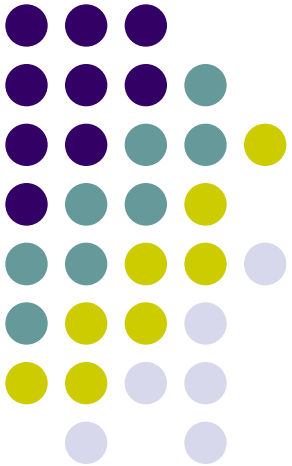
QNET: A tool for querying protein interaction networks

Banu Dost[+], Tomer Shlomi*, Nitin Gupta[+], Eytan Ruppin*, Vineet Bafna[+], Roded Sharan*

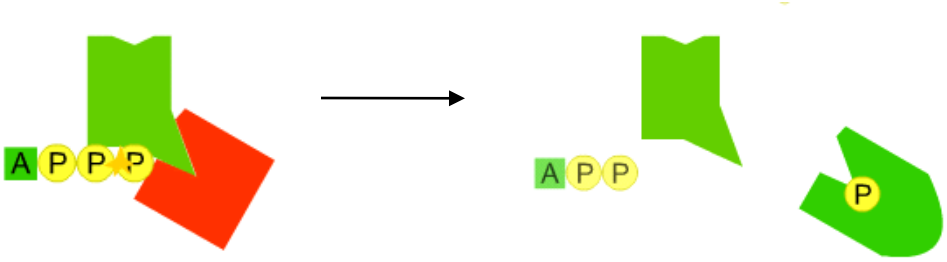
[+]University of California, San Diego
*Tel Aviv University, Israel
contact: bdost@cs.ucsd.edu



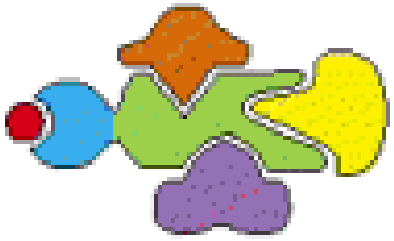
Protein Interaction Networks

- Proteins rarely function in isolation, protein interactions affect all processes in a cell.

- Forms of protein-protein interactions:
  - Modification, complexation [Cardelli, 2005].



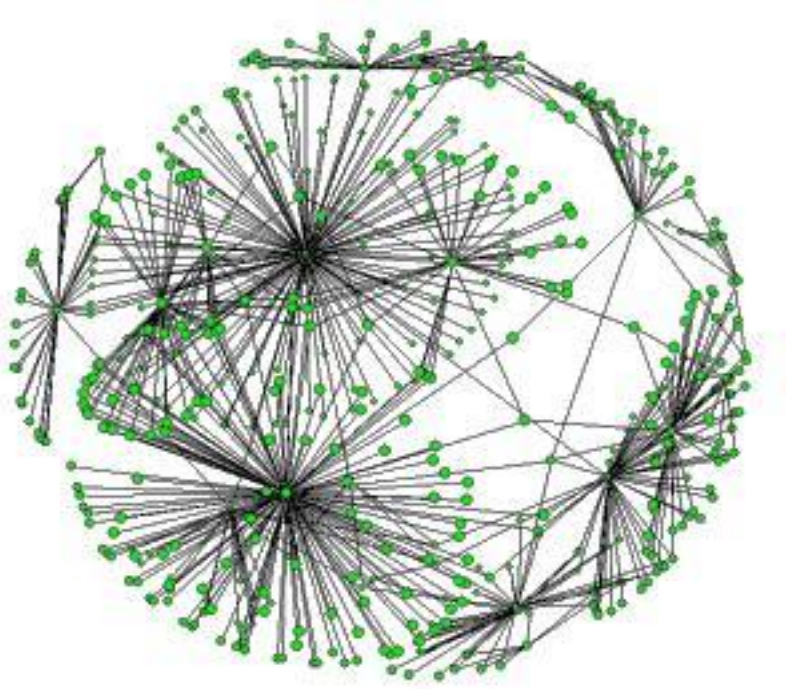
e.g. phosphorylation



e.g. protein complex

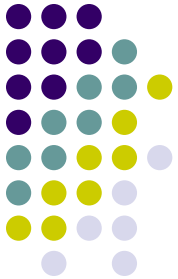
Protein Interaction Networks

- Proteins rarely function in isolation, protein interactions affect all processes in a cell.
- Forms of protein-protein interactions:
  - Modification, complexation [Cardelli, 2005]
- High-throughput methods are available to find all interactions, “PPI network”, of a species.
  - an undirected graph
    - nodes: protein, edges: interactions
  - Yeast DIP network: ~5K proteins, ~18K interactions
  - Fly DIP network: ~7K proteins, ~20K interactions

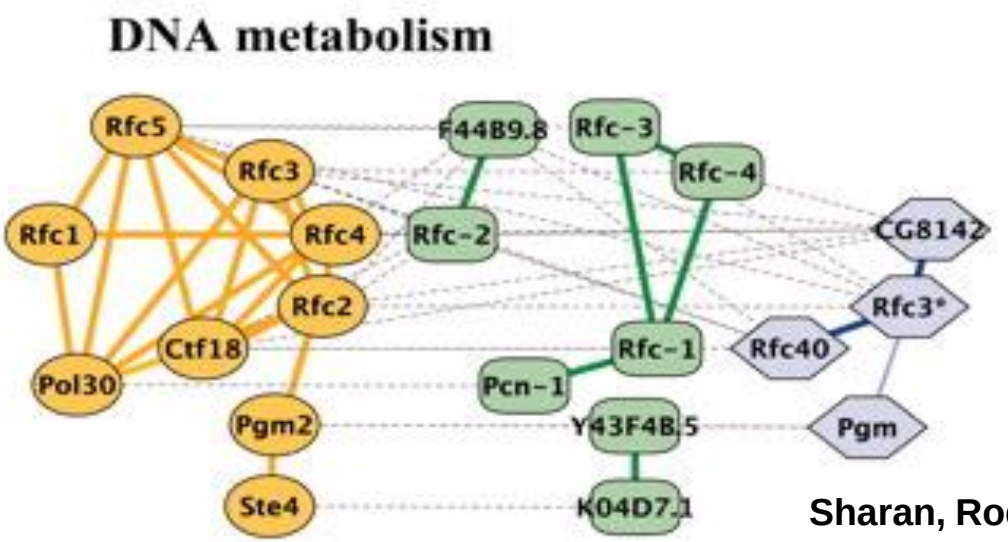


PPI network

Motivation: *Conservation of Subnetworks*



Yeast Worm Fly

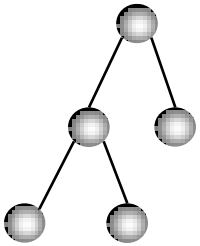


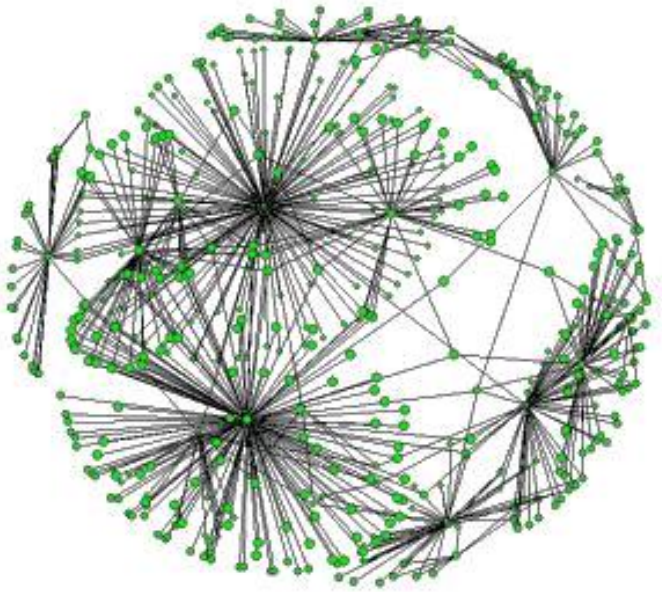
Sharan, Roded et al. (2005), PNAS

- Subnetworks can denote cellular processes, signaling pathways, metabolic pathways, etc.
- Many “subnetworks” are conserved across species.
  - Sequences are conserved
  - Interactions are conserved

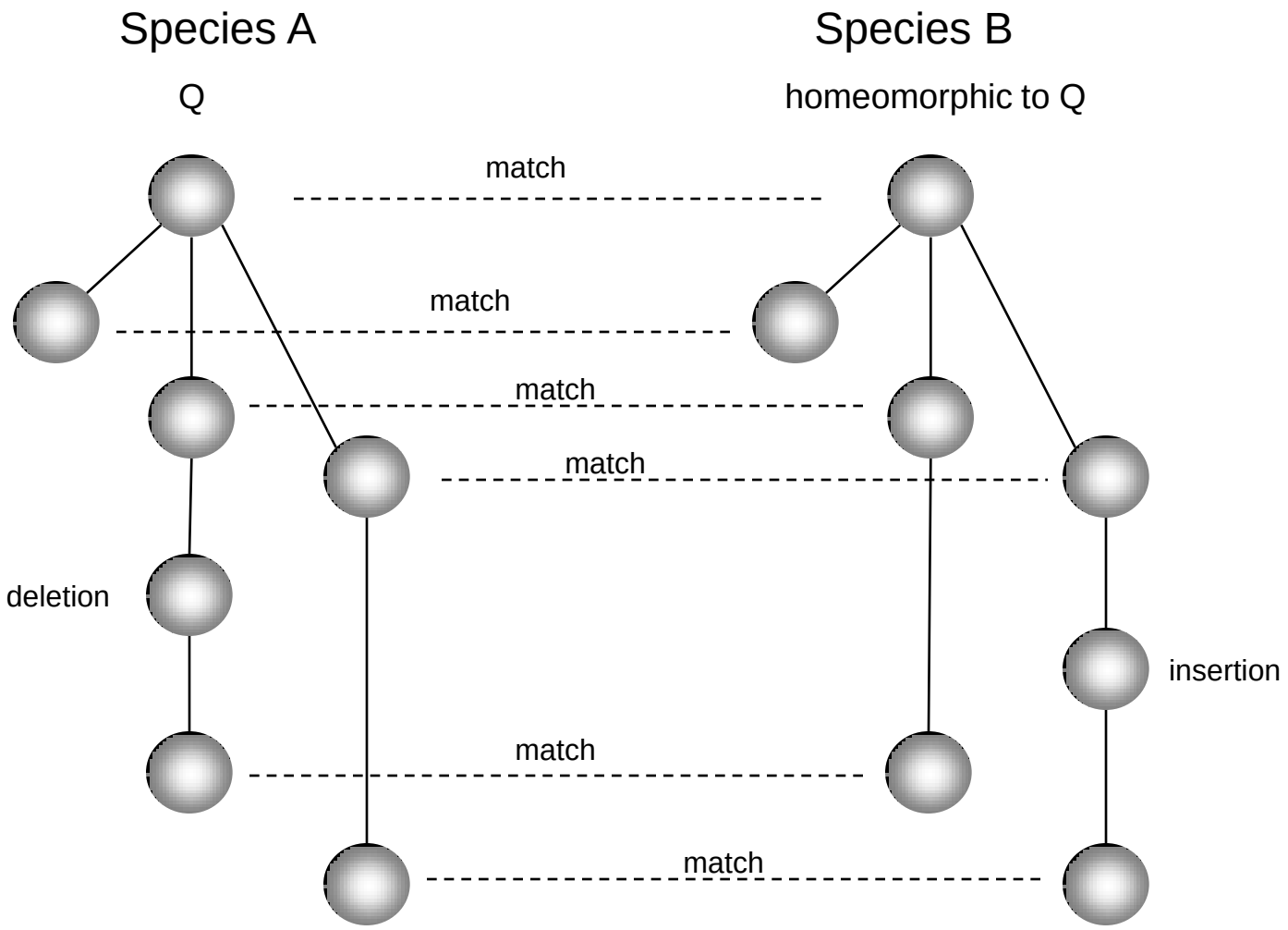
Network Querying Problem

- Species A
  - well studied
  - protein interaction sub-networks defined by extensive experimentation
- Species B
  - less studied
  - little knowledge of sub-networks
  - protein interaction network known using high-throughput technologies
- Can we use the knowledge of A to discover corresponding sub-networks in B if it is "present"?





