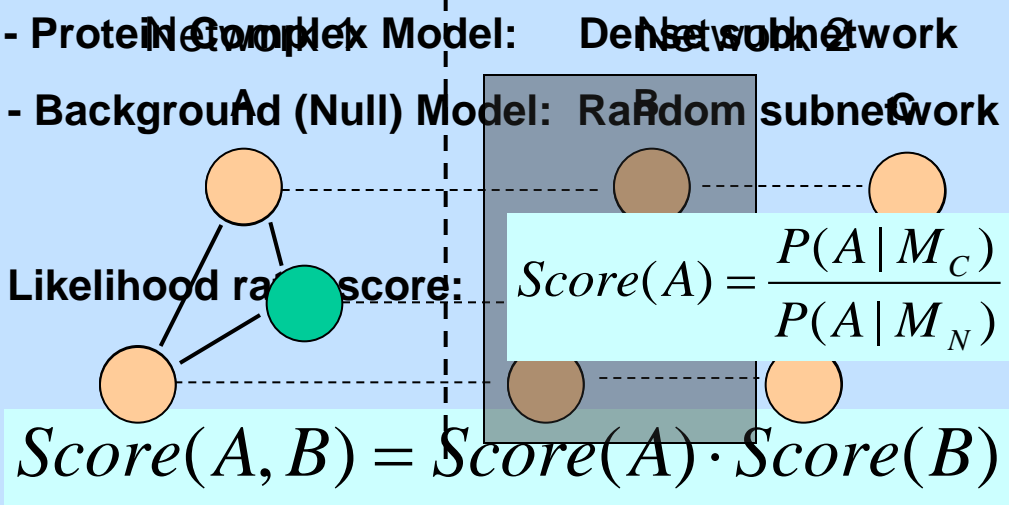
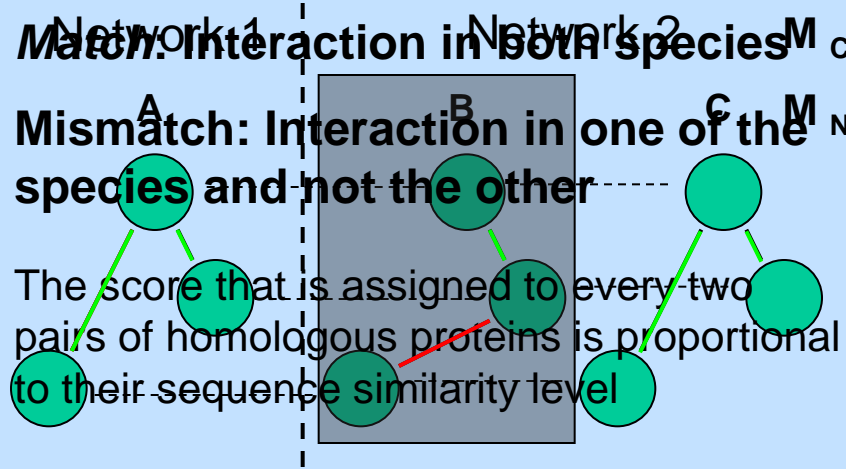
Score improvement

NetworkRI (Pr... '05)

- Does not use evolutionary properties of the PPI network

MaWish (Pr... '05)

No underlying probabilistic model



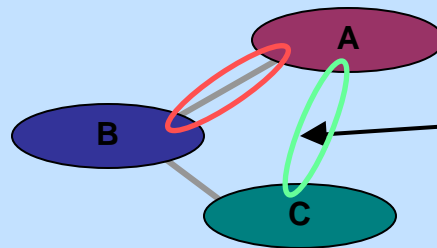
$Score(A \cup B) = Score(A \cup C)$

$Score(A \cup B) = 1 = Score(A \cup C) = 1$

Score improvement (cont.)

$$P(A, B | M_C)$$

Closest Common Ancestor
Hypothetical PPI Network

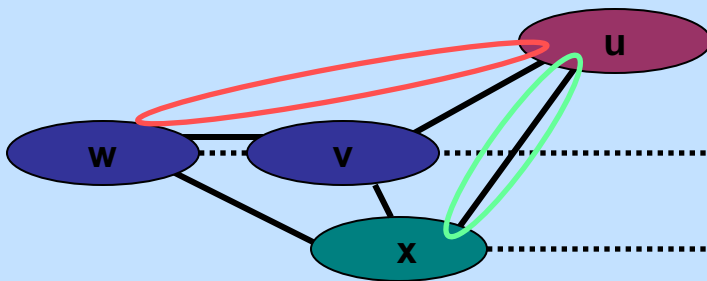


Link Dynamics

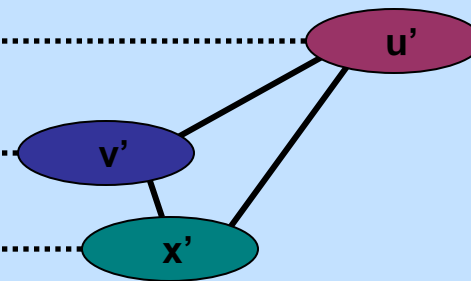
$$P(T_{uv} | M_C) = \beta$$
$$\beta(1 - P_{loss}) + (1 - \beta)P_{gain}$$



Network 1



Network 2



Local multiple alignment

3-way comparison?



