CONSISTENT ALIGNMENT OF METABOLIC PATHWAYS WITHOUT ABSTRACTION

Ferhat Ay*[1], Tamer Kahveci[1], Valerie de Crécy-Lagard[2]
[1] *Department of Computer Science and Engineering, University of Florida,*
[2] *Department of Microbiology and Cell Science, University of Florida,*
Gainesville, FL 32611, USA
Email: {fay, tamer}@cise.ufl.edu, vcrecy@ufl.edu

Pathways show how different biochemical entities interact with each other to perform vital functions for the survival of organisms. Similarities between pathways indicate functional similarities that are difficult to identify by comparing the individual entities that make up those pathways. When interacting entities are of single type, the problem of identifying similarities reduces to graph isomorphism problem. However, for pathways with varying types of entities, such as metabolic pathways, alignment problem is more challenging. Existing methods, often, address the metabolic pathway alignment problem by ignoring all the entities except for one type. This kind of abstraction reduces the relevance of the alignment significantly as it causes losses in the information content. In this paper, we develop a method to solve the pairwise alignment problem for metabolic pathways. One distinguishing feature of our method is that it aligns reactions, compounds and enzymes without abstraction of pathways. We pursue the intuition that both pairwise similarities of entities (homology) and their organization (topology) are crucial for metabolic pathway alignment. In our algorithm, we account for both by creating an eigenvalue problem for each entity type. We enforce the consistency by considering the reachability sets of the aligned entities. Our experiments show that, our method finds biologically and statistically significant alignments in the order of seconds for pathways with ~ 100 entities.

Keywords: metabolic pathway alignment, metabolic reconstruction, alternative enzyme identification

1. INTRODUCTION

One of the fundamental goals of biology is to understand the biological processes that are the driving forces behind organisms' functions. To achieve this goal, interactions between different components that build up metabolism should be examined in detail. These interactions can reveal significant information that is impossible to gather by analyzing individual entities. Recent advances in high throughput technology resulted in an explosion of different types of interaction data which is compiled in databases, such as KEGG[1] and EcoCyc[2]. Analyzing these databases is necessary to capture the valuable information carried by the pathways. An essential type of analysis is the comparative analysis which aims at identifying similarities between pathways of different organisms. Finding these similarities provides valuable insights for drug target identification[3], metabolic reconstruction of newly sequenced genome[4], and phylogenetic tree construction[5].

To identify similarities between two pathways, it is necessary to find a mapping of their entities. This problem is computationally interesting and challenging. Using a graph model for representing pathways, the graph/subgraph isomorphism problems can be reduced to global/local pathway alignment problems in polynomial time. However, since the graph and subgraph isomorphism problems are GI-complete and NP-complete respectively, global/local pathway alignment problems are GI/NP complete. Hence, efficient heuristics are needed to solve these problems in a reasonable time.

In order to reduce the time complexity of the alignment, some existing algorithms restrict the topology of query pathways[6, 7]. For instance, the method proposed by Tohsato *et al.*[7] works for only non-cyclic pathways, whereas the algorithm of Pinter *et al.*[8] restricts the query pathways to multi-source trees. However, those restrictions are far from the reality and they limit the applicability of the methods to only a small percentage of pathways.

A common delusion of existing algorithms for metabolic pathway alignment is to use a model that focuses on only one type of entity and ignores the others. This simplification converts metabolic pathways to the graphs with only compatible nodes. We use the word *compatible* for the entities that are of the same type. For metabolic pathways, two entities are compatible if they both are reactions or enzymes or compounds. We term the conversions that reduces the metabolic pathways to compatible entities as *abstraction.* Previously, reaction based[5], enzyme based[8, 9] and compound based[7] abstractions

*Corresponding author.

are used for representing metabolic pathways. Figure 1 illustrates the problems with the enzyme based abstraction used by Pinter *et al.*[8] and Koyutürk *et al.*[9]. In Figure 1(a), enzymes E_1 and E_2 interact on two different paths. Abstraction in Figure 1(b) loses this information and merges these two paths into a single interaction. After the abstraction, an alignment algorithm aligning the $E_1 \rightarrow E_2$ interactions in Figures 1(a) and 1(b) cannot realize through which path, out of two alternatives, the enzymes E_1 and E_2 are aligned. It is important to note that the amount of information lost due to abstraction grows exponentially with the number of branching entities.

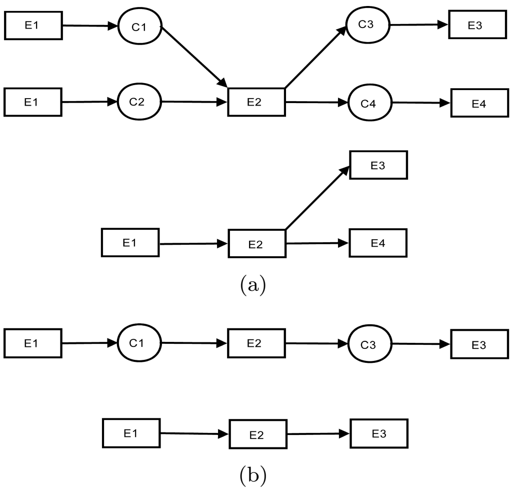


Fig. 1. Top figures in (a) and (b) illustrate two hypothetical metabolic pathways with enzymes and compounds represented by letters E and C, respectively. Bottom figures in (a) and (b) show the same pathways after abstraction when the compounds are ignored.

This paper addresses the pairwise alignment problem for metabolic pathways without any topology restriction or any abstraction. A distinguishing feature of our method is that reported alignments provide the individual mappings for reactions, compounds and enzymes. Furthermore, our algorithm can be extended to work for other pathway types that have entities from different compatibility classes.

In our method, we account for both pairwise and topological similarities of the entities since they both are very crucial for alignment. Singh *et al.*[10] combined homology and topology for protein interaction pathway alignment by creating an eigenvalue problem. A similar approach is previously used for discovery of authoritative information sources on the World Wide Web by Kleinberg[11]. In the case of protein interaction pathways, the alignment problem can be mapped to a single eigenvalue problem since all nodes are of the same type and interactions between them are assumed to be undirected. The algorithm proposed by Singh *et al.*, however, cannot be trivially extended to metabolic pathways as these pathways contain entities of varying types and the interactions are directed.

For metabolic pathway alignment, we first create three eigenvalue problems, one for compounds, one for reactions and one for enzymes. We, also, consider the directions of interactions. We solve these eigenvalue problems using power method. The principal eigenvectors of each of these problems define a weighted bipartite graph. We, then, extract reaction mappings using maximum weight bipartite matching on the corresponding bipartite graph. After that, to ensure consistency of the alignment, we prune the edges in the bipartite graphs of compounds and enzymes which lead to inconsistent alignments with respect to reaction mappings. Finally, we find the enzyme and the compound mappings using maximum weight bipartite matching. We report the extracted mappings of entities as an alignment together with a similarity score that we devise for measuring the similarity between the aligned pathways. Furthermore, we measure the unexpectedness of the resulting alignment by calculating its z-score.