Network Querying Problem: Homeomorphic Alignment



Match of homologous proteins and deletion/insertion of degree-2 nodes

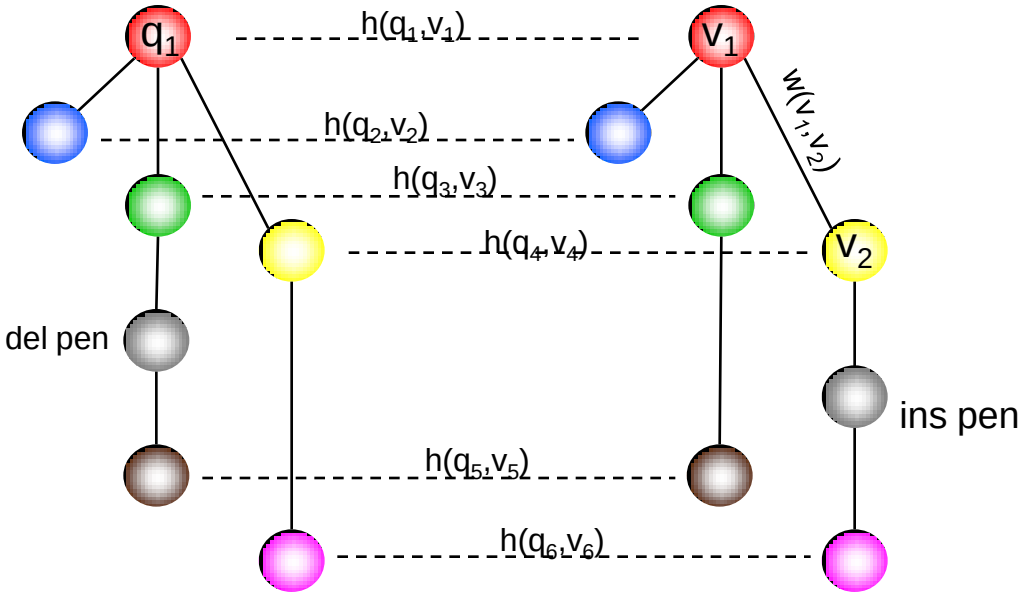
Network Querying Problem: Score of Alignment

$$Score = \sum h(q_i, v_j) + \delta_d(\#Del) + \delta_i(\#Ins) + \sum w(v_i, v_j)$$

Sequence similarity score for matches

Penalty for deletions & insertions

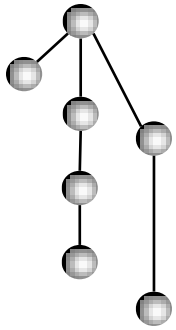
Interaction reliabilities score



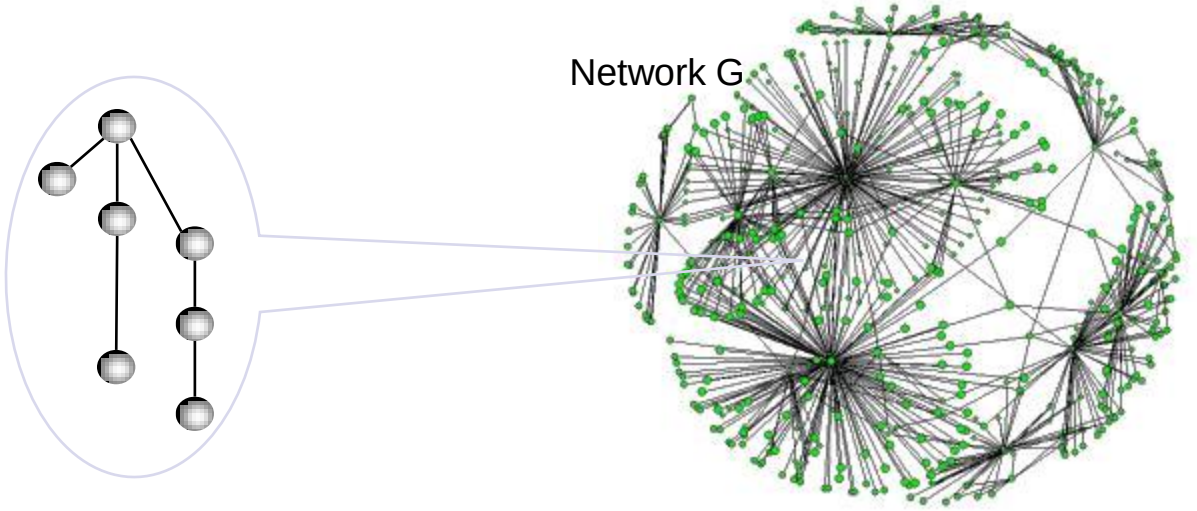
Network Querying Problem

- Given a query graph Q and a network G, find the sub-network of G that is
  - homeomorphic to Q
  - aligned with maximal score

Query Q

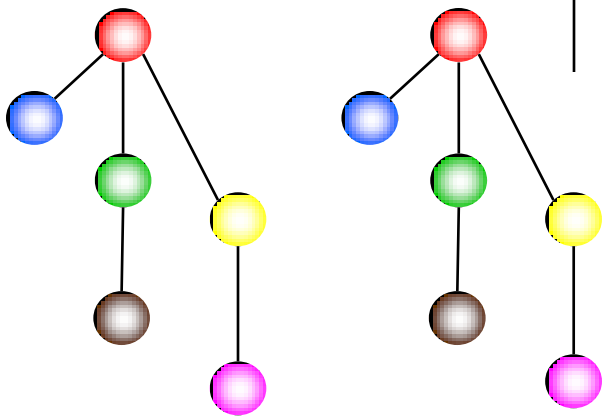


Network G



Complexity

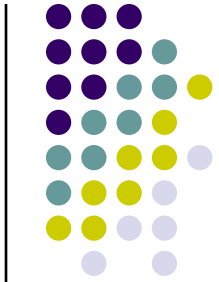
- Network querying problem is NP-complete. (for general n and k)
  - by reduction from sub-graph isomorphism problem



- 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 ~5000, k~10

- We use randomized "color coding" technique developed by [Alon et al, JACM, 1995] to find a tractable solution.
  - Reduces $O(n^k)$ to $n^2 2^{O(k)}$.

Previous Work

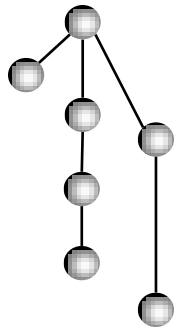


- Current Tools:
  - PathBlast [Kelley et al., 2003]
  - MaWish [Koyuturk et al., 2006]
  - Graemlin [Flannick et al., 2006]

- Different alignment interpretation
- Some heuristics to search for the optimal solution

QNET

- Implemented for tree-like queries.



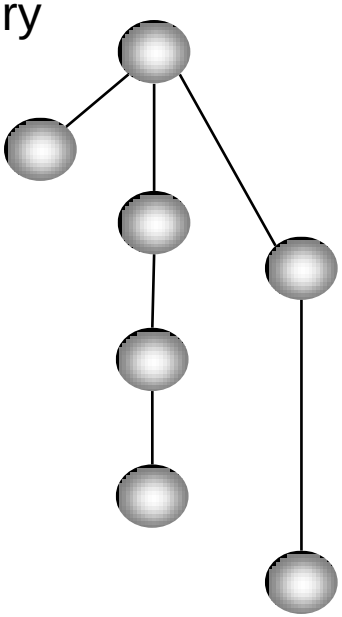
- Color coding approach to search for the global optimal sub-network.
- Extension of QPATH [Shlomi et al., 2006]
  - Solves the problem of querying chains using color coding approach.

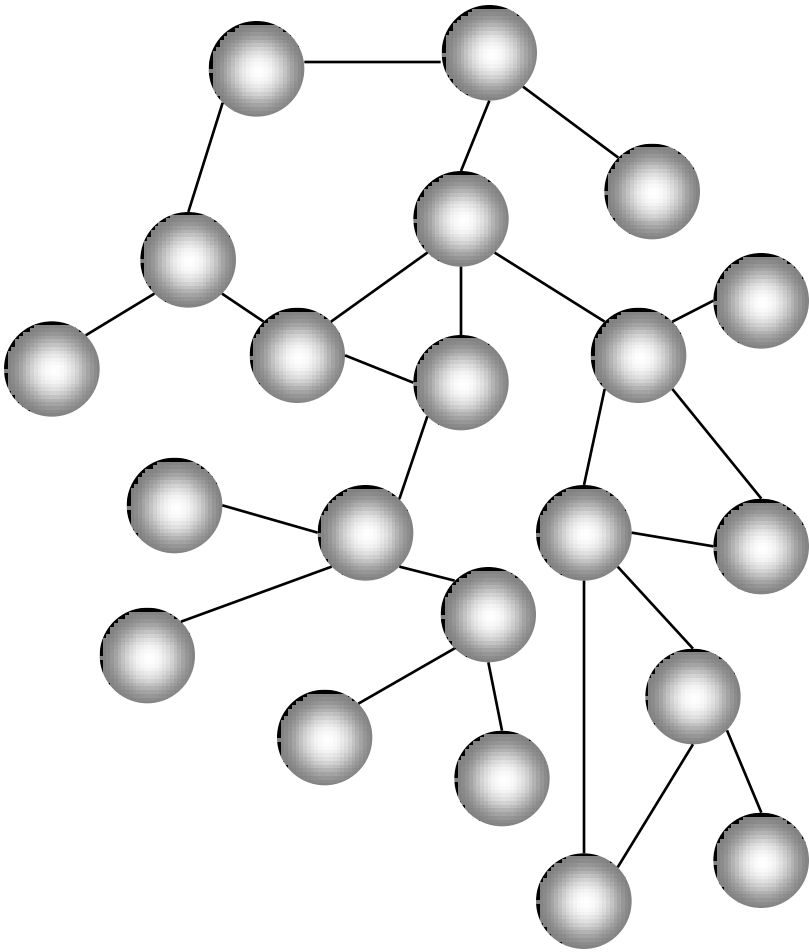
Color Coded Querying - Trees

Network

Query has k nodes.

