Homomorphisms of Multisource Trees into Networks with Applications to Metabolic Pathways

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

Abstract

Network mapping is a convenient tool for comparing and exploring biological networks; it can be used for predicting unknown pathways, fast and meaningful searching of databases, and potentially establishing evolutionary relations. Unfortunately, existing tools for mapping paths into general networks (PathBlast) or trees into tree networks allowing gaps (MetaPathwayHunter) cannot handle large query pathways or complex networks.

In this paper we consider homomorphisms, i.e., mappings allowing to map different enzymes from the query pathway into the same enzyme from the networks. Homomorphisms are more general than homeomorphism (allowing gaps) and easier to handle algorithmically. Our dynamic programming algorithm efficiently finds the minimum cost homomorphism from a multisource tree to directed acyclic graphs as well as general networks.

We have performed pairwise mapping of all pathways for four organisms (E. coli, S. cerevisiae, B. subtilis and T. thermophilus species) and found a reasonably large set of statistically significant pathway similarities. Further analysis of our mappings identifies conserved pathways across examined species and indicates potential pathway holes in existing pathway descriptions.

Availability: The software is available from the authors on request.

1. Introduction

The explosive growth of cellular network databases requires novel analytical methods constituting a new interdisciplinary area of computational systems biology. The main problems in this area are finding conserved subnetworks, integrating interacting gene networks, protein networks and biochemical reactions, discovering critical elements or modules and finding homologous pathways. With the immense increase of good-quality data from high-throughput genomic and proteomic technologies, studies of these questions are more and more challenging from analytical and computational perspectives.

Network mapping is a convenient tool for comparing and exploring biological networks. When mapping metabolic pathways by matching similar enzymes and chemical reactions chains we can match homologous pathways. Network mapping can be used for predicting unknown pathways, fast and meaningful searching of databases, and potentially establishing evolutionary relations.

Let the *pattern* be a pathway for which we are searching for homologous pathways in the *text*, i.e., the known metabolic network of a different species (see Figure 1). This problem includes the Isomorphic Embedding problem, therefore it is NP-hard (see [8]). Given a linear length-ℓ pathway as the pattern and a graph as the text, PathBlast (see [9, 8, 14]) finds the image of the pattern in the text such that no consecutive mismatches or gaps on the pattern and the text are allowed. The path-to-path mapping algorithm builds a global alignment graph and decomposes it to linear pathways mapping.

A single enzyme in one pathway may replace a few sequential enzymes in homologous pathway and vise versa. MetaPathwayHunter [12, 13] finds the optimal homeomorphic tree-to-tree mapping allowing an arbitrary number of gaps. In contrast to the previous approaches, we allow for the mapping of different enzymes from the pattern into the same enzyme from the text while keeping the freedom to map a single edge from the pattern to a path in the text. Such mappings (homomorphisms) are more general than homeomorphisms and easier to handle algorithmically.

Our contributions include: (1) efficient dynamic programming based algorithm and its implementation finding the minimum cost homomorphism from multisource trees into arbitrary networks, (2) a new protein similarity score scheme based on 4-digit EC enzyme hierarchy, (3) experimental pairwise comparison of all pathways in four different organisms (*E. coli*, *S. cerevisiae*, *B. subtilis* and *T. ther-*

*Correspondence author.

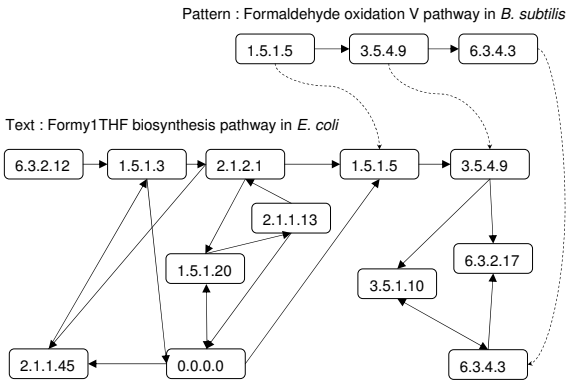

Figure 1. An example of network mapping to find an image of pattern in text.

mophilus) resulting in a reasonably large set of statistically significant pathway similarities, (4) identification of pathways conserved across examined species, potential holes in existing pathway descriptions, and an estimation of the evolutionary relationship between examined species.

