BIOINFORMATICS

Vol. 29 ISMB/ECCB 2013, pages i145–i153
doi:10.1093/bioinformatics/btt235

CAMPways: constrained alignment framework for the comparative analysis of a pair of metabolic pathways

Gamze Abaka[1], Türker Bıyıkoğlu[2] and Cesim Erten[1,*]

[1]Department of Computer Engineering, Kadir Has University, Cibali, Istanbul 34083 and [2]Department of Mathematics, Izmir Institute of Technology, Izmir 35430, Turkey

ABSTRACT

Motivation: Given a pair of metabolic pathways, an alignment of the pathways corresponds to a mapping between similar substructures of the pair. Successful alignments may provide useful applications in phylogenetic tree reconstruction, drug design and overall may enhance our understanding of cellular metabolism.

Results: We consider the problem of providing one-to-many alignments of reactions in a pair of metabolic pathways. We first provide a constrained alignment framework applicable to the problem. We show that the constrained alignment problem even in a primitive setting is computationally intractable, which justifies efforts for designing efficient heuristics. We present our Constrained Alignment of Metabolic Pathways (CAMPways) algorithm designed for this purpose. Through extensive experiments involving a large pathway database, we demonstrate that when compared with a state-of-the-art alternative, the CAMPways algorithm provides better alignment results on metabolic networks as far as measures based on same-pathway inclusion and biochemical significance are concerned. The execution speed of our algorithm constitutes yet another important improvement over alternative algorithms.

Availability: Open source codes, executable binary, useful scripts, all the experimental data and the results are freely available as part of the Supplementary Material at http://code.google.com/p/campways/.

Contact: cesim@khas.edu.tr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Metabolic pathways consisting of metabolites, biochemical reactions transforming a set of metabolites to others and enzymes catalyzing these reactions provide valuable information regarding material processing centers of a functioning cell and cellular metabolism in general. Several online databases including KEGG (Kanehisa *et al*., 2012) and BioCyc (Caspi *et al*., 2008) provide access to metabolic pathways of various organisms. A comparative analysis of pathways from different organisms provides insights for understanding evolution, speciation, phylogenic reconstruction (Mithani *et al*., 2011; Heymans and Singh, 2003) and drug target discovery (Guimerà *et al*., 2007). Pharmaceutical drug testing is usually implemented on animals, most of the time on mice, before human testing. In such an application, it is usually crucial to know whether specific pathway components of the two species exhibit similar properties (Caglic *et al*., 2009). A successful pathway alignment would prove useful for determining whether test results on one species could be transferred to another without incurring complications. Furthermore, such an analysis is not limited to that between different organisms. It may also be applied between pathways of cancer types and those of healthy cell types to enhance our understanding of cancer-specific metabolic features (Agren *et al*., 2012).

A common method for comparative analysis of pathways and biological networks in general is through network alignment. Given a pair of biological networks either from different species or from different tissues within the same species, the goal of network alignment is to map components in one of the networks to their similar counterparts in the other. With regard to alignments targeting specifically metabolic pathways, several methods have been suggested. In Tohsato *et al*. (2000), an alignment method based on enzyme hierarchies and enzyme EC number similarity was suggested for the alignment of possibly more than two pathways. Path matching and graph matching to query certain metabolic pathways in an input graph was provided by Yang and Sze (2007). Sets of reactions in multiple pathways were compared, omitting the connectivity between the reactions in Clemente *et al*. (2007). Heymans and Singh (2003) created an enzyme graph and obtained a one-to-one mapping between the enzymes of two input pathways via maximum weight bipartite matching. Similar enzyme graph construction was used in Pinter *et al*. (2005). An integer quadratic programming-based method was suggested by Zhenping *et al*. (2007). Similar to metabolic pathway alignment is the problem of protein–protein interaction (PPI) network alignment. The graph models used in the latter are undirected, whereas the former usually aligns directed graphs. However, as far as general graph matching and alignment is concerned, most of the time, the techniques can be extended in both directions, and mainly similar approaches are proposed. Two versions of network alignment have been suggested in related work. In local network alignment, the goal is to identify from the input networks, subnetworks that closely match in terms of network topology and/or sequence similarities. Approaches proposed for this version of the problem include PathBLAST (Kelley *et al*., 2004), NetworkBLAST (Sharan *et al*., 2005), MaWISh (Koyutürk *et al*., 2006) and Graemlin (Flannick *et al*., 2006). In global network alignment on the other hand, the goal is to align the networks as a whole, providing unambiguous mappings between the nodes of different networks. Starting with IsoRank (Singh *et al*., 2008), several global network algorithms using similar definitions have been suggested (Aladağ and Erten, 2013; Chindelevitch *et al*., 2010; Kuchaiev and Pržulj, 2011; Zaslavskiy *et al*., 2009).

We provide a constrained alignment framework and a metabolic pathway alignment algorithm, CAMPways. Our algorithm is inspired by the model suggested in Ay *et al*. (2011). Within this

*To whom correspondence should be addressed.

© The Author 2013. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/29/13/i145/197346 by Kadir Has University user on 13 November 2019

general model, the goal is to find a global one-to-many alignment of the pathways such that a node may be mapped to a connected subgraph of many nodes. The model is justified by the fact that biologically meaningful mappings may exist when different organisms perform the same function through varying number of steps. Therefore, it appropriately handles the gaps/mismatches inherent in alignment problems, an issue arising in both sequence-related and network-wise alignment. Such is the motivation behind the PPI network alignment approach of Liao *et al.* (2009) as well. Although this general model of one-to-many alignments is the same, our method diverges from that of Ay *et al.* (2011) after this point. The novelties of the current work are 3-fold. First of all we provide a novel *constrained alignment framework* appropriate for the one-to-many alignments model. This framework has not been used in biological network alignment previously. Second, we show that even the simplest version of the alignment problem within this framework is computationally hard. Based on this computational intractability result, we finally provide a novel algorithm, CAMPways, which appropriately and efficiently implements this framework. Through experimental evaluations based on reverse engineering pathways and biochemical significance measured through functional group conversion hierarchy of KEGG (Kanehisa *et al.*, 2012), we demonstrate that the CAMPways algorithm provides higher quality alignments than the state-of-the-art approaches. Furthermore, a second major advantage of the CAMPways algorithm is in terms of its much faster execution speeds as compared with the alternatives.

2 METHODS AND ALGORITHMS

2.1 Problem definition

The metabolic pathway alignment problem definition we consider is based on the one-to-many alignments of the *reaction-based* pathway representations used in (Ay *et al.*, 2011). Given a metabolic pathway $\mathcal{P}$, we assume a *reaction-based* representation $G_P = (V_P, E_P)$ of $\mathcal{P}$. G_P is a directed graph where each node $u_{r_i} \in V_P$ corresponds to a reaction r_i in $\mathcal{P}$. There exists a directed edge (u_{r_i}, u_{r_j}) if an output compound of r_i is an input compound of r_j. If r_i is reversible, the edge existence condition is extended by considering the case where an input compound of r_i becomes an input compound of r_j. Similar extension applies to r_j as well. Thus, if both reactions are reversible, there are in total four cases for the existence of an edge.

