BIOINFORMATICS

Vol. 25 ISMB 2009, pages i240–i246
doi:10.1093/bioinformatics/btp202

Domain-oriented edge-based alignment of protein interaction networks

Xin Guo* and Alexander J. Hartemink*

Department of Computer Science, Box 90129, Duke University, Durham, NC 27708-0129, USA

ABSTRACT

Motivation: Recent advances in high-throughput experimental techniques have yielded a large amount of data on protein–protein interactions (PPIs). Since these interactions can be organized into networks, and since separate PPI networks can be constructed for different species, a natural research direction is the comparative analysis of such networks across species in order to detect conserved functional modules. This is the task of network alignment.

Results: Most conventional network alignment algorithms adopt a *node-then-edge-alignment* paradigm: they first identify homologous proteins across networks and then consider interactions among them to construct network alignments. In this study, we propose an alternative *direct-edge-alignment* paradigm. Specifically, instead of explicit identification of homologous proteins, we directly infer plausibly alignable PPIs across species by comparing conservation of their constituent domain interactions. We apply our approach to detect conserved protein complexes in yeast–fly and yeast–worm PPI networks, and show that our approach outperforms two recent approaches in most alignment performance metrics.

Availability: Supplementary material and source code can be found at http://www.cs.duke.edu/∼amink/.

Contact: xinguo@cs.duke.edu; amink@cs.duke.edu

1 INTRODUCTION

Understanding complicated networks of interacting proteins is a major challenge in systems biology. Recently, with the rapid progress of high-throughput experimental techniques, protein–protein interaction (PPI) databases have rapidly increased in size, allowing for comparative analysis of PPI networks from which conserved modules can be identified across PPI networks of different species (Sharan and Ideker, 2006; Srinivasan *et al.*, 2007). By analogy to sequence alignment, this problem is called PPI network alignment.

Typically, PPI network alignment algorithms compare PPI networks of two or more species and identify conserved modules, e.g. pathways or protein complexes. Often a PPI network is represented as an undirected graph in which nodes indicate proteins and edges indicate interactions. Hence, the network alignment problem can also be viewed as a graph isomorphism problem.

Many network alignment algorithms have been proposed in recent years and most of them focus on the pairwise alignment of PPI networks. As an early approach, PathBLAST (Kelley *et al.*, 2003) proposed a likelihood-based scoring scheme to search for conserved pathways. Sharan *et al.* (2005a) extended PathBLAST to employ a greedy heuristic to detect conserved protein complexes across species. NetworkBLAST-E (Hirsh and Sharan, 2007) introduced an evolutionary model of networks into the alignment scoring function to extract conserved complexes. MaWISh (Koyutürk *et al.*, 2006) merged pairwise interaction networks into a single alignment graph and treated network alignment as a maximum weight induced subgraph problem. MNAligner (Li *et al.*, 2007) described an integer quadratic programming (IQP) model to identify conserved substructures.

Recently, several network alignment algorithms have been developed that can align more than two species. Graemlin (Flannick *et al.*, 2006) is capable of aligning at least 10 microbial networks at once. NetworkBLAST (Sharan *et al.*, 2005b), another extension of PathBLAST, can align networks of up to three species, and its later version, NetworkBLAST-M (Kalaev *et al.*, 2008), can align 10 networks with tens of thousands of proteins in minutes. In addition, Singh *et al.* (2008) described a method inspired by Google's PageRank to detect global alignments from five eukaryotic PPI networks.

All these network alignment algorithms follow a *node-then-edge-alignment* paradigm. That is, they generally first need to identify homologous proteins across species before they can exploit protein interaction and network topology information to detect conserved subnetworks. The node alignment step essentially acts as a filter, artificially constraining the search space of conserved modules to putatively homologous protein pairs. However, proteins rarely act alone. They interact with each other to carry out their activities, and these interacting proteins are likely to evolve with high correlation during the evolution of species (Goh *et al.*, 2000; Mintseris and Weng, 2005; Pazos *et al.*, 1997). Furthermore, it has been shown recently that such co-evolution is more evident if we focus our attention on interacting domains that are responsible for PPIs (Itzhaki *et al.*, 2006; Jothi *et al.*, 2006; Schuster-Böckler and Bateman, 2007). Based on these observations, we present DOMAIN, an algorithm for *dom*ain-oriented *a*lignment of *i*nteraction *n*etworks, that follows an alternative *direct-edge-alignment* paradigm. DOMAIN does not explicitly restrict its attention to putatively homologous proteins. Instead, it directly aligns PPIs across species by decomposing PPIs in terms of their constituent domain–domain interactions (DDIs) and looking for conservation of these DDIs. We apply DOMAIN to detect conserved protein complexes in yeast–fly and yeast–worm PPI networks, and demonstrate that it achieves better results than two previous techniques in most performance metrics.