Our experiments on KEGG Pathway database show that our algorithm successfully identifies functionally similar entities and sub-paths in pathways of different organisms. Our method produces biologically and statistically significant alignments of pathways very quickly.

Our Contributions:

- We introduce the *consistency* concept for alignment of pathways with different entity types by constructing reachability sets. We develop an algorithm that aligns pathways while enforcing consistency.

- We integrate the graph model that we devised earlier[3] into the context of pathway alignment. Using this model, we develop an algorithm to align pathways when there is no *abstraction.* Unlike existing graph models, this model is a nonredundant representation of pathways without any abstraction.

- We introduce a new scoring scheme for measuring the similarity of two reactions. We also devise a similarity score and a z-score for measuring similarities between two metabolic pathways.

The organization of the rest of this paper is as follows: Section 2 discusses the related work. Section 3 presents our graph model for representing pathways. Section 4 describes the proposed algorithm in detail. Section 5 illustrates the experimental results. Section 6 briefly concludes the paper.

2. BACKGROUND

Pathway alignment problem has been mostly considered for protein interaction networks (PPI). As a result, existing methods often can align two pathways only if all the interacting entities are of the same type[6, 10, 12, 13]. However, metabolic pathways are composed of enzymes, reactions, compounds and interactions between these three types of entities. Therefore, it is not trivial how PPI alignment methods can be extended to align metabolic pathways.

For solving the metabolic pathway alignment problem, existing methods model the pathways as interactions between entities of a single type. This abstraction causes significant information loss as seen in Figure 1. After this abstraction in modeling, a common approach for aligning metabolic pathways is to use graph theoretic techniques. Pinter *et al.*[8] mapped the metabolic pathway alignment problem to the subgraph homomorphism problem. However, they oversimplify the problem by restricting the topology of pathways to multi-source trees. By solely relying on Enzyme Commission (EC)[14] numbers, Tohsato *et al.*[15] proposed an alignment method for metabolic pathways in 2003. Due to the discrepancies in the EC hierarchy, the accuracy of this method is questionable. In 2007, they proposed another method[7], which only considers chemical structures of compounds for alignment. This idea, however, totally ignores the effect of other entities such as enzymes and reactions.

To overcome the above mentioned problems, in this paper, we refuse to use a model that is biased on one entity type. Equipped with a more comprehensive graph model without abstraction and an efficient iterative algorithm, our tool outperforms existing methods for metabolic pathway alignment.

3. MODEL

The first step in developing effective computational techniques to leverage metabolic pathways is to develop an accurate model to represent them. Existing graph models are not sufficient for representing all interactions between different entity types that are present in metabolic pathways. Figure 1 emphasizes the importance of the modeling scheme for pathway alignment. As discussed in Section 2, abstractions in modeling reduce the alignment accuracy dramatically.

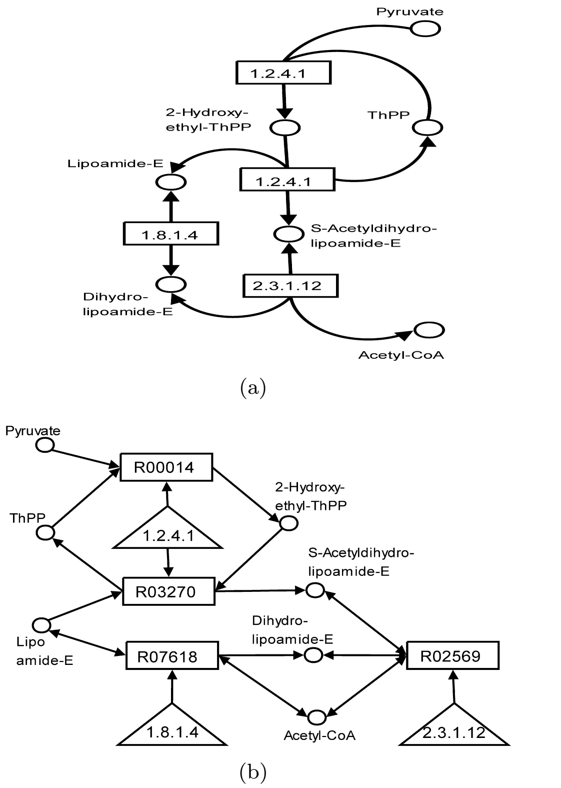
In order to address the insufficiency of existing models, we developed a graph model for representation of metabolic pathways. Our model is a variation of boolean network model and is able to capture all interactions between all types of entities. We discuss this model in the rest of this section.

For the rest of this paper, we will use $\mathcal{P}$, $\mathcal{R}$, $\mathcal{C}$, $\mathcal{E}$ to denote the sets of all pathways, all reactions, all compounds and all enzymes, respectively. Let, $R \subseteq \mathcal{R}, C \subseteq \mathcal{C}, E \subseteq \mathcal{E}$ such that $R = \{R_1, R_2, \ldots, R_{|R|}\}$, $C = \{C_1, C_2, \ldots, C_{|C|}\}$ and $E = \{E_1, E_2, \ldots, E_{|E|}\}$ denote the reactions, compounds and enzymes of the pathway P, respectively. The definition below formalizes our graph model:

Definition 1. A directed graph, $G(V, I)$ for representing the metabolic pathway $P \in \mathcal{P}$, is constructed as follows: The node set, $V = [R, C, E]$, is the union of reactions, compounds and enzymes of P. The edge set, I, is the set of interactions between the nodes. An interaction is represented by a directed edge that is drawn from a node x to another node y, if and only if one of the following three conditions holds:

1) x is an enzyme that catalyzes reaction y.
2) x is an input compound of reaction y.
3) x is a reaction that produces compound y.

Figure 2 illustrates the conversion of a KEGG metabolic pathway to our graph model. As suggested, our model is capable of representing metabolic pathways without losing any type of entities or interactions between these entities. We avoid any kind of abstraction in alignment by using this model. Besides, our model is a nonredundant representation of pathways since it represents each entity using a single node.



(a)

(b)

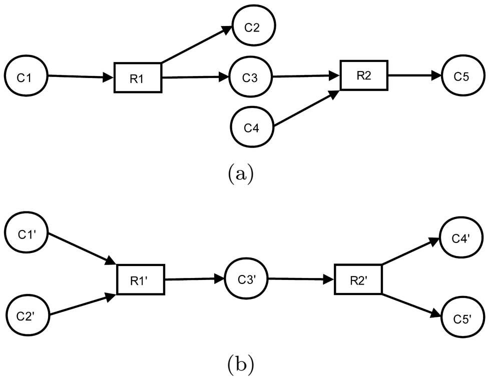
Fig. 2. Graph representation of metabolic pathways: (a) A portion of the reference pathway of Alanine and aspartate metabolism from KEGG database (b) Our graph representation corresponding to this portion. Reactions are shown by rectangles, compounds are shown by circles and enzymes are shown by triangles.

