# Topology-Free Querying of Protein Interaction Networks

Sharon Bruckner Falk Hüffner Richard M. Karp Ron Shamir Roded Sharan

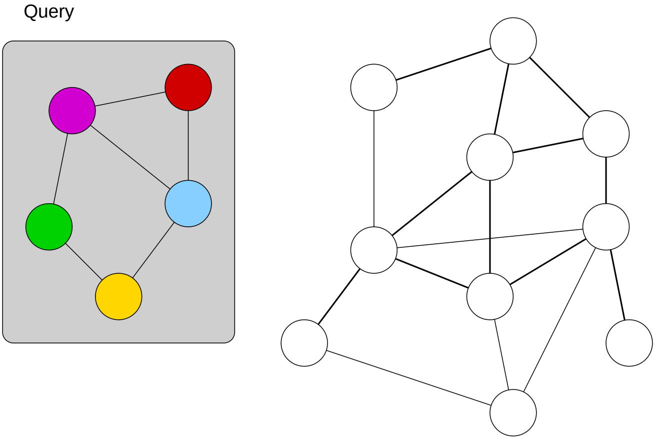
Tel Aviv University

4 May 2009

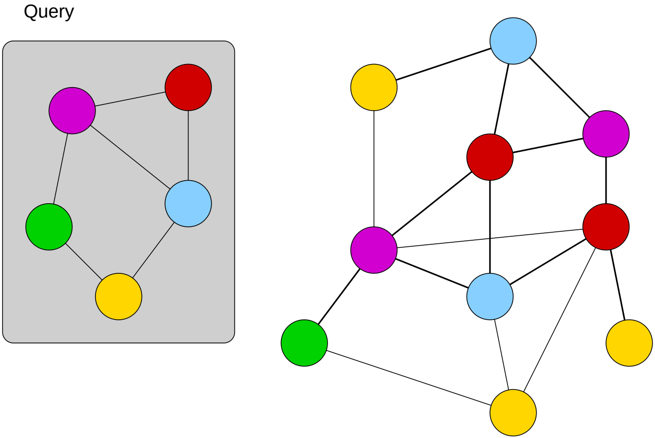
# Protein complexes

- A protein complex is a group of proteins which interact with each other to perform some task.
- Many protein complexes are known, in particular for model organisms like yeast.
- Problem: does a known protein complex also exists in the protein interaction network of another species?

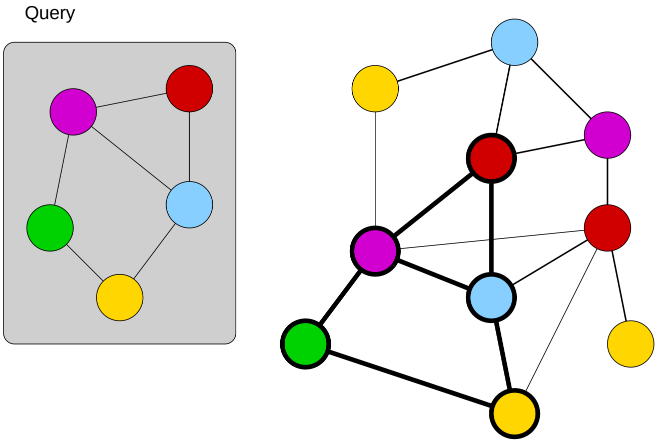
# Complex query as Constrained Subgraph Isomorphism



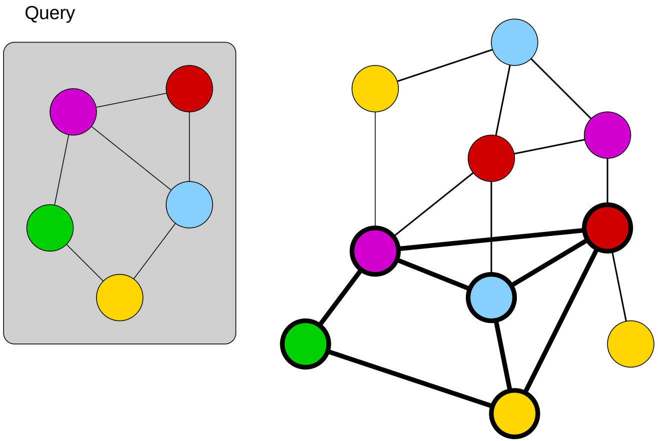
# Complex query as Constrained Subgraph Isomorphism