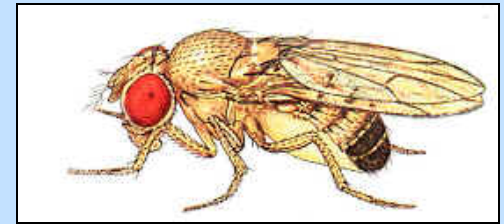
S. cerevisiae

- 4389 proteins
- 14319 interactions



C. elegans

- 2718 proteins
- 3926 interactions

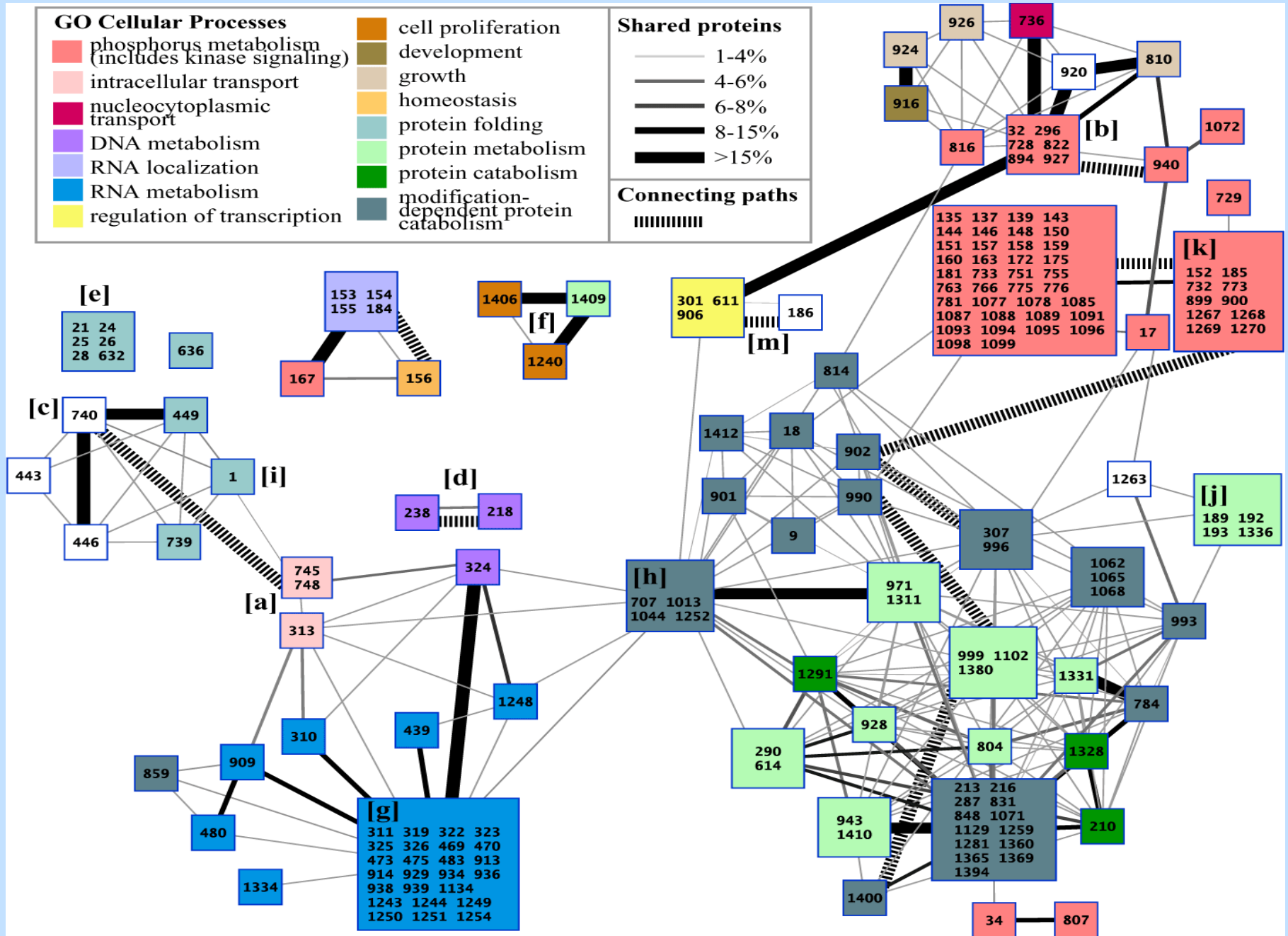


D. melanogaster

- 7038 proteins
- 20720 interactions

Generalizing Network Alignment

- Alignment graph is extensible to multiple species.
- Likelihood scoring is easily extensible, up to sequence similarity terms: require scoring a multiple sequence alignment.
- Ignored till now: need to balance edge and vertex terms.
- Practical solution:
 - Sensible threshold for sequence similarity.
 - Nodes in alignment graph are filtered accordingly.
 - Node terms are removed from score.



71 conserved regions: 183 significant clusters and 240 significant paths.

Interaction Prediction

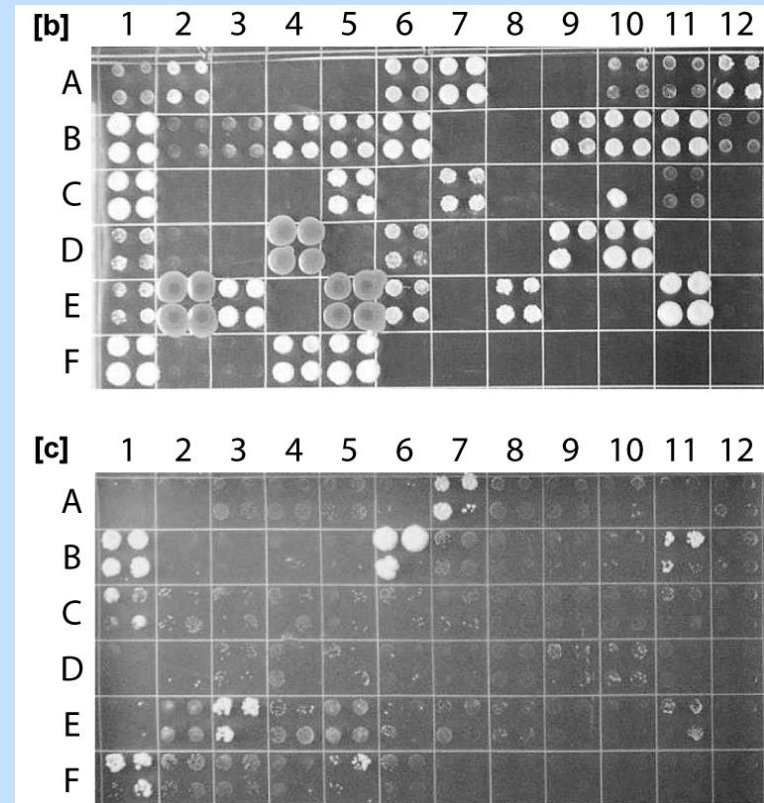
A pair of proteins is predicted to interact if:

1. Sequence-similar proteins interact in the other two species.
2. The proteins co-occur in the same conserved complex.

Species	Sensitivity (%)	Specificity (%)	<i>P-value</i>	Strategy
Yeast	50	77	1-25	[1]
Worm	43	82	1e-13	[1]
Fly	23	84	5e-5	[1]
Yeast	9	99	1e-6	[2]+[1]
Worm	10	100	6e-4	[2]+[1]
Fly	0.4	100	0.5	[2]+[1]

Experimental Validation

- 65 predictions for yeast using strategies [1]+[2] were tested in lab.
- **Success rate: 40-52%.**
- Outperforms the interolog approach (Matthews et al.'01, Yu et al.'04) at 16-31%.



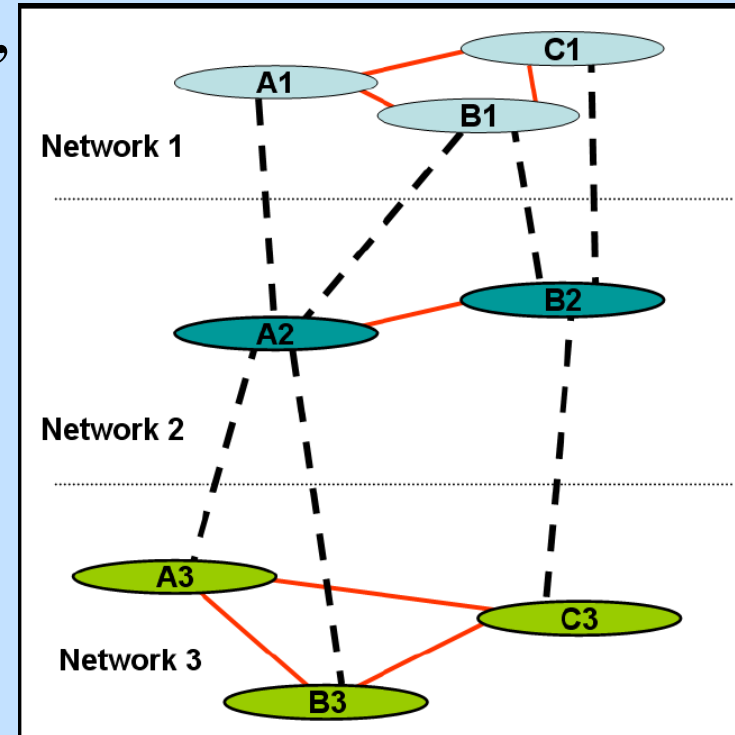
The Scalability Problem

- Network alignment scales as n^k (in time and space) for n proteins and k species, hence practical only for $k=2,3$ (takes several hours).
- Progressive alignment is fast (Graemlin by Flanick et al., GR 2006) but does not perform as well.