4. ALGORITHM

Motivated by previous research on alignment of pathways and growing demand on fast and accurate tools for analyzing biological pathways, in this section we describe our algorithm for pairwise alignment of metabolic pathways. Before going into the details of the algorithm, it is better to formally state the alignment problem. To do this we first need to define an alignment and the consistency of an alignment.

Let, $P, \bar{P} \in \mathcal{P}$ stand for the two query metabolic pathways which are represented by graphs $G(V, I)$ and $\bar{G}(\bar{V}, \bar{I})$, respectively. Using our graph formalization V can be replaced by $[R, C, E]$, where R denotes the set of reactions, C denotes the set of compounds and E denotes the set of enzymes of P. Similarly, $\bar{V}$ is replaced by $[\bar{R}, \bar{C}, \bar{E}]$.

Definition 2. An *alignment* of two metabolic pathways $P = G(V, I)$ and $\bar{P} = \bar{G}(\bar{V}, \bar{I})$, is a *mapping* $\varphi : V' \rightarrow \bar{V}'$ where $V' \subseteq V$ and $\bar{V}' \subseteq \bar{V}$.



(a)

(b)

Fig. 3. *Consistency of an alignment and an example nonsensical matching:* Figures in (a) and (b) are graph representations of two query pathways. Enzymes are not displayed for simplicity. Suppose that our alignment algorithm mapped the reactions R1 to R1' and R2 to R2'. In this scenario, a trivial consistent matching is C1-C1'. An example for a nonsensical matching that cause inconsistency is C2' - C5. When C1 is matched to C1', a consistent matching might be C2' - C4 since they are inputs of two neighbor reactions.

Before arguing the consistency of an alignment, we discuss the reachability concept for entities. Given two entities $v_i, v_j \in V$ which are compatible, v_j is *reachable* from v_i if there is a directed path from v_i to v_j in graph G. As a shorthand notation, $v_i \Rightarrow v_j$ denotes that v_j is reachable from v_i.

Using the definition and the notation above, we define a consistent alignment as follows:

Definition 3. An alignment of two pathways $P = G(V, I)$ and $\bar{P} = \bar{G}(\bar{V}, \bar{I})$ defined by the mapping $\varphi : V' \rightarrow \bar{V}'$ is *consistent* if and only if all the conditions below are satisfied:

- For all $\varphi(v) = \bar{v}$ where $v \in V$ and $\bar{v} \in \bar{V}$, v and $\bar{v}$ are compatible.
- $\varphi(v)$ is one-to-one.
- For all $\varphi(v_i) = \bar{v_i}$, there exists $\varphi(v_j) = \bar{v_j}$ where $v_i, v_j \in V$ and $\bar{v_i}, \bar{v_j} \in \bar{V}$, such that $v_i \Rightarrow v_j$ and $\bar{v_i} \Rightarrow \bar{v_j}$, or $v_j \Rightarrow v_i$ and $\bar{v_j} \Rightarrow \bar{v_i}$.

The first condition filters out matchings of different entity types. The second condition ensures that none of the entities are mapped to more than one entity. The last condition restricts the mappings to the ones which are supported by at least one other mapping. Additionally, it eliminates nonsensical matchings that cause inconsistency as described in Figure 3.

Now, let, $SimP_{\varphi} : \mathcal{P} \times \mathcal{P} \rightarrow \Re \cap [0, 1]$ be a pairwise pathway similarity function, induced by the mapping φ. The maximum score, $SimP_{\varphi} = 1$, is achieved when two pathways are identical. In Section 4.5, we will describe in detail how $SimP_{\varphi}$ is computed after φ is created. In order to restate our problem, it is only necessary to know that there exists such a similarity function for pathways.

Under the light of the above definitions and formalizations, here is the *problem statement* for pairwise metabolic pathway alignment:

Definition 4. Given two metabolic pathways, $P = G(V, I)$ and $\bar{P} = \bar{G}(\bar{V}, \bar{I})$, the alignment problem is to find a consistent mapping $\varphi : V \rightarrow \bar{V}$ that maximizes $SimP_{\varphi}(P, \bar{P})$.

In the following sections, we describe our algorithm for metabolic pathway alignment.

4.1. Pairwise Similarity of Entities

Metabolic pathways are composed of entities which are either enzymes, compounds or reactions. The degree of similarity between pairs of entities of two pathways is a good indicator of the similarity between these pathways.

A number of similarity measures have been devised for each type of entity in the literature. In the rest of this section, we describe the similarity functions we used for enzyme and compound pairs. We also discuss the similarity function we developed for reaction pairs. All pairwise similarity scores are normalized to the interval of $[0, 1]$ to ensure compatibility between similarity scores of different entities.

Enzymes: An enzyme similarity function is of the form $SimE : \mathcal{E} \times \mathcal{E} \rightarrow \Re \cap [0, 1]$. In our implementation, the two options we provide the user for enzyme similarity scoring are:

• *Hierarchical enzyme similarity score*[15] depends only on Enzyme Commission (EC)[14] numbers of enzymes.

• *Information content enzyme similarity score*[8] uses EC numbers of enzymes together with the information content of this numbering scheme.

Compounds: Two different methods we use for compound similarity are:

• *A trivial compound similarity score* returns 1 if two compounds are identical and 0 otherwise.

• *SIMCOMP compound similarity score* for compounds is defined by Hattori *et al.*[16]. This score is assessed by mapping chemical structures of compounds to graphs and then measuring the similarity between these graphs.

Reactions: Our similarity score for reactions depends on the similarities of the components that take place in the reaction process such as enzymes, input compounds and output compounds. It is of the form $SimR : \mathcal{R} \times \mathcal{R} \rightarrow \Re \cap [0, 1]$. Our reaction similarity score employs the maximum weight bipartite matching technique. The following is a brief description of the maximum weight bipartite matching problem:

Definition 5. Let, U and V be two disjoint node sets and S be a $|U| \times |V|$ matrix representing edge weights between all possible pairs with one element from U and one element from V, where existing edges correspond to a nonzero entry in S. Maximum Weight Bipartite Matching problem is to find a list of node pairs, such that the sum of edge weights between the elements of these pairs is maximum. We denote this sum of edge weights by $MBS(U, V, S)$.

Let, R_i and R_j be two reactions from $\mathcal{R}$. Define R_i as a combination of input compounds, output compounds and enzymes and denote it by $[C_{in}^{R_i}, C_{out}^{R_i}, E^{R_i}]$, where $C_{in}^{R_i}, C_{out}^{R_i} \subseteq C$ and $E^{R_i} \subseteq E$. Similarly, define R_j as $[C_{in}^{R_j}, C_{out}^{R_j}, E^{R_j}]$. Additionally, compute the edge weight matrices $S_{C_{out}}$ and $S_{C_{in}}$ using the selected compound similarity score and S_E using the selected enzyme similarity.

