BIOINFORMATICS

Vol. 30 ECCB 2014, pages i438–i444
doi:10.1093/bioinformatics/btu450

HubAlign: an accurate and efficient method for global alignment of protein–protein interaction networks

Somaye Hashemifar and Jinbo Xu*

Toyota Technological Institute at Chicago, Chicago, IL 60637, USA

ABSTRACT

Motivation: High-throughput experimental techniques have produced a large amount of protein–protein interaction (PPI) data. The study of PPI networks, such as comparative analysis, shall benefit the understanding of life process and diseases at the molecular level. One way of comparative analysis is to align PPI networks to identify conserved or species-specific subnetwork motifs. A few methods have been developed for global PPI network alignment, but it still remains challenging in terms of both accuracy and efficiency.

Results: This paper presents a novel global network alignment algorithm, denoted as HubAlign, that makes use of both network topology and sequence homology information, based upon the observation that topologically important proteins in a PPI network usually are much more conserved and thus, more likely to be aligned. HubAlign uses a minimum-degree heuristic algorithm to estimate the topological and functional importance of a protein from the global network topology information. Then HubAlign aligns topologically important proteins first and gradually extends the alignment to the whole network. Extensive tests indicate that HubAlign greatly outperforms several popular methods in terms of both accuracy and efficiency, especially in detecting functionally similar proteins.

Availability: HubAlign is available freely for non-commercial purposes at http://ttic.uchicago.edu/~hashemifar/software/HubAlign.zip

Contact: jinboxu@gmail.com

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

1 INTRODUCTION

High-throughput experimental techniques such as yeast-two-hybrid (Kayarkar, 2009) and protein co-immunoprecipitation (Aebersold and Mann, 2003) have produced a large amount of protein–protein interaction (PPI) data for several organisms such as *Homo sapiens* (Radivojac *et al.*, 2008) and *Saccharomyces cerevisiae* (Collins *et al.*, 2007). PPI networks contain a significant amount of information about modular organization of cells and protein functions. Comparative analysis such as alignment of PPI networks can help identify evolutionarily conserved pathways/complexes that may be structurally or functionally important and species-specific pathways/complexes, and infer protein functions.

Similar to sequence alignment, we can also align PPI networks either locally or globally. Local network alignment (LNA) such as NetworkBlast (Sharan *et al.*, 2005), Mawish (Koyutürk *et al.*, 2006) and AlignNemo (Ciriello *et al.*, 2012) aims to find small isomorphic subnetworks corresponding to pathways and protein complexes (Wang and Gao, 2012) and thus may yield a many-to-many mapping between the proteins. These methods search for conserved subnetworks using an alignment graph, in which nodes correspond to groups of orthologous proteins and edges to conserved interactions. These methods mainly differ in building alignment graphs, the definition of dense clusters and search algorithms.

Different from LNA, global network alignment (GNA) aims to maximize the overall match between the input networks. Such methods such as IsoRank (Singh *et al.*, 2008a,b), Mawish, MI-GRAAL (Kuchaiev and Przulj, 2011), GHOST (Patro and Kingsford, 2012), PISwap (Chindelevitch *et al.*, 2013) and NETAL (Neyshabur *et al.*, 2013) are designed for pairwise alignment while others such as NetworkBlast, Graemlin 2.0 and IsoRankN for multiple alignment. In addition to network topology information, all network alignment algorithms excluding NETAL and MAGNA (Saraph and Milenković, 2013) make use of sequence similarity to help improve alignment accuracy. IsoRank aligns two PPI networks by exploiting the observation that two proteins are good match if their interacting partners can match well. IsoRankN is an extension of IsoRank and mainly for multiple network alignment. It applies IsoRank to compute the alignment score between each pair of networks, and then employs a PageRank-Nibble algorithm to cluster all the proteins by their alignment score (Liao *et al.*, 2009). Graemlin2.0 integrates network topology and phylogeny information and uses a hill-climbing algorithm to generate alignments (Flannick *et al.*, 2008; Kuchaiev *et al.*, 2010). MI-GRAAL, an improved version of GRAAL, integrates network topology information such as graphlet signature and sequence similarity to align two nodes (Kuchaiev and Przulj, 2011). GHOST uses graph spectrum to measure the topological similarity of proteins (Patro and Kingsford, 2012). Both MI-GRAAL and GHOST use a seed-and-extend strategy to build an alignment. MI-GRAAL fulfills this by solving a weighted bipartite matching, while GHOST by solving a quadratic problem. PISwap refines an alignment generated by other tools such as IsoRank. It iteratively swaps the edges in an alignment until reaching an optimum (Chindelevitch *et al.*, 2013). MAGNA uses a genetic algorithm to search for the best alignment (Saraph and Milenković, 2013). NETAL aligns two proteins based upon their interacting partners.

Current global network alignment methods have two major issues. One is that existing algorithms run slowly, especially in aligning very large PPI networks. The other is that the alignment accuracy is still low. This motivates us to develop a new method for global network alignment to significantly improve both alignment accuracy and computational efficiency.

