Mapping and Filling Metabolic Pathways

BINF 2007

Qiong Cheng*, Dipendra Kaur, 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

Network mappings are extensively used for comparing, exploring, and predicting biological networks, are essential for pathway database search. We have applied homomorphisms to mapping metabolic pathways using EC notation based enzyme matching cost. We have performed pairwise mapping of all pathways for four organisms (E.coli,S.cerevisiae, B.subtilis and T.thermophilus species) and found a set of statistically significant pathway similarities. Our experiments show that organism-to-organism pathway mapping identifies pathway holes simultaneously providing possible candidates for filling them. We have proposed a framework of detecting and filling pathway holes that includes database search for missing enzymes and proteins with the matching prosites and high sequence similarity. We have identified conserved pathways across examined species and detected and filled pathway holes in existing pathway descriptions.

Homomorphism 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.

➢ **Enzyme mapping cost Δ**

❖ EC notation - a 4-level hierarchical scheme - a sub-sub-subclass indication of biochemical reaction

❖ Measure Δ [x, y] by the lowest common upper class distribution : $\Delta[X, Y] = \log_2 c(X, Y)$

❖ Measure Δ [x, y] by tight reaction property :

Given Enzyme $X = x_1 \cdot x_2 \cdot x_3 \cdot x_4$ and Enzyme $Y = y_1 \cdot y_2 \cdot y_3 \cdot y_4$, we have
If $(x_1 = y_1$ and $x_2 = y_2$ and $x_3 = y_3$ and $x_4 \neq y_4)$ $\Delta[X, Y] = 1$;
If $(x_1 = y_1$ and $x_2 = y_2$ and $x_3 \neq y_3)$ $\Delta[X, Y] = 10$;
Otherwise $\Delta[X, Y] = +\infty$

❖ Our experimental study indicates that the measurement by tight reaction property results in biochemically more relevant pathway matches

➢ **Graph Homomorphism**

Let pattern $P = \langle V_P, E_P \rangle$ and text $T = \langle V_T, E_T \rangle$ be directed graphs, a homomorphism mapping $f : P \rightarrow G$ satisfies

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

❖ $f_p : E_P \rightarrow$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.

➢ **Homomorphism mapping cost = vertex-to-vertex cost + edge-to-path cost**

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

Problem Formulation

Given: an arbitrary graph pattern $P = \langle V_P, E_P \rangle$ and an arbitrary graph text $T = \langle V_T, E_T \rangle$
Find: min cost homomorphism $f : P \rightarrow T$

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

Previous work

➢ Mapping : Linear pattern→Graph (Kelly et al 2004) ($o((V_P)^{k-2}|V_T|^2)$)
➢ 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 ($o(t!)\ o(|V_P|^{k-1}|V_T|^2)$); Yang et al 2007 ($o(2^{3k}|V_T|^3)$)

Solution

➢ **Optimal homomorphism mapping algorithm**

1. find minimum feedback vertex set MFVS of P
2. construct a multi source tree $P' = \langle V_P, MFVS, E(V_P - MFVS) \rangle$
3. for every possible fixed mapping $f'_v : F(P') \rightarrow V_T$ do
4. obtain min cost homomorphism of P' to T under mapping f'_v
5. choose min cost homomorphism for all possible fixed mappings

➢ **Minimum Feedback vertex set problem**

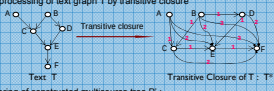
Given: an undirected graph G=(V,E) and a nonnegative weight function w on V
Find: a minimum weight subset of V whose removal leaves an acyclic graph.