The similarity score of (R_i, R_j) is computed as:

$$\begin{aligned} SimR(R_i, R_j) &= \gamma_{C_{in}} MBS(C_{in}^{R_i}, C_{in}^{R_j}, S_{C_{in}}) \\ &+ \gamma_{C_{out}} MBS(C_{out}^{R_i}, C_{out}^{R_j}, S_{C_{out}}) \\ &+ \gamma_E MBS(E^{R_i}, E^{R_j}, S_E) \qquad (1) \end{aligned}$$

Here, $\gamma_{C_{in}}, \gamma_{C_{out}}, \gamma_E$ denote the relative weights of input compounds, output compounds and enzymes on reaction similarity, respectively. Typical values for these parameters are $\gamma_{C_{in}} = \gamma_{C_{out}} = 0.3$ and $\gamma_E = 0.4$. These values are empirically determined after a number of experiments. One more factor that defines reaction similarity is the choice of $SimE$ and $SimC$ functions. Since we have two options for each, we end up having four different options for reaction similarity depending on the choices of $SimE$ and $SimC$.

Now, we can create the pairwise similarity vectors $\overrightarrow{H_R}^0$, $\overrightarrow{H_C}^0$, $\overrightarrow{H_E}^0$ for reactions, compounds and enzymes, respectively. Since, calculation of these vectors is very similar for each entity type we just describe the one for reactions.

The entry ${H_R}^0((i-1)|R|+j)$ of $\overrightarrow{H_R}^0$ vector stands for the similarity score between $R_i \in R$ and $\bar{R}_j \in \bar{R}$, where $1 \le i \le |R|$ and $1 \le j \le |\bar{R}|$. We will use the notation ${H_R}^0(i,j)$ for this entry since $\overrightarrow{H_R}^0$ can, also, be viewed as a $|R| \times |\bar{R}|$ matrix. One thing we need to be careful about is that $\overrightarrow{H_R}^0, \overrightarrow{H_C}^0, \overrightarrow{H_E}^0$ vectors should be of the unit norm. This normalization is crucial for stability and convergence of our alignment algorithm, as we will clarify in Section 4.2. We, therefore, compute an entry of $\overrightarrow{H_R}^0$ as:

$${H_R}^0(i,j) = \frac{SimR(R_i, \bar{R}_j)}{||\overrightarrow{H_R}^0||_1} \qquad (2)$$

In a similar fashion, we compute all entries of $\overrightarrow{H_C}^0, \overrightarrow{H_E}^0$ by using $SimC$ and $SimE$ functions, respectively. We use these three vectors to carry the homology information throughout the algorithm. In Section 4.3, we will describe how they are combined with topology information to produce an alignment.

4.2. Similarity of Topologies

Previously, we discussed why and how we use pairwise similarities of entities. However, although pairwise similarities are necessary, they are not sufficient. The induced topologies of the aligned entities should also be similar. In order to account for topological similarity, in this section, we define the notion of neighborhood for each compatibility class. After that, we create *support matrices* that allow us to exploit this neighborhood information.

To be consistent with our reachability definition, we define our neighborhood relations according to directions of interactions. In other words, we distinguish between *backward neighbors* and *forward neighbors* of an entity.

Let, $BN(x)$ and $FN(x)$ denote the backward and forward neighbor sets of an entity x. We need to show how to construct these sets for each entity type. We start by defining neighborhood of reactions to build backbones for topologies of the pathways. Then, using that backbone we define neighborhood concepts for compounds and enzymes.

Consider two reactions R_i and R_j of the pathway P. If an output compound of R_i is an input compound for R_j, then R_i is a backward neighbor of R_j and R_j is a forward neighbor of R_i. We construct the forward and backward neighbor sets of each reaction in this manner. For instance, in Figure 2(b), R02569 is a forward neighbor of R03270, and R03270 is a backward neighbor of R02569.

A more generalized version of neighborhood definition can be given to include not only instant neighbors but also neighbors of neighbors, and so on. However, it complicates the algorithm unnecessarily, since our method already considers the support of indirect neighbors as described in Section 4.3.

As stated before, neighborhood definitions of compounds and enzymes depend on the topology of reactions. Let, C_i and C_j be two compounds, R_s and R_t be two reactions of the pathway P. If $R_s \in BN(R_t)$ and C_i is an input (output) compound of R_s and C_j is an input (output) compound of R_t then $C_i \in BN(C_j)$ and $C_j \in FN(C_i)$. For example, in Figure 2(b), Lipoamide-E and Dihydro-lipoamide-E are neighbors since they are inputs of two neighbor reactions R02569 and R03270, respectively. For enzymes the construction is similar.

In the light of the above definitions, we create *support matrices* for each compatibility class. These matrices contain the information about topological similarities of pathways. In here, we only describe how to calculate the support matrix for reactions. The calculations for enzymes and compounds is done similarly.

Definition 6. Let, $P = G([R, C, E], I)$ and $\bar{P} = \bar{G}([\bar{R}, \bar{C}, \bar{E}], \bar{I})$ be two metabolic pathways. The support matrix for reactions of P and $\bar{P}$ is a $|R||\bar{R}| \times |R||\bar{R}|$ matrix denoted by A_R. An entry of the form $A_R[(i-1)|R|+j][(u-1)|R|+v]$ identifies the fraction of the total support provided by $R_u, \bar{R}_v$ matching to $R_i, \bar{R}_j$ matching. Let, $N(u,v) = |BN(R_u)||BN(\bar{R}_v)| + |FN(R_u)||FN(\bar{R}_v)|$ denote the number of possible neighbor matchings of R_u and $\bar{R}_v$.

Each entry of A_R is computed as:

$$A_R[(i-1)|R|+j][(u-1)|R|+v] =$$

$$\begin{cases} \frac{1}{N(u,v)} & \text{if } (R_i \in BN(R_u) \text{ and } \bar{R}_j \in BN(\bar{R}_v)) \\ & \text{or } (R_i \in FN(R_u) \text{ and } \bar{R}_j \in FN(\bar{R}_v)) \\ 0 & \text{otherwise} \end{cases}$$

After filling all entries, we replace the zero columns of A_R with $|R||\bar{R}| \times 1$ vector $[\frac{1}{|R||\bar{R}|}, \frac{1}{|R||\bar{R}|} \ldots, \frac{1}{|R||\bar{R}|}]^T$. This way support of the matching indicated by the zero column is uniformly distributed to all other matchings.

For example, in Figure 1(a), $|BN(E2)| = 1$ and $|FN(E2)| = 2$ and in Figure 1(b), $|BN(E2)| = 1$ and $|FN(E2)| = 1$. Hence, the support of matching E2 of Figure 1(a) with E2 of Figure 1(b) should be equally distributed to its 3 (i.e. $1 \times 1 + 2 \times 1$) possible neighbor matching combinations by assigning 1/3 to the corresponding entries of A_E matrix.