This paper presents a novel global network alignment algorithm, denoted as HubAlign, to align two PPI networks using both network topology and sequence homology information, based upon the observation that topologically and functionally

*To whom correspondence should be addressed.

important proteins (such as hubs and bottlenecks) in a PPI network are more conserved and thus, shall be aligned. We use a minimum-degree heuristic method to estimate the relative importance of one protein from the global network topology information. Such a score reflects the topological and functional importance of one protein in a PPI network. Then, we use a greedy algorithm to align two proteins based upon the combination of their importance scores and sequence similarity. That is, we align more important proteins first and then gradually less important. Such a procedure is more biologically meaningful and leads to a much faster and more accurate alignment algorithm. We have tested HubAlign on both prokaryotic and eukaryotic PPI networks, showing that HubAlign greatly outperforms several popular methods such as IsoRank, MI-GRAAL, GHOST and PISwap in terms of both alignment accuracy and running time.

2 METHODS

Main idea. A biological network usually contains some topologically and functionally important proteins such as hubs and bottlenecks. Hub proteins have many connections, may be involved in various biological modules and play a central role in all biological processes. In Han's work (Han *et al.*, 2004), proteins with more than five interactions are defined as hubs, while those with fewer interactions are peripheral nodes. Bottlenecks refer to those proteins with a high betweenness centrality (i.e., the number of shortest paths passing through a node) (Yu *et al.*, 2007). These proteins usually connect functional clusters, so removing them can divide a PPI network into several subnetworks and disrupt the cooperation between functional modules (Dunn *et al.*, 2005). Because hubs and bottlenecks are topologically and functionally important, they tend to mutate more slowly and thus, are more conserved. That is, they are more likely to be aligned. To make use of this observation, we assign a score or weight to each node and edge of a PPI network using an iterative minimum-degree heuristics algorithm, measuring the topological and functional importance of a node (i.e., the likelihood of being a hub or bottleneck) and an edge in the PPI network with respect to the global network topology. Such an importance score reflects the global topological property of a protein. Then we calculate an alignment score for a pair of proteins using two properties: their relative importance scores (i.e., global topological property) and sequence information. Meanwhile, the global topological property is the most important and informative. Finally, we construct a global network alignment by picking those protein pairs with high alignment scores using a greedy method.

Definition. We represent a protein–protein interaction (PPI) network by an undirected graph $G = (V, E)$ where V is the set of vertices (proteins) and E the set of edges (interactions). Let N (u) denote the neighbors of a node $u \in V$ and $|N(u)|$ is the size of N(u). Let deg(u) denote the degree of vertex u, i.e. $\deg(u) = N(u)$. Each edge $e = (u, v) \in E$ may be associated with a score indicating the interaction strength. A global alignment of two networks $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ is a function $g = V_1 \rightarrow V_2$ that maps node set V_1 to V_2. Without loss of generality, we assume $|V_1| \leq |V_2|$ where $|V|$ is the number of vertices in set V.

Computing the topological and functional importance of proteins. We calculate the relative importance of a node or edge based upon only the network topology information of a PPI network. Such a relative importance shows the role of a node or edge in maintaining network structure or function (Zhao *et al.*, 2006). Although high-degree nodes play an important role in maintaining the structure and function of a network (Zotenko *et al.*, 2008), we do not simply use the degree of one node to calculate its relative importance, as the degree is only a local property. We want a global topological property reflecting the structure of the entire network.

We do not use existing measures such as edge-betweenness (Ellens, 2011) either, which defines the number of the shortest paths going through an edge in a network. That is, edge-betweenness takes into consideration only the shortest paths in a graph. Nevertheless, for the robustness of a network the longer alternative paths are also important (Ellens, 2011). In addition, it is also observed that (i) edges connecting high-degree nodes are more important, as they connect many nodes and may be relevant to the global structure property of the network (H., 2006); and (ii) a pair of two nodes with a large number of common neighbors are more likely to be related (Liu, 2009).

Here we use a minimum-degree heuristics algorithm to calculate the topological importance of nodes and edges, starting from the nodes with degree one and stopping at those with degree d. The value of d cannot be very large, as the deletion of very high-degree nodes (e.g. hubs) may destroy the whole network functionally or structurally while random deletion of a fraction of peripheral nodes may cause only a small damage to the network (Wang, 2007; Zhao *et al.*, 2006). Empirically $d = 10$ yields a good result. To calculate the relative importance of nodes, we assign an initial weight to nodes and edges as follows.

$$w(e) = \begin{cases} 1 & e \in E \\ 0 & \text{otherwise} \end{cases}, w(u) = 0 \quad \forall u \in V$$

Where w(e) and w(u) represent the weight of edge e and node u, respectively. We may initialize the edge weight by the PPI confidence score if it is available in the PPI data.