Given two pathway representations G_P, G'_P, we need to formalize the types of mappings that are allowed under the one-to-many mapping restrictions. Let R_x indicate a subset of V_P such that the induced subgraph of the nodes in R_x is connected in the underlying undirected graph. Denote the set of all such subsets of size greater than zero and less than or equal to k with $\mathcal{R}_k$. Let $\mathcal{R}'_k$ denote the analogous set for G'_P. A *legal alignment* $\mathcal{A}$ between G_P, G'_P is a set of mappings (R_x, R'_x) for $R_x \in \mathcal{R}_k$, $R'_x \in \mathcal{R}'_k$ such that the following are satisfied:

(i) For $(R_x, R'_x) \in \mathcal{A}$, $|R_x|$ or $|R'_x|$ is 1.
(ii) For $(R_x, R'_x) \in \mathcal{A}$ and $(R_y, R'_y) \in \mathcal{A}$, $R_x \cap R_y = \emptyset$ and $R'_x \cap R'_y = \emptyset$.

The first condition implies that all mappings in the alignment are one-to-many mappings, whereas the second implies that all mappings are pairwise compatible in the sense that no reaction from a given pathway may belong to more than one mapping. The quality of an alignment is usually defined in terms of two possibly conflicting measures; *homological similarity* and *topological similarity*. The former can be defined as a sum of homology scores of all mappings in the alignment. The homology score of a given mapping (R_x, R'_x) can be defined in terms of the similarities of input compounds, output compounds and enzymes of R_x and R'_x. Such similarity scores are usually determined as a result of sequential similarity analysis of the molecules under consideration (enzymes or input/output compounds). For the current study, we use the homological similarity scores produced by Ay *et al.* (2011). For a given mapping (R_x, R'_x), first E_x, E'_x, which correspond to the unions of all enzymes involved in the reactions subsets R_x and R'_x, respectively, are produced. An enzymatic homological similarity between E_x, E'_x can be computed by creating a bipartite graph where a partition corresponds to the enzymes of E_x and the other to those of E'_x. A similarity score between every pair of enzymes from E_x and E'_x is assigned as the weight of the corresponding edge in the bipartite graph. The homology score between E_x, E_y then corresponds to the maximum weight bipartite matching of the produced graph. Similar constructions can be carried out for the unions of input compounds, I_x, I'_x and the unions of output compounds O_x, O'_x. The homology score of R_x, R'_x is then defined as a convex combination of the scores attained from the scores calculated independently for the enzymes, input compounds and output compounds. Topological similarity on the other hand is a measure of the conservation of network topologies with respect to the given set of mappings in the alignment. Given a pair of mappings $(R_x, R'_x) \in \mathcal{A}$ and $(R_y, R'_y) \in \mathcal{A}$, a *conserved edge* is induced by this pair if there exists an edge from a reaction in R_x to a reaction in R_y and an edge from a reaction in R'_x to a reaction in R'_y, or vice versa. Topological similarity is then defined as a score proportional to the number of conserved edges induced by the pairs of mappings in the alignment. Once both types of similarity scores are resolved, the network alignment problem is usually posed as that of maximizing a convex combination of these two scores.

2.2 Constrained alignment framework

We provide a formal description of our constrained alignment framework within the provided one-to-many pathway alignments model. Rather than posing the problem as one of a simultaneous optimization of two possibly conflicting goals, that is, that of homological similarity and of topological similarity, we propose a framework where the only goal is to maximize topological similarity while satisfying some constraints on homological similarity.

Given a pathway representation $G_P = (V_P, E_P)$ let G_p^k denote the *kth extension* of G_p. It is a directed, edge-weighted graph. Each node u_{R_x} in G_p^k corresponds to a reaction subset $R_x \in \mathcal{R}_k$. There exists a directed edge (u_{R_x}, u_{R_y}) in G_p^k if there exists a directed edge from u_{r_i} to u_{r_j} in G_p, where $r_i \in R_x$ and $r_j \in R_y$. Let $w(u_{R_x}, u_{R_y})$ denote the total number of such edges. We note that $G_p'^k$ can be defined analogously. The set of

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/29/13/i145/197346 by Kadir Has University user on 13 November 2019

constraints of node u_{R_x} in G_p^k, denoted with $Cons(u_{R_x})$, is defined as the subset of nodes of $G_p'^k$ that u_{R_x} can be mapped to. The definition can be extended to the nodes of $G_p'^k$ analogously. Note that this definition is symmetrical in the sense that $u_{R'_y} \in Cons(u_{R_x})$ if and only if $u_{R_x} \in Cons(u_{R'_y})$. Assume $|Cons(u_{R_x})| \leq k_1$ for any node u_{R_x} in G_p^k and $|Cons(u_{R'_y})| \leq k_2$ for any node $u_{R'_y}$ in $G_p'^k$, for fixed constants k_1 and k_2. All constraints can be represented as a bipartite *similarity* graph where the nodes of G_p^k form one partition and those of $G_p'^k$ form the other, and each constraint is represented with an edge in the bipartite graph. The *constrained alignment* problem is that of finding a subset of constraints, that is, a subset of edges from the bipartite similarity graph, such that the subset of edges define a legal alignment and the number of conserved edges induced by the alignment is maximum. It is worth noting that the concept of constrained alignments has appeared in biological network alignment literature before. Zaslavskiy *et al*. (2009) provide a definition of constrained alignments applicable to global one-to-one alignments of PPI networks. We note that our constrained alignment framework may trivially be generalized to undirected PPI networks. Moreover, our framework is more general; it strictly includes the model of Zaslavskiy *et al*. (2009). There are instances that can not be defined using their model, whereas the opposite is never the case. Using our notation, given u_{R_x}, u_{R_y} from one of the networks, if $Cons(u_{R_x}) \cap Cons(u_{R_y}) \neq \emptyset$, their model imposes the condition that $Cons(u_{R_x}) = Cons(u_{R_y})$. Considering the case where the *Cons* definition reflects high-homological similarity, this is restrictive; either long homologically similar chains of nodes are to be created incorrectly or some homologically similar pairs missed completely.

We first state that the constrained alignment problem defined herein is computationally intractable even in a restricted case.

Proposition 2.1. *The constrained alignment problem where* $k = k_1 = 1$ and $k_2 = 3$ *is NP-complete.*

Proof. Because of space considerations, the proof is provided in the Supplementary Document. We simply state that as the proof works for the undirected graphs as well, the same theorem can immediately be applied to the constrained pairwise alignment of PPI networks. ■

To provide further depth to our understanding of the problem within the constrained alignment framework, we next state the following proposition, which may suggest a clue as to the point the computational intractability starts dissolving.

Proposition 2.2. *The constrained alignment problem where* $k = k_1 = 1$ *and* k_2 *any positive integer constant is polynomially solvable if one of the directed graphs* G_p *or* G'_p *is acyclic.*

Proof. Because of space considerations, the proof is left to the Supplementary Document. ■

2.3 The CAMPways alignment algorithm

Although Proposition 2.2 provides a positive result, it is restrictive to be useful in practice. We provide a more general algorithm that although may not find the optimum in all cases, will in general produce high-quality alignments. Assuming G_p^k, $G_p'^k$, the constants k_1, k_2, and a homological similarity value between the pair $(u_{R_x}, u_{R'_y})$ for any node u_{R_x} in G_p^k and any node $u_{R'_y}$ in $G_p'^k$, the algorithm consists mainly of three steps. These major steps are depicted in Figure 1 on a sample input pathway pair.