The remainder of the paper is organized as follows. The next section describes previous work. Section 3 presents the proposed models for mapping with the definition of protein similarity score scheme. Section 4 introduces necessary definitions and graph-theoretical problem formulation. Section 5 presents our dynamic programming algorithm and analyzes its runtime. Section 6 describes our computational study of metabolic pathways of four organisms. The analysis and validation of experimental study is given in Section 7. Finally, we draw conclusions in Section 8.

2 Previous Work

The earlier papers comparing pathways did not take into account their topology and instead focused only on similarity between proteins or genes. Biochemical pathway similarity was defined in terms of sequence similarity of involved genes (see [5]); a multiple pair-wise comparison algorithm utilizing 4-digit EC enzyme hierarchy was proposed in [15]. The alignment of linear pathways was reduced to the sequence alignment problem in [2].

A series of papers [9, 8, 14] has taken into account the nonlinearity of protein network topology and formulated the mapping problem as follows. Given a linear length-ℓ pathway *pattern* $T = (VT, ET)$ and *text* graph $G = (VG, EG)$, find an image of the pattern in the text without consecutive gaps and minimizing mismatches between proteins. A global alignment graph in [9] was built in which each vertex represents a pair of proteins and each edge represents a conserved interaction, gap, or mismatches; their objective is to find the k-highest-scoring path with limited length ℓ and no consecutive gaps or mismatches based on the built global graph. The approach takes $O(|VT|^{\ell+2}|VG|^2))$.

PathBlast with the same problem formulation as [9] was presented in [8]. However, PathBlast's solution is to randomly decompose the text graph into linear pathways which are then aligned against the pattern and then to obtain optimal mapping based on standard sequence alignment algorithms. The algorithm requires $O(\ell!)$ random decompositions to ensure that no significant alignment is missed, effectively limiting the size of the query to about six vertices.

In [13], metabolic pathways were modeled as outgoing trees; the problem was reduced to the approximately labeled tree homeomorphism problem. The proposed bottom-up dynamic programming algorithm has runtime $O(m^2 n/\log m + mn \log n)$ where m and n are the number of vertices in pattern and text, respectively.

In [11] the problem was formulated as an integer quadratic problem to obtain the global similarity score based on the mapping of as many as node-to-node similarity and as many as edge-to-edge similarity. An exhaustive searching approach was employed in [17] to find the vertex-to-vertex and path-to-path mappings with the maximal mapping score under the condition of limited length of gaps or mismatch. The algorithm has worst case time complexity $O(2^m \times m^2)$. A label-diversity back-track algorithm was proposed in [16] to align two networks with cycles based on the mapping of as many as path-to-path similarity.

Finally, a more general approach to aligning of metabolic pathways – homomorphisms allowing edge-to-path mapping – has been first proposed in [3].

3 Modeling Metabolic Pathway Mappings

A metabolic pathway is a series of chemical reactions catalyzed by enzymes that occur within a cell. Metabolic pathways are represented by directed networks in which vertices correspond to enzymes and there is a directed edge from one enzyme to another if the product of the reaction catalyzed by the first enzyme is a substrate of the reaction catalyzed by the second.

Mapping metabolic pathways should capture the similarities of enzymes represented by proteins as well as topological properties that cannot be always reduced to sequential reactions represented by paths. Below we first describe our approach to measure enzyme similarity and then discuss advantages and drawbacks of the homomorphism mappings of metabolic networks.

Our implementation provides two alternative enzyme similarity scores. One approach is to employ the lowest common upper class distribution proposed in [15] and discussed in [13]. The corresponding penalty score for gap is

2.0.

Our new approach makes full use of EC encoding and the tight reaction property classified by EC. The EC number is expressed with a 4-level hierarchical scheme. The 4-digit EC number, $d_1.d_2.d_3.d_4$ represents a sub-sub-subclass indication of biochemical reaction. If $d_1.d_2$ of two enzymes are different, their similarity score is infinite; if d_3 of two enzymes are different, their similarity score is 10; if d_4 of two enzymes are different, their similarity score is 1; or else the similarity score is 0. The corresponding penalty score for gap is 0.5. Our experimental study indicates that the proposed similarity score scheme results in biochemically more relevant pathway matches.

The topology of most metabolic pathways is a simple path, but frequently pathways may branch or have several incoming arcs – all such topologies are instances of a *multisource tree*, i.e., a directed graph which becomes an undirected tree when edge directions are disregarded. The query pathways are usually simple and can be represented as a multisource tree but in some cases they can have a cycle or alternative ways to reach the same vertex. Then we suggest to follow the standard practice of breaking such cycle or paths by removing edges.