We update the weight by always removing one of the nodes with minimum degree. When one node is removed, its adjacent edges are also removed and the weight of the removed node and edges are allocated to their neighboring nodes and edges. In this way, the topological information is propagated from a node to its neighbors. In particular, when removing node $u \in V$, we update the weights as follows.

(1) If $\deg(u) = 1$, $\forall v \in N(u)$, set $w(v) = w(v) + w(u) + w(u, v)$.

(2) If $\deg(u) > 1$, $\forall v_1, v_2 \in N(u)$, set

$$w(v_1, v_2) = w(v_1, v_2) + \frac{w(u) + \sum_{v \in N(u)} w(u, v)}{\frac{|N(u)||N(u) - 1|}{2}}$$

Figure 1 shows for a small example PPI network how an edge gains more weight after the removal of some peripheral nodes. For example, when nodes d, c, e and f are removed, their own weight and those of their adjacent edges are transferred to the edge (a, b), which indicates that this edge is important in maintaining the network connectivity. After calculating the weights,

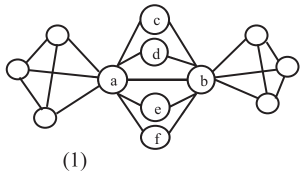

Fig. 1. Illustration of the algorithm for the calculation of topological importance score. (1) the original graph; and (2) the graph resulting from removing nodes d, c, e and f. The thickness of an edge shows its weight

we assign an importance score as follows to each node by combining both node and edge weight to indicate its topological importance in the network.

$$S(v) = w(v) + \lambda \sum_{u \in V} w(u, v)$$

Where S(v) is the score of node v, λ controls the importance of the edge weight relative to the node weight. Empirically $\lambda = 0.2$ yields a biologically more meaningful alignment. Finally, we normalize S(v) as follows to reduce the impact of network size.

$$S(v) = S(v)/\max_{v \in V}\{S(v)\}$$

The way we calculate the relative importance of nodes and edges is inspired by graph tree-decomposition, which is used to simplify a graph as a tree in which each vertex represents a highly connected subgraph component and each edge represents the intersection between two adjacent components. The size of the highly connected components reflects the topological complexity of a graph and also importance of nodes. Several simple heuristics methods such as the minimum-degree heuristic method (Bodlaender and Koster, 2010; Robertson and Seymour, 1984) are developed to tree-decompose a general graph.

Remark. To validate that the resultant importance score (i.e., S) makes biological sense, we examine the top 50 proteins with the highest S scores in the human PPI network. Meanwhile, all the top 10 proteins have a very high degree, which indicates they are vital hubs of the network. The two example proteins are P62993 with degree 663 and Q9H0R8 with degree 491. See Figure 2a for the subnetwork containing P62993. On the other hand, among the top 50 proteins there are also some low-degree proteins, such as Q9UPN3 with degree 7. As shown in Figure 2c, although Q9UPN3 is not a hub, it is a bottleneck connecting several functional modules. This protein is also related to breast cancer disease (Rohan, 2009). Another interesting example is P04156 with degree 52 and betweenness 0.005. As shown in Figure 2b, this protein is a hub connecting several other hubs. Removal of this protein can disrupt the cooperation of the hubs connecting to it.

Building alignment. The normalized S score measures the relative importance of one protein with respect to the whole PPI network. It reflects the global topological properties of one protein in a network. If two nodes have similar S scores, they may be similarly important in their respective networks. Thus, they are more likely to be aligned. We calculate the topological similarity between two nodes $u \in V_1$ and $v \in V_2$ as follows:

$$A(u, v) = \min(S(u), S(v))$$

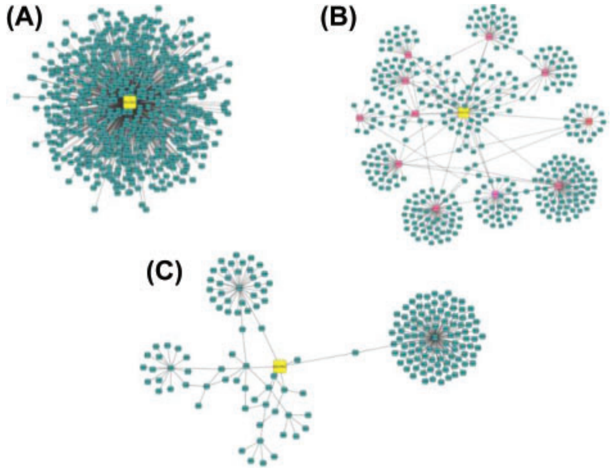

Fig. 2. Three example proteins (in yellow) with high importance scores in the human PPI network. (**a**) Protein P62993, which has the largest degree; (**b**) Protein P04156, which connects to some hubs (in red); (**c**) Protein Q9UPN3 with low degree that performs as a bottleneck

We also incorporate sequence homology information (i.e., sequence similarity) into our alignment score. Let B(u, v) denote the normalized BLAST bitscore for two proteins u and v. The final alignment score is defined as follows.

$$A^*(u, v) = \alpha \times A(u, v) + (1 - \alpha) \times B(u, v)$$

Where $0 \leq \alpha \leq 1$ is a parameter that controls the contribution of sequence similarity relative to topological similarity. Meanwhile, $\alpha = 1$ implies that only topological information is used, while $\alpha = 0$ implies that only sequence information is used. Tuning α allows us to find the relative importance of sequence information in aligning the networks. In our implementation, we set α to 0.7 by default. That is, our method uses much more network topology information than sequence information.