We use the terms A_R, A_C and A_E to represent the support matrices for reactions, compounds and enzymes, respectively. Power of these support matrices is that, they allow us to distribute the support of a matching to other matchings according to the distances between them. This distribution is crucial for favoring the matchings whose neighbors can also be matched as well. The method for distributing the matching scores appropriately is described in the next section.

4.3. Combining Homology and Topology

Both the pairwise similarities of entities and the organization of these entities together with the interactions between them provide us precious information about the functional correspondence and evolutionary similarity of metabolic pathways. Hence, an accurate alignment strategy needs to combine these factors cautiously. In this subsection we describe our strategy for achieving this combination.

From the previous sections, we have $\overrightarrow{H_R}^0$, $\overrightarrow{H_C}^0$, $\overrightarrow{H_E}^0$ vectors containing pairwise similarities of entities and A_R, A_C, A_E matrices containing topological similarities of pathways. Using these vectors and matrices together with a weight parameter $\alpha \in [0, 1]$, for adjusting the relative effect of topology and homology, we transform our problem into three eigenvalue problems as follows:

$$\overrightarrow{H_R}^{k+1} = \alpha A_R \overrightarrow{H_R}^k + (1-\alpha)\overrightarrow{H_R}^0 \quad (3)$$

$$\overrightarrow{H_C}^{k+1} = \alpha A_C \overrightarrow{H_C}^k + (1-\alpha)\overrightarrow{H_C}^0 \quad (4)$$

$$\overrightarrow{H_E}^{k+1} = \alpha A_E \overrightarrow{H_E}^k + (1-\alpha)\overrightarrow{H_E}^0 \quad (5)$$

for $k \geq 0$.

For stability purposes $\overrightarrow{H_R}^k, \overrightarrow{H_C}^k$ and $\overrightarrow{H_E}^k$ are normalized before each iteration.

Lemma 4.1. *A_R, A_C and A_E are column stochastic matrices.*

Proof. Each entry of A_R, A_C and A_E are nonnegative by Definition 6. By construction, entries of each column of these matrices sums up to one. □

Lemma 4.2. *Let, A be an $N \times N$ column stochastic matrix and E be an $N \times N$ matrix such that $E = \overrightarrow{H} e^T$, where $\overrightarrow{H}$ is an N-vector with $||\overrightarrow{H}||_1 = 1$ and e is an N-vector with all entries equal to 1. For any $\alpha \in [0, 1]$ define the matrix M as:*

$$M = \alpha A + (1-\alpha)E \quad (6)$$

The maximal eigenvalue of M is $|\lambda_1| = 1$. The second largest eigenvalue of M satisfies $|\lambda_2| \leq \alpha$.

Proof. Omitted, see Haveliwala *et al.*[17] □

Using an iterative technique called power method, our aim is to find the stable state vectors of the Equations (3), (4) and (5). We know by Lemma 4.1 that A_R, A_C and A_E are column stochastic matrices. By construction of $\overrightarrow{H_R}^0, \overrightarrow{H_C}^0, \overrightarrow{H_E}^0$, we have $||\overrightarrow{H_R}^0||_1 = 1, ||\overrightarrow{H_C}^0||_1 = 1, ||\overrightarrow{H_E}^0||_1 = 1$. Now, by the following theorem, we show that the stable state vectors for Equations (3), (4) and (5) exist and they are unique.

Theorem 4.1. *Let, A be an $N \times N$ column stochastic matrix and H^0 be an N-vector with $||H^0||_1 = 1$. For any $\alpha \in [0, 1]$, there exists a stable state vector H^s, which satisfies the equation:*

$$H = \alpha A H + (1-\alpha)H^0 \quad (7)$$

Furthermore, if $\alpha \in [0, 1)$, then H^s is unique.

Proof. *Existence:* Let, e be the n-vector with all entries equal to 1. Then, $e^T H = 1$ since $||H||_1 = 1$ after normalizing H. Now, Equation 7 can be rewritten as:

$$\begin{aligned} H &= \alpha A H + (1-\alpha)H^0 = \alpha A H + (1-\alpha)H^0 e^T H \\ &= (\alpha A + (1-\alpha)H^0 e^T)H = MH \end{aligned}$$

where $M = \alpha A + (1-\alpha)H^0 e^T$. $H^0 e^T$ is a column stochastic matrix, since its columns are all equal to

H^0 and $||H^0||_1 = 1$. Created as a weighted combination of two column stochastic matrices, M is also column stochastic. Then, by Lemma 4.2, $\lambda_1 = 1$ is an eigenvalue of M. Hence, there exists an eigenvector H^s corresponding to the eigenvalue λ_1, which satisfies the equation $\lambda_1 H^s = MH^s$.

Uniqueness: Applying Lemma 4.2 to the M matrix defined in the existence part, we have $|\lambda_1| = 1$ and $|\lambda_2| \leq \alpha$. If $\alpha \in [0, 1)$, then $|\lambda_1| > |\lambda_2|$. This implies that, λ_1 is *the* principal eigenvalue of M and H^s is the unique eigenvector corresponding to it. □

Convergence rate of power method for Equations (3), (4) and (5) are determined by the eigenvalues of the M matrices (as defined in Equation 6) of each equation. Convergence rate is proportional to $O(\frac{|\lambda_2|}{|\lambda_1|})$, which is $O(\alpha)$, for each equation. Therefore, choice of α not only adjusts the relative importance of homology and topology, but it also affects running time of our algorithm. Our experiments showed that our algorithm performs well and converges quickly with $\alpha = 0.7$.

In Equations (3), (4) and (5), before the first iteration of power method we only have initial pairwise similarity scores. In the k^{th} iteration, the weight of pairwise similarity score stays to be $(1 - \alpha)$, whereas weight of total support given by $(k - t)^{th}$ degree neighbors of $R_i, \bar{R_j}$ is $\alpha^{k-t}(1 - \alpha)$. That way, neighborhood topologies of matchings are thoroughly utilized without ignoring the effect of initial pairwise similarity scores. As a result, stable state vectors calculated in this manner, are convenient candidates for extracting the entity mappings to create the overall alignment for the query pathways.

4.4. Extracting the Mapping of Entities

Having combined homological and topological similarities of query metabolic pathways, now, it only remains to extract the mapping, φ, of entities. However, since we restrict our consideration to consistent mappings, this extraction by itself is still challenging. Figure 3 points out the importance of maintaining consistency of an alignment.

An alignment is described by the mapping φ, that gives the individual matchings of entities. Lets denote φ as $\varphi = [\varphi_R, \varphi_C, \varphi_E]$, where φ_R, φ_C and φ_E are consistent mappings for reactions, compounds and enzymes, respectively.

