SubMAP: Aligning metabolic pathways with subnetwork mappings

Ferhat Ay and Tamer Kahveci

Computer and Information Science and Engineering,
University of Florida, Gainesville, FL 32611
fay@cise.ufl.edu,tamer@cise.ufl.edu

Abstract. We consider the problem of aligning two metabolic pathways. Unlike traditional approaches, we do not restrict the alignment to one-to-one mappings between the molecules of the input pathways. We follow the observation that in nature different organisms can perform the same or similar functions through different sets of reactions and molecules. The number and the topology of the molecules in these alternative sets often vary from one organism to another. In other words, given two metabolic pathways of arbitrary topology, we would like to find a mapping that maximizes the similarity between the molecule subsets of query pathways of size at most a given integer k. We transform this problem into an eigenvalue problem. The solution to this eigenvalue problem produces alternative mappings in the form of a weighted bipartite graph. We then convert this graph to a vertex weighted graph. The maximum weight independent subset of this new graph is the alignment that maximizes the alignment score while ensuring consistency. We call our algorithm **SubMAP** (**Sub**network **M**appings in **A**lignment of **P**athways). We evaluate its accuracy and performance on real datasets. Our experiments demonstrate that SubMAP can identify biologically relevant mappings that are missed by traditional alignment methods and it is scalable for real size metabolic pathways.
Availability: Our software and source code in C++ is available at http://bioinformatics.cise.ufl.edu/SubMAP.html

1 Introduction

Biological pathways show how different molecules interact with each other to perform vital functions. In the literature, the terms "network" and "pathway" are used interchangeably for different types of interaction data. Metabolic pathways, an important class of biological pathways, represent how different compounds are transformed through various reactions. Analyzing these pathways is essential in understanding the machinery of living organisms.

The efforts on analyzing pathways can be classified into two types. The first type takes one pathway into account at a time and explores the important properties of that network such as its robustness [1], steady states [2] and modular structure [3]. The second type is the comparative approach which considers

multiple pathways to identify their frequent subgraphs [4, 5] and their alignments [6–11]. Alignment is a fundamental type of comparative analysis which aims to identify similar parts between pathways. For metabolic pathways, these similarities provide insights for drug target identification [12, 13], metabolic reconstruction of newly sequenced genome [14], phylogenic reconstruction [15, 16] and enzyme cluster and missing enzyme identification [17, 18].

In the literature, alignment is often considered as finding one-to-one mappings of the molecules of two pathways. In this case, the global/local pathway alignment problems are GI/NP complete as the graph/subgraph isomorphism problems can be reduced to them in polynomial time [19]. A number of studies have been done to systematically align different types of biological networks. For metabolic pathways, Pinter *et al.* [6] devised an algorithm that aligns query pathways with specific topologies by using a graph theoretic approach. Tohsato *et al.* proposed two algorithms one relying solely on Enzyme Commission (EC [20]) numbers of enzymes and the other considering only the chemical structures of compounds of the query pathways [9, 10]. Latterly, Cheng *et al.* developed a tool, *MetNetAligner*, for metabolic pathway alignment that allows a certain number of insertions and deletions of enzymes [11]. However, these methods do not integrate different types of information (e.g., topology, homology) and focus on a single similarity score (e.g., enzyme similarity, compound similarity, etc.). Furthermore, some of these methods limit the query pathways to certain topologies, such as trees, non-branching paths or limited cycles, which degrades their applicability to complex pathways. Recently, Singh *et al.* [21] and Ay *et al.* [7, 8] combined both topological features and homological similarity of pairwise molecules to find the alignments of protein interaction networks and metabolic pathways respectively. These two algorithms showed that this integration increases the accuracy of alignment. Additionally, these methods do not restrict the topologies of query pathways and hence are applicable to arbitrarily complex pathways.

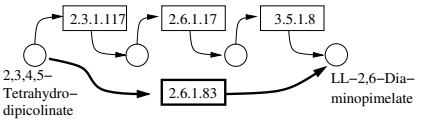


Fig. 1. A portion of Lysine biosynthesis pathway. Each reaction is represented by the Enzyme Commission (EC) number of the enzyme that catalyze it. Each circle represents a compound. Humans use the the path on the top with three reactions, whereas plants and Chlamydia can achieve this transformation directly by a single reaction through the path (**in bold**) at the bottom.

All the methods discussed above limit the possible molecule mappings to only one-to-one mappings. As also pointed out by Deutscher *et al.* [22] considering each molecule one by one fails to reveal its function(s) in complex pathways. This restriction prevents all the above methods from identifying biologically relevant mappings when different organisms perform the same function through varying number of steps. As an example, there are alternative paths for LL-2,6-Diaminopimelate production in different organisms [13, 23]. Figure 1 illustrates two paths both producing LL-2,6-Diaminopimelate starting from

2,3,4,5-Tetrahydrodipicolinate. The bottom path represents the shortcut used by Chlamydia and plants on the path to the synthesis of an important amino acid, L-Lysine. This shortcut is not available for humans since we lack LL-DAP aminotransferase (2.6.1.83). Humans use a three step process shown as the top path in Figure 1 to do this transformation. Thus, a meaningful alignment should match the top path with three reactions to the bottom with a single reaction when the human lysine biosynthesis pathway is aligned to the same pathway of a plant or Chlamydia. However, since these two paths have different number of reactions, traditional alignment methods fail to identify this mapping.

Our aim in this paper is to design an algorithm that can accurately identify such biologically relevant mappings by allowing one-to-many mappings of molecules. Note that, in Figure 1 the topologies of both reaction sets are linear paths. It is possible to have reaction sets with arbitrary topologies. Therefore, we use the term *subnetwork* to include all types of topologies. Also, since we only consider the sets of reactions that are connected, we will simply use the term *subnetwork* instead of *connected subnetwork*.