Query

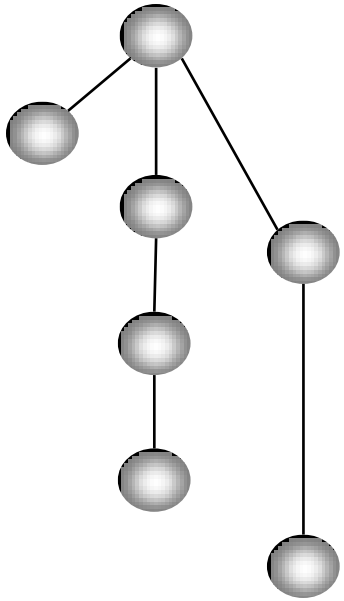


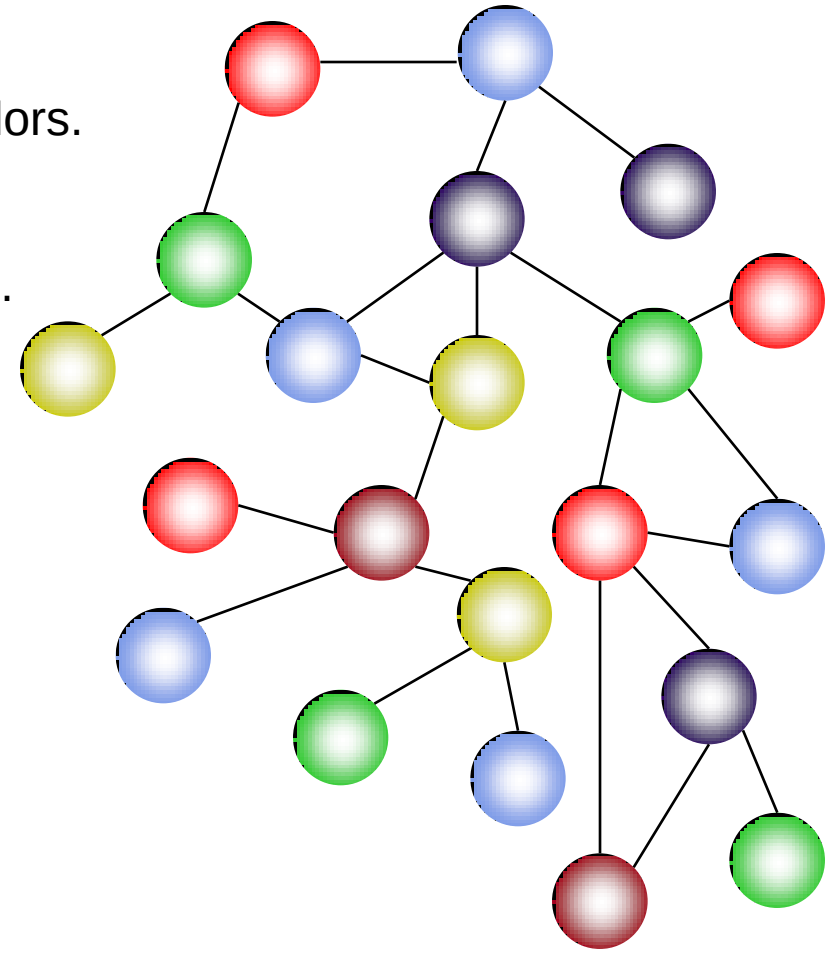


Color Coded Querying - Trees

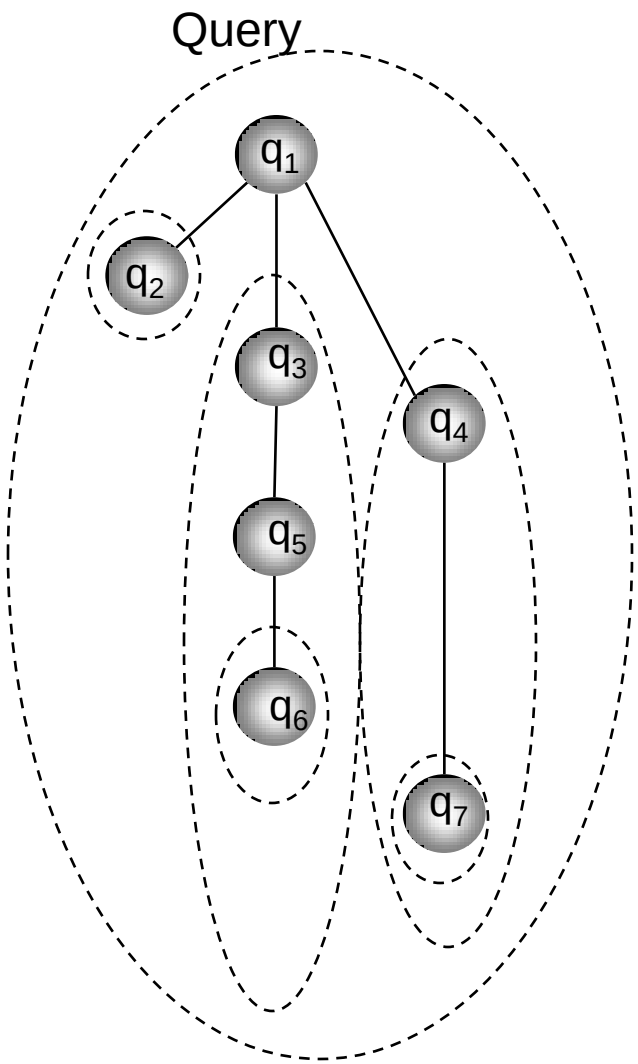
Network

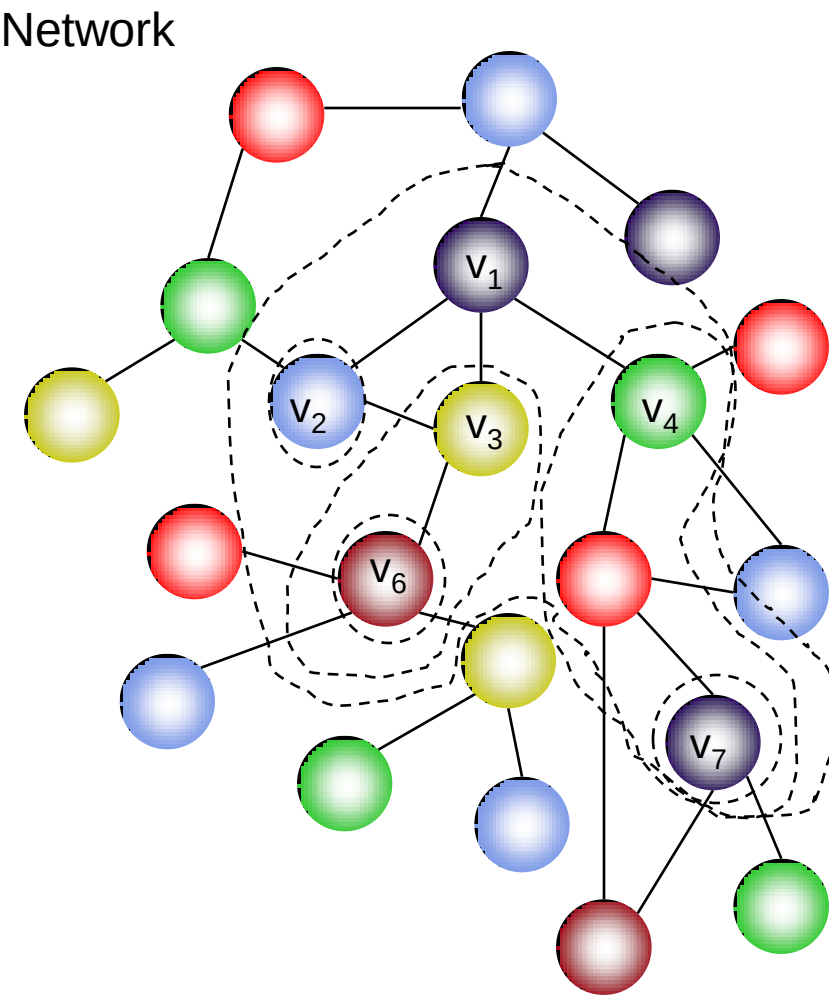
Query has k nodes.
Randomly color the network with k distinct colors.
Suppose optimal sub-network is “colorful”.
(all of its vertices colored with distinct colors)
Use the colors to remember the visited nodes.