Our algorithm identifies the pair of nodes with the highest alignment score as a seed alignment and gradually extends it using a greedy algorithm. After aligning a pair of nodes *u* and *v*, we then consider aligning their neighbors, which is reasonable because functional modules and protein complexes are densely connected and tend to be separated from other subnetwork modules. Algorithm continues to align neighboring nodes until their alignment score is relatively high (more than the average of the alignment scores). When the subnetwork alignment resulting from the initial seed is terminated, the next best unaligned pair is chosen as a new seed. This procedure is repeated until all proteins of the smaller network are aligned with the proteins of the larger network.

Time complexity. Let $n = \max\{|V_1|, |V_2|\}$. At the first step, it takes O(n) to find the node with minimum degree. As we mentioned before, we only remove the nodes with degree less than 10. Thus, updating the weight of the neighbors can be done in O(1). Further, as we can remove up to n nodes from a network, the total time complexity for the first step is $O(n^2)$. At the second step, we calculate the alignment score for each pair of nodes of the input networks. Because there are at most n^2 pair nodes, this step takes $O(n^2)$. At the final step, a seed can be selected in $O(n^2)$. Then for extension, we use a priority queue to save the neighbors of each pair of aligned nodes. Because each node of the graph has at most *n* neighbors, updating the priority queue takes O(nlog(n)). Extracting the pair with highest score from this queue can be done in constant time. That is, the final step for aligning n nodes takes $O(n^2 \log(n))$. As such, the total time complexity is $O(n^2 \log(n))$.

3 RESULTS

We compare our algorithm HubAlign with several popular and publicly available global network alignment methods IsoRank (Singh *et al.*, 2008a,b), MI-GRAAL (Kuchaiev and Przulj, 2011), GHOST (Patro and Kingsford, 2012) and PISwap (Chindelevitch *et al.*, 2013). Following Chindelevitch *et al.* (2013), we use the alignment produced by GRAAL and IsoRank as input of PISwap. We do not compare HubAlign with Graemlin 2.0 (Flannick *et al.*, 2008), because the latter requires the knowledge of phylogenetic relationship among species. We do not compare HubAlign with GRAAL because MI-GRAAL is an improved version of GRAAL. Following Kuchaiev's work (Kuchaiev *et al.*, 2010), we evaluate network alignment quality by five measures including edge correctness (EC), largest common connected subgraph (LCCS), symmetric substructure score (S^3), functional consistency (FC) and average of functional similarity (AFS). Meanwhile, EC, LCCS and S^3 reflect network topological similarity of an alignment, but not biological significance. FC and AFS reflect biological significance by measuring the consistency of the GO (gene ontology) terms assigned to the aligned proteins. FC and AFS shall be more important metrics than EC, LCCS and S^3. The alignment accuracy of two PPI networks depends not only on the evolutionary distance of two respective species, but also on the quality of the PPI data. That is, the closer the two species are or the higher quality the PPIs are, the better alignment we may obtain.

Edge correctness (EC). It is calculated as the percentage of edges in the first network that are aligned to edges in the second network. Here we assume the first network is smaller than the second one. Let (V_1, E_1) and (V_2, E_2) denote two networks under alignment where V and E denote nodes and edges, respectively and $g : V_1 \rightarrow V_2$ be an alignment. Mathematically, EC is defined as follows.

$$EC = \frac{|\{(u, v) \in E_1 : (g(u), g(v)) \in E_2\}|}{|E_1|} \times 100$$

Symmetric substructure score (S^3). The intuition underlying S^3 is to penalize the alignments that map sparse regions of the network to denser ones and vise-versa (Saraph and Milenković, 2013). Let G[V] denote the induced subnetwork of G with node set V and E(G) denote the edge set of network G. Let $f(E_1) = \{(g(u), g(v)) \in E_2 : (u, v) \in E_1\}$ and $f(V_1) = \{g(v) \in V_2 : v \in V_1\}$. Mathematically, S^3 is defined as follows.

$$S^3 = \frac{|f(E_1)|}{|E_1| + |E(G_2[f(V_1)])| + |f(E_1)|} \times 100$$

Largest common connected subgraph (LCCS). It is calculated as the number of edges in the largest connected subgraph in an alignment. Larger and denser subgraphs give more insight into common topology of the network (Kuchaiev and Przulj, 2011). In addition, the larger and denser subgraphs may be more biologically important (Hu et al., 2005), as Bader and Spirin have shown that a dense PPI subnetwork may correspond to a vital protein complex (Bader and Hogue, 2003; Spirin and Mirny, 2003).

