SAGA: A Fast and Flexible Graph Matching Tool

Yuanyuan Tian, Richard C. McEachin and Jignesh M. Patel*
National Center for Integrative Biomedical Informatics, University of Michigan

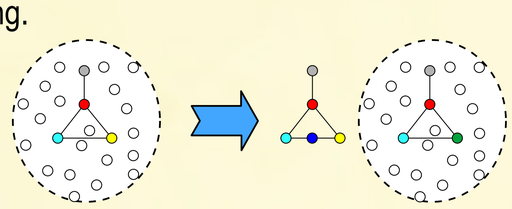
1. Overview

Need for Graph Matching Methods:

- Biological networks (graphs) characterize how individual molecules interact in biological processes. Example networks: pathways and protein interaction networks.
- Graph matching and graph comparison are primary operations for understanding cellular functions encoded in biological networks.

Challenges:

- Large amount of biological graph data – KEGG, GenMAPP, HPRD, BIND, …
- Graph database sizes are large and increasing in size.
- Datasets are noisy/incomplete – ***Approximate*** graph matching is required. Approximate graph matching is computationally more expensive than exact graph matching.



2. Model/Approach

The SAGA Graph Similarity Module:

- Allows approximate matching of node/edge labels, and structural differences (e.g. allow node/edge deletion and addition).
- A powerful mechanism for dealing with noise/partial information.
- Employs an index-based method to efficiently evaluate approximate graph matching.

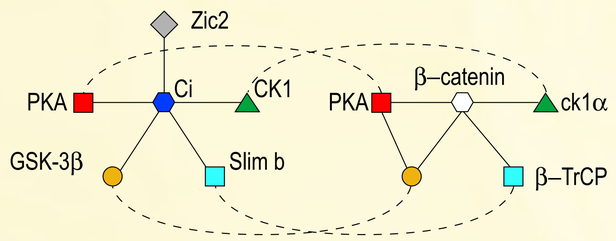
The Database-Centric SAGA Approach:

- 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.

3. Anecdotal Examples

The Wnt pathway is well studied, and its similarity to the hedgehog pathway can be used to understand and predict additional entries in the hedgehog pathway.

Querying KEGG with the Hedgehog Pathway



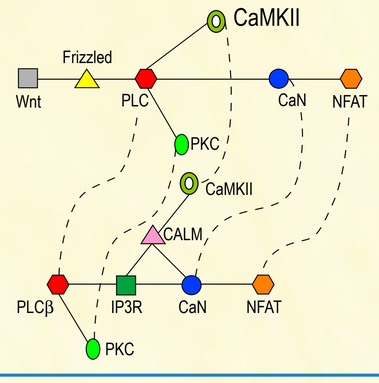
Query: Hedgehog Signaling KEGG id: 04340hsa	Match: Wnt Canonical Signaling KEGG id: 04310hsa

The Calcium pathway has two additional components arguably belonging to the Wnt/Ca2+ pathway.

Querying KEGG with Wnt/Ca2+ Pathway

Query: Wnt/Ca2+ Signaling KEGG id: 04310hsa

Match: Calcium Signaling KEGG id: 04020hsa



Conclusion

- The SAGA model is flexible and powerful
  Produces biological meaningful results
- The indexing method is efficient and scalable
  A query with 8 nodes and 28 edges is evaluated in a few seconds when running against a database of 12,065 graphs (with an average of 172 nodes and 21,312 edges per graph).

Future Work

- Design efficient matching algorithms for very large graph queries.
  e.g. aligning a large protein interaction network against a set of other protein interaction networks.
- Design efficient algorithms for other important graph operations
  e.g. path queries, boolean graph queries (union, intersection, difference)
- Mining frequent subgraphs

Acknowledgements

This research was supported by the National Institutes of Health under grant #1-U54-DA021519-01A1.