DP solution for Color Coded Querying - Trees



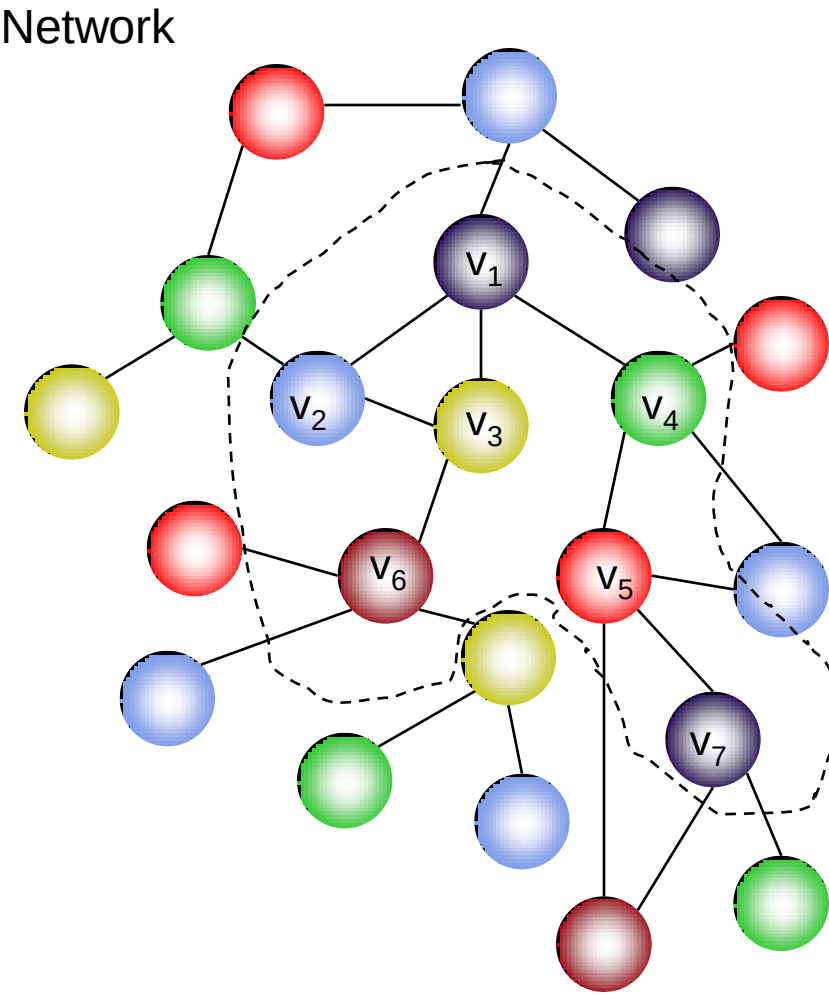


Probability of failure

- The optimal alignment can be found only if the optimal sub-network is “colorful”.

$$P(\,failure) = 1 - \frac{k!}{k^k} \le 1 - e^{-k}$$

- Repeat color-coded search multiple times until probability of failure ≤ ε.



Number of Repeats

- Necessary number of repeats to guarantee a failure ≤ ε?
  - Repeat $N = (\ln 1/\varepsilon).e^{k}$ times, then

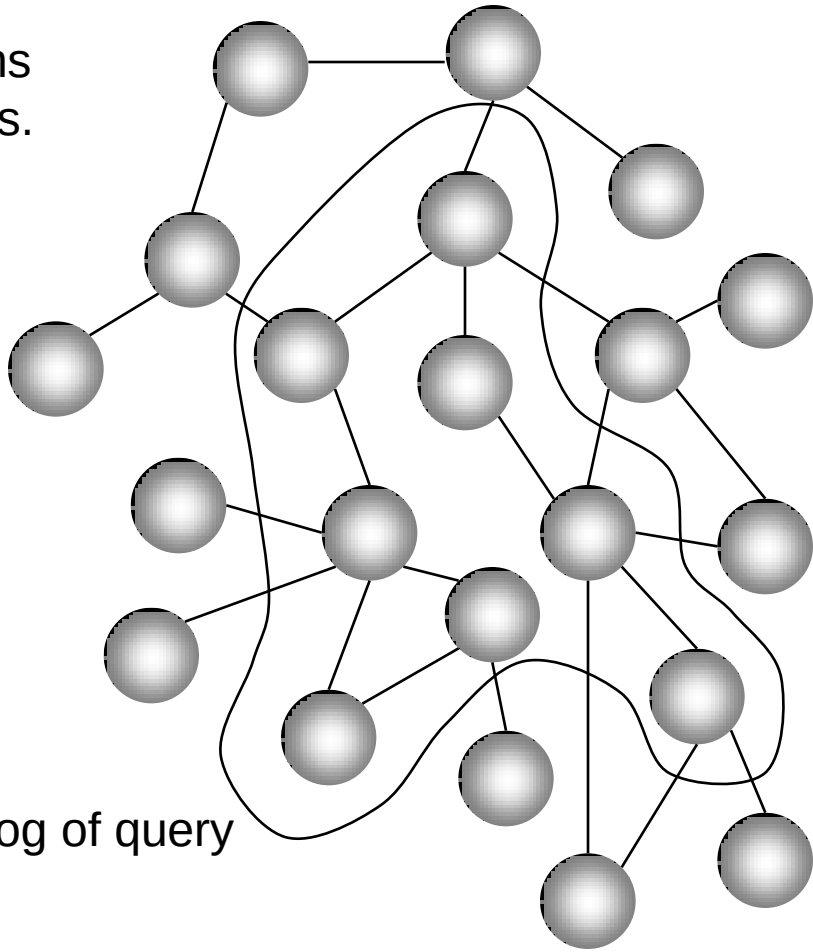
$$\text{P(failure per trial)} \le 1 - e^{-k} \le e^{-e^{-k}}$$

$$\text{P(failure in all trials)} \le e^{-e^{-k}.N} = e^{-\ln(1/\varepsilon)} \le \varepsilon$$

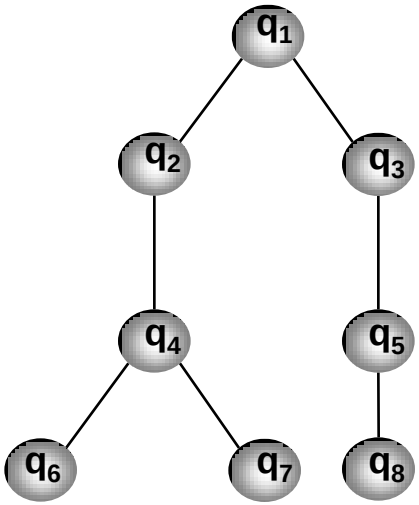
  - k=9 and ε=0.01 => N ~ 30K
  - We reduce N by a new approach “restricted color coding”.

Restricted Color Coding

Network

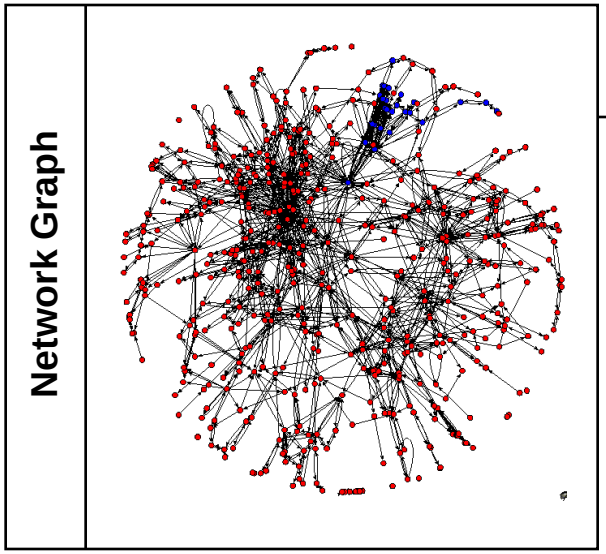


Idea: take advantage of queries whose proteins tend to have non-overlapping sets of homologs.

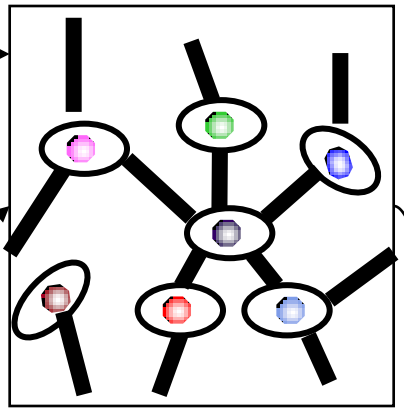


On average, N insertions & 10-fold reduction in homolog of query proteins => P(failure per trial) = 0.

Network Querying with Color Coding Approach

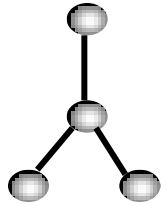


randomly color



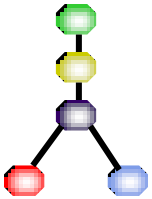
repeat
N times

query

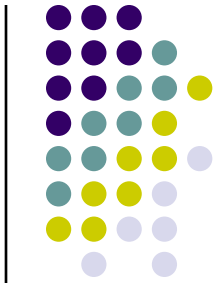


DP
algorithm

high scoring
subnetwork



Querying General Graphs

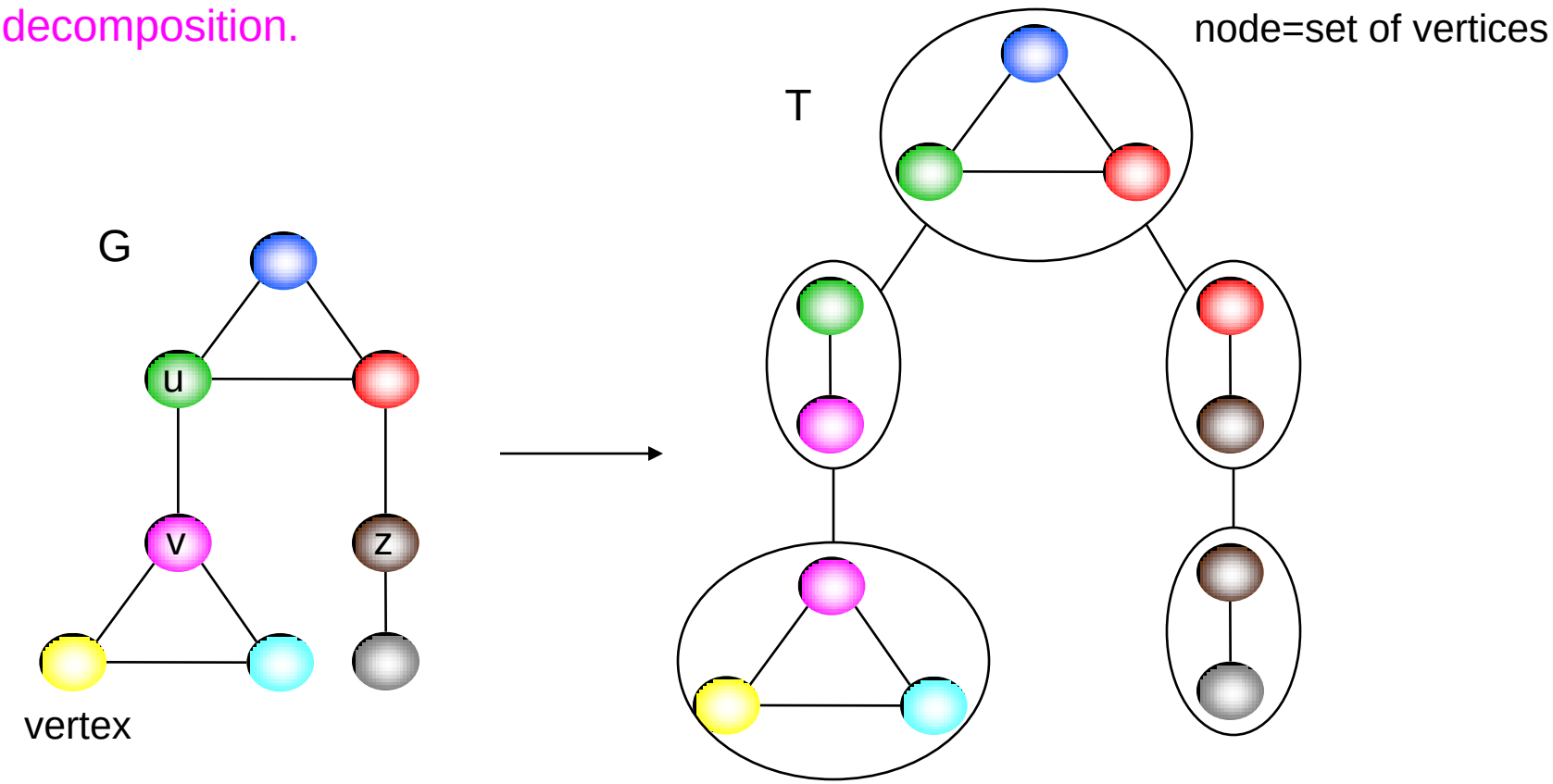


- We have extended the algorithm for also general graphs.