Problem definition: Here, we consider the problem of aligning two metabolic pathways. Unlike traditional alignment approach, we allow aligning a molecule of one pathway to a connected subnetwork of the other. More formally, let $\mathcal{P}$ and $\bar{\mathcal{P}}$ be two query pathways and k be a positive integer. We want to find the mapping between the molecules of $\mathcal{P}$ and $\bar{\mathcal{P}}$ with the largest alignment score, such that (1) each molecule in $\mathcal{P}$ ($\bar{\mathcal{P}}$) can map to a subnetwork of $\bar{\mathcal{P}}$ ($\mathcal{P}$) with at most k molecules and (2) each molecule can appear in at most one mapping.

The first condition above allows one-to-many mappings. The second condition enforces *consistency*. That is, if a molecule is already mapped alone or as a part of a subnetwork, it cannot map to another molecule. We elaborate on consistency and the problem definition later in Section 2. *Note that, allowing one-to-many mappings in alignment introduces new computational challenges that cannot be addressed using existing methods and hence novel methods are needed to tackle this problem.*

Contributions: In this paper, we propose a novel algorithm named **SubMAP** that finds subnetwork mappings in alignment of pathways. SubMAP accounts for both the effect of pairwise similarities (homology) and the organization of pathways (topology). This combination is motivated by its successful applications on pathway alignment by Singh *et al.* [21] and Ay *et al.* [7, 8]. However, allowing one-to-many mappings makes it impossible to trivially extend these methods to our problem. To address this challenge, we map our problem to an eigenvalue problem. We solve this eigenvalue problem using an iterative technique called power method. The result of the power method converges to a principal eigenvector. This eigenvector defines a weighted bipartite graph where each node corresponds to a molecule or a subnetwork. The edges are only between two nodes from different pathways and their weights define the similarity of these nodes. Unlike the problem with only one-to-one molecule mappings, the resulting nodes of the bipartite graph can be intersecting as the same molecule can appear in more than one subnetwork. We term such node pairs as *conflicting*. In order to ensure

that the alignment is consistent, we construct a vertex weighted *conflict graph* with nodes representing a mapping of two subnetworks one from each pathway and edges representing a *conflict* between two mappings (i.e., they create inconsistency). The similarity values in the principal eigenvector are the weights of the nodes in conflict graph. Our algorithm aims to find the set of mappings (nodes) that has no conflicts (edges) and maximizes the total weight of nodes. This problem is equivalent to finding *maximum weight independent subset*. Since the maximum weight independent set problem is NP-hard, we use a heuristic to extract an independent set from the conflict graph which gives us a non-conflicting set of one-to-many mappings. We report these mappings as the *alignment* of the query pathways. *Our experiments on the metabolic pathways from KEGG [24] database suggest that SubMAP finds biologically meaningful alignments efficiently. Also, SubMAP is scalable as it aligns pathways with around 50 reactions while allowing subnetworks of size three in less than a minute.*

The rest of the paper is organized as follows. Section 2 describes our algorithm. Section 3 presents experimental results. Section 4 concludes the paper.

2 Our Algorithm: SubMAP

In this section, we present our algorithm for pairwise metabolic pathway alignment that allows one-to-many molecule mappings. We begin by introducing some notation that we use throughout this section. Then, we formally state the problem and describe the SubMAP algorithm in detail.

Let, $\mathcal{P}$ be a pathway which is represented by a directed unweighed graph $G = (V, E)$. Here, we only use the reactions of the pathway in graph representation. Hence, the vertex set $V = \{r_1, r_2, \ldots, r_n\}$ is the set of all reactions of $\mathcal{P}$. We include a directed edge e_{ij} from r_i to r_j in E if and only if at least one output compound of r_i is an input compound of r_j. We call r_i a *backward neighbor* of r_j and r_j a *forward neighbor* of r_i if $e_{ij} \in E$. Note that reactions can be reversible (bi-directional) and hence both e_{ij} and e_{ji} can exist.

A *subnetwork* of a pathway is a subset of its reaction set such that the induced undirected graph of the elements of this subset forms a connected graph. Let $R_i \subseteq V$ be such a subnetwork of $\mathcal{P}$. We define $\mathcal{R}_k$ as $\mathcal{R}_k = \{R_1, R_2, \ldots, R_N\}$ where $|R_i| \leq k$ for all $i \in [1, N]$. Here, $|R_i|$ denotes the cardinality of the reaction set R_i. Verbally, $\mathcal{R}_k$ is the set of all subnetworks of $\mathcal{P}$ that have at most k reactions. Using this notation, we define a binary relation that maps a reaction of a query pathway to a subnetwork of the other as follows:

Definition 1 *Let $\mathcal{P}$ and $\bar{\mathcal{P}}$ be two pathways and k be a positive integer. Also, let $\mathcal{R}_k = \{R_1, R_2, \ldots, R_N\}$ and $\bar{\mathcal{R}}_k = \{\bar{R}_1, \bar{R}_2, \ldots, \bar{R}_M\}$ be the sets of subnetworks with size at most k of $\mathcal{P}$ and $\bar{\mathcal{P}}$. We define a binary relation between $\mathcal{R}_k$ and $\bar{\mathcal{R}}_k$ that allows one-to-many reaction mappings as $\varphi : \varphi \subseteq (\mathcal{R}_1 \times \bar{\mathcal{R}}_k) \cup (\mathcal{R}_k \times \bar{\mathcal{R}}_1)$.*

Clearly, φ allows one-to-one and one-to-many mappings. The cardinality of φ ($|\varphi|$) is at most $nM + mN - nm$ where n, m are the number of reactions of $\mathcal{P}$ and $\bar{\mathcal{P}}$ respectively. In other words, the number of all possible mappings is

$nM + mN - nm$. The *alignment* of $\mathcal{P}$ and $\bar{\mathcal{P}}$ is a binary relation that is a subset of all these possible mappings and satisfies certain criteria that we describe next.

Recall that for a mapping $(R_i, \bar{R}_j) \in \varphi$ one of the R_i or $\bar{R}_j$ can contain more than one reaction. Reporting this mapping as a part of our alignment implies that all the reactions of the subnetwork with multiple reactions are aligned to a single reaction of the other. To have a *consistent alignment* none of the reactions of these subnetworks can be included in any other mapping. Next, we formally define the term *conflict* to characterize this property.