Greedy algorithm for obtaining MFVS

1. delete degree 1/0 vertices from V and set remaining vertices to V'
2. MFVS←∅
3. while V' ≠ ∅ do
4. pick up the set S of maximal degree vertices
5. MFVS ← MFVS ∪ S
6. Delete degree 1/0 vertices from V'

➢ **Best homomorphism 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

$$DT[i, j] = \begin{cases} \Delta(v_i, u_j) & \text{if } v_i \text{ is a leaf in T} \\ \Delta(v_i, u_j) + \Sigma_{l=1 \text{ to } \deg(v_i)} \min_{k=1 \text{ to } |V_T|} C(l, k) & \text{if } v_i \text{ is a leaf in T} \end{cases}$$

$$C[i_1, j_2] = DT[i_1, j_2] + \lambda(h(j_1, j_2) - 1)$$

λ is penalty for gaps
$h(j_1, j_2)$ = # (hops between u_{j1} and u_{j2} 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 $(v_P, u_T \in T')$ takes $t_v = \deg_{v_P}(v_P)\deg_{u_T}(u_T)$
- Filling DT takes $\Sigma_{v_P \text{ in } V_P} \Sigma_{u_T \text{ in } V_T} t_v = \Sigma_{v_P \text{ in } V_P} \deg_{v_P}(v_P) \Sigma_{u_T \text{ in } V_T} \deg_{u_T}(u_T) = 2|E_P||E_T|$ The total

runtime is $O(|V_P||E_T| + |V_P||V_T|)$.

➢ **Runtime for obtaining optimal homomorphism mapping**

Total number of all possible fixed mappings of MFVS = $O(|V_T|^{(1+|MFVS|)})$
Total runtime for repeating multisource tree to graph mapping procedure with all possible fixed mappings of MFVS = $O(|V_T|^{(1+|MFVS|)}(|V_G||E_G|+|V_G|\log|V_G|))$

Let t(v) denote the number of "reasonable" text images of v,
Total runtime = $O(\Pi_{v \in MFVS} t(v)\ (|VG||EG|+|VG''||VT|))$

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 (see table 1):

➢ 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)

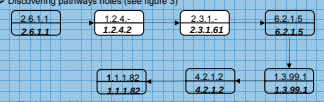


Fig 2. Resolving ambiguity example: Mapping of glutamate degradation VII pathways from B.subtilis to T. thermophilus (p<0.01). The enlighted node reflects enzyme homology.

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.

➢ Filling pathways holes

Three types:

❖Fillings found by same EC number : EC number is in pattern and text organisms

❖ Fillings found by group neighbors

❖ Fillings found by Left/Right Neighbor

Type	Example			
	Pattern pathway (P)	Text pathway (T)	Mapping of interest	Similarity Score & Acc Num
Fillings found by same EC number	Gamma glutamyl cycle	Superpathway of glycolysis pyruvate dehydrogenase TCA and glyoxylate bypass	P: 2.3.2.2; -.-.-.-; 2.3.2.4 T: 2.3.3.9, 1.1.1.37; 2.3.3.1	051 P49814
Fillings found by group neighbors	Alanine biosynthesis I	Superpathway of lysine threonine methionine and S-adenosyl-L-methionine biosynthesis	P: 2.6.1.66;-.-.-.-; -.-.-.- -.-.-.-; -.-.-.-; -.-.-.- -.-.-.-; -.-.-.-; -.-.-.- 5.1.1.1 T: 2.6.1.1; 2.7.2.4; 1.2.1.11; 4.2.1.52; 1.3.1.26; 2.3.1.117; 2.6.1.17; 3.5.1.18; 5.1.1.7	0.8 P39754 EC 2.6.1.16 With statistical significance 0.26
Fillings found by Left/Right Neighbor	Alanine biosynthesis I	Superpathway of lysine threonine methionine and S-adenosyl-L-methionine biosynthesis	P: 2.6.1.66;-.-.-.-; -.-.-.- -.-.-.-; -.-.-.-; -.-.-.- -.-.-.-; -.-.-.-; -.-.-.- 5.1.1.1 T: 2.6.1.1; 2.7.2.4; 1.2.1.11; 4.2.1.52; 1.3.1.26; 2.3.1.117; 2.6.1.17; 3.5.1.18; 5.1.1.7	0.7 P10725 EC 5.1.1.1

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

➢ ....