- Idea:
  - Map the original graph into a tree, i.e. tree decomposition. (Polynomial time for bounded-tree-width graphs)
  - Solve the querying problem on this tree using DP.

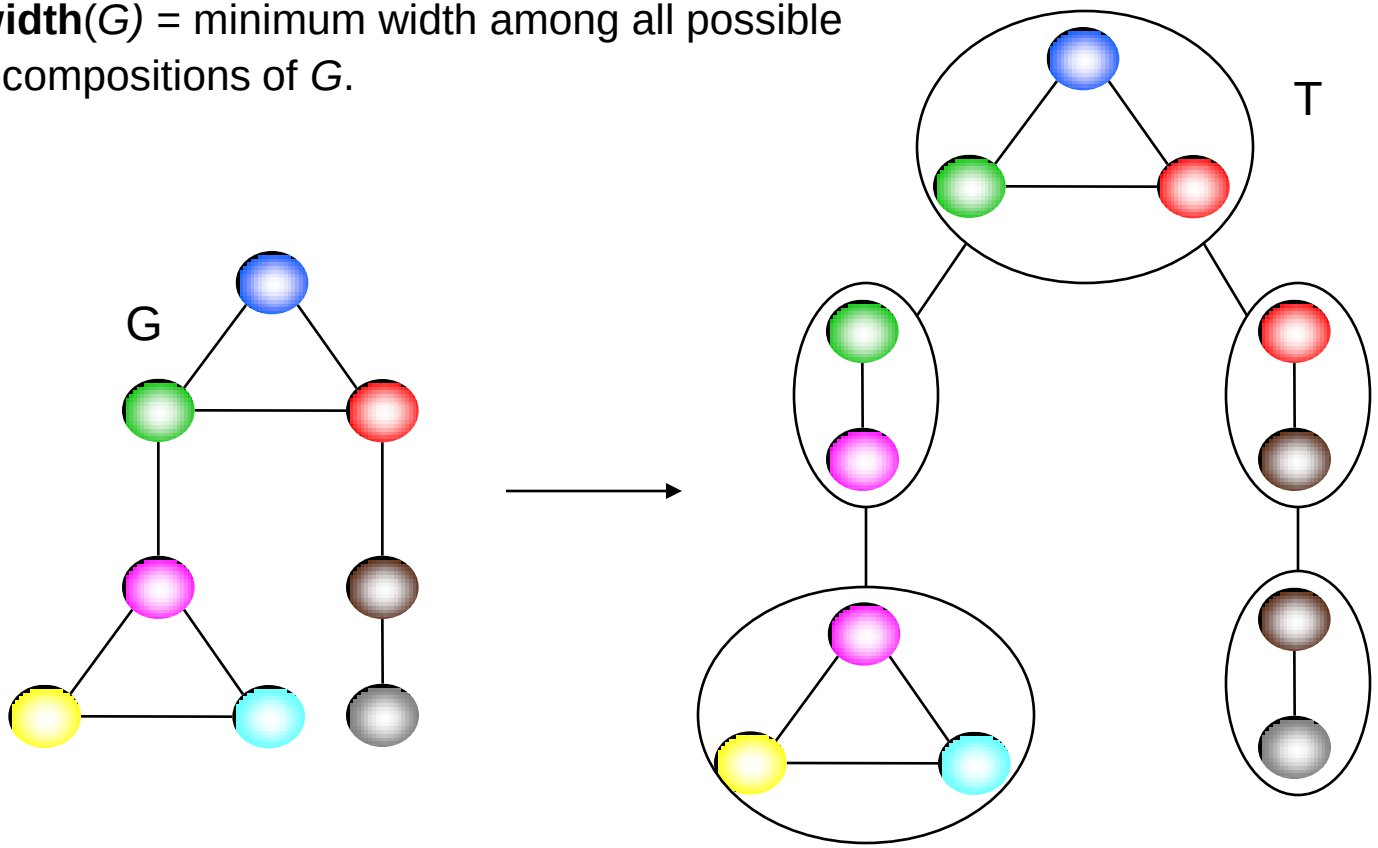
Color Coded Querying – General Graphs

Map the original query into a tree using tree-decomposition.



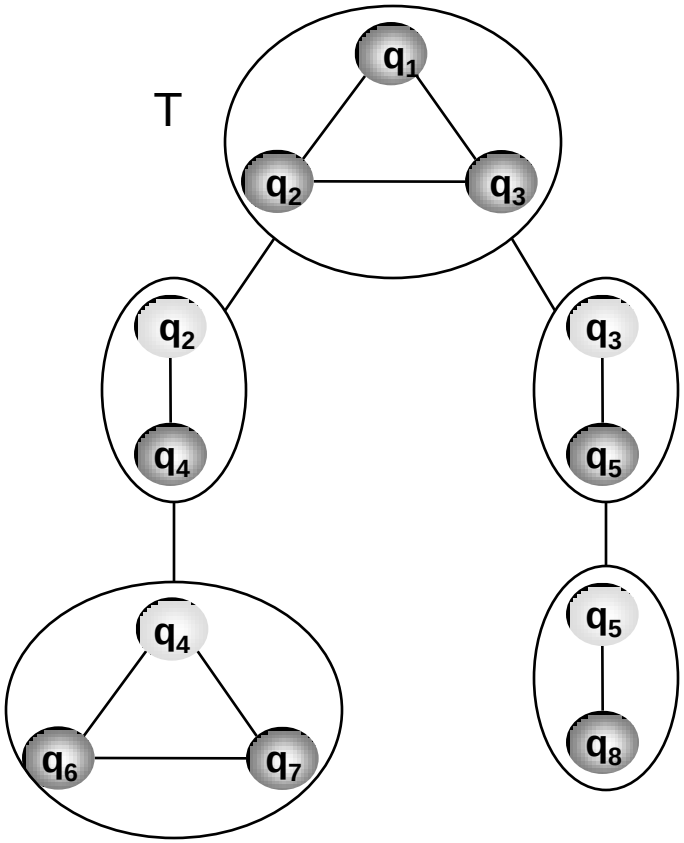
Color Coded Querying – General Graphs

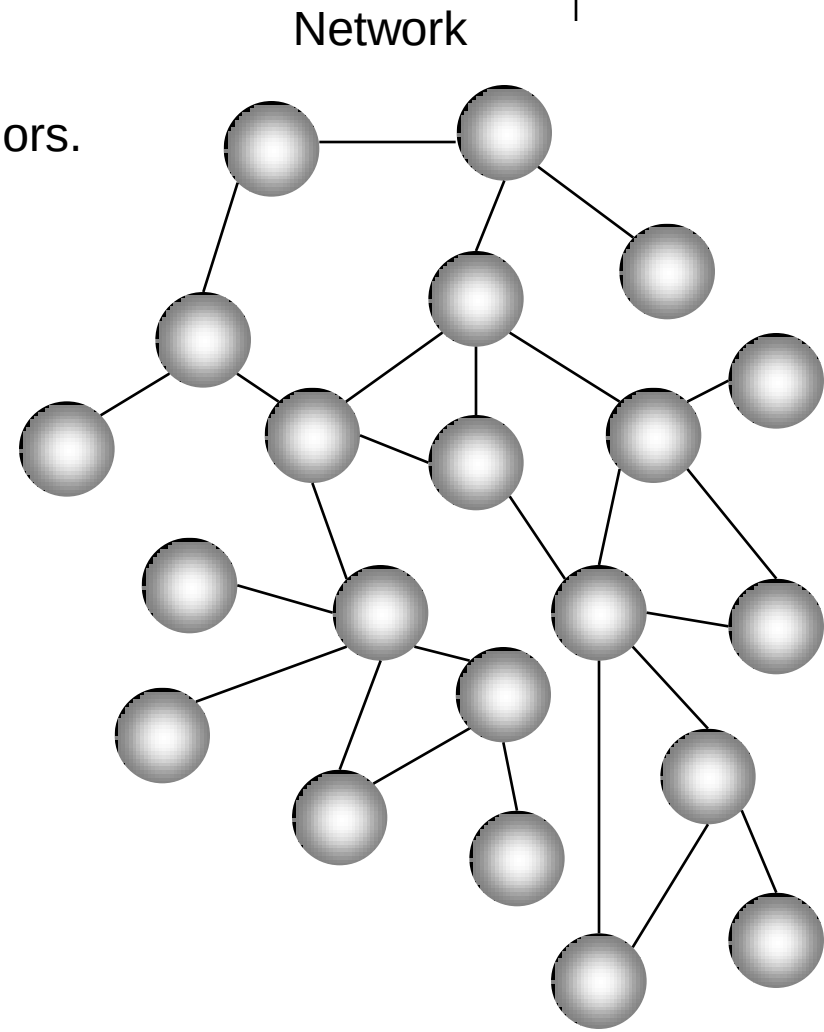
Width(T) = size of its largest node – 1.
Tree-width(*G*) = minimum width among all possible tree decompositions of *G*.