Main idea: imitate the greedy search w/o explicitly constructing the alignment graph.

Scaling Up Network Alignment

- Maintain linear representation.
- Observe: “network alignment node” is a vertical “path”
- Given a current seed, use dynamic programming to identify the vertical “path” which contributes most to the score.
- Complexity reduces to $O(m2^k)$!



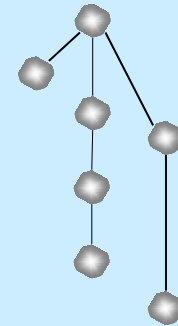
#Species	#Nodes	#PPI edges	#Sequence similarity edges	Restricted order run time (sec)
3	8132	102288	26834	40
5	11945	193843	57142	72
7	17236	301365	103887	83
10	31458	877032	327219	140

Network querying

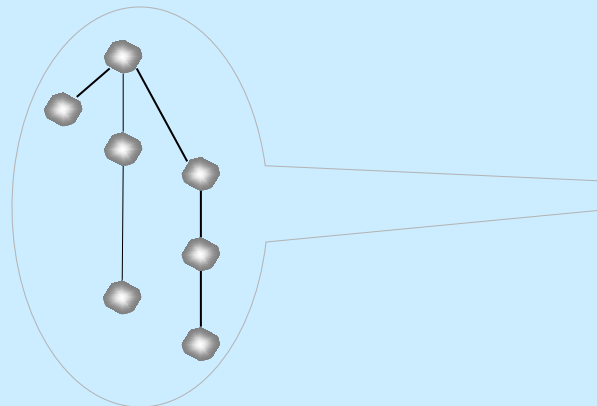
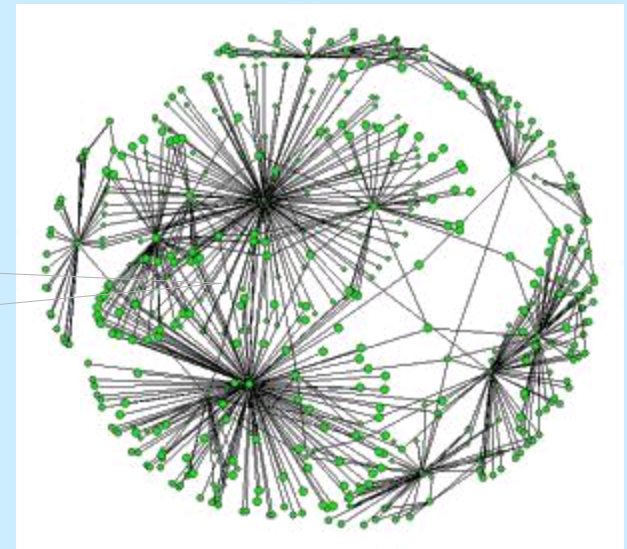
Problem definition

- Given a query graph Q and a network G , find the subnetwork of G that is:
 - **Aligned** with Q
 - The alignment has **maximal score**