If we go back to definition of consistency, there are three conditions that φ should satisfy. The first one is trivially satisfied for any φ of the form $[\varphi_R, \varphi_C, \varphi_E]$, since we beforehand distinguished each entity type. For the second condition, it is sufficient to create one-to-one mappings for each entity type. By using maximum weight bipartite matching we get one-to-one mappings $\varphi_R, \varphi_C, \varphi_E$, which in turn implies φ is one-to-one since intersections of compatibility classes are empty.

The difficult part of finding a consistent mapping is combining mappings of reactions, enzymes and compounds without violating the third condition. For that purpose, we choose a specific ordering between extraction of reaction, enzyme and compound mappings. We create the mapping φ_R first. We extract this mapping by using maximum weight bipartite matching on the bipartite graph constructed by the edge weights in $\overrightarrow{H_R}^S$ vector. Then, using the aligned reactions and the reachability concept, we prune the edges from the bipartite graph of compounds (enzymes) for which the corresponding compound (enzyme) pairs are inconsistent with the reaction mapping. In other words, we prune the edge between two compounds (enzymes), $x, \bar{x}$, if there does not exist any other compound (enzyme) pair $y, \bar{y}$ such that, x is reachable from $\bar{x}$ and y is reachable from $\bar{y}$, or $\bar{x}$ is reachable from x and $\bar{y}$ is reachable from y. By pruning these edges we make sure that for any φ_C and φ_E extracted from the pruned bipartite graphs, $\varphi = [\varphi_R, \varphi_C, \varphi_E]$ is consistent.

Recall that, our aim is to find a consistent alignment which maximizes the similarity score $SimP_\varphi$. The φ defined above satisfies the consistency criteria. For the maximality of similarity score, in the next section, we define $SimP_\varphi$ and then discuss that φ is the mapping that maximizes this score.

4.5. Similarity Score of Pathways

As we present in the previous section, our algorithm guarantees to find a consistent alignment represented by the mappings of entities. One can discuss the accuracy and biological significance of our alignment by looking at the individual matchings that we reported. However, this requires a solid background of the specific metabolism of different organisms. To computationally evaluate the degree of similarity between pathways we devise a similarity score.

We use pairwise similarities of aligned entities to calculate the overall similarity between two query pathways. The definition of similarity function $SimP_{\varphi}$, is as follows:

Definition 7. Let, $P = G([R, C, E], I)$ and $\bar{P} = \bar{G}([\bar{R}, \bar{C}, \bar{E}], \bar{I})$ be two metabolic pathways. Given a mapping $\varphi = [\varphi_R, \varphi_C, \varphi_E]$ between entities of P and $\bar{P}$, similarity of P and $\bar{P}$ is calculated as:

$$SimP_{\varphi}(P, \bar{P}) = \frac{\beta}{|\varphi_C|} \sum_{\forall (C_i, \bar{C_j}) \in \varphi_C} SimC(C_i, \bar{C_j}) + \frac{(1-\beta)}{|\varphi_E|} \sum_{\forall (E_i, \bar{E_j}) \in \varphi_E} SimE(E_i, \bar{E_j})$$

where $|\varphi_C|$ and $|\varphi_E|$ denote the cardinality of corresponding mappings and $\beta \in [0, 1]$ is a parameter that adjusts the relative influence of compounds and enzymes on the alignment score.

Calculated as above, $SimP_{\varphi}$ gives a score between 0 and 1, such that a bigger score implies a better alignment between pathways. We use $\beta = 0.5$ in our experiments, since we do not want to bias our score towards enzymes or compounds. The user can choose $\beta = 0$ to have an enzyme based similarity score or $\beta = 1$ to have a compound based similarity score. Reactions are not considered while calculating this score since reaction similarity scores are already determined by enzyme and compound similarity scores.

Now, having defined the pathway similarity score, we need to show that the consistent mapping $\varphi = [\varphi_R, \varphi_C, \varphi_E]$ found in the previous section, is the one that maximizes this score. But, this follows from the fact that we used maximum weight bipartite matching on the pruned bipartite graphs of enzymes and compounds. In other words, since maximality of the total edge weights of φ_C and φ_E are beforehand assured by the extraction technique, their summation is guaranteed to give the maximum $SimP_{\varphi}$ value for a fixed β.

Complexity Analysis

Let, $P = G([R, C, E], I)$ and $\bar{P} = \bar{G}([\bar{R}, \bar{C}, \bar{E}], \bar{I})$ be two query pathways. The overall time complexity of our algorithm, which is dominated by the power method iterations, is:
$O(|R|^2|\bar{R}|^2 + |C|^2|\bar{C}|^2 + |E|^2|\bar{E}|^2)$.

5. EXPERIMENTS

In order to evaluate the performance of our algorithm we conduct various experiments.

Datasets: We use KEGG Pathway database, which currently has 72,628 pathways which are generated from 360 reference pathways. We convert the pathway data into our graph model.

Parameters: We allow users to change a set of parameters in our implementation. This flexibility is important in some scenarios. For instance, if a user is interested only in enzyme similarities or compound similarities between pathways, then it would be enough to set the parameters accordingly. Due to space limitations, we report the results with only one parameter setting.

α is the parameter that adjusts the relative weight of topology and homology. As we discussed, $\alpha = 0.7$ works well for our method. There is no significant difference between different $SimE$ and $SimC$ scores. We use the information content enzyme similarity score for $SimE$ and the SIMCOMP similarity score for $SimC$ in our experiments. $\gamma_{C_{in}}, \gamma_{C_{out}}, \gamma_E$ are relative weights of each component in reaction similarity calculation. We set $\gamma_{C_{in}} = 0.3, \gamma_{C_{out}} = 0.3, \gamma_E = 0.4$ to balance the effect of compounds and enzymes on reaction similarity. β_C, β_E are relative weights of compounds and enzymes in overall similarity score and they are set to $\beta_C = 0.5, \beta_E = 0.5$.

5.1. Biological Significance

Our first experiment focuses on the biological significance of the found alignments. An alignment should reveal functionally similar entities or sub-paths between different pathways. More specifically, it is desirable to match the entities that can substitute each other or the sub-paths that serve similar functions. We use pathway pairs which are known to contain not identical but functionally similar entities or sub-paths in this experiment.

Alternative Enzymes: Two enzymes are called *alternative enzymes*, if they catalyze two reactions with different input compounds that produce a specific target compound. Similarly, we name these reactions as *alternative reactions* and their inputs as *alternative compounds*. Identifying alternative entities is important and useful for various applica-

Table 1. Alternative enzymes that catalyze the formation of a common product using different compounds. [1]Pathways: 00620-Pyruvate metabolism, 00252-Alanine and aspartate metabolism, 00860-Porphyrin and chlorophyll metabolism. [2]Organism pairs that are compared. [3]KEGG numbers of aligned reaction pairs. [4]EC numbers of aligned enzyme pairs. [5] Aligned compounds pairs are put in the same column. Common target products are underlined. Alternative input compounds are shown in **bold**. Abbreviations of compounds: MAL, malate; FAD, Flavin adenine dinucleotide; OAA, oxaloacetate; NAD, Nicotinamide adenine dinucleotide; Pi, Orthophosphate; PEP, phosphoenolpyruvate; Asp, L-Aspartate; Asn, L-Aspargine; Gln, L-Glutamine; PPi, Pyrophosphate; Glu, L-Glutamate; AMP, Adenosine 5-monophosphate; CPP, coproporphyrinogen III; PPHG, protoporphyrinogen; SAM, S-adenosylmethionine; Met, L-Methionine.