Color Coded Querying – General Graphs

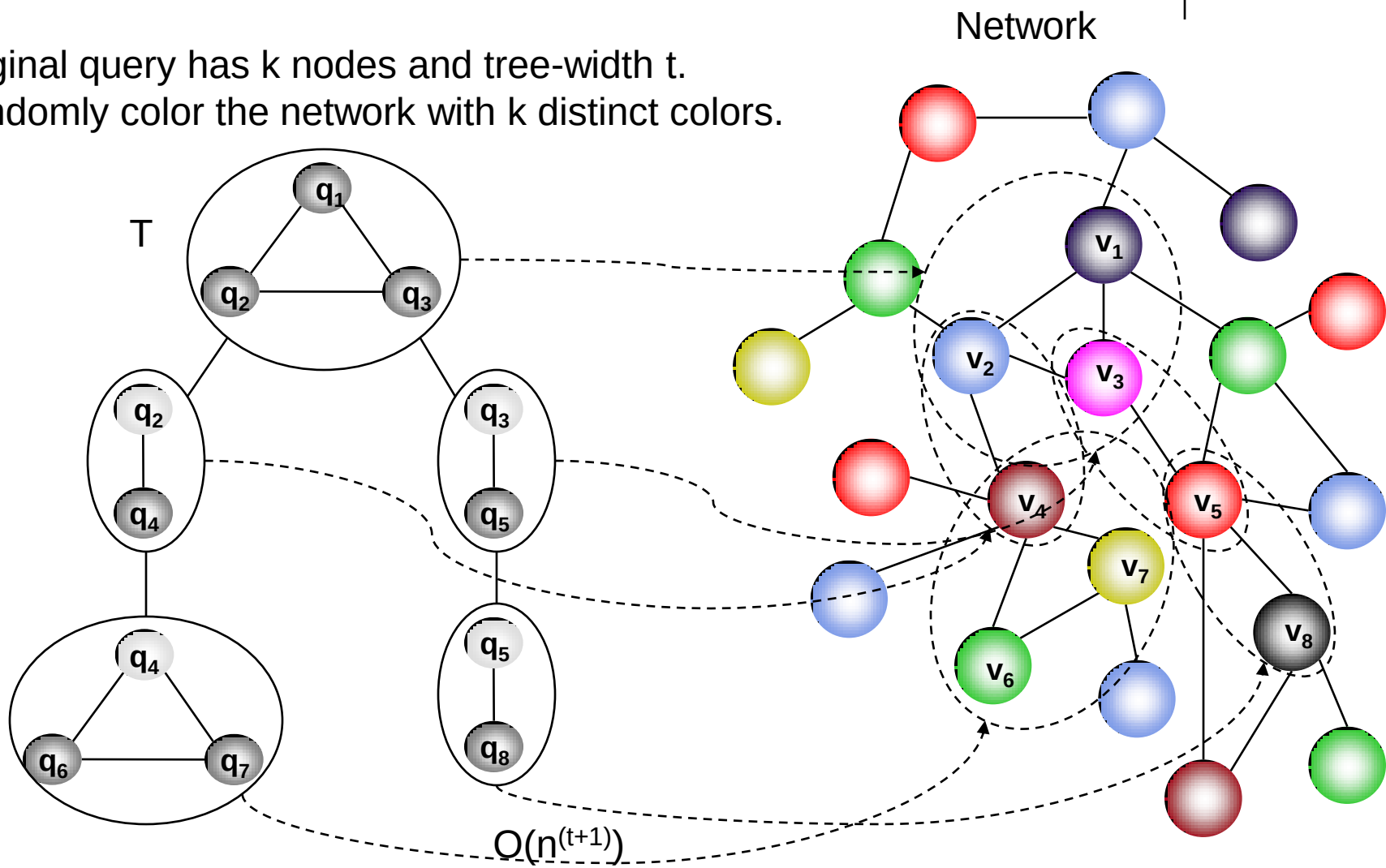
Original query has k nodes and tree-width t.
Randomly color the network with k distinct colors.
.



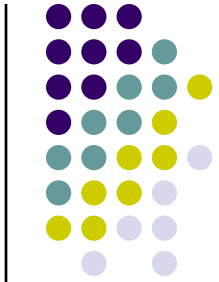


Color Coded Querying – General Graphs

Original query has k nodes and tree-width t.
Randomly color the network with k distinct colors.



Running time



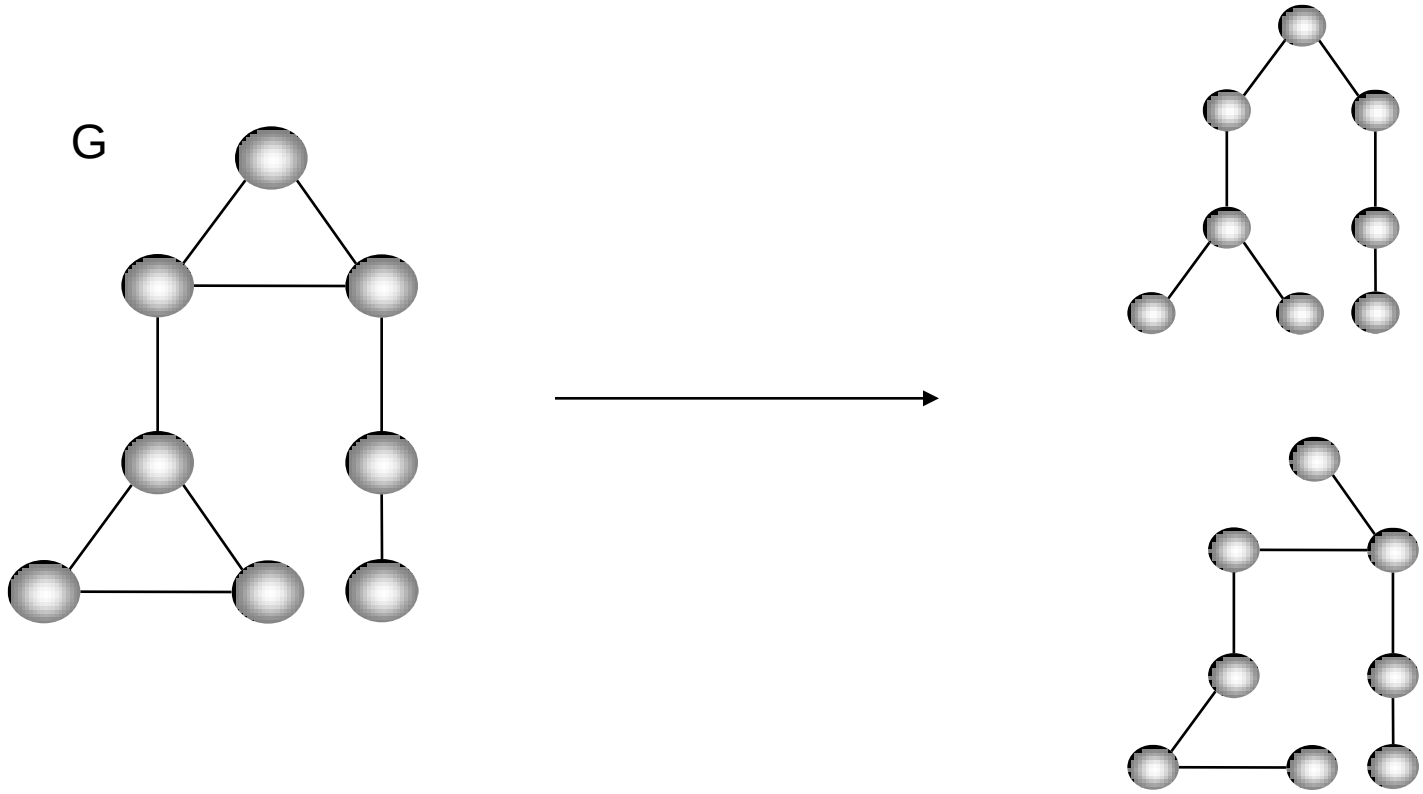
- n=size of network, k=size of query.

- Tree queries:
  - Reduces $O(n^k)$ to $n^2 2^{O(k)}$.
    - Tractable for realistic values of n and k.
    - n ~5000, k~10

- Bounded-tree-width graphs:
  - t : tree-width
  - $n^{(t+1)} 2^{O(k)}$

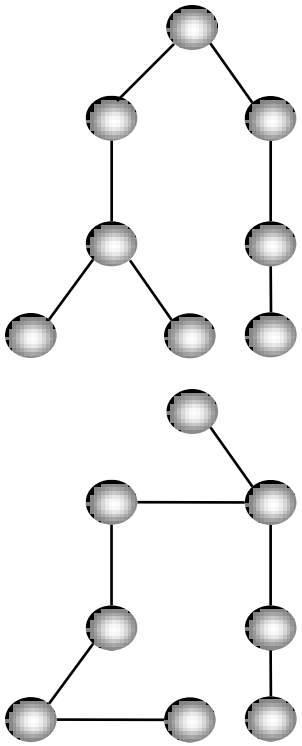
Heuristic for Color Coded Querying - General Graphs

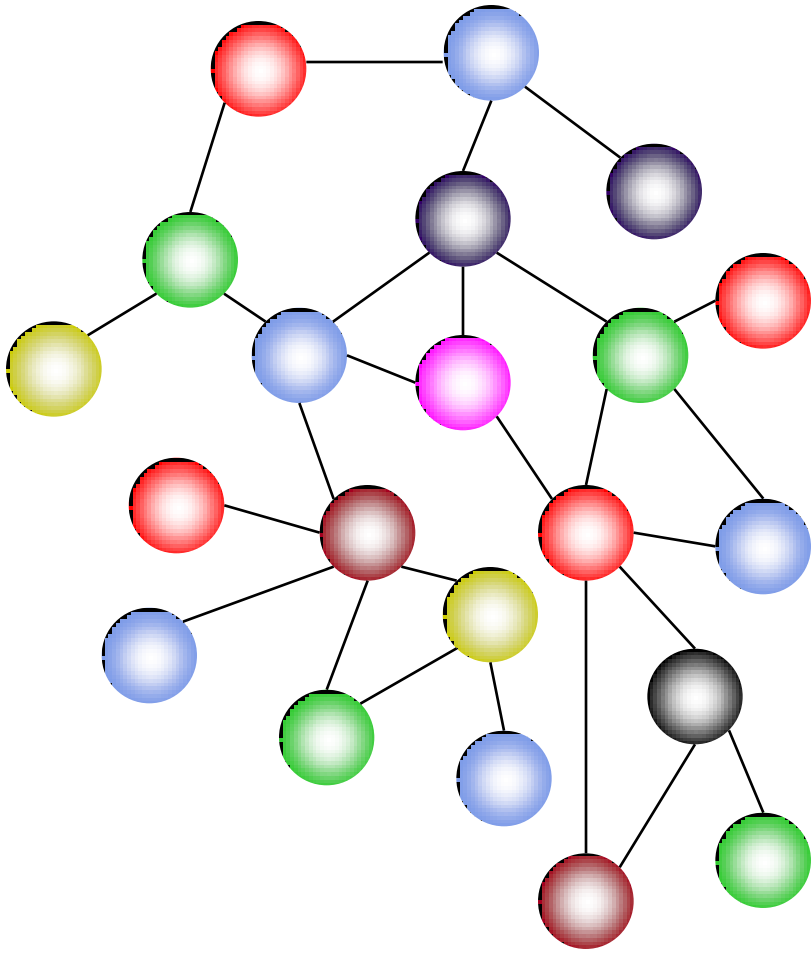
1. Extract several spanning trees from the original query.