Step1-Constructing the bipartite Similarity Graph: This step involves the construction of $Cons(u_{R_x})$ for every node u_{R_x} in G_p^k such that $|Cons(u_{R_x})| \leq k_1$ and $Cons(u_{R'_y})$ for any node $u_{R'_y}$ in $G_p'^k$ such that $|Cons(u_{R'_y})| \leq k_2$. Assuming an edge-weighted bipartite graph on the set of nodes of G_p^k in one partition and those of $G_p'^k$ in the other, where each weight represents the homological similarity of the pair of nodes, a reasonable goal is to find out a subset of edges that satisfies the degree constraints k_1, k_2 and that maximizes the sum of edge weights in the output subset; see Figure 1 where the weight is depicted through the thickness of bipartite graph edges in the similarity graph. The problem then turns into that of *b-matching* (or the *degree constrained subgraph problem*), which has been studied fairly well starting with the pioneering work of Edmonds (1965). Polynomial time solutions, including appropriate modifications of the network flow algorithms (Gabow, 1983) and belief propagation methods (Bayati *et al*., 2011), have been suggested. For efficiency considerations, we choose to use a simple greedy algorithm for this step. Each time the algorithm selects the heaviest edge that does not violate the degree constraints k_1, k_2 for neither of the end points and extends the output set with the edge. The algorithm stops when there are no more edges to consider, and the bipartite graph resulting from the output set of edges is the similarity graph, S.

Step2-Conflict Graph Generation and Conflict Resolution: Assume the bipartite similarity graph S is extended with the directed edges of G_p^k, $G_p'^k$, that is, directed edge (u_{R_x}, u_{R_y}) is inserted in S for u_{R_x} and u_{R_y} in G_p^k, if (u_{R_x}, u_{R_y}) is an edge in G_p^k. Analogous extensions apply to edges of $G_p'^k$. We construct an undirected node-weighted *conflict* graph $\mathcal{C}$, where each node corresponds to a set of four nodes providing a conserved edge in the extended graph S. More precisely, in the conflict graph, there is a node corresponding to 4−tuple $\prec u_{R_x}, u_{R_y}, u_{R'_x}, u_{R'_y} \succ$ if and only if all of the following hold:

(i) $R_x \cap R_y = \emptyset$ and $R'_x \cap R'_y = \emptyset$.
(ii) Either (u_{R_x}, u_{R_y}), $(u_{R'_x}, u_{R'_y})$ are in G_p^k, $G_p'^k$, respectively, or (u_{R_y}, u_{R_x}), $(u_{R'_y}, u_{R'_x})$ are in G_p^k, $G_p'^k$, respectively.
(iii) $\{u_{R_x}, u_{R'_x}\}$, $\{u_{R_y}, u_{R'_y}\}$ are undirected edges in S.

Denote such a 4−tuple with a c_4, as the underlying undirected subgraph induced on the four nodes gives rise to a 4−cycle. A weight of 1 is assigned to the c_4s satisfying only one part of condition *ii,* and a weight of 2 is assigned to those satisfying both parts of *ii*. It should be clear that each c_4 node in the conflict graph represents a pair of reaction subset mappings that gives rise to at least one conserved edge. Furthermore, the weight of the node provides the number of edges conserved as a result of the pair of mappings. The conflict graph depicted in Figure 1 is the exact conflict graph corresponding to the partially depicted extended similarity graph in the figure. Note that although the structure of the 4−tuple $\prec u_{R_9}, u_{R_2}, u_{R'_5}, u_{R'_6} \succ$ resembles that of a c_4, that is, conditions *ii* and *iii* defined earlier in the text are valid, it does not correspond to a node in the conflict graph, as condition *i* is not satisfied. Regarding the weights, it should be noted that the node $\prec u_{R_1}, u_{R_9}, u_{R'_4}, u_{R'_5} \succ$ has weight two, and the rest has weight one in the conflict graph depicted in the figure.

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/29/13/i145/197346 by Kadir Has University user on 13 November 2019

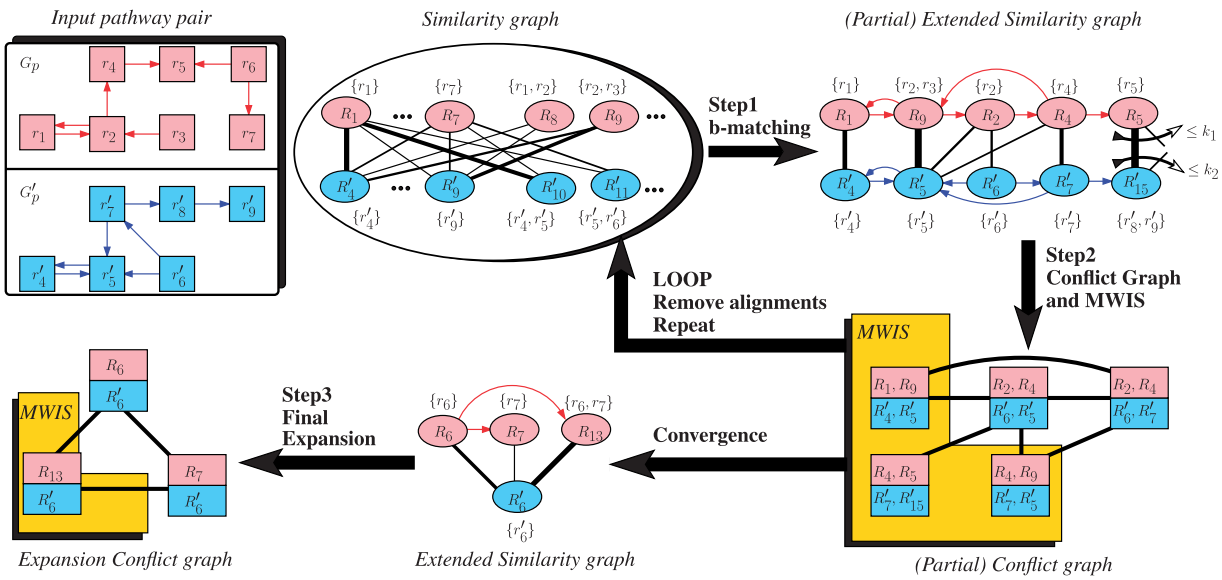

Fig. 1. CAMPways algorithm depicted on a sample input for $k=2$; the final alignment includes 1-to-1 and 1-to-2 mappings of reactions. First step involves *b-matching*; degrees of nodes are bounded by k_1 or k_2 depending on the partition they belong to in the similarity graph. Only a small representative portion of the extended similarity graph is shown. The conflict graph arising from this portion is shown exactly. All the alignments in the *MWIS* boxes of the loops in *Steps 1* and *2* and in the *MWIS* box of the final expansion step are included in the output alignment. Note that the conflict graph definitions within the loops and that of the final expansion phase are different

Let $C_1 = \prec u_{R_x}, u_{R_y}, u_{R'_x}, u_{R'_y} \succ$, $C_2 = \prec u_{R_w}, u_{R_z}, u_{R'_w}, u_{R'_z} \succ$ and let $S_1 \in \{R_x, R_y\}, S_2 \in \{R_w, R_z\}$ and $S'_1 \in \{R'_x, R'_y\}, S'_2 \in \{R'_w, R'_z\}$. For a c_4 C_i, let $\mathcal{M}_{C_i}(u)$ indicate the neighbor of u in C_i from the opposite network. There exists an edge between the nodes corresponding to the two c_4s in the conflict graph if and only if at least one of the following holds:

(i) $\exists S_1, S_2$ such that $S_1 \neq S_2$ and $S_1 \cap S_2 \neq \emptyset$.
(ii) $\exists S'_1, S'_2$ such that $S'_1 \neq S'_2$ and $S'_1 \cap S'_2 \neq \emptyset$.
(iii) $\exists S_1, S_2$ such that $S_1 = S_2$ and $\mathcal{M}_{C_1}(S_1) \neq \mathcal{M}_{C_2}(S_2)$.
(iv) $\exists S'_1, S'_2$ such that $S'_1 = S'_2$ and $\mathcal{M}_{C_1}(S'_1) \neq \mathcal{M}_{C_2}(S'_2)$.