P. Id[1]	Organism[2]	Reaction								
		R. Id[3]	Enzyme[4]	Compounds[5]						
00620	*S. aureus*	R01257	EC:1.1.1.96	MAL	+ ***FAD***		→	OAA	+ $FADH_2$	
	H. sapiens	R00342	EC:1.1.1.37	MAL	+ ***NAD***		→	OAA	+ NADH	
00620	*A. thaliana*	R00345	EC:4.1.1.31	OAA	+ ***Pi***		→	PEP	+ CO_2	
	S. aureus	R00341	EC:4.1.1.49	OAA	+ ***ATP***		→	PEP	+ CO_2	+ ADP
00252	*C. hydro.*	R00578	EC:6.3.5.4	Asp	+ ATP	+ ***Gln***	→	Asn	+ AMP	+ PPi
	C. parvum	R00483	EC:6.3.1.1	Asp	+ ATP	+ $\boldsymbol{NH_3}$	→	Asn	+ AMP	+ Glu
00860	*S. aureus*	R06895	EC:1.3.99.22	CPP	+ $\boldsymbol{O_2}$		→	PPHG	+ CO_2	
	H. sapiens	R03220	EC:1.3.3.3	CPP	+ ***SAM***		→	PPHG	+ CO_2	+ Met

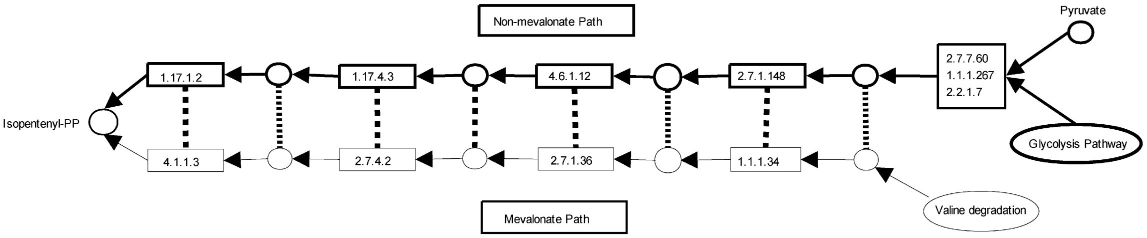


Fig. 4. Identification of alternative sub-paths: A portion of the metabolic pathway of steroid biosynthesis from KEGG. *H.sapiens* produces Isopentenyl-PP via the lower path which is called Mevalonate Path. However, *E.coli* uses a totally different path called Non-mevalonate Path, for producing Isopentenyl-PP which is shown in bold. Using our algorithm, we align the Steroid biosynthesis pathways of *H.sapiens* and *E.coli.* We illustrate the resulting matchings of entities by dashed lines. Compound names are omitted for simplicity.

tions. Some examples are, metabolic reconstruction of newly sequenced organisms[4] and identification of drug targets[3, 18, 19].

We test our tool to search for well-known alternative enzymes presented in Kim *et al.*[20] Table 1 illustrates four cases in which our algorithm successfully identifies alternative enzymes, with the corresponding reaction mappings. Furthermore, resulting compound matchings are consistent with the alternative compounds proposed in Kim *et al.* For instance, there are two different reactions that generate Asparagine (Asn) from Aspartate (Asp) as seen in Table 1. One is catalyzed by aspartate ammonia ligase (EC:6.3.1.1) and uses Ammonium (NH_3) directly, whereas the other is catalyzed by transaminase (EC:6.3.5.4) that transfers the amino group from Glutamine (Gln). We compare the Alanine and aspartate pathways (00252) of two organisms that use the two different routes. Our algorithm aligns the alternate reactions, enzymes and compounds correctly. Our alignment results for the other 3 examples in Table 1 are also consistent with the experimental results, see 20.

Alternative Paths: As metabolic pathways are experimentally analyzed, it is discovered that different organisms may produce the same compounds by totally different paths. Experimental identification provide us well documented examples of such alternative paths. We use our algorithm to identify these known alternative paths in metabolic pathways.

It is shown that, two alternative paths for Isopentenyl-PP production in different organisms exist[21]. Figure 4 illustrates these paths and the entity mappings found by our algorithm. Despite the fact that EC numbers of aligned enzymes are totally different, which indicates that their initial pairwise similarity scores are 0, our algorithm aligns these functionally similar paths since it also accounts for the topological similarities of pathways.

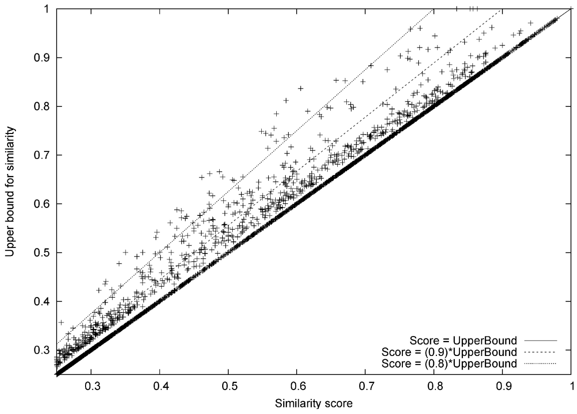


Fig. 5. Effect of consistency restriction on alignment scores: Similarity scores of alignments with consistency restriction and upper bound on the similarity of corresponding pathways without any restriction are shown for pairs of 15,000 randomly selected pathways. Scores below 0.25 are discarded as they indicate dissimilar pathways.

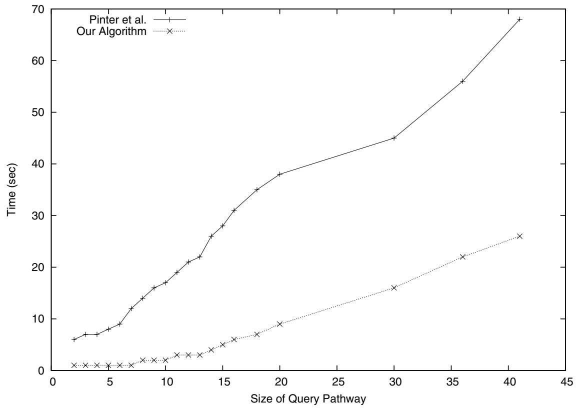


Fig. 6. Running time comparison of our method and the method of Pinter *et al.*: Pathways of varying size are queried against a database of pathways. Total time for each query including IO operations and unexpectedness calculations are plotted for each pathway size. Pathway size is measured as the number of enzymes in the pathway.