The obvious way to preserve the pathway topology is to use isomorphic embedding – one-to-one correspondence between vertices and edges of the pattern and its image in the text. The requirement on edges can be relaxed – an edge in the pattern can be mapped to a path in the text ([12, 13]) and the corresponding mapping is called a *homeomorphism*. The computational drawback of isomorphic embedding and homeomorphism is that the problem of finding optimal mapping is NP-complete and, therefore, requires severe constraints on the topology of the text to become efficient. In [12, 13], the text is supposed to be a tree, the pattern should a directed tree while allowing multisource-tree pattern complicates the algorithm. Their algorithm is complex and slow because it repeatedly finds minimum weight perfect matchings.

In this paper we propose to additionally relax one-to-one correspondence between vertices – instead we allow different pattern vertices to be mapped to a single text vertex. The corresponding mapping is called a *homomorphism*. Such relaxation may sometimes cause confusion – a path can be mapped to a cycle. For instance, if two enzymes with similar functions belong to the same path in the pattern and a cycle with similar enzyme belongs to the text, then the path can be mapped into a cycle (see Figure 2). However, if the text graph is acyclic this cannot happen. Even if there are cycles in the text, still one can expect that functionally similar enzymes are very rare in the same path.

Computing minimum cost homomorphisms is much simpler and faster than homeomorphisms. We will show that a fast dynamic programming algorithm can find the minimum cost homomorphism that allows edge-to-path mapping for the multisource tree pattern and an arbitrary text graph.

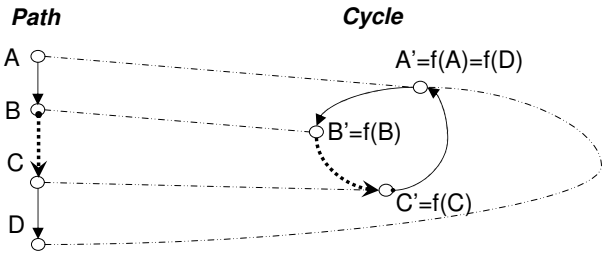

Figure 2. Homomorphism of a path $(A, B, \ldots, C, D)$ in the pattern onto a cycle $(A', B', \ldots, C', D' = A')$ in the text.

4 Graph-Theoretical Problem Formulation

We first give notations and definitions and then formulate the corresponding graph-theoretical problem.

A *pattern* $T = (VT, ET)$ is a directed graph with vertex set VT and edge set ET. We only consider the case of T being a *multisource tree*. Following [13], a multi-source tree is a directed graph, whose underlying undirected graph is a tree. It is not necessarily a directed tree – each node can have several incoming as well as several outgoing edges.

A *text* $G = (VG, EG)$ is a directed graph with vertex set VG and edge set EG. We further distinguish the case when G is a general network and the case when G is a directed acyclic graph.

A mapping $f : T \to G$ from pattern $T = (VT, ET)$ to text $G = (VG, EG)$ is called a *homomorphism* if

(1) every vertex in VT is mapped to a vertex in VG ;

(2) every edge $e = (u, v) \in ET$ is mapped to a directed path $f(e) = (u_0 = f(u), u_1, u_2, ..., u_k = f(v))$ in G.

We will now introduce the cost of a homomorphism. Let $\Delta(u, v)$, $u \in VT$, $v \in VG$, be the cost of mapping an enzyme corresponding to the pattern vertex u into an enzyme corresponding to the text vertex v.

Following [13], the rule (2) allows edge-to-path mapping, but edge-to-edge mapping is still preferable. Therefore, the homomorphism cost should increase proportionally to the number of extra hops in the images of edges, i.e.,

$$\sum_{e \in ET} (|f(e)| - 1)$$

where $|f(e)| = k$ is the number of hops in the path $f(e) = (u_0 = f(u), u_1, u_2, ..., u_k = f(v))$.

Following [13], the cost of a homomorphism $f : T \to G$ takes in account cost of enzyme mapping and edge-to-path mapping as follows

$$cost(f) = \sum_{v \in VT} \Delta(v, f(v)) + \lambda \sum_{e \in ET} (|f(e)| - 1)$$

where λ is the cost of a single extra hop in an edge-to-path mapping. This parameter balances enzyme mapping and edge-to-path costs. If $\lambda = 0$, then only the enzyme mapping cost is taken into account. If λ is very large, then the enzyme mapping cost contribution is negligible. In our computational experiments we use $\lambda = 0.5$.