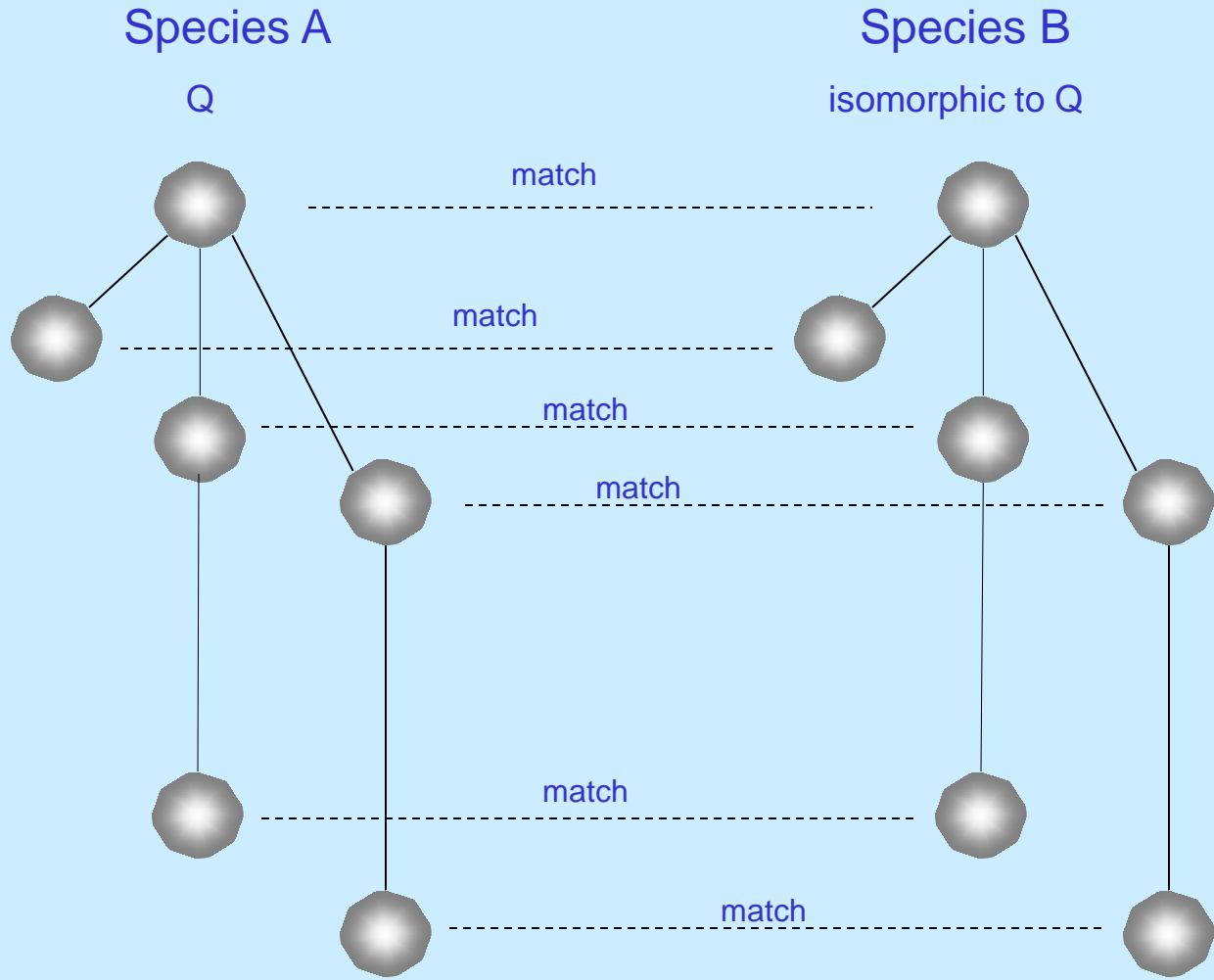
Query Q



Network G

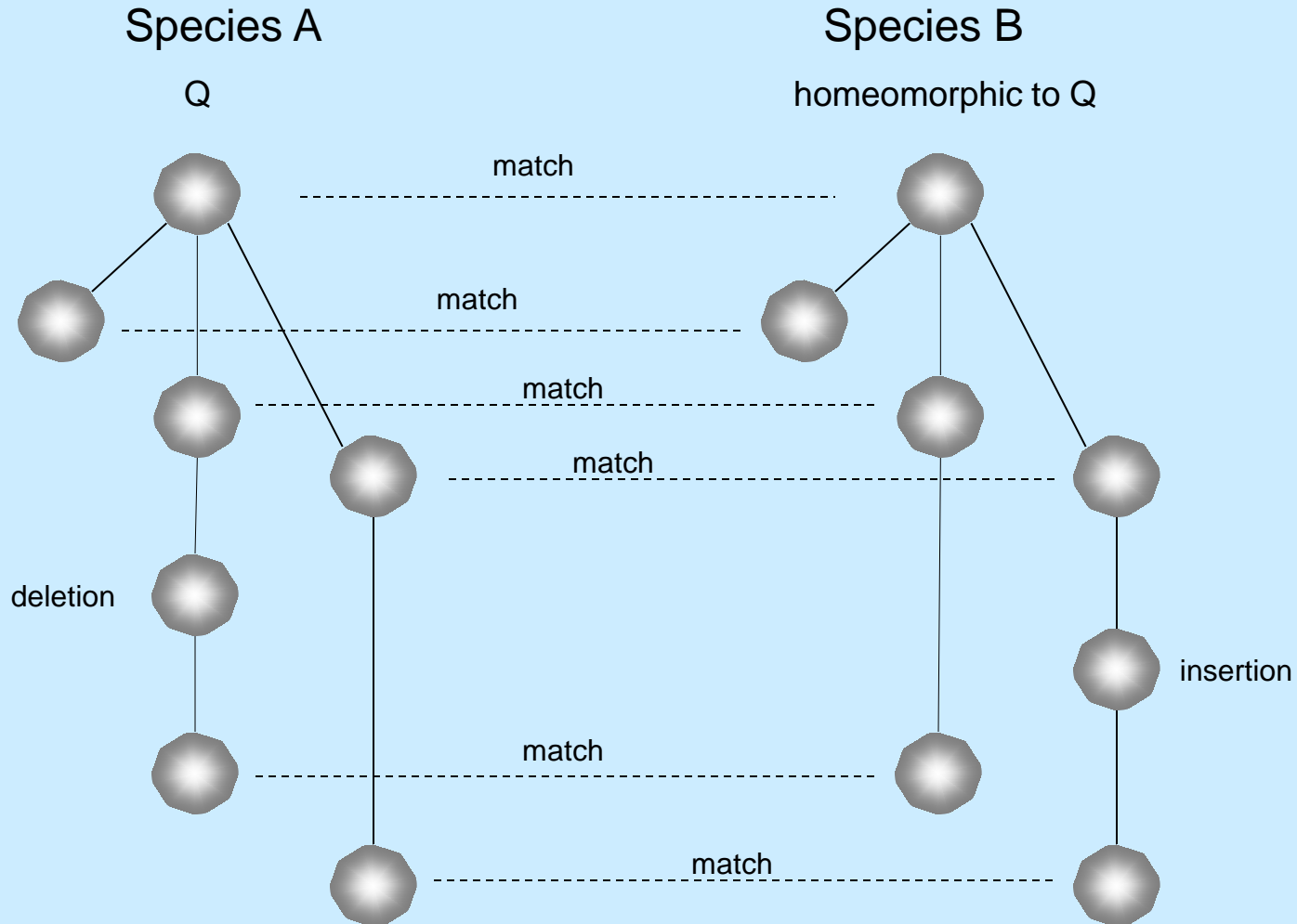


Isomorphic Alignment



Match of sequence-similar proteins

Homeomorphic Alignment

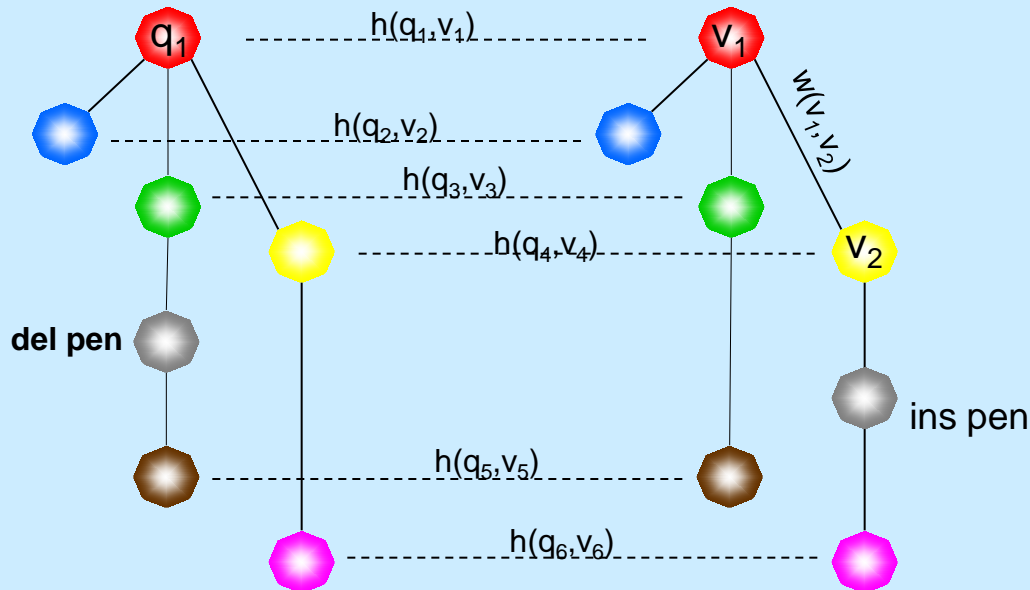


Match of sequence-similar proteins and deletion/insertion of degree-2 nodes

Score of Alignment

$$\text{Score} = \sum_{\text{matches}} h(q_i, v_j) + \delta_d (\# \text{Del}) + \delta_i (\# \text{Ins}) + \sum w(v_i, v_j)$$

Sequence similarity score for matches + Penalty for deletions & insertions + Interaction reliability scores