The article is organized as follows: Section 2 presents the details of DOMAIN. Section 3 describes the quality assessment measures, as well as the experimental results of DOMAIN compared with two extant methods. In Section 4, we discuss implications of the results, along with further directions.

*To whom correspondence should be addressed.

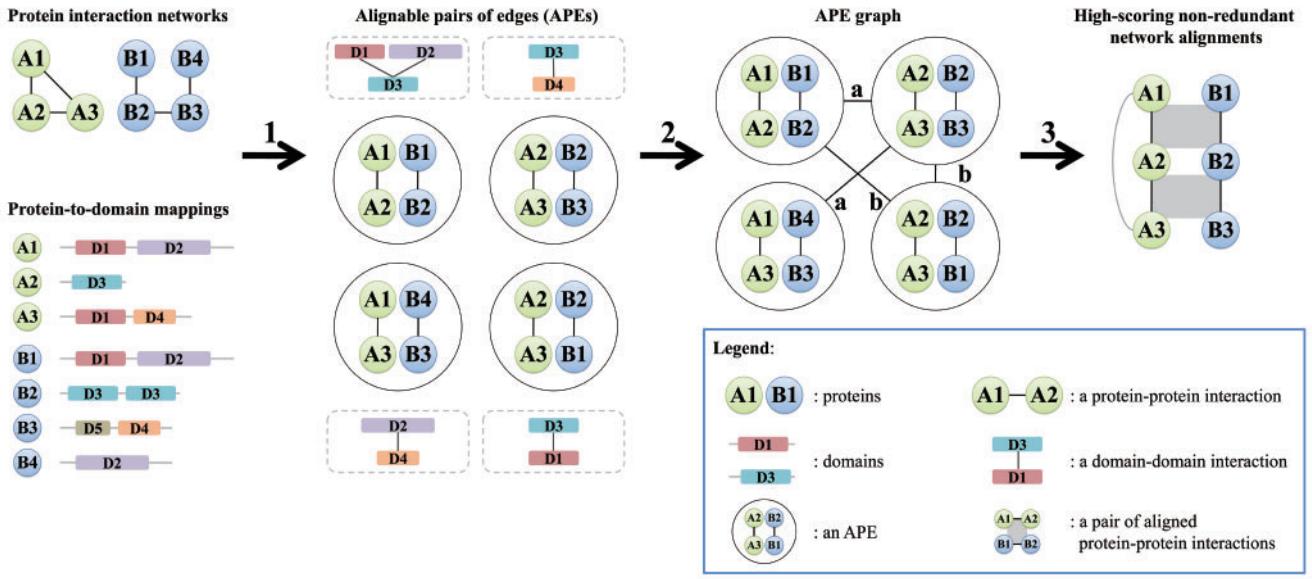

Fig. 1. Method overview. (1) Constructing APEs. The input of DOMAIN includes two PPI networks and the constituent domains of the proteins. Using this information, DOMAIN calculates species-specific DDI probabilities, and then identifies a set of APEs across networks. (2) Building an APE graph. An APE graph is a merged representation of the PPI networks, in which each node represents an APE and each edge represents one of four network connectivities connecting two APEs: (a) alignment extension, (b) node duplication, (c) edge indel (insertion/deletion), or (d) edge jump. The details of these connectivities are given in Section 2.2. (3) Searching for high-scoring non-redundant subgraphs within the APE graph. We use a greedy heuristic to carry out this task.