Finally, the graph-theoretical problem formulation is as follows.

Minimum Cost Homomorphism Problem. Given a multi-source pattern tree T and a text graph G, find the minimum cost homomorphism $f : T \to G$.

5 Dynamic Programming Algorithm

We will first describe preprocessing of the text graph G and ordering of vertices of the pattern graph T. Then we define the dynamic programming table and show how to fill that table in a bottom-up manner. We conclude with the runtime analysis of the entire algorithm.

Text Graph Preprocessing. In order to compute the cost of a homomorphism it is necessary to know the number of hops for any shortest path in the text graph G. Although finding single-source shortest paths in general graphs is slow, in our case it is sufficient to run breadth-first-search with runtime $O(|EG| + |VG|)$. Assuming that G is connected, i.e., $|EG| \geq |VG|$, we conclude that the total runtime of finding all shortest paths is $O(|VG||EG|)$. In the resulting transitive closure $G' = (VG, EG')$ of the graph G, each edge $e \in EG'$ is supplied with the number of hops $h(e)$ in the shortest path connecting its ends.

Pattern Graph Ordering. We will further need a certain fixed order of vertices in VT as follows. Let $T' = (VT, ET')$ be the undirected tree obtained from T by disregarding edge directions. Let us choose an arbitrary vertex $r \in VT$ as a root and run depth-first search (DFS) in T' from r. Let $\{r = v_1, \ldots, v_{|VT|}\}$ be the order of the DFS traversal of VT and let $e'_i = (v_i, v) \in ET'$ (corresponding to directed edge $e_i \in ET$) be the unique edge connecting v_i to the set $\{v_1, \ldots, v_{i-1}\}$. The vertex $v \in \{v_1, \ldots, v_{i-1}\}$ is called a *parent* of v_i and v_i is called a *child* of v.

DP Table. Now we will describe our dynamic programming table $DT[1, \ldots, |VT|][1, \ldots, |VG|]$. Each row and column of this table corresponds to a vertex of T and G, respectively. While the columns $u_1, \ldots, u_{|VG|}$ of DT are in no particular order, the rows $\{r = v_1, \ldots, v_{|VT|}\}$ of DT are sorted according to the DFS traversal of T'. Each element $DT[i, j]$ is equal to the best cost of a homomorphism from the subgraph of T induced[1] by vertices $\{v_{|VT|}, v_{|VT|-1}, \ldots, v_i\}$ into G' which maps v_i into u_j.

Filling DP Table. The table DT is filled bottom-up for $i = |VT|, |VT| - 1, \ldots, 1$ as follows. If v_i is not a parent for any vertex in T, then v_i is a leaf and

$$DT[i, j] = \Delta(v_i, u_j)$$

In general, let v_i be a parent for the vertices $v_{i_1}, \ldots, v_{i_k}$. In order to compute $DT[i, j]$, we should find the cheapest mapping of each of the children $v_{i_1}, \ldots, v_{i_k}$ subject to v_i being mapped to u_j. The mappings of the children do not depend on each other since the only connection between them in the tree T is through v_i. Therefore, each child v_{i_l}, $l = 1, \ldots, k$ should be mapped into u_{j_l} minimizing the contribution of v_{i_l} to the total cost

$$C[i_l, j_l] = DT[i_l, j_l] + \lambda(h(j, j_l) - 1)$$

where $h(j, j_l)$ depends on direction of e_{i_l}, i.e., $h(j, j_l) = h(u_j, u_{j_l})$ if $e_{i_l} = (v_i, v_{i_l})$ and $h(j, j_l) = h(u_{j_l}, u_j)$ if $e_{i_l} = (v_{i_l}, v_i)$. Finally,

$$DT[i, j] = \Delta(v_i, u_j) + \sum_{l=1}^{k} \min_{j'=1,\ldots,|VG|} C[i_l, j']$$

Runtime Analysis. As we mentioned earlier, the runtime for constructing the transitive closure $G' = (VG, EG')$ is $O(|VG||EG|)$. The runtime to fill a cell $DT[i, j]$ is proportional to

$$t_{ij} = deg_T(v_i) deg_{G'}(u_j)$$

where $deg_T(v_i)$ and $deg_{G'}(u_j)$ are degrees of v_i and u_j in graphs T and G', respectively. Indeed, the number of children of v_i is $deg_T(v_i) - 1$ and for each child v_{i_l} of v_i there are at most $deg_{G'}(u_j)$ feasible positions in G' since $f(v_i)$ and $f(v_{i_l})$ should be adjacent. The runtime to fill the entire table DT is proportional to

$$\sum_{j=1}^{|VG|} \sum_{i=1}^{|VT|} t_{ij} = \sum_{j=1}^{|VG|} deg_{G'}(u_j) \sum_{i=1}^{|VT|} deg_T(v_i) = 2|EG'||ET|$$

Thus the total runtime is $O(|VG||EG| + |EG'||VT|)$. Even though G is sparse, $|EG'|$ may be as large as $O(|VG|^2)$, i.e., the runtime is $O(|VG|(|EG| + |VG||VT|))$.

