Studies in mining conserved subgraph among metabolic pathways

SIAM Data Mining

2008

Qiong Cheng*(cscqxcx@cs.gsu.edu), **Robert Harrison, Alexander Zelikovsky (**alexz@cs.gsu.edu**)**
Department of Computer Science, Georgia State University, Atlanta, GA 30303

*** Partially supported by GSU Molecular Basis of Disease (MBD) and Brains & Behavior (B&B)**

Abstract

Graph is a popular theoretic formalism for modeling multidimensional data such as metabolic pathways. Mining conserved subgraph frequently includes comparison of unknown networks with well-understood networks accounting for the similarity of the vertices as well as their corresponding topological structures.

Traditional graph comparisons are classified as two types of problems: subgraph isomophism and subgraph homeomorphism. Both problems are known to be NP-complete. However different application domains on graph comparison require different problem formulations and the corresponding solutions.

Based on the specific application in metabolic pathways, we focused on mapping of multi-source tree onto multi-source tree, onto DAG(directed acyclic graph) and further mapping among arbitrary graphs. Arising from innate biochemical reaction observation property that several enzymes can catalyze the same reaction and a single enzyme can catalyze several reactions, we formulated the problem to be homeomorphism embedding and allowed the occurrence of different vertices in query graph to be mapped to a vertex in well-known target graph.

Our experimental study finds significant conserved pairs in pairwise mapping of all pathways for four organisms (E.coli, S.cerevisiae, B. subtilis and T. thermophilus species).

Homeomorphism embedding of metabolic pathways

➢ **Metabolic pathway model** – a directed graph in which vertices correspond to enzymes and there is a directed edge between two enzymes if the product of the reaction catalyzed by the first enzyme is a substrate of the reaction catalyzed by the second.

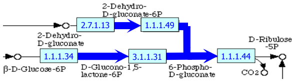
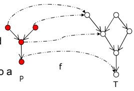


Fig 1. A portion of pentose phosphate pathway

➢ **Mapping metabolic pathways** - should capture the similarities of enzymes represented by proteins as well as topological properties.

➢ **Graph Homeomorphism embedding**



Let pattern P =<V_P, E_P> and text T =<V_T, E_T> be directed graphs, a homomorphism mapping f : P→G satisfies

❖ $f_v : V_P \rightarrow V_T$: Different pattern vertices can be mapped to a single text vertex;

❖ $f_e : E_P \rightarrow$paths of T : An edge in the pattern can be mapped to a path in the text;

We allow different enzymes to be mapped to the same enzyme.

➢ **Homeomorphism embedding cost = vertex-to-vertex cost + edge-to-path cost**

Edge-to-path cost = l ($|f_e(e)|-1$) , which proportionally increase with the number of extra hops the images of edge
Homomorphism cost = $\Sigma_{v \text{ in } V_P} \Delta(v, f_v(v)) + \lambda\, \Sigma_{e \text{ in } E_P} (|f_e(e)|-1)$

Problem Formulation

Given: an arbitrary graph pattern P =<V_P, E_P> and an arbitrary graph text T =<V_T,E_T>
Find: min cost homomorphism f : P → T

minimize $\Sigma_{v \text{ in } V_P} \Delta(v, f_v(v)) +\lambda\, \Sigma_{e \text{ in } E_P} (|f_e(e)|-1)$
s.t. $\forall v$ in V_P, $\exists f_v(v)$ in V_T
$\forall e$ in E_p , $\exists f_e(e)$ in Path_T

Solution

➢ **Optimal homeomorphism embedding algorithm**

```
find minimum feedback vertex set MFVS of P
construct a multi source tree P' = <$V_p$- MFVS, E($V_p$- MFVS)>
for every possible fixed mapping $f'_v$: F(P) →$V_T$ do
    obtain min cost homomorphism of P' to T under mapping $f'_v$
choose min cost homomorphism for all possible fixed mappings
```

➢ **Minimum Feedback vertex set problem**

Given: an undirected graph G=(V,E) and a nonnegative weight function w on V
Find: a minimum weight subset of V whose removal leaves an acyclic graph.

Greedy algorithm for obtaining Minimum Feedback vertex set MFVS

```
delete degree 1/0 vertices from V and set remaining vertices to V'
MFVS← $\phi$
while V' $\neq \phi$ do
    pick up the set S of maximal degree vertices
    MFVS ← MFVS U S
    Delete degree 1/0 vertices from V'
```

➢ **Best homeomorphism embedding of multisource tree to arbitrary graph**

❖ Preprocessing of text graph T by transitive closure

❖Ordering of constructed multisource tree P' :

1. DFS traversal of P'
2. Processing order in opposite way with the DFS traversal

Key: Each edge e_i in P' is the unique edge connecting v_i with the previous vertices in the order