# Complex query as Constrained Subgraph Isomorphism



# Complex query as CONSTRAINED SUBGRAPH ISOMORPHISM



# Problems with Constrained Subgraph Isomorphism

- Not error tolerant

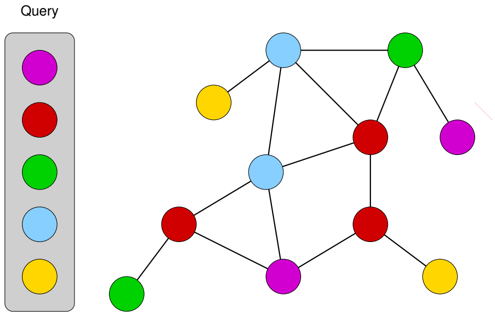
# Problems with Constrained Subgraph Isomorphism

- Not error tolerant
- Interactions between query proteins (*topology*) might not be available

# Problems with Constrained Subgraph Isomorphism

- Not error tolerant
- Interactions between query proteins (*topology*) might not be available
- Computationally very hard

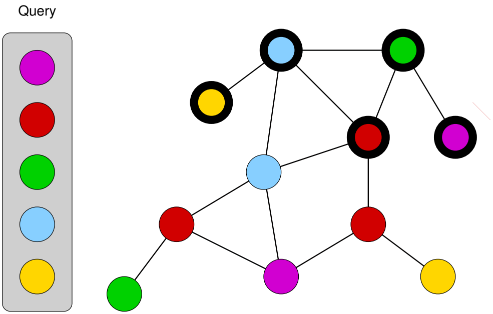
# Complex Query as Colorful Connected Subgraph



### Colorful Connected Subgraph

**Input:** An undirected, vertex colored graph $G$.
**Output:** Find a connected subgraph of $G$ whose vertices use each color exactly once (*colorful subgraph*).

# Complex Query as Colorful Connected Subgraph



## Colorful Connected Subgraph

**Input:** An undirected, vertex colored graph $G$.

**Output:** Find a connected subgraph of $G$ whose vertices use each color exactly once (*colorful subgraph*).

# Dynamic Programming

### Idea

Instead of looking at all $O(n^k)$ possible subgraphs, look only at $O(2^k)$ color sets.

# Dynamic Programming

**Idea**

Instead of looking at all $O(n^k)$ possible subgraphs, look only at $O(2^k)$ color sets.

$T[v, S]$ for $v \in V$ and $S$ a set of colors: true if there is a connected subgraph of $|S|$ vertices containing $v$ with exactly the colors in $S$

$$T[v, S] = \bigvee_{\substack{u \in N(v) \\ S_1 \uplus S_2 = S}} T[v, S_1] + T[u, S_2] + w(u, v)$$

# Dynamic Programming

## Idea

Instead of looking at all $O(n^k)$ possible subgraphs, look only at $O(2^k)$ color sets.

$T[v, S]$ for $v \in V$ and $S$ a set of colors: true if there is a connected subgraph of $|S|$ vertices containing $v$ with exactly the colors in $S$

$$T[v,S] = \bigvee_{\substack{u \in N(v) \\ S_1 \uplus S_2 = S}} T[v, S_1] + T[u, S_2] + w(u,v)$$

## Theorem

*Colorful Connected Subgraph with $k$ colors can be solved in $O(3^k|E|)$ time.*

# Fixed-parameter tractability

### Theorem

Colorful Connected Subgraph *with $k$ colors can be solved in $O(3^k|E|)$ time.*

### Corollary

Colorful Connected Subgraph *is fixed-parameter tractable with respect to $k$.*

# Integer Linear Programming

An Integer Linear Program (ILP) can maximize a linear function under linear constraints and integrality constraints.

# Integer Linear Programming

An Integer Linear Program (ILP) can maximize a linear function under linear constraints and integrality constraints.

## Binary variables

$c_v, v \in V$: $v = 1 \iff v$ is part of the complex

## Constraints

For each color $\gamma$: $\sum_{v \in V:\text{color}(v)=\gamma} c_v = 1$

# Integer Linear Programming

An Integer Linear Program (ILP) can maximize a linear function under linear constraints and integrality constraints.