Heuristic for Color Coded Querying - General Graphs

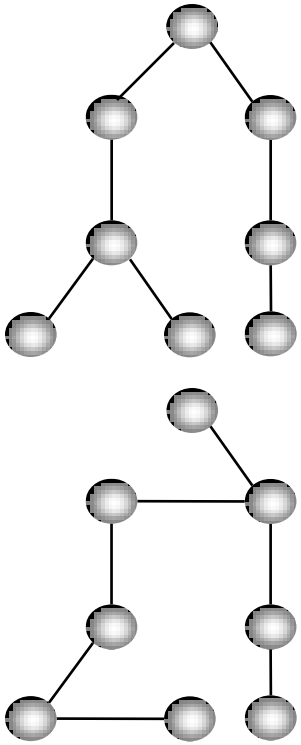
1. Extract several spanning trees from the original query.
2. Query each spanning tree in the network.

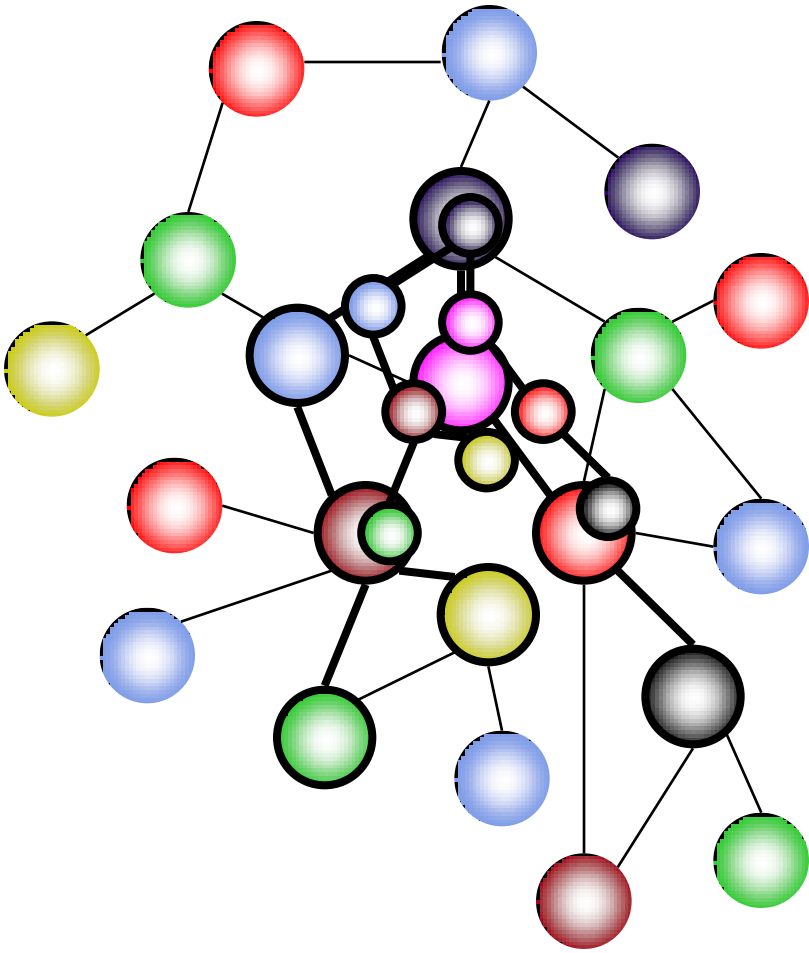




Heuristic for Color Coded Querying - General Graphs

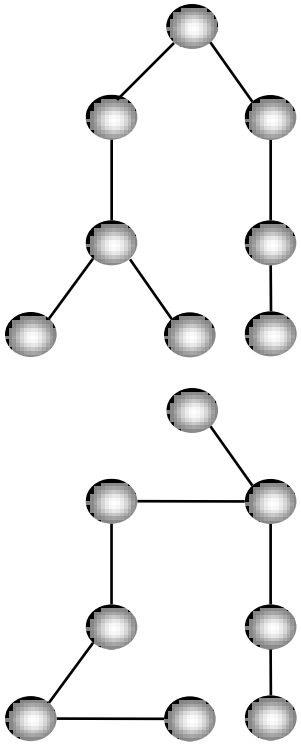
1. Extract several spanning trees from the original query.
2. Query each spanning tree in the network.

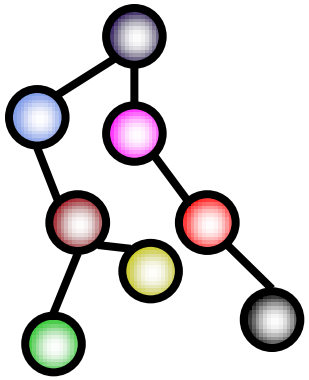


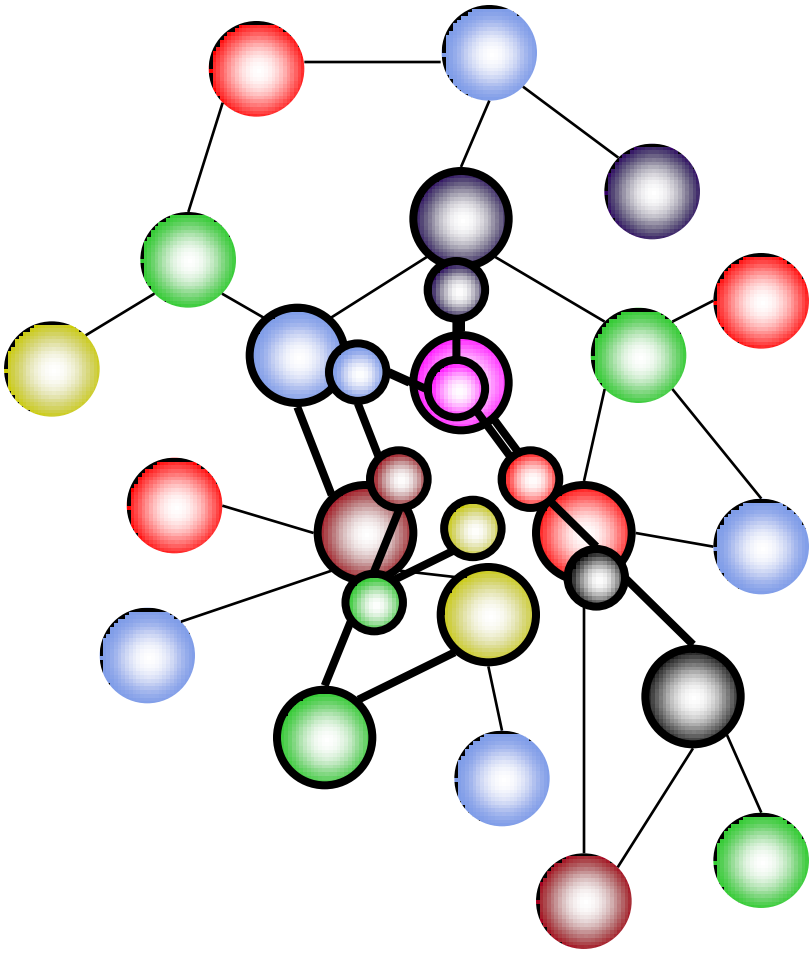


Heuristic for Color Coded Querying - General Graphs

1. Extract several spanning trees from the original query.
2. Query each spanning tree in the network.

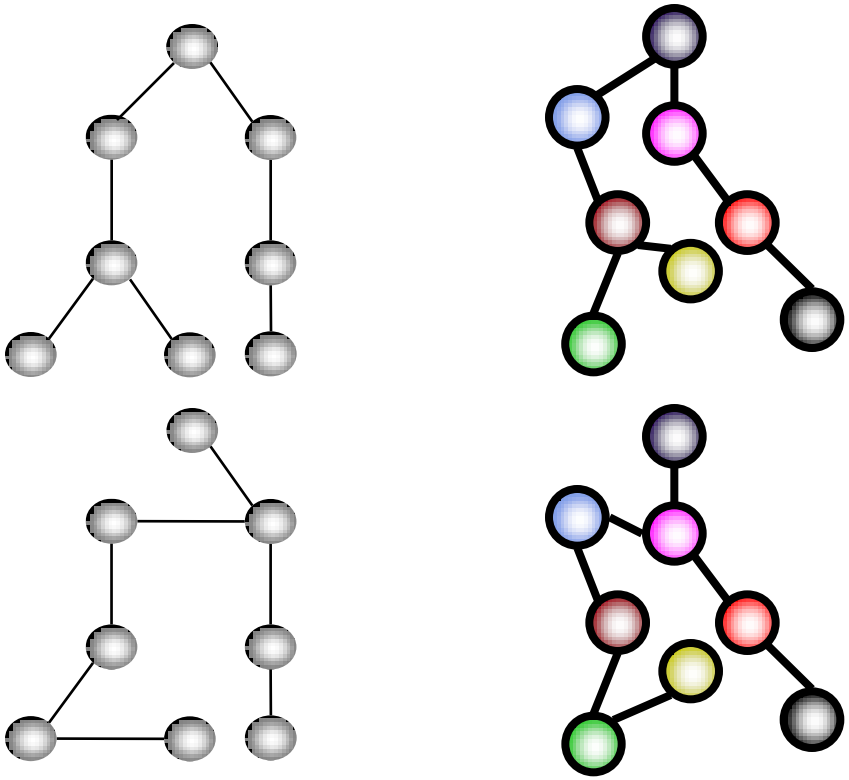


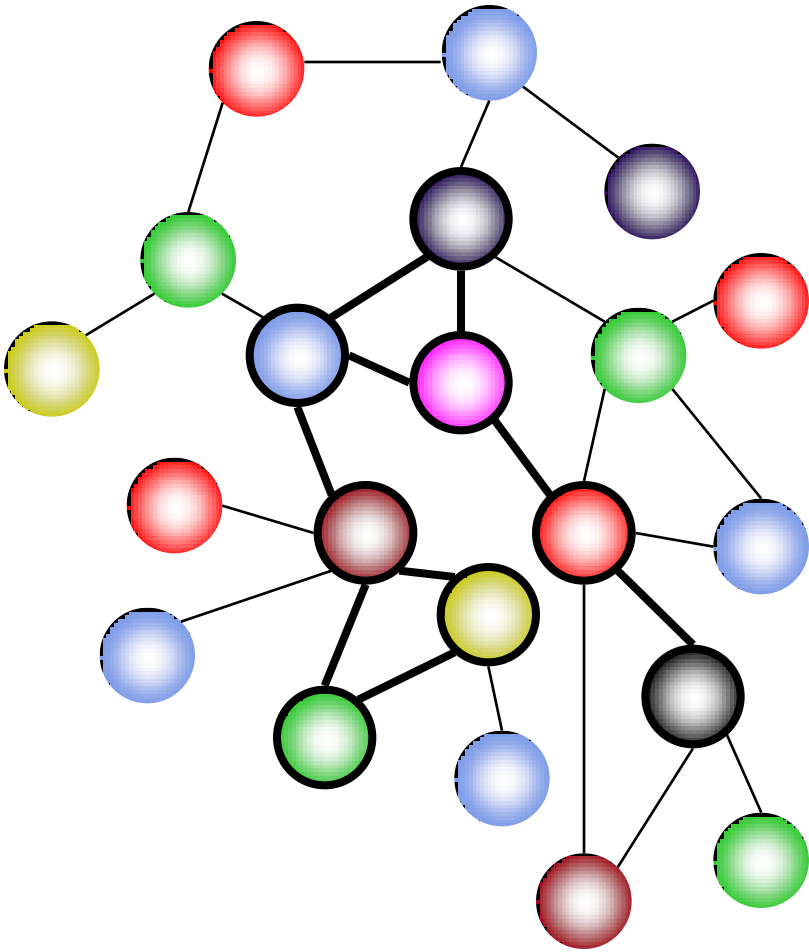




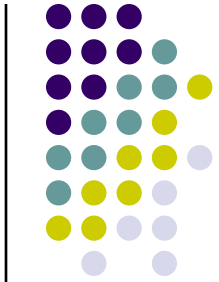
Heuristic for Color Coded Querying - General Graphs

1. Extract several spanning trees from the original query.
2. Query each spanning tree in the network.
3. Merge the matching trees to obtain matching graph.