Complexity

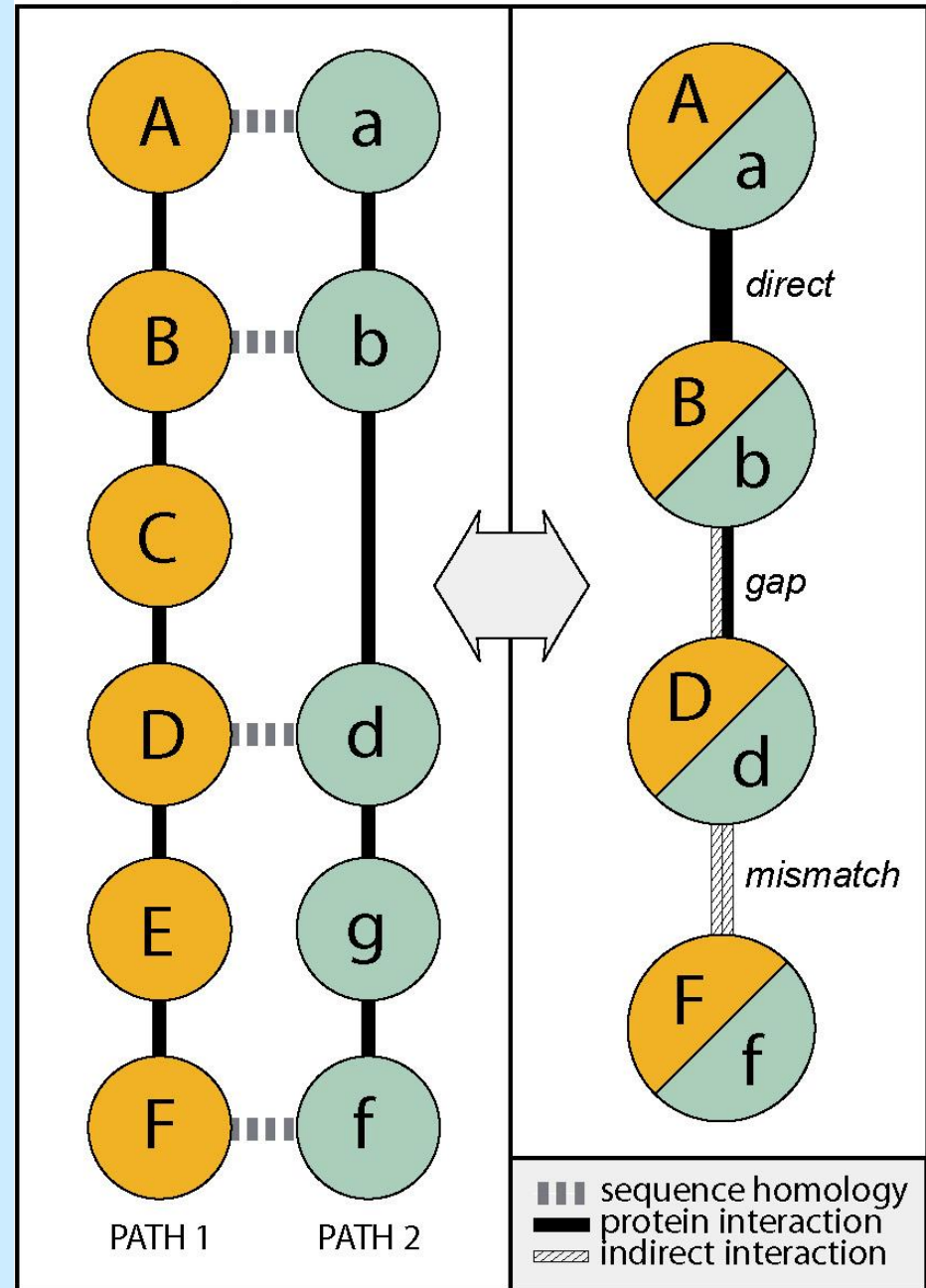
- Network querying problem is NPC by reduction from subgraph isomorphism (in contrast to sequence querying!!!)
- Naïve algorithm has $O(n^k)$ complexity
 - n = size of the PPI network, k = size of the query
 - Intractable for realistic values of n and k
 - $n \sim 5000$, $k \sim 10$
- Reduction in complexity can be achieved by:
 - Constraining the network [Pinter et al., Bioinformatics'05]
 - Allowing vertex repetitions
 - Constraining the query (fixed parameter algs.)

PathBLAST

Reduction to finding paths in an “alignment” graph.

- Repetitions are possible.
- No general handling of insertions/deletions

[a] Pathway alignment [b] Alignment graph

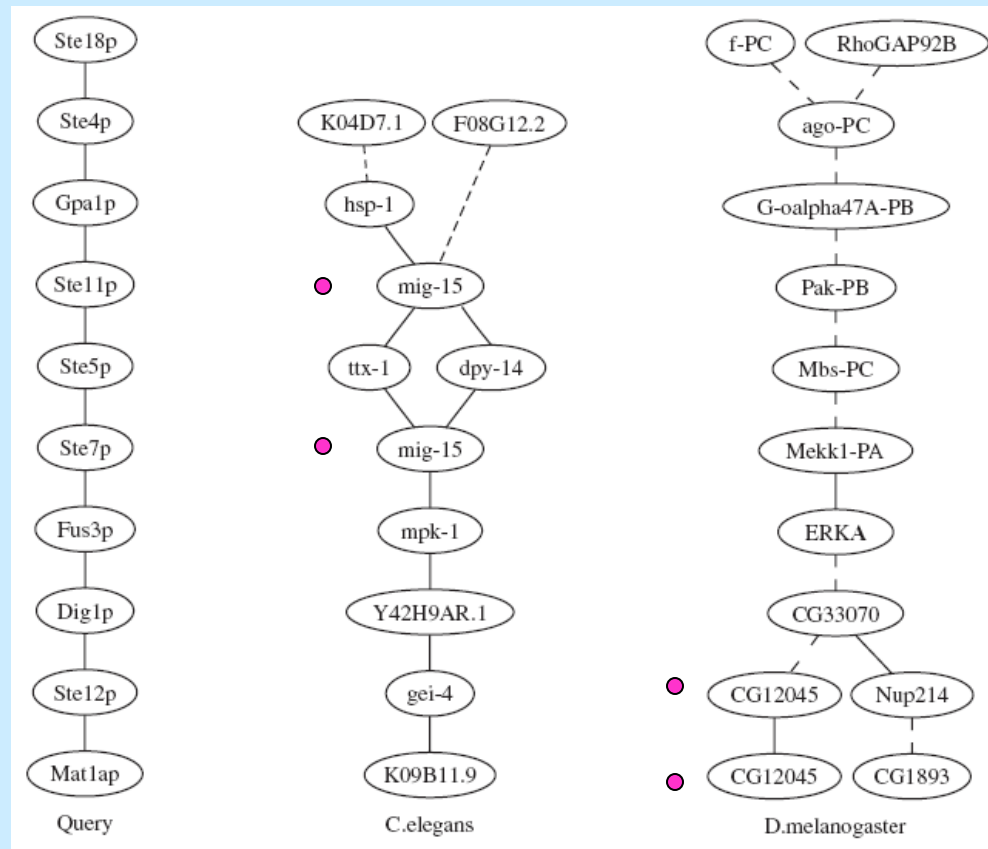


DP-Based Approach

- Use dynamic programming (a la sequence alignment):
 $W(i, j)$ is the maximal score of a partial alignment of query nodes $\{1 \dots i\}$ that ends at vertex j of the network.