This construction implies that an edge exists between a pair of c_4s if and only if the pair of conserved edges represented by the c_4s can not coexist in any legal alignment. For the conflict graph of Figure 1 for instance, the edge between the c_4s $\prec u_{R_1}, u_{R_9}, u_{R'_4}, u_{R'_5} \succ$ and $\prec u_{R_2}, u_{R_4}, u_{R'_6}, u_{R'_7} \succ$ is due to condition *i*; reaction subsets R_9 and R_2 share a reaction. Therefore, no legal alignment can include both of the corresponding conserved edges. On the other hand, the edge between $\prec u_{R_4}, u_{R_5}, u_{R'_7}, u_{R'_{15}} \succ$ and $\prec u_{R_2}, u_{R_4}, u_{R'_6}, u_{R'_5} \succ$ is due to *iii*. Simultaneously conserving both edges corresponding to both c_4s, R_4 would have to be mapped to two different reaction subsets, which is not possible in any legal alignment by definition. The discussion regarding the conflict graph construction leads to the following proposition:

PROPOSITION 2.3. *The maximum weight independent set (MWIS) of $\mathcal{C}$ provides an optimum solution to the constrained alignment problem.*

However, some modifications are necessary to make our conflict graph model more useful in practical applications of the constrained alignment framework. First, each node in the conflict graph may not necessarily have an exact binary contribution, that is, 1 or 2 to the quality of the final alignment. Therefore, we propose appropriate generalizations for the weights of conflict graph nodes. We provide two alternative weighting schemes. For a given edge e in the similarity graph S, let $w_S(e)$ denote the weight of e, which reflects the homological similarity of the reaction subsets corresponding to the end points of e. For $C_1 = \prec u_{R_x}, u_{R_y}, u_{R'_x}, u_{R'_y} \succ$, the first scheme, denoted with $\mathcal{W}_1$, assigns a weight of $\alpha \times H(C_1) + (1-\alpha) \times I(C_1)$, where

$$H(C_1) = \frac{1}{2} \times (w_S(u_{R_x}, u_{R'_x}) + w_S(u_{R_y}, u_{R'_y}))$$

$$I(C_1) = \frac{1}{2(k^2+1)} \times \sum_{\substack{i,j \in \{u_{R_x}, u_{R_y}\}, i \neq j \\ i',j' \in \{u_{R'_x}, u_{R'_y}\}, i' \neq j'}} w(i,j) + w(i',j')$$

For the computation of $I(C_1)$, the total number of directed edges between R_x, R_y and between R'_x, R'_y is normalized with the maximum number of possible directed edges $G_p^k, G_p'^k$ in any c_4. The parameter α is a balancing parameter between the weight of homological similarity and that of conserved interactions. Our second weighting scheme does not check the number of conserved edges; as long as there is at least one conserved edge, the contribution of edge conservation remains the same. On the other hand, depending on the evolutionary distance of the organisms providing the input pathways, it might be more meaningful do differentiate between the alignments yielding *one-to-many* mappings as opposed to those providing *one-to-few* mappings. Therefore, for the second scheme, denoted with $\mathcal{W}_2$, we introduce additional input parameters $\alpha_1, \alpha_2 \ldots \alpha_k$ such that $\alpha_1 + \alpha_2 + \cdot s + \alpha_k = 1$. Each α_i reflects the relative importance of the *one-to-i* mappings in the complete alignment. Without loss of

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/29/13/i145/197346 by Kadir Has University user on 13 November 2019

generality, let $|R_x| >= |R'_x|$ and $|R_y| >= |R'_y|$. The weight of $C_1 = \prec u_{R_x}, u_{R_y}, u_{R'_x}, u_{R'_y} \succ$ is defined as $\alpha_{|R_x|} \times |R_x| + \alpha_{|R_y|} \times |R_y|$.

A second issue is related to resolving conflicts, that is, the computation of the MWIS of the conflict graph. The problem is NP-complete in general (Garey and Johnson, 1979). Several greedy heuristics have been investigated in Sakai *et al.* (2003). We implemented each and applied extensive tests to determine their performances. The *GWMIN2* heuristic, which selects the node u in the conflict graph $\mathcal{C}$ that maximizes $\mathcal{W}(u)/\sum_{v \in N^+_{\mathcal{C}}(u)} \mathcal{W}(v)$, where $N^+_{\mathcal{C}}(u)$ denotes the neighborhood of u in $\mathcal{C}$ together with the node u itself, provided better results than the rest. Furthermore, it provides a theoretical guarantee that the weight of the output independent set is at least $\sum_{u \in V_{\mathcal{C}}} [\mathcal{W}(u)^2 / \sum_{v \in N^+_{\mathcal{C}}(u)} \mathcal{W}(v)]$, where $V_{\mathcal{C}}$ denotes the vertex set of the conflict graph $\mathcal{C}$. Therefore, we chose to implement this part of our algorithm using this heuristic.

Finally, we note that the resulting mappings are those limited to the edges of the bipartite similarity graph S constructed after *Step1*. To enlarge the alignment, we remove all mapped nodes from G^k_p, G'^k_p after the execution of *Step1* and *Step2*, restore all the homological similarity edges and repeat both steps. This whole process is iterated until convergence, that is, the conflict graph $\mathcal{C}$ generated after *Step2* becomes empty. For the example pathway alignment of Figure 1, the loop iterates only once; the remaining extended similarity graph contains nodes defined on reaction subsets R_6, R_7, R_{13} and R'_6, which gives rise to an empty conflict graph.

Step3-Final Alignment Expansion: The iterative process involving the first two steps aforementioned produces mappings based on 4−tuples because of the conserved interaction maximization goal of the constrained alignment framework. The convergence of the process implies that no more conserved interactions can be attained. However, there may still exist potential mappings with high-homological similarity that might be added to the alignment. To implement such an expansion, we first remove all the mapped nodes from G^k_p, G'^k_p and restore all homological similarity edges. Considering the resulting similarity graph S, we create a new type of a conflict graph, called the *expansion conflict graph*. Each node in the expansion conflict graph corresponds to a 2−tuple $\prec u_{R_x}, u_{R'_x} \succ$ such that $\{u_{R_x}, u_{R'_x}\}$ is an edge in S. There is an edge between two nodes of this conflict graph if and only if the intersection of their reaction subsets coming from the same pathway is non-empty; see Figure 1 for the expansion conflict graph generation on the sample pathways. Note that the conflict graph defined in *Step 2* is conceptually different from the expansion conflict graph of this step. We finally apply the *GWMIN2* heuristic to resolve the conflicts in the expansion conflict graph, and the alignment is expanded with the mappings corresponding to the resulting nodes.