2 METHODS

As illustrated in Figure 1, DOMAIN consists of three stages: (1) it constructs a complete set of alignable pairs of edges (APEs); (2) it builds an APE graph; and (3) it employs a heuristic search to identify conserved protein complexes across species. The three subsections that follow elaborate upon these three stages.

2.1 Constructing and scoring APEs

Domains are structural and functional units of proteins. Many studies (Bernard *et al.*, 2007; Deng *et al.*, 2002; Riley *et al.*, 2005) have revealed that direct PPIs are often mediated by interactions between the constituent domains of the two interacting proteins. These studies have made two particular assumptions that we adopt as well: (1) DDIs are independent of each other, and (2) two proteins interact if at least one pair of domains from two proteins interact. These assumptions allow us to formulate the probability of an interaction between two proteins in terms of a 'noisy-or' over the DDIs that might possibly mediate the interaction between those two proteins. In our network alignment scenario where we seek to align edges directly, we additionally assume that a pair of cross-species PPIs can be aligned to one other only if they are plausibly mediated by at least one common DDI.

We represent the input PPI networks from two species as undirected graphs $G_1(V_1, E_1)$ and $G_2(V_2, E_2)$, where nodes indicate proteins and edges indicate the observed PPIs. We first wish to construct a complete set of APEs. We say that a pair of edges, $e_1 \in E_1$ and $e_2 \in E_2$, is *alignable* if there exists a DDI that can plausibly mediate the two PPIs represented by that pair of edges. We say that a DDI can *plausibly mediate* a PPI if the corresponding interaction probability between the two domains is above some value $\epsilon > 0$. Using a non-zero value for ϵ allows us to filter out domains between which there is negligible evidence of a DDI.

For an edge $e \in E_1$ or E_2, we define $\mathcal{D}(e)$ to be all the possible interactions between the constituent domains of the two proteins. Given the species-specific probabilities of DDIs that mediate PPIs, we can then write the score of an APE $c = (e_1, e_2)$ using a 'noisy-or' formulation:

$$f(c) = \Pr(e_1, e_2 | \Theta^1, \Theta^2) = 1 - \prod_{d_{\alpha,\beta} \in \mathcal{D}(e_1) \bigcap \mathcal{D}(e_2)} (1 - g(\theta^1_{\alpha,\beta}, \theta^2_{\alpha,\beta}))$$

where $d_{\alpha,\beta}$ denotes an interaction between domains α and β, and $\theta_{\alpha,\beta} = \Pr(d_{\alpha,\beta})$, and $\Theta = \{\theta_{\alpha,\beta}\}$. The function $g(\theta^1_{\alpha,\beta}, \theta^2_{\alpha,\beta})$ measures the probability of aligning the PPI e_1 to the PPI e_2 mediated by interactions between domains α and β. In this work, we have chosen to set $g(\theta^1_{\alpha,\beta}, \theta^2_{\alpha,\beta}) = (\theta^1_{\alpha,\beta} \cdot \theta^2_{\alpha,\beta})^{1/2}$.

As previous authors have also done, to estimate the species-specific DDI probabilities Θ, we applied the EM (expectation–maximization) algorithm of Deng *et al.* (2002) for each given network.

2.2 Building an APE graph

The APE graph is motivated by the evolutionary model of PPI networks suggested by Berg *et al.* (2004). The model indicates that PPI networks are shaped primarily by two kinds of evolutionary events, *link dynamics* and *gene duplication*. Link dynamics events are primarily caused by sequence mutations of a gene and affect the connectivities of the protein whose coding sequence undergoes mutations. Gene duplication, the second kind of evolutionary event, is often followed by either silencing of one of the duplicated genes or by functional divergence of the duplicates. From the perspective of protein domains, a link dynamics event may result from switching a constituent domain of a protein to another, or a change in a domain's interaction partners; a gene duplication event consists of duplication of one protein, followed by a domain switching or being removed in one or both of the duplicates, or followed by progressive small changes from point mutations that cause a change in domain interaction partners.

With this motivation in place, we define an APE graph to be an undirected weighted graph, where nodes correspond to the APEs identified above, and edges correspond to one of four evolutionary relationships that we consider between two APEs, as illustrated in Figure 2 and as listed below:

(a) Alignment extension: two APEs are connected if they share two proteins, one per species.