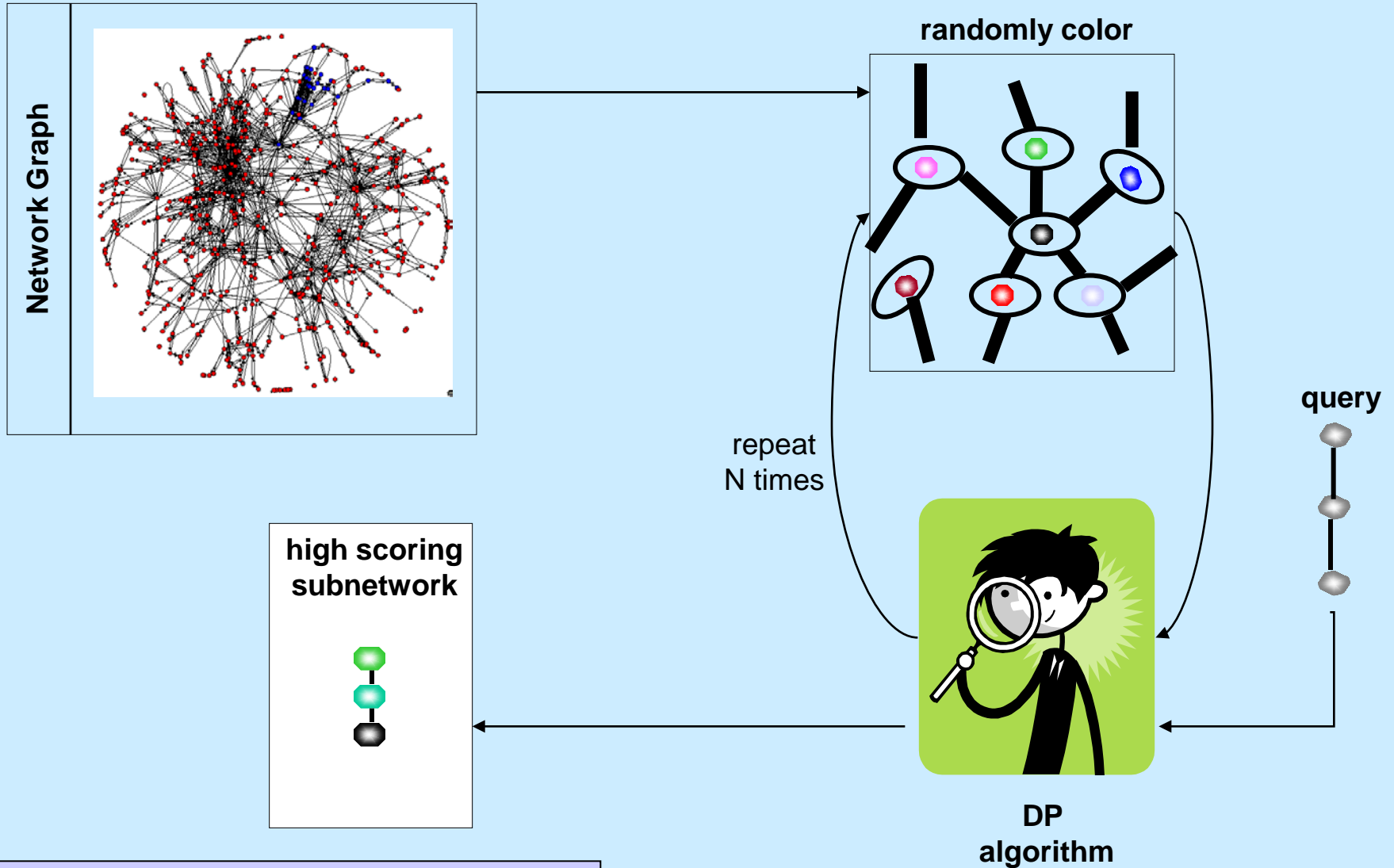
$$W(i, j) = \max \begin{cases} W(i-1, m) + h(i, j) + w(m, j), (m, j) \in E & \text{match} \\ W(i, m) + w(m, j) + \delta_i, (m, j) \in E & \text{insertion} \\ W(i-1, j) + \delta_d & \text{deletion} \end{cases}$$

Cross-Species Comparison of Signaling Pathways



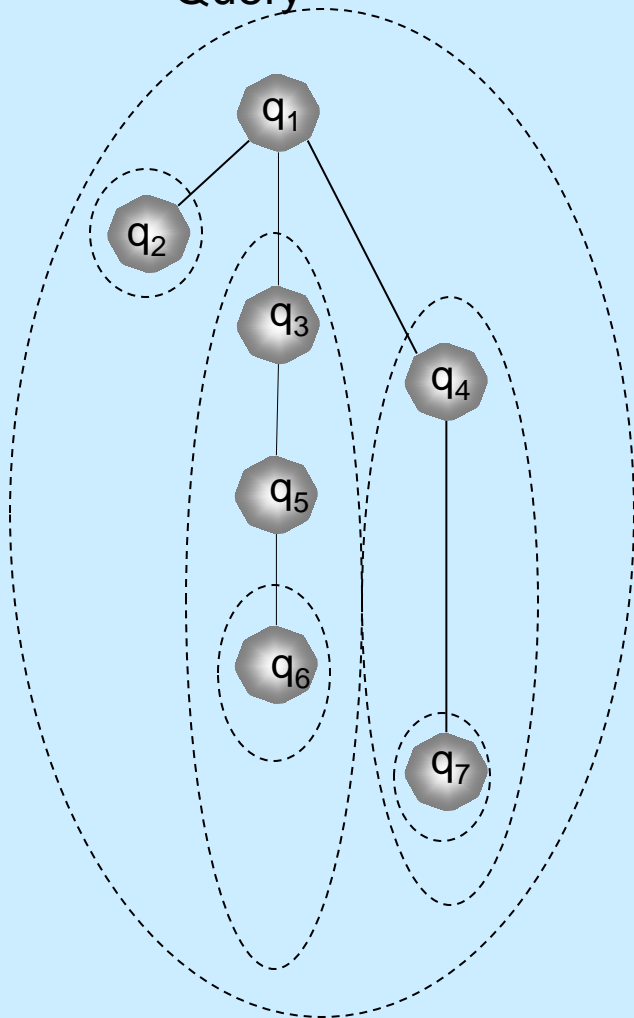
- But DP may introduce protein repetitions along the path.