6 Mapping Metabolic Pathways

In this section we first describe the metabolic pathway data, then explain how we measure statistical significance

[1] A subgraph *induced* by the subset of vertices S includes only edges that have both ends in S.

pattern network (tree pathways)		text network (number of pthways)			
		T. thermophilus(208)	*B. subtilis*(226)	*E. coli*(113)	*S. cerevisiae*(151)
T. thermophilus	# of mapping pairs	38	21	18	18
	# of mapped pattern pathways	28	14	12	13
	# of mapped text pathways	35	20	18	17
B. subtilis	# of mapping pairs	162	217	121	58
	# of mapped pattern pathways	80	143	85	39
	# of mapped text pathways	106	153	92	40
E. coli	# of mapping pairs	9	5	38	3
	# of mapped pattern pathways	2	3	3	2
	# of mapped text pathways	9	5	14	5
S. cerevisiae	# of mapping pairs	24	12	12	14
	# of mapped pattern pathways	9	7	6	13
	# of mapped text pathways	21	12	12	14

Table 1. Pairwise statistical homomorphisms among *T. thermophilus*, *B. subtilis*, *E. coli* and *S. cerevisiae*.

of homomorphisms and report the results of pairwise mappings between four species.

Data. The genome-scale metabolic network data in our studies were drawn from BioCyc [1, 7, 10], the collection of 260 Pathway/Genome Databases, each of which describes metabolic pathways and enzymes of a single organism. We have chosen metabolic networks of *E. coli*, the yeast *S. cerevisiae*, the eubacterium *B. subtilis* and the archeabacterium *T. thermophilus* so that they cover major lineages Archaea, Eukaryotes, and Eubacteria. The bacterium *E. coli* with 113 pathways is the most extensively studied prokaryotic organism. *T. thermophilus* with 208 pathways belongs to Archaea. *B. subtilis* with 226 pathways is one of the best understood Eubacteria in terms of molecular biology and cell biology. *S. cerevisiae* with 151 pathways is the most thoroughly researched eukaryotic microorganism.

Statistical Significance of Mapping. Although the cost of a homomorphism reflects the similarity of pathways, it alone cannot assure us that such cost is not obtained by chance. Only statistically significant cost values can be taken in account. Statistical significance is measured by p-value, i.e., the probability of the null hypothesis that the cost value is obtained by pure chance. Following a standard randomization procedure, we randomly permute pairs of edges (u, v) and (u', v') if no other edges exist between these 4 vertices u, u', v, v' in the text graph by reconnecting them as (u, v') and (u', v). This allows us to keep the incoming and outgoing degree of each vertex intact. We find the minimum cost homomorphism from the pattern graph into the fully randomization of the text graph and check if its cost is at least as big as the minimum cost before randomization of the text graph. We say that the homomorphism is statistically significant with $p < 0.01$ if we found at most 9 better costs in 1000 randomization of the text graph.

Experiments. For each pair of four species (*B. subtilis*, *E. coli*, *T. thermophilus* and *S. cerevisiae*), using our algorithm we find the best homomorphism from each pathway of one species to each pathway of the other and check if this homomorphism is statistically significant, i.e., if $p < 0.01$. We have run our experiments on a Pentium 4 processor, 2.99 GHz clock with 1.00 GB RAM. The total runtime was 1.5h for the input/output of pathways and computing the optimal pattern-to-text mapping and its p-value for every pair of pathways (there are in total 516052 pattern-text pathway pairs).

Results. The results of our experiments are reported in Table 1. The first column contains the name of the species from whose metabolic network the pattern pathways have been chosen. Note that if a pathway is not a multisource tree or degenerate (i.e., has less than 3 nodes), then it is omitted. We did not omit any pathway from the text species since our algorithm supports any network as a text. For every species-to-species mapping, we compute the number of mapped pairs with $p < 0.01$, the number of the pattern pathways that have at least one statistically significant homomorphic image and the number of the text pathways that have at least one statistically significant homomorphic pre-image.