Definition 2 *Let φ be a binary relation and $R_i, R_u \in \mathcal{R}_k$ and $\bar{R}_j, \bar{R}_v \in \bar{\mathcal{R}}_k$. The distinct pairs $(R_i, \bar{R}_j) \in \varphi$ and $(R_u, \bar{R}_v) \in \varphi$ **conflict** if and only if $(R_i \cap R_u) \cup (\bar{R}_j \cap \bar{R}_v) \neq \varnothing$.*

Conflicts can cause inconsistencies about which reaction subset of one pathway should be aligned to the one of the other pathway. If φ has a conflicting pair of elements, we say φ is *inconsistent*. Since this is not a desirable property, *we limit our alignment to the consistent relations only.*

In order to find biologically relevant alignments we also need a meaningful scoring scheme. One standard scoring scheme for this purpose incorporates the homology of the aligned molecules with their topologies [7, 8, 21]. Here, we generalize this scheme to one-to-many mappings. We will elaborate on this similarity score later in Section 2.4. Next, we state our problem formally.

Problem formulation: Given k and two pathways $\mathcal{P}$ and $\bar{\mathcal{P}}$, let $\mathcal{R}_k$ and $\bar{\mathcal{R}}_k$ be the sets of subnetworks with size at most k of $\mathcal{P}$ and $\bar{\mathcal{P}}$ respectively. We want to find the *consistent* binary relation $\varphi \subseteq (\mathcal{R}_1 \times \bar{\mathcal{R}}_k) \cup (\mathcal{R}_k \times \bar{\mathcal{R}}_1)$ that *maximizes* the summation of the similarity scores of the aligned subnetworks.

In the following, we present our algorithm SubMAP. Section 2.1 explains how we enumerate the subnetworks of query pathways. Section 2.2 and 2.3 discuss homological and topological similarities respectively. Section 2.4 describes the eigenvalue formulation and extraction of the alignment.

2.1 Enumeration of Connected Subnetworks

The first step of SubMAP is to create the sets of all connected subnetworks of size at most k for each query pathway. Here, we describe the enumeration process for a single query pathway. Let $G = (V, E)$ represent a pathway and k be a positive integer. We construct the set of subnetworks $\mathcal{R}_k$ as follows. For $k = 1$ $\mathcal{R}_k = \mathcal{R}_1 = V$. For $k > 1$ we define $\mathcal{R}_k$ recursively by using $\mathcal{R}_{k-1}$. At each recursive step we check for each reaction in V if it can be added to already enumerated subnetworks of size $k-1$ to create a new connected subnetwork of size k. This way the kth recursive step takes $O(|V|.(|\mathcal{R}_{k-1}| - |\mathcal{R}_{k-2}|))$ time.

The size of the set $\mathcal{R}_k$ can be exponential in k when G is dense. However, metabolic pathways are usually sparse (on the average there are 2.5 forward neighbors per reaction). We observe that the number of subnetworks of real metabolic pathways for $k = 3$ is around $5|V|$ and for $k = 4$ it is $10|V|$ on the average. In Section 3.2, we provide a detailed discussion of how $|\mathcal{R}_k|$ changes with different pathway sizes and different k values.

2.2 Homological Similarity of Subnetworks

Recall that the relation φ maps a reaction to a subnetwork that can contain multiple reactions. This necessitates computing the similarity between reaction sets. Since reactions are defined by their input and output compounds (i.e., substrates and products) and the enzymes that catalyze them, we measure the homological similarity between reactions using the similarities of these components.

In the literature, there are alternative pairwise similarity scores for compounds, enzymes and reactions. Particularly, two well known measure are information content similarity for enzyme pairs [6] and SIMCOMP [25] for compound pairs. We denote these measures by $SimE$ and $SimC$ respectively. We defer the readers to Ay *et al.* [8] for details on computing these similarities. Here, we utilize these similarity measures to compute the homological similarity between two reaction sets. To calculate this, we first construct the sets of the unions of input compounds (I_i), output compounds (O_i) and enzymes (E_i) of the reactions in each subnetwork R_i. For instance, in Figure 1 if we take upper path as the subnetwork R_i, then $E_i = \{2.3.1.117, 2.6.1.17, 3.5.1.8\}$. Let $\gamma_e, \gamma_i, \gamma_o$ denote the relative weights of the similarities of enzymes, input compounds and output compounds respectively. We define $SimRSet$ as:

$$SimRSet(R_i, \bar{R}_j) = \gamma_e W(E_i, \bar{E}_j, SimE) + \gamma_i W(I_i, \bar{I}_j, SimC) + \gamma_o W(O_i, \bar{O}_j, SimC)$$

Here W denotes the sum of edge weights of the pairs returned by the maximum weight bipartite matching (MWBM) of the two sets. MWBM finds an assignment between the nodes of two sets such that maximizes the sum of the weights of these assignments specified by the similarity score. We use $\gamma_i = \gamma_o = 0.3$ and $\gamma_e = 0.4$ as they provide a good balance between enzymes and compounds.

We calculate $SimRSet$ for all possible one-to-many mappings between the subnetworks of two pathways. The number of possible pairings is $nM+mN-nm$ where n, m are the number of reactions of $\mathcal{P}$, $\bar{\mathcal{P}}$ and $N = |\mathcal{R}_k|$ and $M = |\bar{\mathcal{R}}_k|$. Therefore, in this step, we calculate $SimRSet$ function $nM + mN - nm$ times. This way, we assess the homological similarities between all possible subnetwork mappings. Even though this scoring is a good measure of similarity, relying solely on this score ignores the topology similarity which we explain next.

2.3 Topological Similarity of Subnetworks

The motivation for utilizing topological similarity is that the induced topologies of two aligned subnetworks should also be similar. In other words, if R_i is mapped to $\bar{R}_j$, then their neighbors in the corresponding pathways should also be similar. Motivated by this, we first extend the neighborhood definition of reactions to reaction subnetworks. Then, we introduce the notion of *support* between two mappings.