3 DISCUSSION OF RESULTS

The CAMPways implementation is in C++ using the LEDA library (Mehlhorn and Naher, 1999). Source code, useful scripts for testing and evaluations, all the data and output results are available as part of the Supplementary Material. We experimented on data from the KEGG database (Kanehisa *et al.*, 2012) as retrieved and reformatted by Ay *et al.* (2011). Our comparative performance evaluations presented in this section are with regards to those achieved in SubMAP (Ay *et al.*, 2011), as the used problem definitions are the same; the goal being one-to-many mappings for an input pair of pathways. We note that although a version of SubMAP using network compression to speed-up the original algorithm has appeared recently (Ay *et al.*, 2012), lack of publicly available implementation made further extensive comparisons with the new version impossible. Nevertheless, it is suggested that the compression-based version is provided mainly for execution performance at the expense of output alignment qualities. Therefore, in terms of alignment qualities, it is sensible to compare CAMPways with SubMAP. According to the reported results of Ay *et al.* (2012), attaining considerable runtime efficiency could cost an accuracy loss of almost 50%, where accuracy is measured in terms of the Pearson's correlation coefficient between the alignment outputs of the compressed version and the original version of SubMAP. Our experimental results on the other hand indicate that not only does our algorithm provide superior runtime efficiency but also achieves this without incurring any cost on accuracy; to the contrary, the alignment outputs provided by CAMPways provide better accuracies than those of the original SubMAP algorithm.

Although the KEGG database provides pathways under detailed metabolism categories, such as *Glycerolipid metabolism* and *Tryptophan metabolism* among many others, directly using these pathways in a network alignment study does not reveal enough information. The most important reason is the lack of a gold standard to be the basis of an objective evaluation of the alignment qualities. Although less serious, the small pathway sizes constitute yet another problem. Predicting the behavior of a possible alignment method at this scale may not lead to reliable conclusions. A mechanism to handle both of these issues is to merge all pathways from detailed metabolism categories that are categorized under the same more general metabolism categories provided in KEGG. Considering the first 11 of the listed high-level categories, we merged all pathways specified under each into a larger metabolic network. This way we obtained 11 metabolic networks in total, each corresponding to one of the following metabolisms: 1.1 carbohydrate metabolism, 1.2 energy metabolism, 1.3 lipid metabolism, 1.4 nucleotide metabolism, 1.5 amino acid metabolism, 1.6 metabolism of other amino acids, 1.7 glycan biosynthesis and metabolism, 1.8 metabolism of cofactors and vitamins, 1.9 metabolism of terpenoids and polyketides, 1.10 biosynthesis of other secondary metabolites and 1.11 xenobiotics biodegradation and metabolism. The number of pathways contained in each larger metabolic network changes between 2 and 15. The subjects of all experimental evaluations of this section are these metabolic networks from pairs of different species.

The next two subsections provide our comparative experimental evaluations with regards to the accuracies of output alignments produced by CAMPways and SubMAP. We used two types of accuracy parameters for this purpose. The first one is based on reverse engineering successes of the output alignments, whereas the second one is based on their biochemical significances in terms of coherence with regards to the functional group conversion categorizations as provided by KEGG. We finally conclude our evaluations by providing a running time analysis of CAMPways and a discussion of experimental results

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/29/13/i145/197346 by Kadir Has University user on 13 November 2019

on observed execution speeds of both algorithms running on networks under consideration.

3.1 Reverse engineering metabolic pathways

The large metabolic networks under consideration can be regarded as networks *engineered* out of small pathways on detailed metabolism categories. A natural accuracy measure is then the reverse engineering capabilities of the provided output alignments; intuitively an alignment mapping reactions that belong back to the same original KEGG pathway is considered to be of high quality. Thus, the pathways on detailed metabolism categories provided by KEGG become our gold standard. Note that this approach assumes the retrieved pathways are noise-free, that is, all pathways in KEGG are considered perfectly correct without any missing data or incorrect pathway associations. Let X, X' denote two species and G_X, G'_X be their metabolic networks corresponding to some metabolism $1.m$, listed earlier in the text. Let $\prec u_{R_x}, u_{R'_x} \succ$ be a mapping from an alignment of G_X, G'_X, where R_x is a subset of reactions from X and R'_x is a subset of reactions from X'. Without loss of generality, let $R_x = \{r_x\}$, that is, it is the subset containing a single reaction in the one-to-many mapping. Let $P_1 \ldots P_x$ be the pathways that include reaction r_x in the set of pathways associated with metabolism $1.m$ in the species X. We call the mapping *correct* if every reaction in the subset R'_x is included in at least one of the pathways $P'_1, \ldots P'_x$ where each P'_i is a pathway in metabolism $1.m$ of species X', corresponding to P_i of X. We divide the experimental evaluations into two; those regarding the alignments between species within the same domain and those between species from different domains. We pick *Homo sapiens (hsa)* and *Mus musculus (mmu)* as the two representative species from the eukaryota domain, and the *Escherichia coli (eco)* and *Agrobacterium tumefaciens (atc)* from bacteria. The value of $k = 3$ is fixed, that is, each reaction from one of the networks may be mapped to at most three reactions from the other. For the CAMPways alignments, we pick $k_1 = k_2 = 3$.

3.1.1 Same-domain alignments The evaluations of the output alignments of *hsa* versus *mmu* and *atc* versus *eco* with regards to all 11 high-level metabolism categories are presented in Table 1. Each multi-row in the table provides the results for the alignments of two pairs of networks for metabolisms 1.1 through 1.11 from top to bottom; the top row at the *mth* multi-row lists the alignment results of the hsa-mmu network pair pertaining to metabolism category $1.m$, and the bottom row lists those of the atc-eco network pair for the same metabolism category. The *TR* column in the table provides the number of total reactions of the network pair. The *coverage* column provides the total number of reactions covered by the mappings in the alignment. The *correct mappings* column provides the number of correct mappings in the alignment, whereas the *ratio* column provides the ratio of the number of correct mappings to the total number of mappings produced by the alignment. In each subcolumn, we indicate the name of the algorithm providing the alignment scores with respect to the parameter provided in the column including it. The subcolumn marked with S provides the corresponding column scores of the alignments produced by SubMAP and the one marked with C_1 provides those of the alignments produced by CAMPways with weighting scheme $\mathcal{W}_1$ and $\alpha = 0.3$. Alignments obtained for other settings of α provide almost the same results as this setting. The subcolumn marked with C_2 provides the corresponding column scores of CAMPways with weighting scheme $\mathcal{W}_2$ and $\alpha_1 = 0.4$, $\alpha_2 = 0.5$ and $\alpha_3 = 0.1$. The coverages of both algorithms are similar; in some instances, coverages of SubMAP are better, whereas in others, both versions of CAMPways provide higher coverage, although in neither case the differences are large. With regard to the number of correct mappings, CAMPways results are overwhelmingly superior to those of SubMAP. For the *atc-eco* alignment of 1.11 xenobiotics biodegradation and metabolism for instance, even though SubMAP provides a much larger coverage than CAMPways (153 versus 134), the number of correct mappings of CAMPways is still better (60 versus 53). This implies that although in some cases SubMAP aggressively creates mappings in favor of covering many reactions, in a lot of the mappings, it provides the mapped reactions that do not share the same pathway. Over all 22 instances, in five instances, SubMAP does not execute until completion because of excessive memory consumption; shown with empty entries in Table 1. For 16 instances, CAMPways provides a larger number of correct mappings, whereas only in one instance, both algorithms provide equal number of correct mappings. The provided ratios also confirm