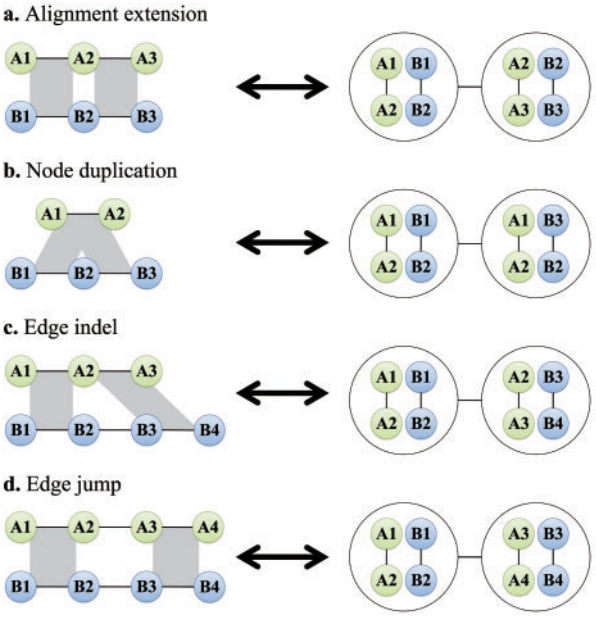

Fig. 2. Four connectivities in an APE graph. The details of these connectivities are given in the text, and the legend is the same as in Figure 1.

(b) Node duplication: two APEs are connected if they share a protein in one species and a PPI in the other.

(c) Edge indel (insertion/deletion): two APEs are connected if they share a protein in one species and the graph distance between the two PPIs in the other network is 1.

(d) Edge jump: in this case, all proteins within the two APEs are distinct, but for each species, the graph distance between the two PPIs in their corresponding network is 1. We consider this case because our current knowledge of both PPIs and DDIs is noisy and incomplete. Thus, if there exists a pair of PPIs that can make two APEs connected in each network, we treat the pair as a potential APE. Note that some insignificant DDIs (probabilities of DDIs $<\epsilon$) are shared in such potential APEs.

Given this definition of an APE graph, we note that every subgraph in an APE graph corresponds to a network alignment.

Each node in an APE graph contributes the score $f(c)$ of its corresponding APE, and each edge is scored by a positive number according to its connection relationship. Using these edge scores, we want to reward alignment extension and penalize both node duplication and edge indel. Let $\gamma_a, \gamma_b, \gamma_c$ and γ_d be the edge scores of alignment extension, node duplication, edge indel and edge jump, respectively. We thus need to assign $\gamma_a > 1$ and $\gamma_b, \gamma_c < 1$. Because we neither wish to reward nor penalize an edge jump, we simply assign $\gamma_d = 1$. For a subgraph $G_s(V_s, E_s)$ in an APE graph, the overall score for its corresponding network alignment is calculated as

$$S(G_s) = \prod_{e \in E_s} \gamma(e) \cdot \prod_{c \in V_s} f(c)$$

where $\gamma(e)$ is the edge score for $e \in E_s$, and $f(c)$ is the score of the APE $c \in V_s$.

2.3 Detecting protein complexes

Network alignment methods generally require a search algorithm to detect high-scoring subgraphs from a single or several weighted graphs. Such tasks are computationally difficult, so a number of search heuristics have been proposed: for example, PathBLAST uses randomized dynamic programming to search for conserved pathways across networks, while NetworkBLAST-E implements a greedy heuristic to search for conserved protein complexes. As many pairwise network methods aim to identify conserved protein complexes, for comparative purposes, we devise a greedy heuristic for finding conserved protein complexes across species.

The heuristic aims to identify high-scoring non-redundant subgraphs from the resultant APE graph. Specifically, exhaustively starting from each APE, we iteratively expand the subgraph by introducing a new APE that increases the alignment score the most, until any of the following empirical stopping conditions occur: (i) the number of proteins in either species exceeds an upper limit (we used 15); (ii) the score of the next expanding APE is smaller than a threshold (we used 10^{-2}); (iii) the overall alignment score of the subgraph is smaller than a threshold (we used 10^{-3}); or (iv) the graph distance of the next expanding APE exceeds an upper limit (we used 4). At the end, small and redundant subgraphs are removed if the number of proteins in a subgraph is less than four, or if there exists a higher scoring subgraph overlapping $>80\%$ of proteins in either species.

