miBLAST and SAGA: Two Scalable NCIBI Bioinformatics Tools

Jignesh M. Patel

Department of EECS

University of Michigan

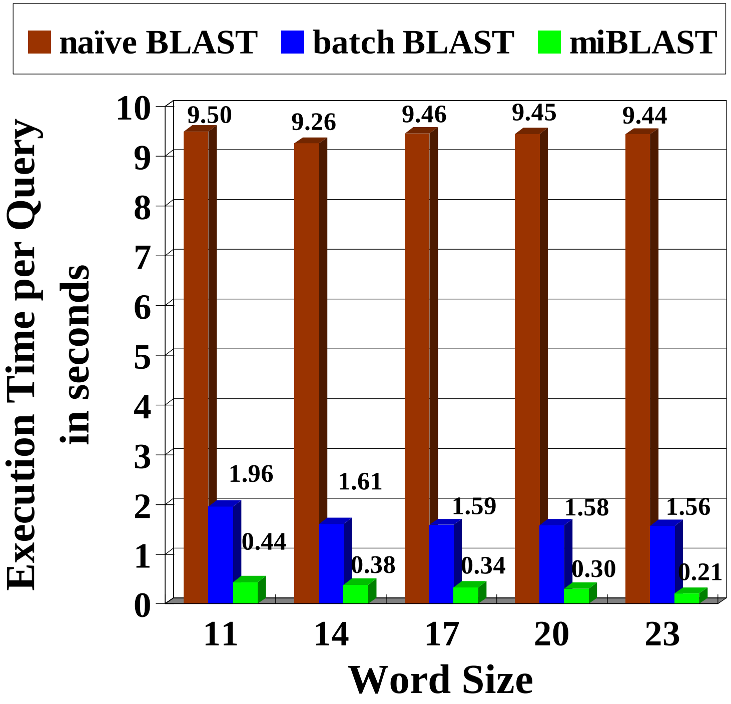
Contact: jignesh@eecs.umich.edu

Kim, Boyd, Athey, Patel, NAR 2005

miBLAST: Scalable BLAST for Batch Workloads

- A common task is to search a large sequence database using a “set” of query sequences.
  - Validation of the Affymetrix probe set against UniGene.
- Running BLAST repeatedly for each query is inefficient.
- Approach: A novel database-inspired “join” algorithm which indexes both the data and the query sets.
- Free download at www.eecs.umich.edu/miblast
- Modifications for MPSS underway (for actual deployment at ISB)

Query the Affymetrix probe set against Human UniGene



miBLAST is 22X faster than BLAST

Tian, McEachin, States, and Patel

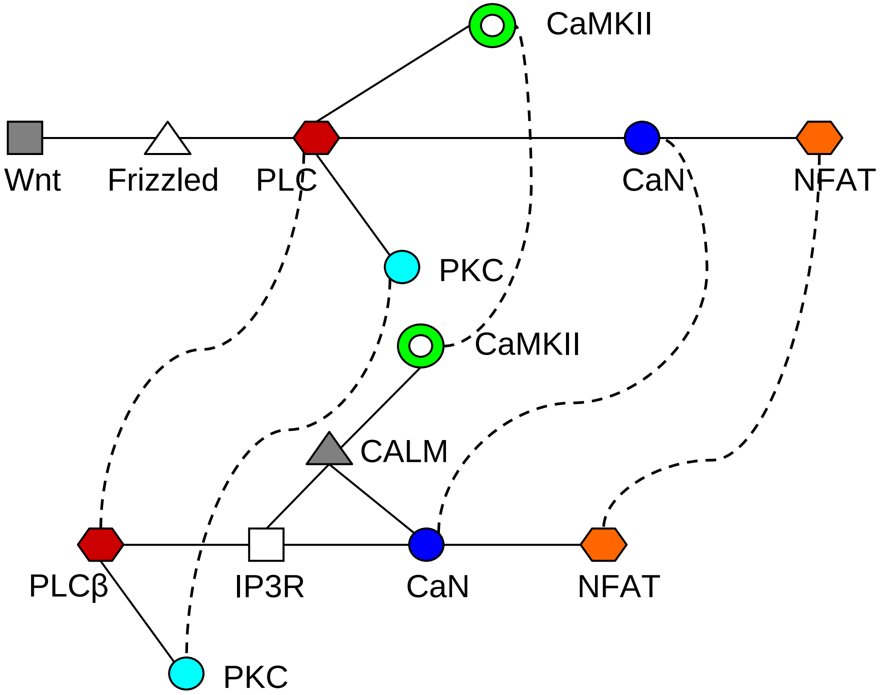
SAGA: A Fast and Flexible Graph Matching Tool