Table 1. Same-domain reverse engineering experiment

TR	Coverage			Correct mappings			Ratio		
	S	C1	C2	S	C1	C2	S	C1	C2
437	—	*435*	*435*	—	*211*	*213*	—	*0.99*	*0.98*
458	—	416	416	—	166	171	—	0.82	0.83
62	*62*	*62*	*62*	*29*	*31*	*31*	*0.96*	*1*	*1*
116	105	110	110	45	51	51	0.93	0.94	0.94
745	—	*726*	*726*	—	*361*	*361*	—	*0.99*	*0.99*
264	244	254	254	96	105	103	0.82	0.82	0.83
320	—	*320*	*320*	—	*159*	*159*	—	*0.99*	*0.99*
296	280	262	262	110	128	128	0.90	0.98	0.98
496	*491*	*481*	*481*	*221*	*239*	*239*	*0.96*	*0.99*	*0.99*
369	352	340	339	122	143	143	0.79	0.86	0.86
134	*128*	*130*	*130*	*59*	*64*	*64*	*0.96*	*0.98*	*0.98*
108	102	97	97	37	39	39	0.78	0.82	0.82
168	*148*	*168*	*168*	*73*	*76*	*76*	*1*	*0.90*	*0.90*
73	69	64	64	31	31	31	0.96	0.96	0.96
307	—	*306*	*307*	—	*150*	*151*	—	*0.98*	*0.98*
334	325	324	326	129	143	144	0.87	0.89	0.90
31	*28*	*28*	*28*	*12*	*14*	*14*	*1*	*1*	*1*
51	43	43	44	15	17	17	0.78	0.80	0.77
35	*34*	*34*	*34*	*16*	*17*	*17*	*1*	*1*	*1*
23	21	20	20	8	9	9	0.8	0.9	0.9
207	*201*	*200*	*200*	*87*	*100*	*100*	*0.92*	*1*	*1*
175	153	134	134	53	60	60	0.81	0.89	0.89

Note: In each multi-row, the top row lists the hsa-mmu alignment results and the bottom row lists the atc-eco results. The entries of the rows corresponding to the hsa-mmu network pair are italicized for readability purposes. Each multi-row itself provides the results for the alignments of networks for metabolisms 1.1 through 1.11 from top to bottom.

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/29/13/i145/197346 by Kadir Has University user on 13 November 2019

the superiority of CAMPways over SubMAP. Note that the ratio does not normalize the number of correct of mappings with coverage but rather with the total number of output mappings. Thus, it is a measure of the percentage of the correct mappings in the alignment.

3.1.2 Across-domains alignments We repeated the same tests for every pair of species under consideration such that members in the pair belong to different domains giving rise to four pairwise alignment instances per metabolism. Two noteworthy observations arise. First, both the number of correct mappings and the correctness ratios decrease for all alignments as compared with those presented in Table 1. This is in accordance with the intuition that as the divergence of the pair of species increase, any global alignment starts providing more dissimilar mappings, that is, mappings that match reactions from different pathways of the given species. Second, comparing the alignment qualities of the algorithms, the trend is the same as with the same-domain experiments; in almost all cases, CAMPways provides more correct mappings and better correctness ratios. Over all 44 instances, SubMAP is unable to produce results in 20 of them. In seven instances, both algorithms provide equal number of correct mappings. For 16 instances, CAMPways alignments induce more correct mappings, whereas only for a single instance, the correct mapping count of SubMAP is better. The complete table with detailed results of the across-domains setting can be found in the Supplementary Document.

We note that we implemented several tests to determine how the correctness values and the number of 1-to-i mappings for each $i = 1, 2, 3$ in the output alignments of CAMPways change with respect to various $\alpha_1, \alpha_2, \alpha_3$ settings in the $\mathcal{W}_2$ version of the algorithm. Because of space constraints, we provide a detailed discussion regarding these results in the Supplementary Document.

3.2 Biochemical significance of the alignments

To compare the alignment qualities of both algorithms in terms of biochemical significance, we use the functional group conversion (FGC) hierarchy data provided as part of the RCLASS database of KEGG (Kanehisa *et al.*, 2012). The reactions in the database are classified into hierarchically organized functional group categories. The same functional group undergoes the same or similar chemical reaction(s) regardless of the size of the molecule it is a part of (March, 1985). Thus, an inter-species alignment of a pair of pathways is considered biochemically validated if the alignment maps reaction subsets classified under the same FGC category. There are five levels of the KEGG hierarchy where the initial root level consists of eight high-level FGC categorizations: carbon-related, hydrogen-related, isomerization-related, nitrogen-related, oxygen-related, phosphorus-related, sulfur-related and halogen-related. The correctness measure is defined analogous to that used in the previous section; for a fixed level i of the hierarchy, a mapping is called *correct* if there exists at least one category at the *ith* level of the FGC hierarchy that includes all the reactions involved in the mapping. We compare and evaluate the correctness values provided by the alignments of CAMPways and SubMAP algorithms for the first five levels of the hierarchy starting with the root level at $i = 1$.

Table 2. Same-domain biochemical significance experiments

Level 1		Level 2		Level 3		Level 4		Level 5	
S	C	S	C	S	C	S	C	S	C
—	*193*	—	*193*	—	*193*	—	*192*	—	*192*
—	154	—	154	—	151	—	144	—	138
23	*23*	*22*	*23*	*22*	*23*	*21*	*23*	*21*	*22*
32	41	32	41	32	39	32	39	32	39
323	*343*	*323*	*343*	*323*	*343*	*318*	*340*	*316*	*338*
97	105	97	105	97	104	93	103	92	102
—	*103*	—	*103*	—	*101*	—	*101*	—	*101*
66	84	66	84	64	80	64	80	63	80
209	*229*	*209*	*229*	*208*	*229*	*205*	*227*	*205*	*227*
117	143	110	139	104	132	97	130	93	127
53	*57*	*53*	*57*	*52*	*57*	*52*	*57*	*52*	*56*
37	35	37	35	34	33	33	33	33	32
5	*6*	*5*	*6*	*5*	*6*	*5*	*6*	*5*	*6*
20	21	20	21	20	21	20	21	19	21
—	*123*	—	*123*	—	*123*	—	*123*	—	*123*
96	115	94	114	93	111	93	110	90	109
9	*13*	*9*	*13*	*9*	*13*	*9*	*13*	*9*	*13*
16	17	16	16	16	16	15	15	14	15
14	*16*	*14*	*16*	*13*	*16*	*13*	*16*	*13*	*16*
7	9	7	9	7	9	6	8	6	8
79	*97*	*78*	*97*	*76*	*97*	*76*	*97*	*76*	*97*
44	59	44	58	42	55	42	55	42	54

Note: The correspondence of the rows and multi-rows are the same as in Table 1.