## Binary variables

$c_v, v \in V$: $v = 1 \iff v$ is part of the complex

## Constraints

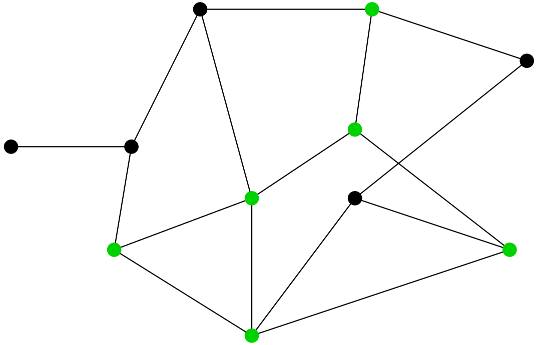
For each color $\gamma$: $\sum_{v \in V:\text{color}(v)=\gamma} c_v = 1$

## Central problem

Given a graph $G = (V, E)$ and binary variables $c_v, v \in V$, find linear constraints such that $G[\{v \mid c_v = 1\}]$ is connected.

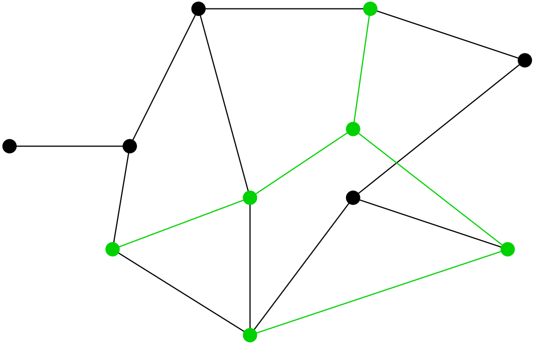
# Connected Subgraph als ILP

Sufficient: find a directed tree



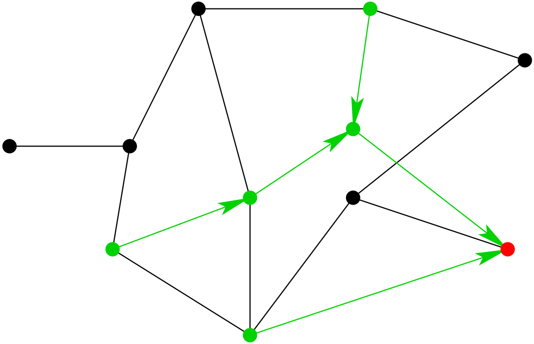
# Connected Subgraph als ILP

Sufficient: find a directed tree



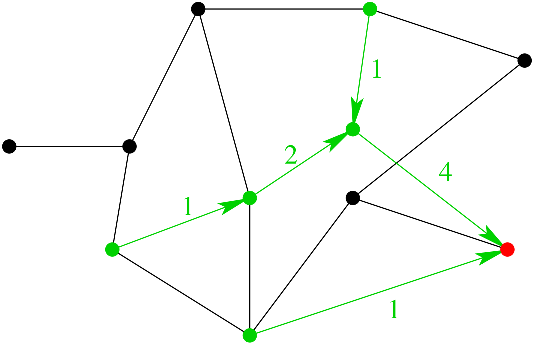
# Connected Subgraph als ILP

Sufficient: find a directed tree



# Connected Subgraph als ILP

Sufficient: find a directed tree



- An (arbitrary) selected vertex servers as sink
- Each other selected vertex is source of a flow of 1
- Only selected vertices take part in flow

# Model extensions

- More than one color per vertex

# Model extensions

- More than one color per vertex
- Insertions/Deletions

# Model extensions

- More than one color per vertex
- Insertions/Deletions
- Maximize edge weight of complex