Definition 3 *Let R_i, $R_u \in \mathcal{R}_k$. Then, R_u is a **forward neighbor** of R_i ($R_u \in FN(R_i)$) if and only if there exists $r_a \in R_i$ and $r_b \in R_u$ such that r_b is a forward neighbor of r_a or $R_i \cap R_u \neq \varnothing$. R_i is a **backward neighbor** of R_u ($R_i \in BN(R_u)$) if and only if R_u is a forward neighbor of R_i.*

Definition 4 *Let $R_i, R_u \in \mathcal{R}_k$ and $\bar{R}_j, \bar{R}_v \in \bar{\mathcal{R}}_k$. The mapping ($R_i$, $\bar{R}_j$) **supports** the mapping (R_u, $\bar{R}_v$) if and only if both $R_j \in FN(R_i)$ and $\bar{R}_v \in FN(\bar{R}_u)$ or both $R_j \in BN(R_i)$ and $\bar{R}_v \in BN(\bar{R}_u)$.*

Definition 4 states that the mapping of R_i to $\bar{R}_j$ favors all possible mappings of forward (backward) neighbors of R_i to those of $\bar{R}_j$. For instance, if $FN(R_i) = 2$, $FN(\bar{R}_j) = 2$, $BN(R_i) = 1$ and $BN(\bar{R}_j) = 2$, then the mapping $(R_i, \bar{R}_j)$ supports $2 \times 2 + 1 \times 2 = 6$ mappings. We distribute the support of $(R_i, \bar{R}_j)$ equally to these six mappings. There can be cases when one mapping does not provide support to any others. In such cases, we simply distribute its support equally to all possible mappings ($nM + mN - nm$). Conceptually, we consider the support of each mapping (R_i, $\bar{R}_j$) on the other mappings as a matrix. We call it *the support matrix* (S) since it stores the topological support between different mappings. Notice that we are setting the entries of S in a way that for each mapping the sum of the relative weights of its support is 1. In other words, the sum of all the entries in each column of S is one. This ensures the stability and convergence of our algorithm as we explain in Section 2.4. Interested reader can find detailed description of the support matrix in a previous work of ours [8].

Trivial but costly way of creating S matrix is to check each mapping against all the others to calculate the support values. However, such an exhaustive strategy will require computing a huge matrix S of size $(nM + mN - nm) \times (nM + mN - nm)$. Since the creation of S will incur prohibitive computational costs, we do not construct this matrix literally. Instead, for each mapping $(R_i, \bar{R}_j)$, we take the sets $FN(R_i)$, $FN(\bar{R}_j)$ and $BN(R_i)$, $BN(\bar{R}_j)$ to generate only the pairs supported by $(R_i, \bar{R}_j)$. In other words, we use the sparse matrix form of the support matrix S.

2.4 Aligning Two Pathways

Both the homological similarities of subnetworks and their topological organization provide us significant information for the alignment of metabolic pathways. A good alignment algorithm needs to combine these two factors in an efficient and accurate way. Here, we describe how we achieve this combination in SubMAP by using an iterative technique called *power method*.

Let k be a given parameter and $\mathcal{P}$, $\bar{\mathcal{P}}$ be two pathways with connected subnetwork sets $\mathcal{R}_k = \{R_1, R_2, \ldots, R_N\}$ and $\bar{\mathcal{R}}_k = \{\bar{R}_1, \bar{R}_2, \ldots, \bar{R}_M\}$ respectively. We represent the homological similarity of all subnetwork pairs by the column vector $\overrightarrow{H}$ of size $nM + mN - nm$, where n, m are the number of reactions of $\mathcal{P}$, $\bar{\mathcal{P}}$ respectively and $N = |\mathcal{R}_k|$, $M = |\bar{\mathcal{R}}_k|$. Each entry of $\overrightarrow{H}$ denotes the homological similarity between two subnetworks one from each pathway which corresponds to a mapping.

Let S be the $(nM + mN - nm) \times (nM + mN - nm)$ support matrix as described in Section 2.3. Given a parameter $\alpha \in [0, 1]$ to adjust the relative weights of homology and topology, we combine homology and topology through power method iterations as follows:

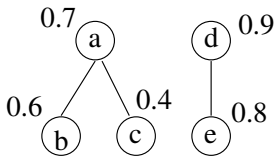
$$\overrightarrow{H^{k+1}} = \alpha S \overrightarrow{H^k} + (1 - \alpha) \overrightarrow{H^0} \tag{1}$$

In this equation, $\overrightarrow{H^0} = \overrightarrow{H}$. We iterate this equation till $\overrightarrow{H^{k+1}} = \overrightarrow{H^k}$ (i.e., it converges). The resulting vector $\overrightarrow{H^p}$ is the principal eigenvector of the matrix $\alpha S + (1-\alpha)\overrightarrow{H^0}e$ where e is a row vector of size $nM + mN - nm$ with all entries equal to 1. This system converges to a unique principal eigenvector when the matrices S and $\overrightarrow{H^0}e$ are both column stochastic. We assure this as each column of S as well as $\overrightarrow{H^0}$ itself adds up to one. Each entry of $\overrightarrow{H^p}$ gives us a combination of homological and topological similarities for the corresponding mapping. We use $\alpha = 0.6$ in this paper since in our previous work we observed that this value provides a good combination of the two similarities [7, 8].

Recall that our aim is to find the relation φ that *maximizes* the summation of the similarity scores defined by $\overrightarrow{H^p}$ while preserving the consistency between mappings. Using $\overrightarrow{H^p}$ and the definition of conflict between two mappings (Definition 2), we create a vertex weighted undirected graph $G_c = (V_c, E_c, w)$, which we name as *the conflict graph* as follows. Each mapping $(R_i, \bar{R}_j) \in \varphi$ corresponds to a vertex in V_c. We set the weight of each vertex $a = (R_i, \bar{R}_j)$ (i.e., $w(a)$) to the similarity between R_i and $\bar{R}_j$ as computed in $\overrightarrow{H^p}$. Since the number of possible one-to-many mappings is $nM + mN - nm$, the conflict graph has $nM + mN - nm$ vertices (i.e., $|V_c| = nM + mN - nm$). We draw an undirected edge between two vertices $a = (R_i, \bar{R}_j)$ and $b = (R_u, \bar{R}_v)$ if $(R_i \cap R_u) \cup (\bar{R}_j \cap \bar{R}_v) \neq \varnothing$ (i.e., a and b conflict). For instance, in Figure 2 there is an edge between a and b representing that they conflict since reaction r_1 is common to both a and b.