As with the experiments of the previous section, we use two types of evaluations; those pertaining to the same-domain alignments and those of the across-domains alignments. The results of the former are presented in Table 2. The used network pairs and the correspondence of rows, multi-rows are the same as in Table 1. The subcolumns marked with S indicate the results of SubMAP alignments and those marked with C indicate results of CAMPways' $\mathcal{W}_1$ version. The $\mathcal{W}_2$ version provides results similar to those of $\mathcal{W}_1$; therefore, they are not included in the table. The main column titles indicate all five levels of the FGC hierarchy that provide the categories relevant for the correctness definition of a mapping. Each table entry in these columns corresponds to the number of correct mappings. It can easily be verified that in all the experimental instances, the CAMPways alignments are superior to those of the SubMAP. As the network pairs under consideration are those of the same-domain species, going from more abstract categorizations of the root level 1 to the less abstract levels deeper in the FGC hierarchy, the number of correct mappings does not decrease significantly. We also note that for the 1.7 glycan biosynthesis and metabolism, although there are an average of 80 mappings for the *hsa-mmu* pair, both algorithms produce few correct mappings. The ratio of the correct mappings to the total number of mappings of the alignment is almost 6%. This is in contrast with the 90% correctness ratio of the same pair under the reverse engineering results of the previous section presented in Table 1. The prime reason for the

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/29/13/i145/197346 by Kadir Has University user on 13 November 2019

low correctness values is the lack of FGC categorizations for most of the reactions involved in the mentioned network. This in turn provides a potential application for the network alignment; the FGC category of a reaction can be transferred to those with unknown categorizations if they belong to the same mapping in the alignment. With regards to the results of the across-domains setting, it can be stated that similar to the results of Table 2, the alignment outputs of the CAMPways algorithm provide more correct mappings than those of the SubMAP in almost all network instances under all hierarchy levels; the only exception is the *hsa-atc* metabolism 1.10, in which case the correctness values of both algorithms are already low to bear any significance. The complete table providing results under the across-domains setting is provided in the Supplementary Document.

The aforementioned analysis based on functional group conversion hierarchies is extended to include the RPAIR data provided by KEGG on a sample mapping pair provided by both algorithms executed on the amino acid metabolism networks of the *atc-eco* pair. A *reactant pair* is defined as a pair of a substrate and a product that preserve chemical substructures through enzymatic reactions. In fact, the RCLASS database classification also provides information regarding reactant pairs. The difference is that the classifications of RCLASS are produced by computerized methods based on chemical structure comparison or molecular alignment, whereas those of RPAIR are produced by manually compiled reactant pairs and molecular alignments incorporating biochemical knowledge. The sample mapping pair provided by the CAMPways alignment is depicted in Figure 2. The *atc* reactions R01374 [D-phenylalanine: acceptor oxidoreductase (deaminating)] and R01582 (D-phenylalanine: 2-oxoglutarate aminotransferase) are together mapped to reaction R01374 of *eco*. Additionally, reactions R00694 (L-phenylalanine: 2-oxoglutarate aminotransferase) and R01372 [phenylpyruvate: oxygen oxidoreductase (hydroxylating,decarboxylating)] of *atc* are together mapped to the reaction R00694 of *eco*. The output compound C00166 (phenylpyruvate) of the reactions R01374 and R01582 is an input compound of the reactions R00694 and R01372. As a result, there is a directed edge from the node corresponding to the subset of reactions R01374, R01582 to the node corresponding to the subset of reactions R00694, R01372 in the *atc* pathway. Similarly, a directed edge exists from the node of reaction R01374 to the node of R00694 in the *eco* pathway. This implies a conserved edge resulting from the provided mappings. With regards to the classifications, it is worth noting that the FGC categories of the reactions R01374 and R01582 are the same for all five levels of the hierarchy, which further strongly validates the mapping involving these reactions based on the RCLASS classification. Both reactions are co-categorized even at the furthest level, which signifies identical RCLASS entry, RC00006. Further validation is observed when the manually compiled and biochemically more reliable RPATH data are examined; both reactions correspond to the identical reactant pair, RP00289 within RPATH. In contrast, the SubMAP mapping, including R01582, maps this reaction and the reaction R01373 [prephenate hydro-lyase (decarboxylating; phenylpyruvate-forming)] of *atc* to the single reaction R01373 of *eco*. The FGC categories of reactions R01373 and R01582 separate starting with the second level of the hierarchy and thus belong to separate RCLASS entries. Furthermore, there are no connections between the two as far as the RPAIR database is of concern.

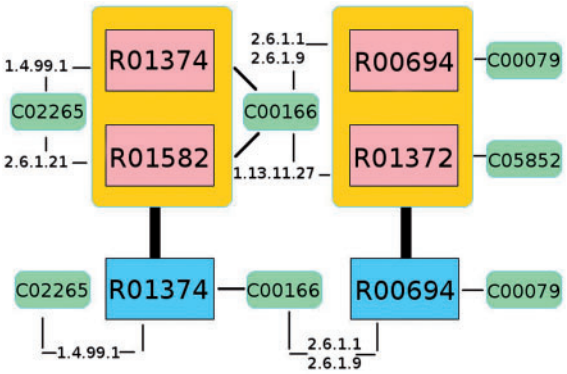

Fig. 2. Sample mapping from the CAMPways alignment of the amino acid metabolism networks. The reactions at the top are part of the *atc* network, whereas those at the bottom are part of the *eco* network. The mapped reactions (reaction subsets) are shown with the vertical edge. Enzymes are shown using EC numbers. The compounds are depicted within small rectangles

Table 3. The TR subcolumns provide the number of reactions in the network pair

TR	S	C	TR	S	C	TR	S	C	TR	S	C	TR	S	C	TR	S	C
62	3.04	0.30	116	62.81	2.26	264	454.21	13.39	296	1620	15.73	496	975.31	39.87	369	121.43	25.23
134	48.09	1.42	108	17.99	0.94	168	0.32	2.94	73	0.50	0.28	334	1788.84	25.17	31	0.06	0.04
51	0.15	0.09	35	0.09	0.04	23	0.04	0.02	207	3.25	1.00	175	0.67	5.39			
93	33.16	2.79	85	6.64	0.82	85	6.51	0.72	93	34.68	2.72	128	40.46	1.67	114	21.52	1.17
118	20.7	1.13	124	42.0	1.45	125	0.44	10.25	116	0.3	6.64	116	0.38	6.08	125	0.41	10.19
39	0.07	0.09	43	0.09	0.05	46	0.10	0.11	36	0.08	0.07	30	0.04	0.03	28	0.05	0.02
27	0.06	0.03	31	0.05	0.03	174	1.26	10.95	208	1.85	20.03	215	1.77	13.24	167	1.27	9.56

Note: CPU times in seconds are provided under the S and C subcolumns.