Functional consistency (FC). We use GO (gene ontology) terms to measure the functional consistency of two aligned proteins. GO terms describe some biological properties of a protein such as Cellular Component (CC), Molecular Function (MF) and Biological Process (BP). We exclude root GO terms from the analysis. Proteins with similar GO terms are supposed to be functionally similar. To analyze the biological significance of an alignment, we calculate the fraction of aligned proteins sharing common GO terms. The fraction is calculated with respect to the size of the smaller network because in a global alignment all nodes of smaller network are aligned to nodes of larger network. The larger the fraction, the more biologically meaningful the alignment is.

Average of functional similarity (AFS). It is calculated based on the semantic similarity of the GO terms, which depends on the distance between them in the ontology. We can use semantic similarity measures to calculate the functional similarity in each category of BP, MF and CC. Schlicker's similarity, based on the Resnik ontological similarity, is one of the best performing methods for computing the functional similarity between proteins (Pesquita *et al.*, 2009; Schlicker *et al.*, 2006). Let $s_c(u, v)$ denote the GO functional similarity of proteins u and v in category c (i.e., BP, MF or CC). AFS is defined as follows.

$$AFS_c = \frac{1}{|V_1|} \sum_{u \in V_1} s_c(u, g(u))$$

3.1 Alignment of the yeast and human PPI networks

We apply our algorithm HubAlign to align the yeast and human PPI networks, which are taken from IntAct (Kerrien *et al.*, 2012). The yeast PPI network has 5673 nodes and 49 830 edges and the human network consists of 9002 nodes and 34 935 edges. We ran IsoRank and PISwap with the default parameters. MI-GRAAl was run using the degree, signature similarity and sequence similarity. The parameters for GHOST are automatically determined or set to default.

As shown in Table 1, our algorithm HubAlign produces an alignment with much larger EC, LCCS and S^3 than the other

Table 1. The EC, LCCS and S^3 of the human–yeast alignments generated by six methods

Method	EC	LCCS	S^3	AFS_{BP}	AFS_{MF}	AFS_{CC}
IsoRank	2.12	44	1.23	0.76	0.63	0.77
MIGRAAL	13.87	4832	8.12	0.63	0.52	0.72
GHOST	17.04	7000	13.59	0.82	0.66	0.83
PISwap	2.16	62	1.23	0.77	0.63	0.77
NETAL	28.65	9695	20.16	0.58	0.46	0.71
HubAlign	21.59	7240	14.67	0.95	0.81	0.88

methods except NETAL. To measure the FC and AFS of an alignment, we extract the GO annotations for all the involved proteins from the Gene Ontology database (Ashburner *et al.*, 2000). Some proteins may not have any GO annotations, so we just take into consideration the aligned pairs in which both proteins have GO annotations. Table 1 show that HubAlign yields alignments with significantly higher AFS than the other methods, especially when BP and MF are considered. We also calculate the percentage of aligned pairs in which two proteins share at least one, two, three, four and five GO terms, respectively. As shown in Table 2, HubAlign greatly outperforms the others in terms of FC. The advantage of HubAlign becomes larger when more shared GO terms are required to determine FC. NETAL yields more aligned proteins and interactions, but many aligned proteins are not functionally similar.

Table 2. Functional consistency of the yeast–human alignments generated by HubAlign and the others

No. of shared GO terms	IsoRank	MI-GRAAL	GHOST	PISwap	NETAL	HubAlign
≥1	33.98	29.02	35.42	34.03	26.03	47.56
≥2	15.02	7.02	15.74	14.84	2.95	28.23
≥3	8.73	2.81	8.69	8.65	0.67	17.41
≥4	4.49	1.06	4.04	4.46	0.24	9.52
≥5	1.97	0.26	1.77	2.00	0.14	4.77

3.2 Alignment of PPI networks of human, yeast, fly, worm and mouse

We also apply HubAlign to align PPI networks of *H.sapiens* (human), *S.cerevisiae* (yeast), *Drosophila melanogaster* (fly), *Caenorhabditis elegans* (worm) and *Mus musculus* (mouse). All these networks are obtained from IntAct (Kerrien *et al.*, 2012). Table 3 lists the AFS of all the pairwise alignments generated by five different methods: HubAlign, IsoRank, PISwap, MI-GRAAL and GHOST. The human–yeast alignment is already analyzed in the preceding section, so it is not included here. Table 3 shows that the alignments produced by HubAlign outperform those by the other methods in term of AFS under all three categories BP, MF and CC. HubAlign also produces alignments with higher FC than all the other methods (see Supplementary Table S1). Specifically, the more common GO terms required to determine FC, the more advantage HubAlign has over the other methods. For example, if only one shared GO term is required, HubAlign greatly outperforms the second best method GHOST for five of nine alignments (i.e., human–fly, fly–yeast, mouse–worm, mouse–fly and