	Subnetwork 1 ($\mathcal{P}$)	Subnetwork ($\bar{\mathcal{P}}$)	$\overrightarrow{H^p}$
a:	$R_1 = \{r_1, r_2\}$	$\bar{R}_1 = \{\bar{r_1}\}$	0.7
b:	$R_2 = \{r_1\}$	$\bar{R}_2 = \{\bar{r_2}\}$	0.6
c:	$R_3 = \{r_3\}$	$\bar{R}_3 = \{\bar{r_1}\}$	0.4
d:	$R_4 = \{r_4\}$	$\bar{R}_4 = \{\bar{r_3}, \bar{r_4}, \bar{r_5}\}$	0.9
e:	$R_5 = \{r_4, r_5\}$	$\bar{R}_5 = \{\bar{r_5}\}$	0.8

(a)



(b)

Fig. 2. (a) Each row corresponds to a possible mapping between subnetworks from two hypothetical metabolic pathways. The first column is the unique label for each mapping. Second and third columns are the reactions in the two subnetworks that can map. Last column is the similarity between the two subnetworks. (b) The conflict graph G_c for the mappings in (a).

Extracting the subset of vertices that do not conflict (i.e., no edges) and maximize the sum of the similarity score from the conflict graph is equivalent to finding its *the maximum weight independent set (MWIS)*. MWIS problem can be reduced to our problem of finding the consistent alignment by simply mapping each vertex to a mapping and each undirected edge to a conflict between two mappings. The MWIS problem is NP-hard [26] and there is no constant factor approximation to the optimal solution unless $P = NP$ [27]. Therefore, we need a heuristic algorithm to find the MWIS of G_c and hence our alignment.

We adopt the greedy heuristic described by Sakai *et al.* [28]. Let $N(v)$ denote the set of vertices that are connected to v. At each iteration of this algorithm, we

pick the vertex v that maximizes $f(v) = \sum_{\forall u_i \in N(v)} \frac{w(v)}{w(u_i)}$. This strategy implies that a vertex is more likely to be picked if the mapping it represents has large similarity score and conflicts with small number of other mappings with small similarity scores. After picking a vertex v, we put v into the result set and remove v and all the vertices connected to it ($v \cup N(v)$). We also remove all the edges incident to at least one of the vertices in ($v \cup N(v)$). When there are no more vertices to remove from G_c, the result set contains the vertices of a maximal weight independent set. For our alignment problem, this vertex set corresponds to a set of non-conflicting subnetwork mappings. As an example, in Figure 2, d is the first vertex to be picked. Then, we remove d and $e \in N(d)$ from the graph and put d in the result set. Next, we pick the vertex b as $f(b) = \frac{0.6}{0.7} > f(a) = \frac{0.7}{0.6+0.4} > f(c) = \frac{0.4}{0.7}$. We remove b and $a \in N(b)$ and put b in the result set. Finally, only c is left and taking it into our result set, we have our consistent alignment as the mappings $b = (r_1, \bar{r_2})$, $c = (r_3, \bar{r_1})$ and $d = (r_4, \{\bar{r_3}, \bar{r_4}, \bar{r_5}\})$.

3 Experimental Results

In this section, we experimentally evaluate the performance of SubMAP.
Dataset: We use the metabolic pathways of 20 organisms taken from the KEGG database. Our dataset contains 1,842 pathways in total. The average number of reactions per pathway is 21 and the largest pathway has 72 reactions.

3.1 Alternative Subnetworks

Different organisms can perform the same function through different subnetworks. We name such altered parts that have similar functions as *alternative subnetworks*. An accurate alignment should reveal alternative subnetworks in different pathways. In our first experiment we evaluate whether SubMAP can find them in real metabolic pathways. We align the pathway pairs which are known to contain functionally similar parts with different reaction sets and topologies. Table 1 presents a subset of mappings that are found by our algorithm.

The first row of Table 1 corresponds to alternative subnetworks in Figure 1. The reaction R07613 represents the bottom path in Figure 1 that plants and Chlamydia use to produce LL-2,6- Diaminopimelate from 2,3,4,5- Tetrahydrodipicolinate. This path is discovered and reported as a shortcut on the L-Lysine synthesis path for plants and Chlamydia which is not present in humans [13, 23]. Watanabe *et al.* [13] also suggest that since humans lack this path and hence the catalyzer of the reaction R07613, namely LL-DAP aminotransferase (EC:2.6.1.83), this is an attractive target for the development of new drugs (antibiotics and herbicides). When we align the Lysine biosynthesis pathways of *H.sapiens* and *A.thaliana* (a plant), our algorithm mapped the reaction R07613 of *A.thaliana* to the three reactions that *H.sapiens* has to use to transform 2,3,4,5- Tetrahydrodipicolinate to LL-2,6- Diaminopimelate (R02734, R04365, R04475). In other words, SubMAP successfully identified the alternative subnetworks of different size (1 for *A.thaliana* and 3 for *H.sapiens*) that perform the same function.

Table 1. *Alternative subnetworks that produce same or similar output compounds from the same or similar input compounds in different organisms.* [1] Main input compound utilized by the given set of reactions. [2] Main output compound produced by the given set of reactions. [3] Reactions mappings that corresponds to alternative paths. Reactions are represented by their KEGG identifiers.