Since our method finds one-to-one mappings, only four of seven enzymes in the Non-mevalonate path are mapped to four enzymes of the Mevalonate path. A future work would be to relax the restriction that mappings should be one-to-one. That way alternative paths with different numbers of entities would be aligned without individual entity mappings.

5.2. Effect of Consistency

In order to output meaningful alignments, we report the alignments that are induced by consistent mappings. We ensure the consistency of an alignment by restricting entity mappings to reachable entities. This restriction is necessary for filtering out nonsensical mappings that degrade the accuracy of the alignment. We compute an upper bound to the loss of similarity score due to consistency restriction. We find upper bounds on similarities for each alignment by removing the consistency restriction. This is done by ignoring the pruning phase, which is described in Section 4.4.

Figure 5 demonstrates the effect of consistency restriction on similarity score. For 91 % of the alignments the similarity score found by consistency restriction is not less than 90 % of the upper bound score. Alignments with similarity scores not less than 80 % of the upper bound score constitute 98.5 % of all pathways. Hence, the loss of similarity score due to consistency restriction is not significant.

5.3. Running Time

As discussed theoretically in Section 4.3, our algorithm is guaranteed to find entity mappings with a high convergence rate. We implement the proposed algorithm in C programming language and compare its performance with an existing metabolic pathway alignment tool designed by Pinter *et al.*[8].

The graph model of Pinter *et al.* oversimplifies the metabolic pathways in two ways. First, they totally discard the compounds and reactions from the pathway and use only enzymes. Second, they ignore some interactions between enzymes to have acyclic graphs. Generally, they map a pathway with n enzymes to a graph with n nodes and $n-1$ edges. Since we refuse to have any kind of abstraction, the graph size for the same pathway is considerably larger in our model. For example, Folate biosynthesis pathway of *E.coli* has 12 enzymes. Their simplified model represent this pathway as a graph with 12 nodes and 11 edges, whereas in our graph model the *same* pathway is represented by 55 nodes (22 reactions, 12 enzymes, 21 compounds) and 84 edges. Since we measure the pathway size by the number of enzymes in this experiment, these two pathways are considered to be of the *same size.* Although our algorithm builds a larger graph, Figure 6 shows that our algorithm still runs significantly faster for all pathway sizes. Our method is at least three times faster than the method of Pinter *et al.* for all test cases.

5.4. Statistical Significance

To evaluate the statistical significance of the alignments found by our method, we calculate z-score for each alignment. We generate a number of random pathways for each alignment by shuffling the labels of the entities of query pathways. Label shuffling corresponds to randomly switching the rows of support matrices of each entity type.

Our experiments show that alignment of same metabolic pathways in different organisms create higher z-scores than different pathways in the same or different organisms. In a specific organism pathways that have similar functions, such as different amino acid metabolisms, give higher z-scores than pathways that belong to different functional groups. Due to space constraints we do not present any results for this part.

6. CONCLUSION

In this paper, we considered the pairwise alignment problem for metabolic pathways. We developed a method that aligns reactions, compounds and enzymes. In our algorithm, we considered both the homology and the topology of pathways. We enforced the consistency of the alignment by considering the reachability sets of the aligned entities. Using maximum weight bipartite matching, we first extracted reaction mappings. Then, we enforced the consistency by applying a pruning technique and we extract the mappings for enzymes and compounds. Our experiments showed that, our method is capable of finding biologically and statistically significant alignments very quickly.

References

1. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, and Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *NAR*, 27(1):29–34, 1999.
2. Keseler IM, Collado-Vides J, Gama-Castro S, Ingraham J, Paley S, Paulsen IT, Peralta-Gil M, and Karp PD. EcoCyc: a comprehensive database resource for Escherichia coli. *NAR*, 33:334–337, 2005.
3. Sridhar P, Kahveci T, and Ranka S. An iterative algorithm for metabolic network-based drug target identification. In *PSB*, volume 12, pages 88–99, 2007.
4. Francke C, Siezen RJ, and Teusink B. Reconstructing the metabolic network of a bacterium from its genome. *Trends in Microbiology*, 13(11):550–558, November 2005.
5. Clemente JC, Satou K, and Valiente G. Finding Conserved and Non-Conserved Regions Using a Metabolic Pathway Alignment Algorithm. *Genome Informatics*, 17(2):46–56, 2006.
6. Dost B, Shlomi T, Gupta N, Ruppin E, Bafna V, and Sharan R. QNet: A Tool for Querying Protein Interaction Networks. In *RECOMB*, pages 1–15, 2007.
7. Tohsato Y and Nishimura Y. Metabolic Pathway Alignment Based on Similarity between Chemical Structures. *IPSJ Digital Courier*, 3, 2007.
8. Pinter RY, Rokhlenko O, Yeger-Lotem E, and Ziv-Ukelson M. Alignment of metabolic pathways. *Bioinformatics*, 21(16):3401–8, 2005.
9. Koyutürk M, Grama A, and Szpankowski W. An efficient algorithm for detecting frequent subgraphs in biological networks. In *ECCB*, pages 200–207, 2004.
10. Singh R, Xu J, and Berger B. Pairwise Global Alignment of Protein Interaction Networks by Matching Neighborhood Topology. In *RECOMB*, pages 16–31, 2007.
11. Kleinberg JM. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604.
12. Dutkowski J and Tiuryn J. Identification of functional modules from conserved ancestral protein interactions. *Bioinformatics*, 23(13):i149–158, 2007.
13. Koyutürk M, Grama A, and Szpankowski W. Pairwise Local Alignment of Protein Interaction Networks Guided by Models of Evolution. In *RECOMB*, pages 48–65, 2005.
14. Webb EC. *Enzyme nomenclature 1992* . Academic Press, 1992.
15. Tohsato Y, Matsuda H, and Hashimoto A. A Multiple Alignment Algorithm for Metabolic Pathway Analysis Using Enzyme Hierarchy. In *ISMB*, pages 376–383, 2000.
16. 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. *J Am Chem Soc*, 125(39):11853–11865, 2003.
17. Haveliwala TH and Kamvar SD. The Second Eigenvalue of the Google Matrix. *Stanford University Technical Report*, March 2003.
18. Sridhar P, Song B, Kahveci T, and Ranka S. Mining metabolic networks for optimal drug targets. pages 291–302, 2008.
19. Song B, Sridhar P, Kahveci T, and Ranka S. Double Iterative Optimization for Metabolic Network-Based Drug Target Identification. *International Journal of Data Mining and Bioinformatics*, 2007.
20. Kim J and Copley SD. Why Metabolic Enzymes Are Essential or Nonessential for Growth of Escherichia coli K12 on Glucose. *Biochemistry*, 46(44), 2007.
21. 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*, November 2006.