Table 3. Performance of HubAlign and the other methods in terms of AFS of the alignments in categories BP, MF and CC. MI-GRAAL and GRAAL do not produce any result for human–fly and yeast–fly alignment

Alignment	AFS	IsoRank	MI-GRAAL	GHOST	PISwap	NETAL	HubAlign
Human–mouse	BP	1.32	0.84	1.58	1.32	0.73	**2.02**
	MF	1.23	0.84	1.50	1.23	0.70	**1.74**
	CC	1.08	0.76	1.20	1.08	0.66	**1.49**
Mouse–fly	BP	0.73	0.62	0.84	0.73	0.50	**1.07**
	MF	0.61	0.50	0.75	0.61	0.33	**0.97**
	CC	0.53	0.42	0.54	0.53	0.34	**0.72**
Mouse–yeast	BP	0.71	0.60	0.85	0.70	0.47	**0.96**
	MF	0.64	0.54	0.80	0.64	0.36	**0.91**
	CC	0.77	0.67	0.84	0.40	0.57	**0.91**
Fly–yeast	BP	0.48	0	0.54	0.48	0.38	**0.68**
	MF	0.35	0	0.42	0.35	0.23	**0.58**
	CC	0.40	0	0.44	0.40	0.36	**0.50**
Human–fly	BP	0.53	0	0.61	0.53	0.41	**0.72**
	MF	0.43	0	0.54	0.43	0.28	**0.65**
	CC	0.38	0	0.41	0.37	0.30	**0.48**
Mouse–worm	BP	0.63	0.50	0.67	0.63	0.42	**0.76**
	MF	0.64	0.46	0.67	0.64	0.31	**0.81**
	CC	0.40	0.31	0.41	0.40	0.25	**0.49**
Human–worm	BP	0.52	0.43	0.60	0.52	0.40	**0.64**
	MF	0.34	0.25	0.40	0.34	0.23	**0.70**
	CC	0.34	0.27	0.40	0.34	0.25	**0.44**
Worm–fly	BP	0.51	0.34	0.55	0.50	0.31	**0.57**
	MF	0.48	0.22	0.52	0.47	0.18	**0.54**
	CC	0.26	0.14	0.28	0.25	0.13	**0.31**
Worm–yeast	BP	0.38	0.31	0.41	0.37	0.26	**0.43**
	MF	0.34	0.25	0.40	0.34	0.23	**0.41**
	CC	0.30	0.25	0.31	0.30	0.24	**0.32**

mouse–yeast) and slightly outperforms GHOST for the remaining four alignments. If at least two shared GO terms are used to determine FC, HubAlign greatly outperforms GHOST for all the alignments. Moreover, HubAlign produces alignments with larger EC, LCCS and S^3 than the others except NETAL (see Supplementary Fig. S1). The NETAL alignments again have very low FC and AFS. These results indicate that HubAlign is able to align more functionally similar proteins and find larger complexes that are significant either topologically or biologically.

Table 4. The EC, LCCS and AFS of the alignments by different algorithms for the bacterial PPI networks

Method	EC	LCCS	S^3	AFSBP	AFSMF	AFSCC
IsoRank	8.50	11	1.51	0.20	0.16	0.07
MI-GRAAL	23.86	400	15.89	0.14	0.12	0.04
GHOST	23.86	440	15.03	0.19	0.14	0.06
PISwap	17.87	289	1.83	0.11	0.08	0.02
NETAL	32.36	661	19.54	0.10	0.07	0.02
HubAlign	24.56	474	16.51	0.25	0.22	0.07

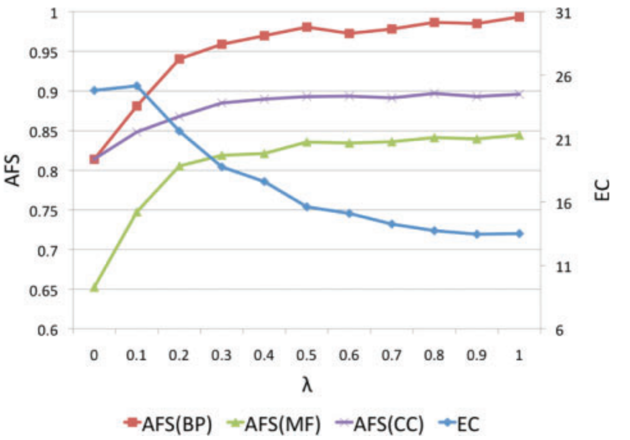

Fig. 3. Performance of HubAlign in terms of AFS and EC with respect to λ. Each curve consists of 11 points corresponding to 11 different λ values: 0, 0.1, . . . , 1 from top to bottom

3.3 Alignment of bacterial PPI networks

We also apply HubAlign to align the PPI networks of two bacterial species *Campylobacter jejuni* and *Escherichia coli*, which have the most complete PPI networks among all bacteria. The PPI network for Bacterium *C.jejuni* has 1111 nodes and 2988 edges (Parrish *et al.*, 2007). *Escherichia coli* is a model organism for studying the fundamental cellular processes such as gene expression and signaling. The *E.coli* PPI network has 1941 nodes and 3989 edges (Peregrín-Alvarez *et al.*, 2009). As shown in Table 4, HubAlign produces an alignment with larger EC, LCCS and S^3 than the other methods except NETAL. In terms of AFS, HubAlign outperforms the other methods although all the AFS values are pretty small due to insufficient GO annotations of the bacterial proteins. The average number of GO terms associated with the proteins of *E.coli* and *C.jejuni* is much smaller than that of the other species. In addition, HubAlign produces alignments with larger FC (see Supplementary Table S2).