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/29/13/i145/197346 by Kadir Has University user on 13 November 2019

3.3 Execution speed and memory requirements

Assuming the degree of every node in G_p, G'_p is bounded by a constant, the running time of CAMPways is $O(|V_p|^2 \log^2 |V_p|)$, where $|V_p|$ is assumed without loss of generality to be larger than $|V'_p|$. We provide a detailed analysis of this running time bound in the Supplementary Document. In comparison, no explicit running time analysis of the SubMAP algorithm is provided. All experimental results in this section are obtained by running the algorithms on an Intel(R) Xeon(R) CPU 2.67 GHz with 24 GB of memory. The required CPU times for all the tested networks are listed in Table 3. The first three rows correspond to the experiments within the same-domain setting and the rest to those within the across-domains setting. The total number of reactions for each instance is listed at the subcolumns marked with *TR*. The columns provide the abbreviations of algorithm names as in Table 2. An important limitation of the SubMAP algorithm is its excessive memory consumption; the SubMAP code could not be executed until completion for some network pairs. For the hsa-mmu alignment of the 1.1 carbohydrate metabolism for instance, the CAMPways algorithm completed in <3 min, whereas the SubMAP code after 2 h of execution consumed all memory resources before crashing. In 15 of the 17 instances within the same-domain setting, CAMPways runs faster than SubMAP. For the across-domains setting in 14 of 28 instances, CAMPways provides better execution time. An important point worth emphasizing is that for the instances where CAMPways run faster, the differences between the execution times of CAMPways and SubMAP are large, whereas for the instances favoring SubMAP, both algorithms provide more or less similar execution times. The difference between the computational efficiency trends of the algorithms under the same-domain and the across-domains settings is interesting. It actually pinpoints the main reason behind the computational efficiency differences of the two algorithms. Within the same-domain setting, the pair of species that the metabolic networks belong to are evolutionarily close. Therefore, the aligned networks induce many conserved edges. In fact, these are the instances for which application of network alignment is sensible; simultaneous nature of the problem in terms of optimizing both homological (high-sequence alignment scores) and topological similarity (high-edge conservation) is most apparent in this setting. Most of the reactions in the pair of networks are aligned throughout the main loop of the CAMPways algorithm, as the generated conflict graphs are large because of high-edge conservation. When the pair of species is evolutionarily apart, the edge conservation is naturally low in which case the main task of both algorithms reduces to that of producing alignments that achieve only high-homological similarity.

ACKNOWLEDGEMENT

The authors thank Türkan Haliloğlu and Kemal Yelekçi for their valuable comments and Aykut Çay for his help in testing.

Funding: TUBITAK (112E137). T.B. is supported by TÜBA GEBIP 2009 and ESF EUROCORES TUBITAK (210T173).

Conflict of Interest: none declared.

REFERENCES

Agren,R. *et al.* (2012) Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using init. *PLoS Comput. Biol.*, **8**, e1002518.

Aladağ,A.E. and Erten,C. (2013) Spinal: scalable protein interaction network alignment. *Bioinformatics*, **29**, 917–924.

Ay,F. *et al.* (2011) Submap: aligning metabolic pathways with subnetwork mappings. *J. Comput. Biol.*, **18**, 219–235.

Ay,F. *et al.* (2012) Metabolic network alignment in large scale by network compression. *BMC Bioinformatics*, **13 (Suppl. 3)**, S2.

Bayati,M. *et al.* (2011) Belief propagation for weighted b-matchings on arbitrary graphs and its relation to linear programs with integer solutions. *SIAM J. Discrete Math.*, **25**, 989–1011.

Caglic,D. *et al.* (2009) Murine and human cathepsin B exhibit similar properties: possible implications for drug discovery. *Biol. Chem.*, **390**, 175–179.

Caspi,R. *et al.* (2008) The MetaCyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic Acids Res.*, **36**, D623–31.

Chindelevitch,L. *et al.* (2010) Local optimization for global alignment of protein interaction networks. In: *Pacific Symposium on Biocomputing*. Hawaii, USA, pp. 123–132.

Clemente,J.C. *et al.* (2007) Phylogenetic reconstruction from non-genomic data. *Bioinformatics*, **23**, e110–e115.

Edmonds,J. (1965) Maximum matching and a polyhedron with 0 1-vertices. *J. Res. Natl Bur. Stand. B*, **69**, 125–130.

Flannick,J. *et al.* (2006) Graemlin: general and robust alignment of multiple large interaction networks. *Genome Res.*, **16**, 1169–1181.

Gabow,H.N. (1983) Scaling algorithms for network problems. In *Proceedings of the 24th Annual Symposium on Foundations of Computer Science, SFCS '83. IEEE Computer Society*, Washington, DC, USA, pp. 248–258.

Garey,M.R. and Johnson,D.S. (1979) *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman, New York.

Guimerà,R. *et al.* (2007) A network-based method for target selection in metabolic networks. *Bioinformatics*, **23**, 1616–1622.

Heymans,M. and Singh,A. (2003) Deriving phylogenetic trees from the similarity analysis of metabolic pathways. *Bioinformatics*, **19**, 138–146.

Kanehisa,M. *et al.* (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, **40**, 109–114.

Kelley,B.P. *et al.* (2004) Pathblast: a tool for alignment of protein interaction networks. *Nucleic Acids Res.*, **32**, 83–88.

Koyutürk,M. *et al.* (2006) Pairwise alignment of protein interaction networks. *J. Comput. Biol.*, **13**, 182–199.

Kuchaiev,O. and Pržulj,N. (2011) Integrative network alignment reveals large regions of global network similarity in yeast and human. *Bioinformatics*, **27**, 1390–1396.

Liao,C.S. *et al.* (2009) Isorankn: spectral methods for global alignment of multiple protein networks. *Bioinformatics*, **25**, i253–i258.

March,J. (1985) *Advanced Organic Chemistry: Reactions, Mechanisms, and Structure*. Wiley, New York.

Mehlhorn,K. and Naher,S. (1999) *Leda: A Platform for Combinatorial and Geometric Computing*. Cambridge University Press, Cambridge.

Mithani,A. *et al.* (2011) Comparative analysis of metabolic networks provides insight into the evolution of plant pathogenic and non-pathogenic lifestyles in Pseudomonas. *Mol. Biol. Evol.*, **28**, 483–499.

Pinter,R.Y. *et al.* (2005) Alignment of metabolic pathways. *Bioinformatics*, **21**, 3401–3408.

Sakai,S. *et al.* (2003) A note on greedy algorithms for the maximum weighted independent set problem. *Discrete Appl. Math.*, **126**, 313–322.

Sharan,R. *et al.* (2005) Conserved patterns of protein interaction in multiple species. *Proc. Natl Acad. Sci. USA*, **102**, 1974–1979.

Singh,R. *et al.* (2008) Global alignment of multiple protein interaction networks. In: *Pacific Symposium on Biocomputing*. Hawaii, USA, pp. 303–314.

Tohsato,Y. *et al.* (2000) A multiple alignment algorithm for metabolic pathway analysis using enzyme hierarchy. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*. AAAI, pp. 376–383.

Yang,Q. and Sze,S.H. (2007) Path matching and graph matching in biological networks. *J. Comput. Biol.*, **14**, 56–67.

Zaslavskiy,M. *et al.* (2009) Global alignment of protein-protein interaction networks by graph matching methods. *Bioinformatics*, **25**, 259–267.

Zhenping,L. *et al.* (2007) Alignment of molecular networks by integer quadratic programming. *Bioinformatics*, **23**, 1631–1639.

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/29/13/i145/197346 by Kadir Has University user on 13 November 2019