For example, for homomorphism from *T. thermophilus* to *B. subtilis*, there are in total 21 statistically significant mapped pairs, 14 non-degenerate tree *T. thermophilus* pathways have statistically significant homomorphic images in *B. subtilis* and 20 out of 226 *B. subtilis* pathways have statistically significant homomorphic pre-images.

7 Implications of Pathway Mappings

In this section we identify pathways conserved across multiple species, show how one can resolve enzyme ambiguity and identify potential holes in pathways, and phylogenetically validate our pathway mappings.

Identifying Conserved Pathways. We first have identified the pathways that are conserved across all 4 species under consideration. Table 2 contains a list of all 20 pathways in

Pathway name
alanine biosynthesis I
biotin biosynthesis I
coenzyme A biosynthesis
fatty acid beta
fatty acid elongation saturated
formaldehyde oxidation V (tetrahydrofolate pathway)
glyceraldehyde 3 phosphate degradation
histidine biosynthesis I
homoserine biosynthesis
lysine biosynthesis I
ornithine biosynthesis
phenylalanine biosynthesis I
phenylalanine biosynthesis II
polyisoprenoid biosynthesis
proline biosynthesis I
quinate degradation
serine biosynthesis
superpathway of gluconate degradation
tyrosine biosynthesis I
UDP galactose biosynthesis
alanine biosynthesis
biotin biosynthesis
fatty acid oxidation pathway
fructoselysine and psicoselysine degradation

Table 2. The list of all 20 pathways in *B. subtilis* that have statistically significant homomorphic images simultaneously in all 3 other species *E. coli*, *T. thermophilus* and *S. cerevisiae*. The lower part contains 4 more different pathways with statistically significant images in all 4 species.

Pathway name
triple: *B. subtilis*, *E. coli*, and *T. thermophilus*
4 aminobutyrate degradation I
de novo biosynthesis of pyrimidine deoxyribonucleotides
de novo biosynthesis of pyrimidine ribonucleotides
enterobacterial common antigen biosynthesis
phospholipid biosynthesis I
PRPP biosynthesis II
salvage pathways of pyrimidine deoxyribonucleotides
ubiquinone biosynthesis
flavin biosynthesis
glycogen biosynthesis I (from ADP D Glucose)
L idonate degradation
lipoate biosynthesis and incorporation I
menaquinone biosynthesis
NAD biosynthesis I (from aspartate)
triple: *B. subtilis*, *E. coli*, and *S. cerevisiae*
oxidative branch of the pentose phosphate pathway
S adenosylmethionine biosynthesis
triple: *B. subtilis*, *T. thermophilus*, and *S. cerevisiae*
tyrosine biosynthesis I
fatty acid elongation unsaturated I

Table 3. The list of 14 pathways conserved across *B. subtilis*, *E. coli*, and *T. thermophilus*; 2 more pathways conserved across *B. subtilis*, *E. coli*, and *S. cerevisiae*; 2 more pathways conserved across *B. subtilis*, *T. thermophilus*, and *S. cerevisiae*.

B. subtilis that have statistically significant homomorphic images simultaneously in all species. The lower part of Table 2 contains 4 more pathways with different names in *E. coli*, *T. thermophilus* and *S. cerevisiae*, which have simultaneous statistically significant images in all species.

Besides 24 pathways conserved across all 4 species we have also found 18 pathways only common for triples of these species. Table 3 gives the pathway names for each possible triple of species (the triple *E. coli*, *T. thermophilus* and *S. cerevisiae* does not have extra conserved pathways).

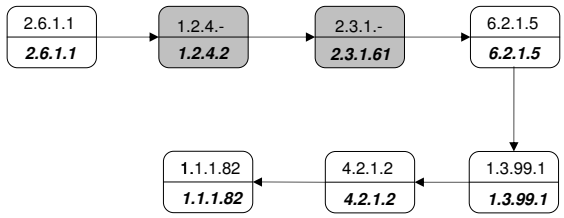

Figure 3. Mapping of glutamate degradation VII pathways from *B. subtilis* to *T. thermophilus* ($p < 0.01$). The node with upper part and lower part represents a vertex-to-vertex mapping. The upper part represents the query enzyme and the lower part represents the text enzyme. The shaded node reflects enzyme homology.