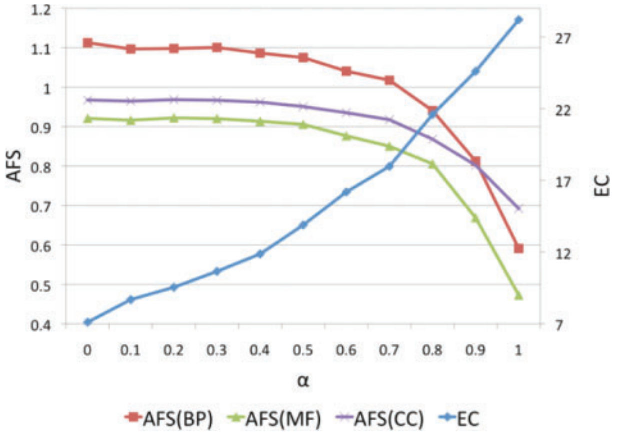

Fig. 4. Performance of HubAlign in terms of AFS and EC with respect to α. Each curve consists of 11 points corresponding to 11 different α values: 0, 0.1, . . . , 1 from bottom to top

3.4 Evaluation of parameters λ and α

Our algorithm makes use of two parameters λ and α. λ determines the relative importance of edge and node weight, while α determines the relative importance of sequence and topological similarity. In this section, we study the relationship between these two parameters and network alignment quality. We apply HubAlign to PPI networks of yeast and human and report EC, LCCS, S^3 and AFS of their alignment for different values of parameter λ between 0 and 1. As shown in Figure 3, AFS increases as λ gets close to 1. The underlying reason could be that the higher values of λ give more importance to the edge weights which in turn, makes the proteins with important interactions align together. On the other hand, by increasing the value of λ, we put less emphasis on the node weight and therefore, it is less likely that the hubs be aligned together. As a result, the topological qualities (i.e. EC, LCCS and S^3) decrease. Figure 3 shows that increasing λ from 0 to 0.2 improves the AFS significantly but does not change the EC much. However, as we continue to increase λ further, the EC decreases sharply. We also observe a slight increase in the biological quality. There are similar plots for the S^3 and the LCCS (see the Supplementary Fig. S2). Thus, we can achieve a good trade-off between the topological and the biological quality by setting λ in the range (0.1, 0.2).

We also compute the yeast–human alignment for different values of α. As shown in Figure 4, increasing α from 0 to 1 decreases AFS. This is because a larger value of α reduces the effect of sequence information. Moreover, in line with our expectations as α goes up, so does the topological quality of the alignment. Figure 4 shows that increasing α from 0 to 0.7 does not change AFS much but improves the EC significantly. However, as we continue to increase α further, the AFS decreases sharply. There are similar plots for the S^3 and the LCCS (see the Supplementary Fig. S3). Thus, we can achieve a good trade-off between the topological and the biological quality by setting α in the range (0.7, 1).

3.5 Running time

Our method HubAlign is much more computationally efficient than the others. Tested on the yeast–human alignment on a 1400 MHz Linux system with 2GB RAM, it takes NETAL, HubAlign, IsoRank, MI-GRAAL and GHOST 80, 412, 7610, 78 525 and 3037 s, respectively, to terminate. PISwap has almost the same running time as IsoRank because the former only slightly refines the result generated by the latter.

4 CONCLUSION

This paper has presented a new method HubAlign for global alignment of two PPI networks by making use of topological importance of proteins in a PPI network. We have implemented and tested HubAlign using quite a few PPI networks and evaluated the resultant alignments using different performance metrics. We have also compared HubAlign with currently popular global network alignment algorithms such as IsoRank, MI-GRAAL, NETAAL, GHOST and PISwap. Experimental results indicate that our algorithm greatly outperforms the others in terms of both alignment accuracy and running time. In particular, our algorithm can align many more functionally similar proteins.

Funding: This work is financially supported by the NSF grant DBI-1262603 (to J.X.) and the NSF CAREER award CCF-1149811 (to J.X.).

Conflict of Interest: none declared.

REFERENCES

Aebersold,R. and Mann,M. (2003) Mass spectrometry-based proteomics. *Nature*, **422**, 198–207.

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

Bader,G.D. and Hogue,C.W. (2003) An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, **4**, 2.