QPath

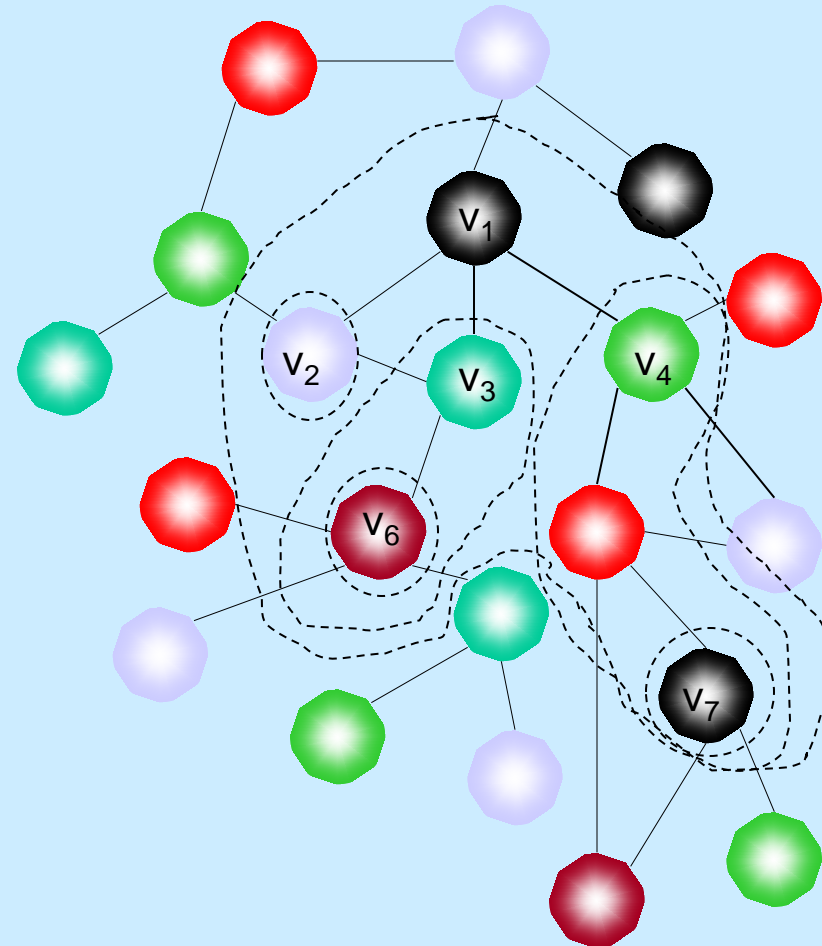


Ideas can be generalized to tree queries and beyond (QNet)