# Complete ILP

$$\text{maximize} \sum_{(v,w)\in E} \omega_{vw} e_{vw} \tag{1}$$

subject to

$$\sum_{v\in V} c_v = t \tag{2}$$

$$\sum_{v\in V} r_v = 1 \tag{3}$$

$$e_{vw} \leq c_v \wedge e_{vw} \leq c_w \qquad \forall (v,w) \in E \tag{4}$$

$$e_{vw} \geq 1/2 c_v + 1/2 c_w - 1/2 \qquad \forall (v,w) \in E \tag{5}$$

$$f_{vw} = -f_{wv} \qquad \forall (v,w) \in E \tag{6}$$

$$\sum_{w\in N(v)} f_{vw} = c_v - t r_v \qquad \forall v \in V \tag{7}$$

$$f_{vw}, f_{wv} \leq (t-1) e_{vw} \qquad \forall (v,w) \in E \tag{8}$$

$$\sum_{\gamma\in\Gamma(v)} g_{v\gamma} \leq 1 \qquad \forall v \in V \tag{9}$$

$$\sum_{v\in V} g_{v\gamma} \leq 1 \qquad \forall \gamma \in C \tag{10}$$

$$\sum_{v\in V} \sum_{\gamma\in\Gamma(v)} g_{v\gamma} = t - N_{\text{ins}} \tag{11}$$

$$g_{v\gamma} \leq c_v \qquad \forall v \in V, \gamma \in \Gamma(v) \tag{12}$$

# Implementation

- First do data reduction

# Implementation

- First do data reduction
  - only 5 % of the vertices are associated with one or more colors

# Implementation

- First do data reduction
  - only 5 % of the vertices are associated with one or more colors
  - many non-colored vertices are too far from any colored vertex to be useful

# Implementation

- First do data reduction
  - only 5 % of the vertices are associated with one or more colors
  - many non-colored vertices are too far from any colored vertex to be useful
- For each remaining connected component:

# Implementation

- First do data reduction
  - only 5 % of the vertices are associated with one or more colors
  - many non-colored vertices are too far from any colored vertex to be useful
- For each remaining connected component:
  - Try a heuristic that does not allow indels

# Implementation

- First do data reduction
  - only 5 % of the vertices are associated with one or more colors
  - many non-colored vertices are too far from any colored vertex to be useful
- For each remaining connected component:
  - Try a heuristic that does not allow indels
  - If this fails:

# Implementation

- First do data reduction
  - only 5 % of the vertices are associated with one or more colors
  - many non-colored vertices are too far from any colored vertex to be useful
- For each remaining connected component:
  - Try a heuristic that does not allow indels
  - If this fails:
    - If few colors, but large instance, use dynamic programming

# Implementation

- First do data reduction
  - only 5 % of the vertices are associated with one or more colors
  - many non-colored vertices are too far from any colored vertex to be useful
- For each remaining connected component:
  - Try a heuristic that does not allow indels
  - If this fails:
    - If few colors, but large instance, use dynamic programming
    - Otherwise, use ILP

# Data

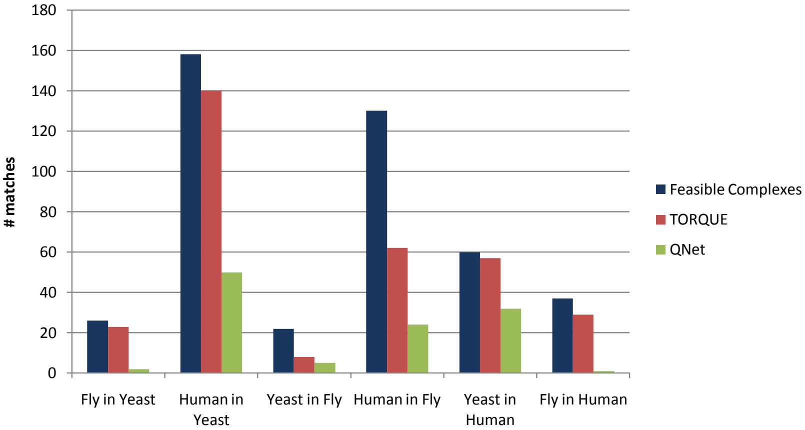
Protein–protein interaction networks:

- yeast (5 430 proteins, 39 936 interactions)
- fly (6 650 proteins, 21 275 interactions)
- human (7 915 proteins, 28 972 interactions)

Query several hundred complexes of size 4–25 from:

- yeast, fly, human (interaction information available)
- bovine, mouse, and rat (not enough interaction information available)