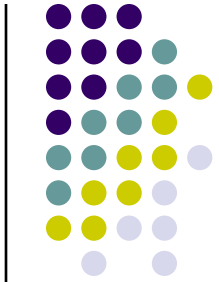


Testing



- Time
- Quality of solutions

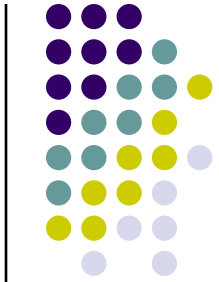
QNET: timing



- Handles queries with upto 9 proteins in seconds.

Query size (k)	#Iterations		Avg. time (sec)	
	Standard color coding	Restricted color coding	Standard color coding	Restricted color coding
5	752	603	1.71	1.58
6	1916	917	6.36	4.73
7	4916	1282	20.46	6.24
8	12690	1669	61.17	9.08
9	32916	2061	173.88	11.03

Test 1: Importance of Topology

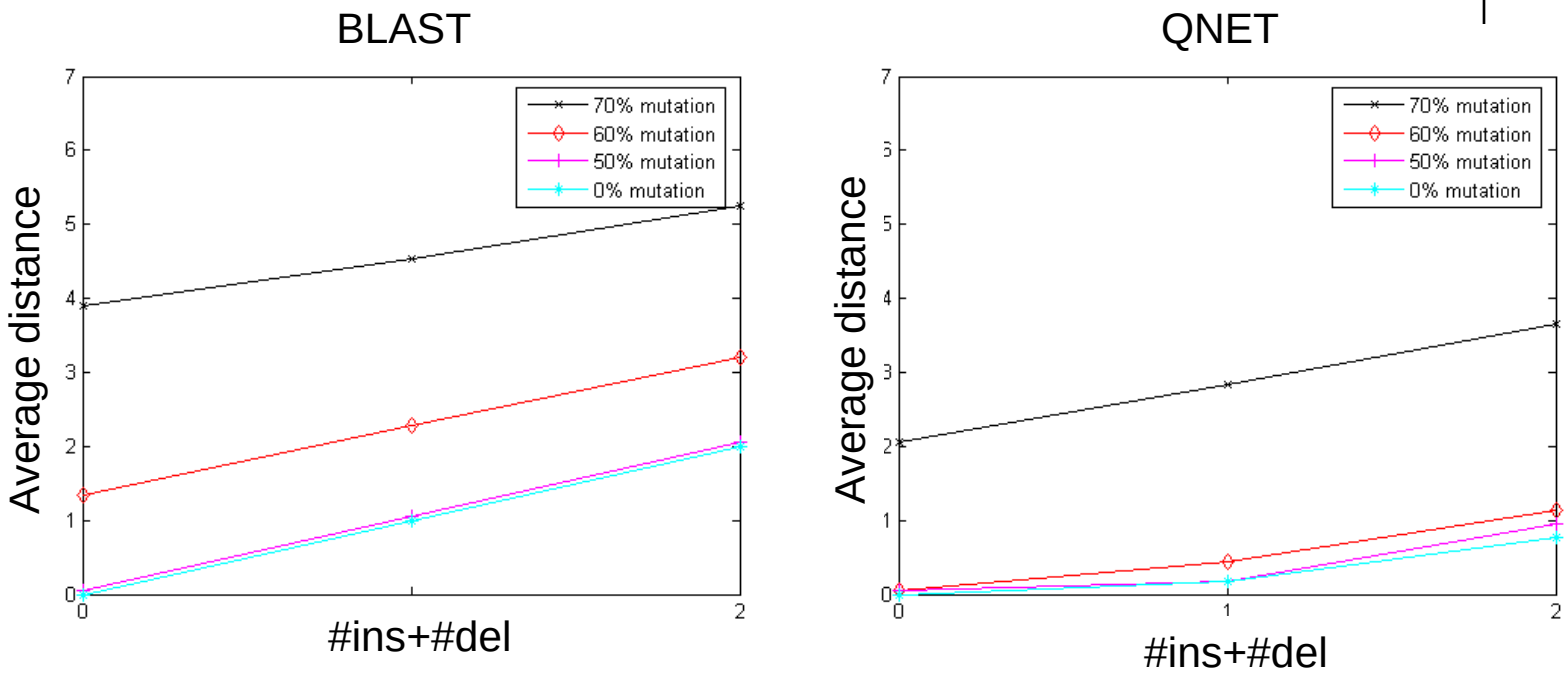


- Motivation: Is sequence similarity enough to find corresponding sub-network?

- Queries:
  - Random tree queries from yeast DIP network [Salwinski, 2004]
  - Topology perturbed (≤2 ins-dels).
- Network:
  - Yeast PPI
  - Protein sequences mutated (50-70 percent)

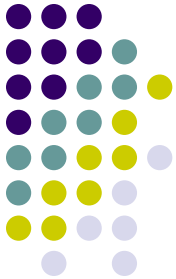
- How distant is the result from the original extracted tree?

Test 1: Importance of Topology



- Distance = #missing proteins + #extra proteins
- Outperforms sequence-based searches.

Test 2: Cross-species comparison of MAPK pathways



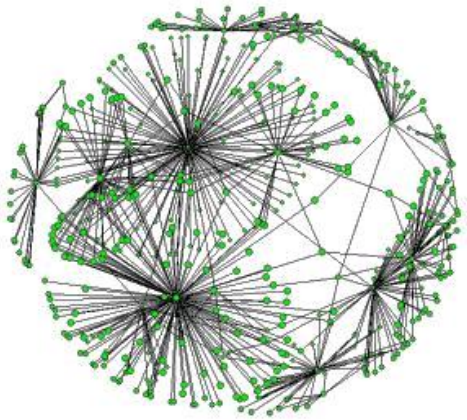
- Motivation: finding conserved pathways.

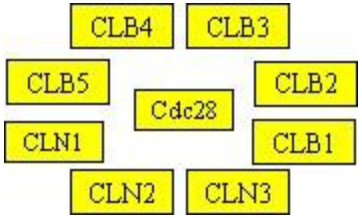
- Query: human MAPK pathway involved in cell proliferation and differentiation.
- Network: fly PPI network
  - ~7K proteins
  - ~20K interactions
- Match: a known fly MAPK pathway involved in dorsal pattern formation.

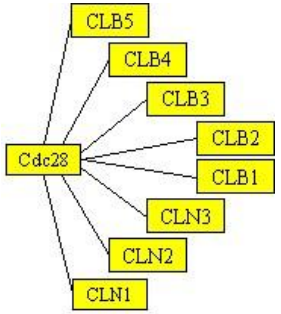
Test 3: Cross-species comparison of protein complexes

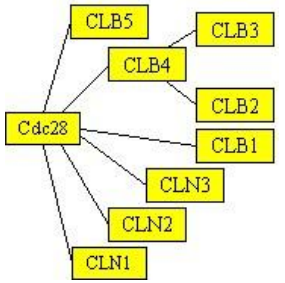
- Motivation: conserved protein complexes between yeast and fly.

- Queries:
  - Hand-curated yeast MIPS complexes [].
  - Project onto yeast DIP network
  - Extract several spanning trees







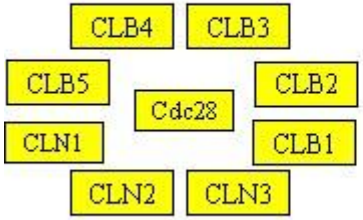


Test 3: Cross-species comparison of protein complexes

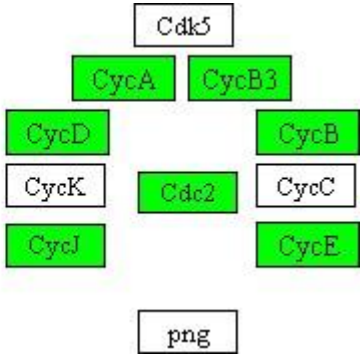
- Motivation: conserved protein complexes between yeast and fly.
- Queries:
  - Hand-curated yeast MIPS complexes [].
  - Project onto yeast DIP network
  - Extract several spanning trees
- Network:
  - Fly DIP network
- Match
  - Consensus matching graph for each query complex.

Test 3: Cross-species comparison of protein complexes

Yeast
Cdc28p complex

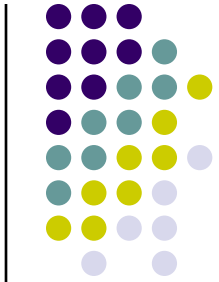


Fly



- Result:
  - ~40 of the queries resulted in a match with >1 protein.
  - 72% of the consensus matches are functionally enriched. (p-value < 0.05)
    - 17% of the random trees extracted from network are functionally enriched.

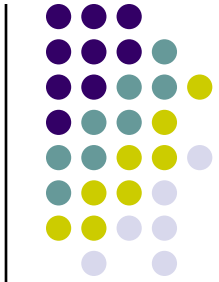
Summary



- QNET: a tool for querying protein interaction networks
  - Tree-like queries

- Randomized algorithm and heuristic proposed for querying general graphs.

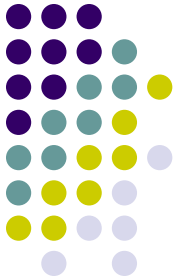
Future Work



- Development of appropriate score functions to better identify conserved pathways.

- Extending QNET for queries with more general structure.
  - bounded-tree-width graphs.

Thank you



- University of California, San Diego
- bdost@cs.ucsd.edu