- Motivation
  - Graph querying is a common requirement for many DBPs.
  - Examples of graph datasets: KEGG, bioNLP, MiMI, …
  - Datasets are noisy/incomplete, so exact matching is inadequate.
- Challenge: Graph Matching is a Hard Problem
  - Subgraph isomorphism is NP-complete!
  - *Approximate* Subgraph Matching: Allow approximate matching of node/edge labels, and structural differences.
- The database-centric SAGA approach
  - Build an index on small database graph substructures.
  - Use the index to match database and query graph fragments.
  - Assemble larger matches by detecting graph cliques.

Results: Query KEGG with Wnt/CA2+ Pathway

Query: Wnt/Ca2+ Signaling

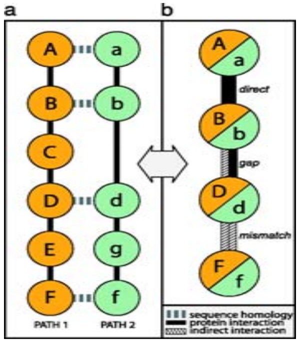
KEGG id: 04310hsa



Match: Calcium Signaling

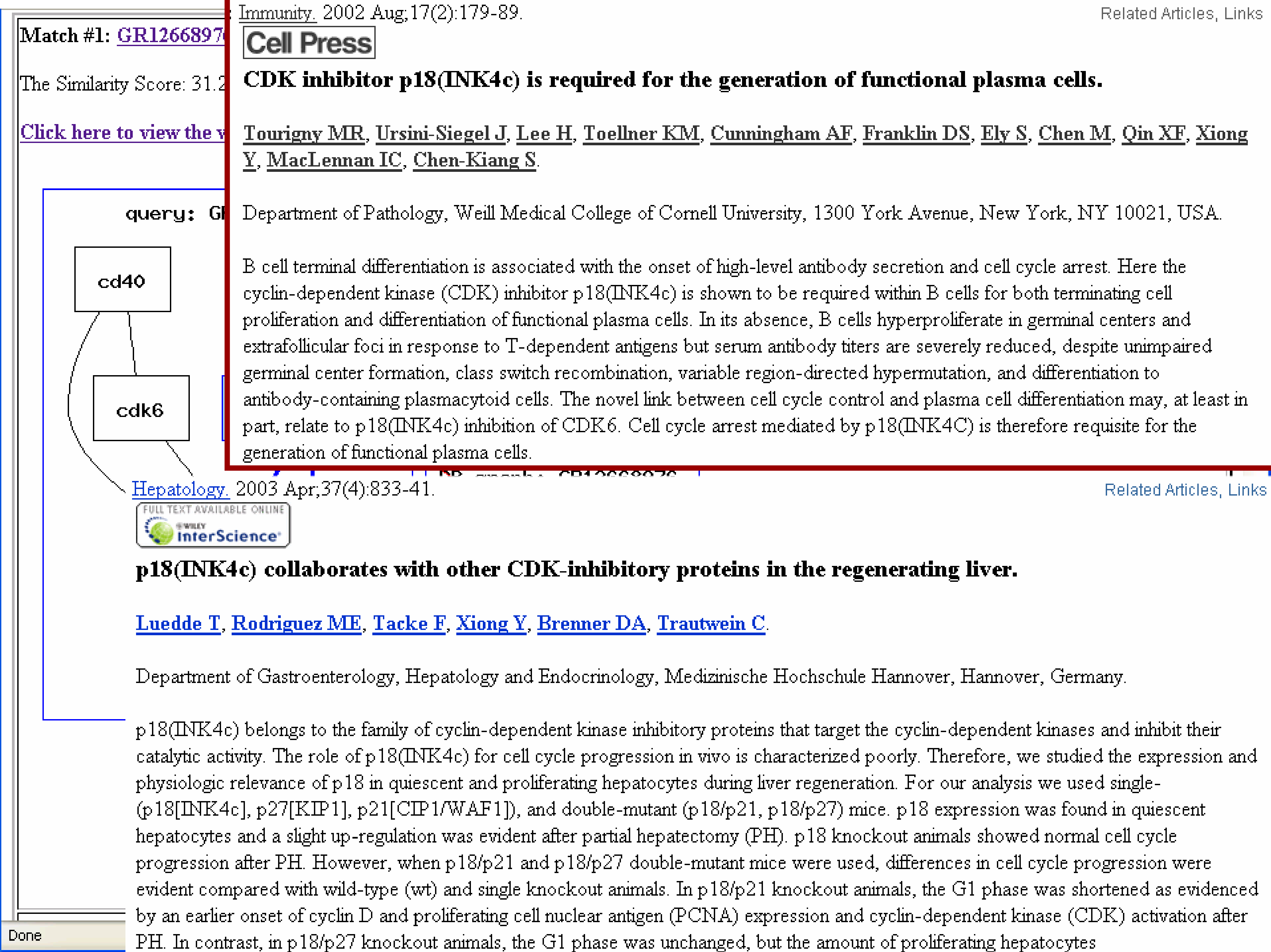
KEGG id: 04020hsa

- **None of the existing methods can detect this match!**
- Limitations of Existing Methods:
  - Gindex & GraphGrep: only perform exact matching
  - Grafil & PIS: no gap nodes are allowed
  - PathBlast: only matches paths; edge alignment only tolerates one gap nodes, e.g. (B,D) with (b,d) and (D, F) with (d, f)



Kelley et. al. PNAS(2003)

PathBlast Example



Match #1: GR1266897

The Similarity Score: 31.2

Click here to view the v

Immunity. 2002 Aug;17(2):179-89.

Related Articles, Links

Cell Press

CDK inhibitor p18(INK4c) is required for the generation of functional plasma cells.

Tourigny MR, Ursini-Siegel J, Lee H, Toellner KM, Cunningham AF, Franklin DS, Ely S, Chen M, Qin XF, Xiong Y, MacLennan IC, Chen-Kiang S.

Department of Pathology, Weill Medical College of Cornell University, 1300 York Avenue, New York, NY 10021, USA.

B cell terminal differentiation is associated with the onset of high-level antibody secretion and cell cycle arrest. Here the cyclin-dependent kinase (CDK) inhibitor p18(INK4c) is shown to be required within B cells for both terminating cell proliferation and differentiation of functional plasma cells. In its absence, B cells hyperproliferate in germinal centers and extrafollicular foci in response to T-dependent antigens but serum antibody titers are severely reduced, despite unimpaired germinal center formation, class switch recombination, variable region-directed hypermutation, and differentiation to antibody-containing plasmacytoid cells. The novel link between cell cycle control and plasma cell differentiation may, at least in part, relate to p18(INK4c) inhibition of CDK6. Cell cycle arrest mediated by p18(INK4C) is therefore requisite for the generation of functional plasma cells.

Hepatology. 2003 Apr;37(4):833-41.