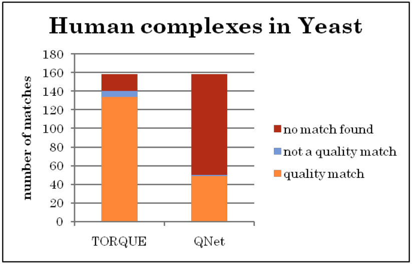
# Number of complexes found

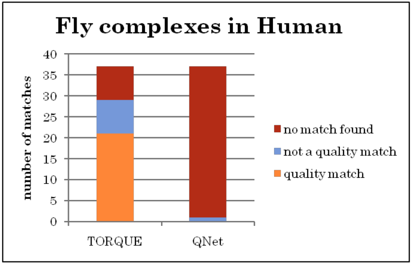


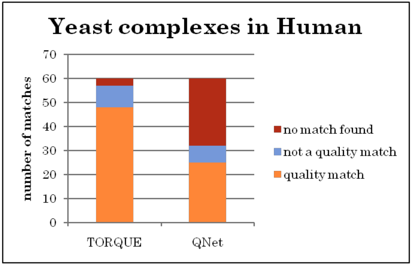
# Evaluation of results

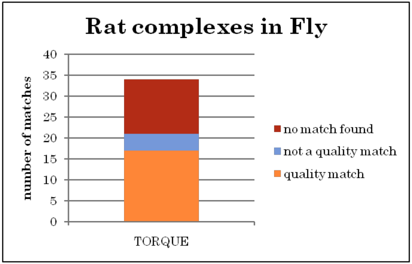
- Functional coherence: Percentage of proposed complexes that are significantly enriched with “GO-Terms”
- Specificity: Percentage of proposed complexes that overlap significantly with known complexes

# Quality of matches

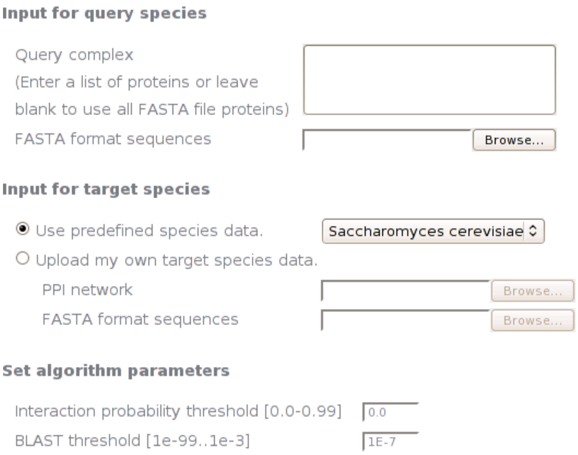








# Torque Web-Server



**Input for query species**

Query complex
(Enter a list of proteins or leave
blank to use all FASTA file proteins)

FASTA format sequences

**Input for target species**

● Use predefined species data. Saccharomyces cerevisiae

○ Upload my own target species data.

PPI network

FASTA format sequences

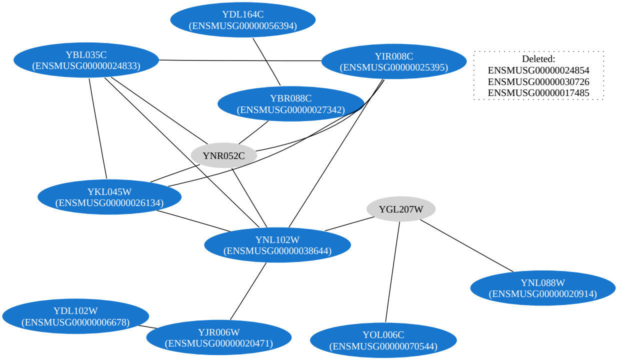
**Set algorithm parameters**

Interaction probability threshold [0.0-0.99] 0.0

BLAST threshold [1e-99..1e-3] 1E-7

http://www.cs.tau.ac.il/˜bnet/torque.html

# TORQUE Web-Server



Blue: matched nodes in the target species. Within each node, top: target protein, bottom: the matching query protein.
Grey: insertions of target proteins. The box lists the deleted query proteins, if any.

Best match for the DNA synthesome complex of the mouse in the network of yeast

# Summary

- A topology-free querying model yields significant complex query results.
- With a combination of dynamic programming and ILP, even difficult instances can be solved optimally.