❖ Dynamic programming

❖ Build DP table - DP = $V_{P'} \times V_{T^*}$

Each row and column corresponds to a vertex of $V_{P'}$ and V_{T^*} respectively;
The columns $u_1, \ldots, u_{|V_{T^*}|}$ are sorted in arbitrary order;
The rows $v_1, \ldots, v_{|V_{P'}|}$ are sorted in the special order
Every item is the min cost homomorphism from P' subgraph induced by previous vertices in the order into T^*

❖ Filling DP table by recursive function

$$\underset{\substack{i<|V_{P'}|\\ j<|V_T|}}{DT[i, j]} = \begin{cases} \Delta(v_i , u_j) & \text{if } v_i \text{ is a leaf in T} \\ \Delta(v_i, u_j) + \Sigma_{l=1 \text{ to } Adj(vi)} \text{Min}_{j'=1 \text{ to } |V_T|} C(i_l, j') & \text{if } v_i \text{ is a leaf in T} \end{cases}$$

$C[i_l, j_l] = DT[i_l, j_l] + \lambda(h(j, j_l) - 1)$
$\lambda$is penalty for gaps
$h(j, j_l) = \#(\text{hops between } u_j \text{ and } u_{j_l} \text{ in T})$

❖ Runtime

Transitive closure takes $O(|V_T||E_T|)$

Pattern graph ordering takes $O(|V_P| + |E_P|)$

Dynamic programming :

- Calculate min contribution of all child pairs of node pair ($v_i \in P, u_j \in T^*$) takes $t_{ij} = deg_P (v_i) deg_{G^*}(u_j)$
- Filling DT takes $\Sigma_{j=1 \text{ to } |V_G|} \Sigma_{i=1 \text{ to } |V_T|} t_{ij} = \Sigma_{j=1 \text{ to } |V_G|} deg_{G^*}(u_j) \Sigma_{i=1 \text{ to } |V_T|} deg_T(v_i) = 2|E_{G^*}||E_T|$The total runtime is $O(|V_G||E_G|+|V_{G^*}||V_T|)$.

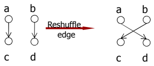
➢ **Runtime for obtaining optimal homeomorphism embedding**

$O(|V_T|^{(1+|MFVS|)} (|V_G||E_G|+|V_{G^*}||V_T|))$

Let t(v) denote the number of “reasonable” text images of v, total runtime = $O(\prod_{v \in MFVS} t(v) (|VG||EG|+|VG^*||VT|))$

Previous work

➢ Mapping : Linear pattern→Graph (Kelly et al 2004) ($o(|V_P|^{\iota+2}|V_T|^2)$)
➢ Mapping : Tree → Tree (Pinter [2005] $o(|V_T|^2|V_P|/\log|V_T|+|V_T||V_P|\log|V_P|)$)
➢ Exhaustively search : Sharan et al 2005 ($o(\iota!|V_P|^{\iota+2}|V_T|^2)$), Yang et al 2007 ($o(2^{|V_T|}|V_T|^2)$)



Experiments results

➢ **Computing P-Value of homomorphism mapping**

❖ Random degree-conserved graph generation by :

Reshuffle nodes

Reshuffle edges

❖ Randomized P-Value computation (P-Value cutoff : 0.01 for 100 randomized graphs)

➢ All-against-all mappings among 4 species :

➢ Identifying conserved pathways

24 pathways that are conserved across all 4 species

18 more pathways that are conserved across at least three of these species

➢Resolving ambiguity (see figure 2)

➢Discovering pathways holes (see figure 3)

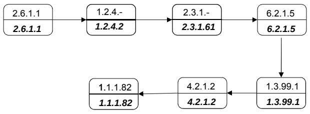


Fig 2. Resolving ambiguity example: Mapping of glutamate degradation VII pathways from B.subtilis to T. thermophilus (p<0.01). The enlighted node reflects enzyme homology.

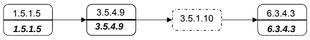


Fig 3. Pathway holes' example: Mapping of formaldehyde oxidation V pathway in B. subtilis to formy1THF biosynthesis pathway in E. coli (p<0.01) (only vertices in the image of the pattern in the text are shown.

References :

➢ Q. Cheng, D. Kaur, R. Harrison, and A. Zelikovsky,"Mapping and Filling Metabolic Pathways ", *RECOMB Satellite Conference on Systems Biology 2007*

➢ Q. Cheng, R. Harrison, and A. Zelikovsky,"Homomorphisms of Multisource Trees into Networks with Applications to Metabolic Pathways", *Proc. of IEEE 7-th International Symposium on BioInformatics and BioEngineering* (BIBE'07)

➢ Ron Y Pinter, Oleg Rokhlenko, Esti Yeger-Lotem, Michal Ziv-Ukelson: Alignment of metabolic pathways. *Bioinformatics. LNCS 3109.* Springer-Verlag.(Aug 2005)21(16): 3401-8

➢ …..

Georgia Tech Poster/Reception session at SIAM Data Mining (2008)