Related Articles, Links

FULL TEXT AVAILABLE ONLINE WILEY InterScience

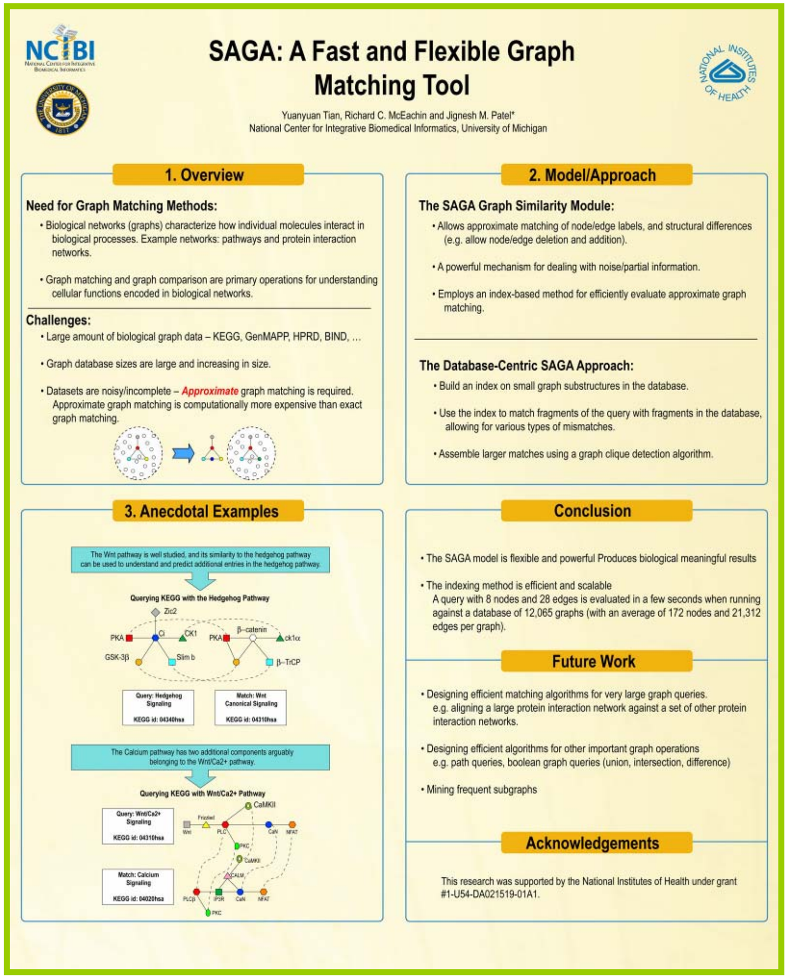
p18(INK4c) collaborates with other CDK-inhibitory proteins in the regenerating liver.

Luedde T, Rodriguez ME, Tacke F, Xiong Y, Brenner DA, Trautwein C.

Department of Gastroenterology, Hepatology and Endocrinology, Medizinische Hochschule Hannover, Hannover, Germany.

p18(INK4c) belongs to the family of cyclin-dependent kinase inhibitory proteins that target the cyclin-dependent kinases and inhibit their catalytic activity. The role of p18(INK4c) for cell cycle progression in vivo is characterized poorly. Therefore, we studied the expression and physiologic relevance of p18 in quiescent and proliferating hepatocytes during liver regeneration. For our analysis we used single- (p18[INK4c], p27[KIP1], p21[CIP1/WAF1]), and double-mutant (p18/p21, p18/p27) mice. p18 expression was found in quiescent hepatocytes and a slight up-regulation was evident after partial hepatectomy (PH). p18 knockout animals showed normal cell cycle progression after PH. However, when p18/p21 and p18/p27 double-mutant mice were used, differences in cell cycle progression were evident compared with wild-type (wt) and single knockout animals. In p18/p21 knockout animals, the G1 phase was shortened as evidenced by an earlier onset of cyclin D and proliferating cell nuclear antigen (PCNA) expression and cyclin-dependent kinase (CDK) activation after PH. In contrast, in p18/p27 knockout animals, the G1 phase was unchanged, but the amount of proliferating hepatocytes

SAGA Demo and Poster



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 for 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

- Designing efficient matching algorithms for very large graph queries.
  e.g. aligning a large protein interaction network against a set of other protein interaction networks.
- Designing 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.