3 RESULTS

3.1 Experimental setup

We compare our method to two extant pairwise network alignment algorithms, NetworkBLAST and MaWISh. We do not include NetworkBLAST-M and Graemlin in our comparisons because they mainly focus on alignment of multiple networks, and because Graemlin requires the unavailable in-house SRINI algorithm (Srinivasan *et al.*, 2006) to assign weights to PPIs. The ISOrank algorithm aims at resolving a different problem of aligning networks globally, while NetworkBLAST-E performs similarly to NetworkBLAST and is not available online. We thus exclude these methods from the comparisons as well.

We apply DOMAIN on yeast–fly and yeast–worm PPI networks taken from DIP (Database of Interacting Proteins, Oct 2008) (Xenarios *et al.*, 2002), as they were widely used in pairwise network alignment studies as benchmarks. The protein-to-domain mappings are taken from Pfam (Pfam 23.0) (Finn *et al.*, 2008), and we only consider high-quality Pfam-A entries. Because not all proteins contain significant Pfam domains, we generate a so-called 'backbone' network, a subnetwork of DIP in which all proteins contain at least one Pfam-A domain. As summarized in Table 1, 78.2% of MIPS annotated proteins and over 70% of GO annotated proteins are contained in backbone networks. To simplify the setting of the four γ parameters, we reduced the parameter space to one dimension by insisting that $\gamma_a = k$, $\gamma_b = \gamma_c = 1/k$ and $\gamma_d = 1$, for some value of $k > 1$. We found that DOMAIN was not sensitive to changes in k. In the results that follow, we used $k = 10$.

3.2 Experimental results

We employ three measures to evaluate the biological significance of the alignments: sensitivity/specificity, MIPS purity and GO enrichment. These measures are also suggested in several other network alignment studies (Dutkowski and Tiuryn, 2007; Hirsh and Sharan, 2007; Kalaev *et al.*, 2008).

The first two measures use the known yeast protein complexes cataloged in MIPS (May 2006) (Mewes *et al.*, 2002) as a gold standard. We exclude category 550 (obtained from high-throughput experiments) and only use complexes at level 3 or lower. In consequence, there exist 122 MIPS complexes spanning 519 yeast proteins in the yeast backbone network, 62 of which contain at

Table 1. Summary of backbone networks

	DIP			Backbone DIP		
	Yeast	Fly	Worm	Yeast	Fly	Worm
Number of PPIs	17 528	22 381	4038	11 426	11 013	2213
Number of proteins	4928	7446	2644	3300	4500	1620
Number of GO annotated proteins[a]	4625	4477	1566	3280	3253	1145
Number of MIPS annotated proteins[b]	1100	–	–	860	–	–

[a]With respect to the biological process annotation of Gene Ontology.
[b]Excluding MIPS category 550.

least three proteins and span 438 proteins. For each identified yeast alignment, we try to find a complex from MIPS that maximizes the hypergeometric score and calculate an empirical enrichment *P*-value. The significance level is obtained from sampling 10 000 random sets of proteins of the same size, and the *P*-values are corrected for multiple testing using the false discovery rate (FDR) (Benjamini and Hochberg, 1995). Then, the specificity is defined as the percent of yeast alignments that have a significant match in MIPS ($P<0.05$), and the sensitivity is defined as the percent of MIPS alignments that have significant matches in the resulting alignments. Moreover, an alignment is called a pure alignment if it satisfies two conditions: (i) it contains at least three MIPS annotated proteins and (ii) there exists a complex in MIPS that covers >75% of its MIPS annotated proteins. We report purity, calculated by the number of pure alignments divided by the total number of alignments with at least three MIPS annotated proteins, as an alternative measure of the sensitive identification of specific complexes.

GO enrichment measures the functional coherence of the proteins in an identified alignment with respect to the biological process annotation of GO, for each species separately. We use the tool GO TermFinder (Boyle *et al.*, 2004) to compute empirical enrichment *P*-values, and correct for multiple testing using FDR. For each species, we report the fraction of process-coherent alignments with *P*-value <0.05 (considering only the alignments with at least one GO annotated protein).