Bodlaender,H.L. and Koster,A.M.C.A. (2010) Treewidth computations I. Upper bounds. *Inform. Comput.*, **208**, 259–275.

Chindelevitch,L. *et al.* (2013) Optimizing a global alignment of protein interaction networks. *Bioinformatics*, **29**, 2765–2773.

Ciriello,G. *et al.* (2012) Alignnemo: a local network alignment method to integrate homology and topology. *PLoS One*, **7**, e38107.

Collins,S.R. *et al.* (2007) Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. *Mol. Cell. Proteomics*, **6**, 439–450.

Dunn,R. *et al.* (2005) The use of edge-betweenness clustering to investigate biological function in protein interaction networks. *BMC Bioinformatics*, **6**, 39.

Ellens,W. *et al.* (2011) Effective resistence. *Linear Algebr. Appl.*, **435**, 24–91.

Flannick,J. *et al.* (2008) Automatic parameter learning for multiple network alignment. In: *Research in Computational Molecular Biology*. Springer, Berlin, Heidelberg, pp. 214–231.

Luo,H. (2006) Modeling and simulation of large-scale complex networks. PhD Thesis, School of mathematical and Geospatial Sciences, RMIT University, Melbourne, Australia.

Han,J.-D.J. *et al.* (2004) Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature*, **430**, 88–93.

Hu,H. *et al.* (2005) Mining coherent dense subgraphs across massive biological networks for functional discovery. *Bioinformatics*, **21**, i213–i221.

Kayarkar,N.A. *et al.* (2009) Protein network in diseases. *Int. J. Drug Discov.*, **1**, 10–17.

Kerrien,S. *et al.* (2012) The IntAct molecular interaction database in 2012. *Nucleic Acids Res.*, **40**, D841–D846.

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

Kuchaiev,O. *et al.* (2010) Topological network alignment uncovers biological function and phylogeny. *J. R. Soc. Interface*, **7**, 1341–1354.

Kuchaiev,O. and Przulj,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.

Liu,D. (2009, June) Protecting neighbor discovery against node compromises in sensor networks. In: *Distributed Computing Systems, 2009. ICDCS'09. 29th IEEE International Conference on*. IEEE Montreal, QC, Canada, pp. 579–588.

Neyshabur,B. *et al.* (2013) NETAL: a new graph-based method for global alignment of protein–protein interaction networks. *Bioinformatics*, **29**, 1654–1662.

Parrish,J.R. *et al.* (2007) A proteome-wide protein interaction map for Campylobacter jejuni. *Genome Biol.*, **8**, R130.

Patro,R. and Kingsford,C. (2012) Global network alignment using multiscale spectral signatures. *Bioinformatics*, **28**, 3105–3114.

Peregrín-Alvarez,J.M. *et al.* (2009) The modular organization of protein interactions in *Escherichia coli*. *PloS Comput. Biol.*, **5**, e1000523.

Pesquita,C. *et al.* (2009) Semantic similarity in biomedical ontologies. *PLoS Comput. Biol.*, **5**, e1000443.

Radivojac,P. *et al.* (2008) An integrated approach to inferring gene–disease associations in humans. *Proteins*, **72**, 1030–1037.

Robertson,N. and Seymour,P.D. (1984) Graph minors. III. Planar tree-width. *J. Comb. Theory Ser. B*, **36**, 49–64.

Rohan,T.E. (2009) *Proteomic Prediction of Breast Cancer Risk: A Cohort Study*. Albert Einstein Coll of Medicine of (Yeshiva Univ) Bronx NY.

Saraph,V. and Milenković,T. (2013) MAGNA: Maximizing Accuracy in Global Network Alignment. In: *arXiv:1311.2452 [q-bio.MN]*.

Schlicker,A. *et al.* (2006) A new measure for functional similarity of gene products based on gene ontology. *BMC Bioinformatics*, **7**, 302.

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.* (2008a) Global alignment of multiple protein interaction networks. In: *Proceeding Pacific Symposium Biocomputing*. *Citeseer*. Hawaii, USA, pp. 303–314.

Singh,R. *et al.* (2008b) Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proc. Natl Acad. Sci. USA*, **105**, 12763–12768.

Spirin,V. and Mirny,L.A. (2003) Protein complexes and functional modules in molecular networks. *Proc. Natl Acad. Sci. USA*, **100**, 12123–12128.

Wang,B. and Gao,L. (2012) Seed selection strategy in global network alignment without destroying the entire structures of functional modules. *Proteome Sci.*, **10**, S16.

Wang,E. *et al.* (2007) Cancer system biology: exploring caner-associated genes on cellular networks. *Cell Mol. Life Sci.*, **64**, 1752–1762.

Yu,H. *et al.* (2007) The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput. Biol.*, **3**, e59.

Zhao,J. *et al.* (2006) Complex networks theory for analyzing metabolic networks. *Chinese Sci. Bull.*, **51**, 1529–1537.

Zotenko,E. *et al.* (2008) Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality. *PloS Comput. Biol.*, **4**, e1000140.