Pathway	Organisms	Input Comp.[1]	Output Comp.[2]	Reaction Mappings[3]
Lysine biosynthesis	***A.thaliana***	2,3,4,5-Tetra-	LL-2,6-Di-	**R07613 ⇔ R02734 + R04365 + R04475**
	E.coli	hydrodipico.	aminopimelate	
Lysine biosynthesis	***A.thaliana***	L-Saccharopine	L-Lysine	**R00451 + R00715 + R00716 ⇔ R00451**
	E.coli	meso-2,6-Di.		
Pyruvate metabolism	***E.coli***	Pyruvate	Oxaloacetate	**R00199 + R00345 ⇔ R00344**
	H.sapiens			
Pyruvate metabolism	***E.coli***	Oxaloacetate	Phosphoenol-	**R00341 ⇔ R00431 + R00726**
	H.sapiens		pyruvate	
Pyruvate metabolism	***T.acidophilum***	Pyruvate	Acetyl-CoA	**R01196 ⇔ R00472 + R00216 + R01257**
	A.tumefaciens			
Glycine, serine,	***H.sapiens***	Glycine	Serine	**R00945 ⇔ R00751 + R00945 + R06171**
threonine metabolism	***R.norvegicus***		L-Threonine	
Fructose and	***E.coli***	L-Fucose	L-Fucose 1-p	**R03163 + R03241 ⇔ R03161**
mannose metabolism	***H.sapiens***		L-Fuculose 1-p	
Citrate cycle	***S.aureus N315***	Isocitrate	2-Oxoglutarate	**R00268 + R01899 ⇔ R00709**
	S.aureus COL			
Citrate cycle	***H.sapiens***	Succinate	Succinyl-CoA	**R00432 + R00727 ⇔ R00405**
	A.tumefaciens			
Citrate cycle	***H.sapiens***	Isocitrate	2-Oxoglutarate	**R00709 ⇔ R00362**
	A.tumefaciens	Citrate	Oxaloacetate	

Another interesting example is the second row that is extracted from the same alignment described above. In this case, the three reactions that can produce L-Lysine for *A.thaliana* are aligned to the only reaction that produces L-Lysine for *H.sapiens*. R00451 is common to both organisms and it utilizes meso-2,6-Diaminopimelate to produce L-Lysine. The reactions R00715 and R00716 take place and produce L-Lysine in *A.thaliana* in the presence of L-Saccharopine [29].

For the alignment of Pyruvate metabolisms of *E.coli* and *H.sapiens*, the third and fourth rows show two mappings that are found by SubMAP. The first one maps the two step process in *E.coli* that first converts Pyruvate to Orthophosphate (R00199) and then Orthophosphate to Oxaloacetate (R00345) to the single reaction that directly produces Oxaloacetate from Pyruvate (R00344) in *H.sapiens*. The second one shows another mapping in which a single reaction of *E.coli* is replaced by two reactions of *H.sapiens*. The first two rows for Citrate cycle also report similar mappings for other organism pairs.

Note that all the above examples are one-to-many reaction mappings and hence a merit of the new algorithm we propose here. Our algorithm SubMAP also reports one-to-one mappings. The last row of Table 1 is an example in which one reaction of an organism is replaced by exactly one reaction of another organism. Aligning Citrate cycles of *H.sapiens* and *A.tumefaciens* reveals that even though both the input and output compounds of two reactions R00709 and R00362 are different SubMAP maps these reactions. Also, if we look at the EC numbers of the enzymes catalyzing these reactions (1.1.1.41 and 4.1.3.6) their similarity is zero (see Information content enzyme similarity [8]). If we were to consider only the homological similarities, these two reactions could not have been mapped to each other. However, both these reactions are the neighbors of two other reactions R01325 and R01900 that are present in both organisms. The mappings of R01325 to R01325 and R01900 to R01900 support

the mapping of their neighbors R00709 to R00362. Therefore, by incorporating the topological similarity our algorithm is able to find meaningful mappings with similar topologies and distinct homologies. An algorithm not considering pathway topologies would fail to identify such mappings.

These results suggest: (i) By allowing one-to-many mappings, our method identifies functionally similar subnetworks even if they have different number of reactions. (ii) The incorporation of topological similarity makes it possible to find mappings that can be missed by only considering homological similarity.

3.2 Number of Connected Subnetworks

Given the parameter k, our algorithm enumerates all connected reaction subnetworks of size at most k for each query pathway. One question that we need to answer is: How many such subnetworks exist? Figure 3 plots this number for all the pathways in our dataset. When $k = 1$, the figure shows the number of reactions in each pathway. For $k > 1$ the results demonstrate that the number of subnetworks increase exponentially with k. However, the increase is significantly lower than the theoretical worst case $\sum_{i=1}^{k} \binom{n}{i}$ (i.e., n choose i). For instance, the largest number of subnetworks we obtained for $n = 72$ and $k = 5$ is around 750 times less than the theoretical worst case.

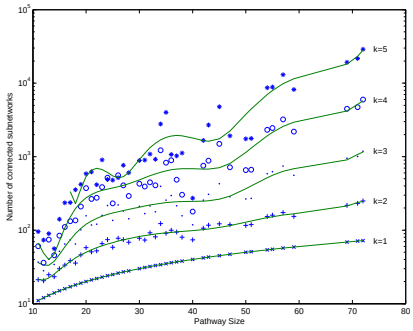


Fig. 3. The number subnetworks with at most k nodes for pathways of different sizes.

The figure also suggests that the number of subnetworks increase linearly with the size of the pathway. This is mainly because the average number of edges (i.e., neighbors) of a node (i.e., subnetwork) remains roughly same as the size of the network increases. As a result, we conclude that for $k \leq 4$, we can enumerate and store all the subnetworks in our dataset. The number of subnetworks for $k = 5$ is still small enough to handle. However, in practice it is unlikely for a single reaction to replace a subnetwork with such a large number of reactions. We expect that $k \leq 4$ would be sufficient to find most of the alternative subnetworks. Hence, we use $k \leq 4$ in our experiments.

3.3 One-to-many Mappings within and across Major Clades

In Section 3.1, we demonstrated that our algorithm can find alternative subnetworks on a number of examples. An obvious question that follows is: How frequent are such alternative subnetworks and what are their characteristics? In other words, is there really a need to allow one-to-many mappings in alignment. In this experiment we aim to answer these questions.