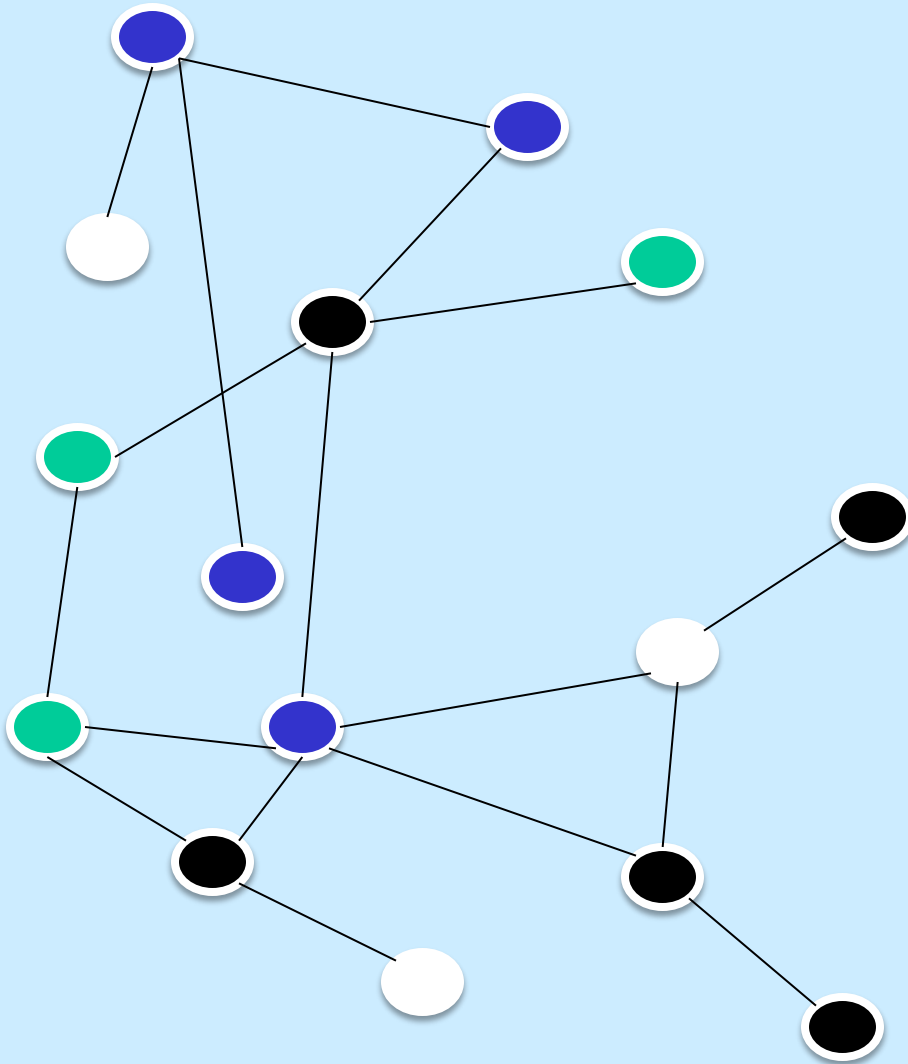
Query



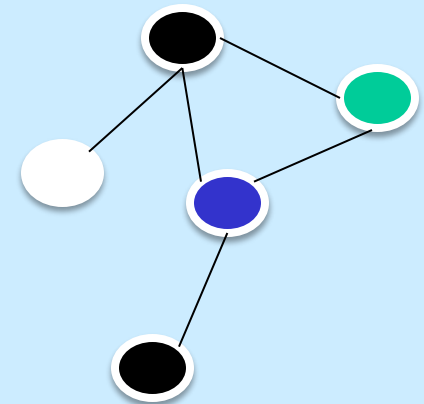
Network



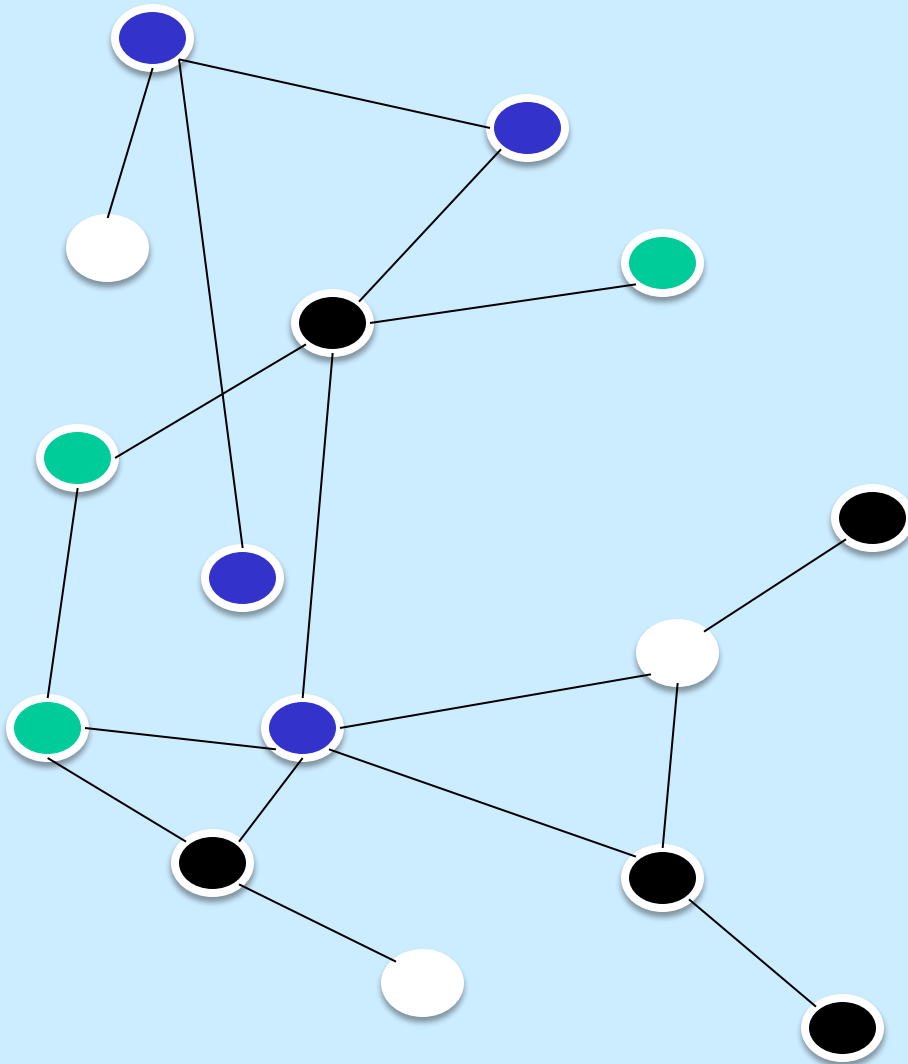
Is topology needed?



?



TORQUE: Topology-free querying



Input:

- ❖ Graph $G=(V,E)$
- ❖ Color set $\{1,2,\dots,k\}$
- ❖ A coloring of network vertices


Output: a connected subgraph that is colorful.

Algorithmic idea

Every connected subgraph has a spanning tree



Every colorful connected subgraph will have a colorful spanning tree

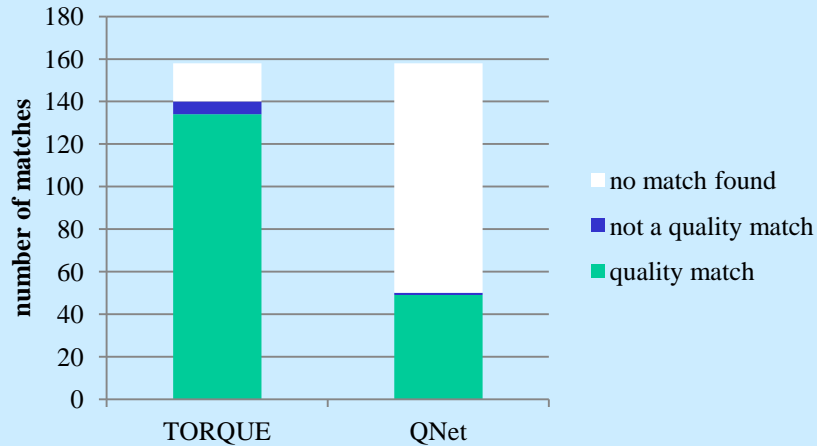


Instead of looking for a colorful subgraph, look for a colorful tree

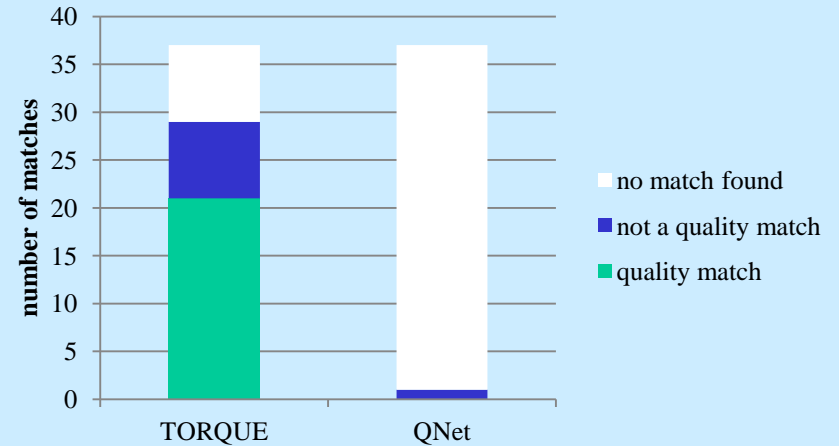
- Two implemented approaches:
 - Dynamic programming (color coding)
 - ILP

Comparison with QNet

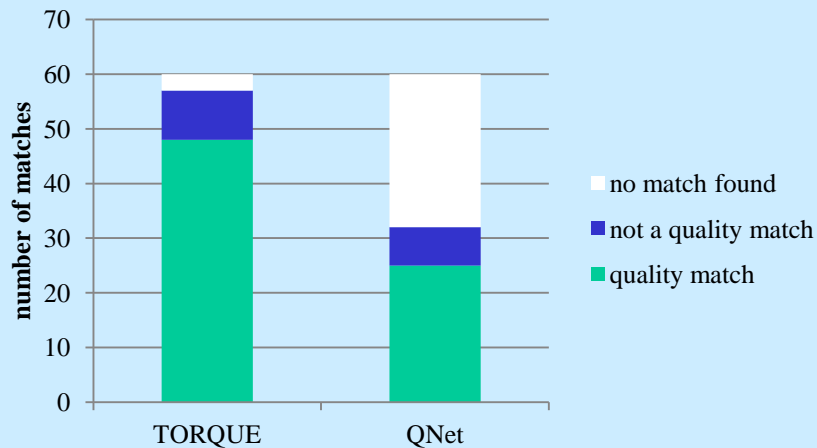
Human complexes in Yeast



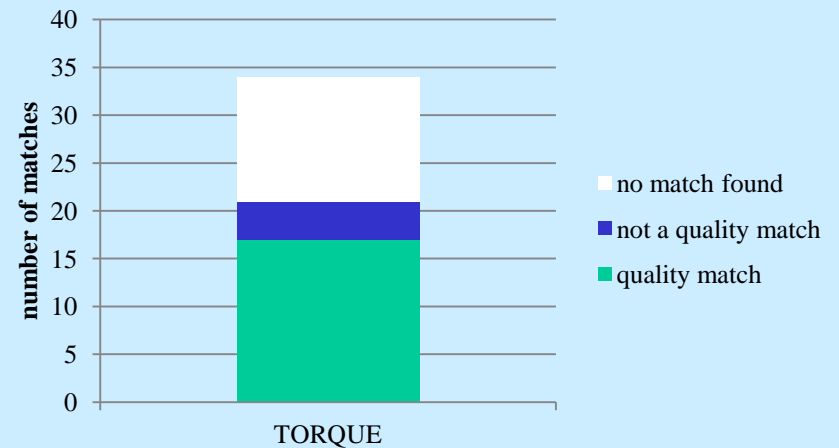
Fly complexes in Human



Yeast complexes in Human



Rat complexes in Fly



Summary & the road ahead...

