



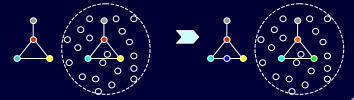
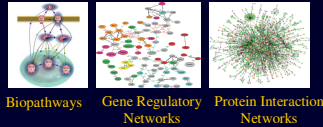
SAGA: An Effective and Efficient Approximate Graph Matching Tool

Yuanyuan Tian, Richard C. McEachin, Calos Santos, David J. States, Jignesh M. Patel
University of Michigan

<http://www.eecs.umich.edu/saga>, jignesh@eecs.umich.edu

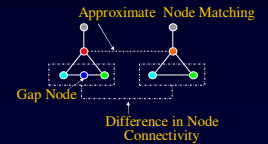
Motivation

- Graphs provide a powerful primitive for modeling real life data, especially biological data.
- With the rapid increase in the availability of graph data, there is a growing need for effective and efficient graph matching methods.
- Most real life datasets are noisy and incomplete in nature: so exact matching does not produce useful results.
- Need *approximate* graph matching.



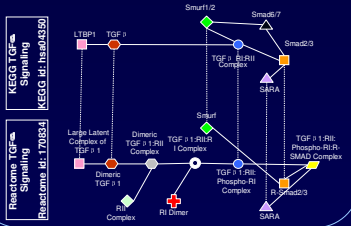
The SAGA Technique

- Goal: Approximate Subgraph Matching**
 - Find subgraphs in the database that are similar to the query graph.
- Graph Similarity Model**
 - Existing methods only allow exact subgraph matching or have very limited approximate matching models.
 - SAGA's graph similarity model allows node approximate matching, gap nodes, as well as differences in node connectivity.
- Index-based Matching Algorithm**
 - Build an index on small graph substructures in the database.
 - Use the index to match fragments of the query with fragments in the database, allowing for various types of mismatches.
 - Assemble larger matches using a graph clique detection algorithm.



Application: Comparing Different Pathway Databases

Comparing KEGG with Reactome

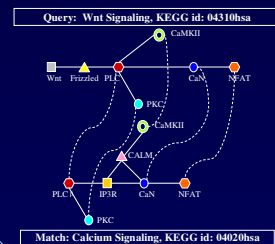


Comparing pathways from different databases can be a precursor to pathway data integration.

SAGA is able to find matches for similar information in different databases, even if they organize pathways in different ways.

Application: Pathway Analysis

Querying KEGG with Wnt Pathway



The Calcium pathway has two additional components arguably belonging to the Wnt pathway.

- Significant similarities between Wnt/CA and Wnt/Hedgehog.
- CA signaling has known Bipolar Disorder (BD) association for > 40 years.
- Hypothesis: The three pathways may share disease association in BD.**

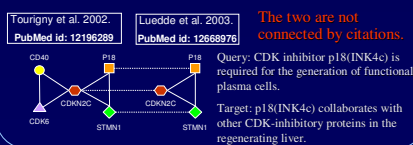
Literature Search:

Signaling Pathway	BD Association References
Calcium	335
Wnt	15
Hedgehog	0

- Hedgehog signaling pathway has been largely overlooked in BD research, although it uses BD-associated components.

Application: Querying Parsed Literature Graphs

Querying PubMed:12196289



The two are not connected by citations.

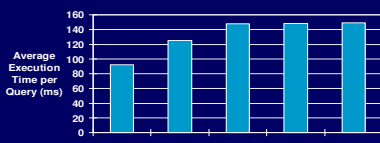
Query: CDK inhibitor p18(INK4) is required for the generation of functional plasma cells.

Target: p18(INK4) collaborates with other CDK-inhibitory proteins in the regenerating liver.

PubMed documents are summarized by genes and gene associations using natural language processing.

SAGA can identify documents addressing similar topics, even if they are published in different research areas.

Efficiency of SAGA



10 disease pathway queries: 25.5 nodes and 19.0 edges per graph on average.

SAGA is orders of magnitude faster than existing tools.

Querying Wnt signaling pathway (73 nodes and 92 edges) against D1 dataset.

NetworkBlast: > 20 hours

SAGA: ~ 8 minutes

dataset	D1	D2	D3	D4	D5
species	human	D1+mouse	D2+rat	D3+worm	D4+yeast
#graphs	162	320	470	567	654
avg #nodes	86.0	86.3	86.6	89.0	91.3
avg #edges	35.3	34.8	31.7	28.5	27.3

Conclusions and Future Work

- The SAGA graph similarity model is flexible and powerful.**
 - Produces biological meaningful results that cannot be found by existing methods.
- The index-based matching algorithm is efficient and scalable (for small query graphs).**
- Future Directions:**
 - Scale to handle large query graphs.
 - Apply to other domains.