We conduct an experiment as follows. We first pick 9 different organisms 3 from each major phylogenic clade. These organisms are *T.acidophilum*, *Halobacterium sp.*, *M.thermoautotrophicum* from Archaea; *H.sapiens*, *R.norvegicus*, *M.musculus* from Eukaryota; and *E.coli*, *P.aeruginosa*, *A.tumefaciens* from Bacteria. We then extract 10 common pathways for these 9 organisms from KEGG. For each of these common pathways, we choose all possible pairs of the 9 organisms ($\binom{9}{2} = 36$) and align that specific pathway for all organism pairs. In these alignments we exclude the self alignments and the alignment with parameter $k = 1$ since those will definitely incur a bias favoring the number of one-to-one alignments. We computed all possible alignments ($10 \times 36 = 360$) for $k = 2$, 3 and 4 ($360 \times 3 = 1{,}080$ alignments in total). Finally, we calculated the number of four possible types of subnetwork mappings which are 1-to-1, 1-to-2, 1-to-3 and 1-to-4. We hypothesize that the metabolisms of the organisms within a clade will tend to perform the same function through the same (or similar) sized sets of reactions while those across different clades will perform from alternative subnetworks of varying sizes.

Table 2. Percentages of 1-to-1, 1-to-2, 1-to-3 and 1-to-4 mappings in between and across three major clades (**A**: Archaea, **E**: Eukaryota, **B**: Bacteria).

	1-to-1	**1-to-2**	**1-to-3**	**1-to-4**
E-E	89.6	8.8	1.1	0.5
B-B	80.1	16.0	3.1	0.8
A-A	78.3	15.7	4.7	1.3
B-E	69.1	23.1	6.3	1.5
A-B	60.5	28.3	8.5	2.7
A-E	55.8	31.0	10.4	2.8

Table 2 summarizes the results of this experiment. The percentages of each mapping type between two clades is shown as a row in this table. The first three rows corresponds to alignments within a clade and the last three represents alignments across two different clades. An important outcome of these results is that there are considerably large number of one-to-many mappings between organisms of different clades. In the extreme case (last row), nearly half of the mappings are one-to-many. The results also support our hypothesis that one-to-one mappings is more frequent for alignments within the clades compared to across clades due to high similarity between the organisms of the same clade. For instance, for both the first and last row one side of the query set is the Eukaryota. However, going from first row to last, we see around 40% decrease in the number of one-to-one mappings and 250%, 850% and 450% increase in the number of 1-to-2, 1-to-3 and 1-to-4 mappings respectively. Considering Archaea are single-celled microorganisms (e.g., Halobacteria) and Eukaryota are complex organisms with cell membranes (e.g., animals and plants), these jumps in the number of one-to-many mappings suggest that the individual reactions in Archaea are replaced by a number of reactions in Eukaryota. These results have two major implications. *(i) One-to-many mappings are frequent in nature. To obtain biologically meaningful alignments we need to allow such mappings. (ii) The characteristics of the alterative subnetworks can help in inferring the phylogenic relationship among different organisms.*

3.4 Evaluation of Running Time and Memory Utilization

SubMAP allows one to many mappings to find biologically relevant alignments. This however comes at the expense of increased computational cost. Theoret-

ically, this increase can be exponential in k in the worst case. The worst case happens when the pathway is highly connected. Metabolic pathways however are sparse and their connectivity follows power law distribution [30]. In order to understand the capabilities and limitations of our method we examine its performance on real datasets in terms of its running time and memory usage.

We evaluate the performance of our method for querying a database of pathways as follows. We create a query set by selecting 50 pathways of varying sizes from our dataset described at the beginning of this section. We then select another 50 pathways of different sizes to use as our database set for this experiment. We pick the latter 50 pathways such that the average reactions per pathway is 21.4, which is very close to that of the entire database. We then align each query pathway with all the database pathways one by one for different values of k. We measure the average running time and the average memory usage for each query pathway and k value combination. Note that we do not present any performance comparison with an existing method as the existing methods do not allow one-to-many mappings. However, our results for $k = 1$, shows the performance of our algorithm when we restrict it to one-to-one mappings.

Figure 4 shows the average running time of SubMAP for query pathways with increasing number of reactions. When $k = 1$ (i.e., only one-to-one mappings as in existing methods), it runs in less than 0.2 seconds even for the largest query pathway in our query set. As k increases, the running time increases as well. This is because the number of subnetworks and the average numbers of forward and backward neighbors of subnetworks increase with k. However, we observe that our method can perform alignments in practical time even when $k = 4$. It aligns pathways with around 50 reactions in less than one minute and 20 minutes for $k = 3$ and 4 respectively. It runs in less than 15 minutes for the largest query pathway (72 reactions) in our query set for $k = 3$.

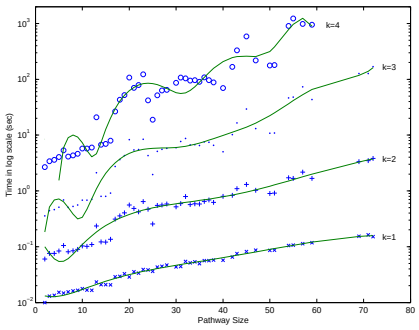


Fig. 4. The average running time of SubMAP when a query pathway is aligned with all the pathways in a pathway database. The selected pathway database contains 50 pathways. X-axis is the number of the reactions of the query pathways.

We also measure the actual memory usage of our algorithm for real pathways of varying sizes and k values (Figure omitted). For $k = 1$ or 2, the memory usage is negligible (1 MB or less) for all pathways. Although the memory usage increases with k, it remains feasible even for query pathways with around 50 reactions for $k = 4$. Our algorithm uses less than 300 MB for the largest query when $k = 3$. For two query pathways both with around 50 reactions and $k = 4$, the memory requirement is around 600 MB. *Thus, our algorithm can run on a standard computer for aligning real-sized metabolic pathways.*

4 Conclusion

In this paper, we considered the problem of aligning two metabolic pathways. The distinguishing feature of our work from the literature is that we allow mapping one molecule of one pathway to a set of molecules of the other. To address this problem, given two metabolic pathways $\mathcal{P}$ and $\bar{\mathcal{P}}$ and an upper bound k on the size of the connected subnetworks, we developed the SubMAP algorithm that can find the consistent mapping of the subnetworks of $\mathcal{P}$ and $\bar{\mathcal{P}}$ with the maximum similarity. We transformed the alignment problem to an eigenvalue problem. The solution to this eigenvalue problem produced a good mixture of homological and topological similarities of the subnetworks. Using these similarity values, we constructed a vertex weighted graph that connects conflicting mappings with an edge. Then, our alignment problem transformed into finding the maximum weight independent subset of this graph. We employed a heuristic method that is used to solve maximum weight independent set problem. The result of this method provided us an alignment that has no conflicting pair of mappings (i.e., consistent). Our experiments on real datasets suggested that our method can identify biologically relevant alignments of alternative subnetworks that are missed by traditional methods. Furthermore, even though SubMAP does not restrict the topologies of query pathways, it is still scalable for real size metabolic pathways when the reaction subsets of size at most four are considered.