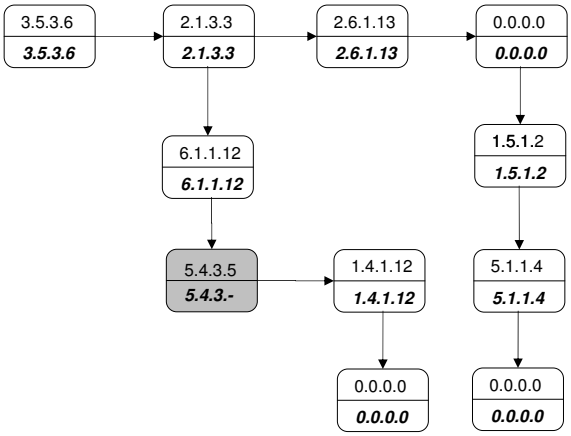

Figure 4. Mapping of interconversion of arginine, ornithine and proline pathway from *T. thermophilus* to *B. subtilis* ($p < 0.01$). The node with upper part and lower part represents a vertex-to-vertex mapping. The upper part represents the query enzyme and the lower part represents the text enzyme. The shaded node reflects enzyme homology.

Resolving Ambiguity. Currently multiple pathways contain unresolved enzymes. Completely unresolved enzymes have EC notation 0.0.0.0/-.-.-.- and partially unresolved enzymes have less "-"'s, e.g., EC notation 1.2.4.-. We can use our mapping tool to suggest possible resolution of these ambiguities as follows.

Let us consider two examples of homomorphism – the mapping of glutamate degradation VII pathway in *B. subtilis* to glutamate degradation VII pathway in *T. thermophilus* (shown in Figure 3), and the mapping of interconversion of arginine, ornithine and proline pathway in *T. thermophilus* to interconversion of arginine, ornithine and proline pathway in *B. subtilis* (shown in Figure 4). When some enzymes in pathway are labeled with the end of ".-", it denotes their exact reactions are not explicit. The mapping results indicate that a potential similar enzyme with similar functions of the unclear enzyme can be found in some other species.

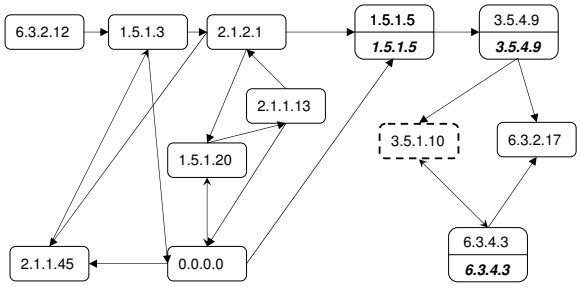

Figure 5. Mapping of formaldehyde oxidation V pathway in *B. subtilis* to formy1THF biosynthesis pathway in *E. coli* ($p < 0.01$). The node with upper part and lower part represents a vertex-to-vertex mapping. The upper part represents the query enzyme and the lower part represents the text enzyme. The node with dashed box represents a gap.

Holes in Pathways. Pathway holes happen when a genome appears to lack the enzymes needed to catalyze reactions in a pathway [6]. We can use our mapping tool to identify potential pathway holes as shown in the following example (see Figure 5).

There is a statistically significant mapping from formaldehyde oxidation V (tetrahydrofolate pathway) in *B. subtilis* to formyITHF biosynthesis in *E. coli*. The correspondence between enzymes shows a gap – enzyme 3.5.1.10 is missing. We found that the enzyme 3.5.1.10 exists in *B. subtilis* according to Enzyme and Swiss-prot databases.

Phylogenetic Validation. One can measure similarity between species based on the number of conserved pathways. The largest amount of conserved pathways is found between *B. subtilis* and *T. thermophilus* – two species-to-species mappings have in total 183 statistically significant pairs of pathways. The next closest two species are *E. coli* and *B. subtilis* which have 126 statistically significant pairs of

pathways. This agrees with fact that *B. subtilis*, *T. thermophilus*, and *E. coli* are prokaryote and *S. cerevisiae* is a eucaryote.

8 Conclusions

In this paper we have introduced a new method of mapping metabolic pathways based on homomorphism. The proposed mapping approach allows to map different enzymes of the pattern pathway into a single enzyme of a text network. We have also define a novel scoring scheme for computing similarity between enzymes based on their EC notation.

We have formulated the graph-theoretical problem corresponding to finding optimal homomorphism from the pattern network to a text network. We give an efficient dynamic-programming method for exactly solving this problem when the pattern is multisource tree an the text is an arbitrary network.