References

1. Edwards JS and Palsson BO. Robustness analysis of the Escherichia coli metabolic network. *Biotechnology Progress*, 16:927–939, 2000.
2. Ay F, Xu F, and Kahveci T. Scalable Steady State Analysis of Boolean Biological Regulatory Networks. *PLoS ONE*, 4(12):e7992, 2009.
3. Schuster S, Pfeiffer T, Koch I, Moldenhauer F, and Dandekar T. Exploring the Pathway Structure of Metabolism: Decomposition into Subnetworks and Application to Mycoplasma pneumoniae. *Bioinformatics*, 18:351–361, 2002.
4. Koyuturk M, Grama A, and Szpankowski W. An efficient algorithm for detecting frequent subgraphs in biological networks. In *ISMB*, pages 200–207, 2004.
5. Qian X and Yoon B. Effective Identification of Conserved Pathways in Biological Networks Using Hidden Markov Models. *PLoS ONE*, 4(12):e8070, 2009.
6. Pinter RY, Rokhlenko O, Yeger-Lotem E, and Ziv-Ukelson M. Alignment of metabolic pathways. *Bioinformatics*, 21(16):3401–8, 2005.
7. Ay F, Kahveci T, and de Crecy-Lagard V. Consistent alignment of metabolic pathways without abstraction. In *Computational Systems Bioinformatics Conference (CSB)*, volume 7, pages 237–48, 2008.
8. Ay F, Kahveci T, and de Crecy-Lagard V. A fast and accurate algorithm for comparative analysis of metabolic pathways. *Journal of Bioinformatics and Computational Biology (JBCB)*, 7(3):389–428, 2009.
9. Tohsato Y and Nishimura Y. Metabolic Pathway Alignment Based on Similarity of Chemical Structures. *Information and Media Technologies*, 3:191–200, 2008.
10. Tohsato Y, Matsuda H, and Hashimoto A. A Multiple Alignment Algorithm for Metabolic Pathway Analysis Using Enzyme Hierarchy. In *ISMB*, pages 376–383, 2000.

11. Cheng Q, Harrison R, and Zelikovsky A. MetNetAligner: a web service tool for metabolic network alignments. *Bioinformatics*, 25(15):1989–90, 2009.
12. Sridhar P, Kahveci T, and Ranka S. An iterative algorithm for metabolic network-based drug target identification. In *Pacific Symposium on Biocomputing (PSB)*, volume 12, pages 88–99, 2007.
13. Watanabe N, Cherney MM, van Belkum MJ, Marcus SL, Flegel MD, Clay MD, Deyholos MK, Vederas JC, and James MN. Crystal structure of LL-diaminopimelate aminotransferase from Arabidopsis thaliana: a recently discovered enzyme in the biosynthesis of L-lysine by plants and Chlamydia. *Journal of Molecular Biology*, 371(3):685–702, 2007.
14. Francke C, Siezen RJ, and Teusink B. Reconstructing the metabolic network of a bacterium from its genome. *Trends in Microbiology*, 13(11):550–558, 2005.
15. Clemente JC, Satou K, and Valiente G. Reconstruction of Phylogenetic Relationships from Metabolic Pathways Based on the Enzyme Hierarchy and the Gene Ontology. *Genome Informatics*, 16(2):45–55, 2005.
16. Heymans M and Singh A. Deriving phylogenetic trees from the similarity analysis of metabolic pathways. *Bioinformatics*, 19:138–146, 2003.
17. Ogata H, Fujibuchi W, Goto S, and Kanehisa M. A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters. *Nucleic Acids Research*, 28:4021–8, 2000.
18. Green ML and Karp PD. A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics*, 5:76, 2004.
19. Damaschke P. Graph-Theoretic Concepts in Computer Science. *Lecture Notes in Computer Science*, 484:72–78, 1991.
20. Webb EC. *Enzyme nomenclature 1992* . Academic Press, 1992.
21. Singh R, Xu J, and Berger B. Pairwise Global Alignment of Protein Interaction Networks by Matching Neighborhood Topology. In *RECOMB*, pages 16–31, 2007.
22. Deutscher D, Meilijson I, Schuster S, and Ruppin E. Can single knockouts accurately single out gene functions? *BMC Systems Biology*, 2:50, 2008.
23. McCoy AJ, Adams NE, Hudson AO, Gilvarg C, Leustek T, and Maurelli AT. L,L-diaminopimelate aminotransferase, a trans-kingdom enzyme shared by Chlamydia and plants for synthesis of diaminopimelate/lysine. *PNAS*, 103(47):17909–14, 2006.
24. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, and Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 27(1):29–34, 1999.
25. Hattori M, Okuno Y, Goto S, and Kanehisa M. Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *Journal of the American Chemical Society (JACS)*, 125(39):11853–65, 2003.
26. LovGasz L. Stable set and polynomials. *Discrete Mathematics*, 124:137–53, 1994.
27. Austrin P, Khot S, and Safra M. Inapproximability of Vertex Cover and Independent Set in Bounded Degree Graphs. In *IEEE Conference on Computational Complexity*, pages 74–80, 2009.
28. Sakai S, Togasaki M, and Yamazaki K. A note on greedy algorithms for the maximum weighted independent set problem. *Discrete Applied Mathematics*, 126:313–22, 2003.
29. Saunders PP and Broquist HP. Saccharopine, an intermediate of aminoadipic acid pathway of lysine biosynthesis. *Journal of Biological Chemistry*, 241:3435–40, 1966.
30. Jeong H, Tombor B, Albert R, Oltvai ZN, and Barabasi AL. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–4, 2000.