We have applied our mapping tool pairwise mapping of all pathways for four organisms (*E. coli*, *S. cerevisiae*, *B. subtilis* and *T. thermophilus* species) representing main different lineages and found a reasonably large set of statistically significant pathway similarities.

We report 24 pathways that are conserved across all 4 species as well 18 more pathways that are conserved across at least three of these species. We show that our mapping tool can be used for identification of potential pathway holes as well resolving enzyme notation ambiguities in existing pathway descriptions.

Acknowledgments

We thank Professor Pinter and Oleg Rokhlenko for giving access to their software package MetaPathwayHunter, Professor Karp and Dr. Kaipa for providing the pathway databases for *T. thermophilus*, *B. subtilis*, *E. coli* and *S. cerevisiae*. We also thank Amit Sabnis, Dipendra Kaur, Kelly Westbrooks for helpful discussions.

References

[1] http://www.biocyc.org/.

[2] M. Chen and R. Hofest. An algorithm for linear metabolic pathway alignment. *In silico biology (In silico biol.) ISSN, 1386-6338: 111-128*, 2005.

[3] Q. Cheng and A. Zelikovsky. Optimal mapping of multi source trees into dag in biological network. *ISBRA'07Poster*, May 2007.

[4] T. Dandekar, S. Schuster, B. Snel, M. Huynen, and P. Bork. Pathway alignment: application to the comparative analysis of glycolytic enzymes. *Biochem. J.*, 1: 115-124, 1999.

[5] C. V. Forst and K. Schulten. Evolution of metabolism: a new method for the comparison of metabolic pathways using genomics information. *J. Comput. Biol.*, 6: 343-360, 1999.

[6] M. L. Green and P. D. Karp. A bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics*, Sep. 2004.

[7] I. M. Keeler, V. J. Collard, C. S. Gama, J. Ingrafts, S. Palely, I. T. Paulson, M. Peralta-Gil, and P. D. Karp. Ecocyc: a comprehensive database resource for escherichia coli. *Nucleic Acids Research, 33(1):D334-337*, 2006.

[8] B. P. Kelly, R. Sharan, R. M. Karp, T. Sittler, D. E. Root, and B. R. Stockwell. Pathblast: a tool for alignment of protein interaction networks. *Nucleic Acids Research*, Vol.32 : W83-W88, 2004.

[9] B. P. Kelly, R. Sharan, R. M. Karp, T. Sittler, D. E. Root, and B. R. Stockwell. Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *PNAS*, 11394-11399, Sep. 30 2003.

[10] C. J. Krieger, P. Zhang, L. A. Mueller, A. Wang, S. Paley, M. Arnaud, J. Pick, S. Rheme, and P. Karp. Metacyc: a microorganism database of metabolic pathways and enzymes. *Nucleic Acids Research, 32(1):D438-42*, 2006.

[11] Z. Li, Y. Wang, S. Zhang, X.-S. Zhang, and L. Chen. Alignment of protein interaction networks by integer quadratic programming. *EMBS '06. 28th Annual International Conference of the IEEE*, 5527-5530, Aug. 2006.

[12] R. Pinter, O. Rokhlenko, D. Tsur, and M. Ziv-Ukelson. Approximate labeled subtree homeomorphism. *In Proceedings of 15th Annual Symposium of Combinatorial Pattern Matching*.

[13] R. Y. Pinter, O. Rokhlenko, E. Yeger-Lotem, and M. Ziv-Ukelson. Alignment of metabolic pathways. *Bioinformatics*.

[14] R. Sharan, S. Suthram, R. M. Kelley, T. Kuhn, S. McCuine, P. Uetz, T. Sittler, R. M. Karp, and T. Ideker. Conserved patterns of protein interaction in multiple species. *PNAS*, Vol.102 : 1974-1979, 2005.

[15] Y. Tohsato, H. Matsuda, and A. Hashimoto. A multiple alignment algorithm for metabolic pathway analysis using enzyme hierarchy. *Proc. 8th International Conference on Intelligent Systems for Molecular Biology*, 376-383, ISMB 2000.

[16] S. Wernicke. Combinatorial algorithms to cope with the complexity of biological networks. *Dissertation*, December 2006.

[17] Q. Yang and S.-H. Sze. Path matching and graph matching in biological networks. *Journal of Computational Biology*, Vol. 14, No. 1: 56-67 : 5527-5530, 2007.