We chose to set the probability threshold of DDIs ϵ to the low but non-zero value of 10^{-20} so as to take into account as much DDI information as possible. For yeast–fly alignment, DOMAIN generated an APE graph consisting of 6918 APEs with 47 964 alignment extension links, 24 549 node duplication links, 5573 edge indel links and 1149 edge jump links; for yeast–worm alignment, it returned a 1410-node APE graph with 4230 alignment extension links, 4087 node duplication links, 140 edge indel links and 37 edge jump links. For accurate comparison, we applied NetworkBLAST and MaWISh on backbone networks with their suggested parameter settings [see Koyutürk *et al.* (2006) and Sharan *et al.* (2005b) for details]. As summarized in Tables 2 and 3, DOMAIN identified more significant non-redundant alignments than NetworkBLAST and MaWISh in both alignments—explaining the good scores on the sensitivity metric—but also managed to outperform the other methods on the specificity and purity metrics. Indeed, it achieved the highest performance on almost every evaluation metric, and in the instances in which it was bested, the difference is slight.

The running time of DOMAIN is comparable with MaWISh and NetworkBLAST. DOMAIN is currently implemented in Perl, and its running time on yeast–fly and yeast–worm backbone networks is <1 min (Intel Core 2 CPU 6600@2.4 GHz, 2 GB RAM). Because the running time is so small, we were able to exhaustively expand from all APEs. If for some reason we needed to further reduce computational complexity, we could instead consider an alternative expansion strategy where we would expand only from 'seed' APEs. The idea would be that if a protein complex is conserved in many species, the PPIs in this complex are likely to be conserved as well, and therefore the corresponding subgraph in the APE graph should contain many alignment extension links. With this in mind, we could rank the APEs by counting the number of their surrounding alignment extension links and select, say, the top 25% as seeds for expansion. We tested this, and the results were nearly identical to those listed in Tables 2 and 3, but the running time for yeast–fly and yeast–worm alignments reduces to 30 and 15 s, respectively. In our case, the running time was not a problem, but it is reassuring that a seed-based expansion strategy seems to be effective at reducing the running time without affecting the results.

3.3 Case studies

DOMAIN is sensitive at detecting small network alignments that might be deemed by other algorithms to be topologically insignificant. For example, DOMAIN reported a network alignment between the yeast NEF1 complex and the fly proteins mei-9, Ercc1 and Xpac with high confidence (Fig. 3). The GO process coherence of these three fly proteins is significant: nucleotide excision repair ($P\simeq10^{-8}$), DNA repair ($P\simeq10^{-6}$), cellular response to DNA damage stimulus ($P\simeq10^{-6}$), etc. However, neither MaWISh nor NetworkBLAST reports any alignment involving the yeast NEF1 complex. They are likely to miss such alignments because (i) the sequence similarity between Rad10 and Ercc1 is insignificant (BLAST *E*-value $\simeq10^{-8}$) and may be ignored if using a restrictive BLAST *E*-value threshold [e.g. 10^{-10} suggested in Hirsh and Sharan (2007)], and (ii) this alignment consists of only three matched proteins and two conserved interactions, so it may not be sufficiently topologically significant for some aligners to detect. On the other hand, the DDIs within this alignment are well-conserved across

Table 2. Performance comparisons of DOMAIN with NetworkBLAST and MaWISh on yeast–fly backbone networks

Method	No. of complexes	No. of proteins		Specificity (%)	Sensitivity (%)	MIPS purity (%)	GO enrichment	
		Yeast	Fly				Yeast (%)	Fly (%)
DOMAIN	**100**	**338**	**313**	**34.0**	**9.0**	**66.7**	**89.0**	78.0
NetworkBLAST	82	299	213	31.7	7.4	40.6	87.8	**79.3**
MaWISh	54	193	142	18.5	4.1	30.0	75.9	66.7

The largest value in each column is indicated in bold.

Table 3. Performance comparisons of DOMAIN with NetworkBLAST and MaWISh on yeast–worm backbone networks

Method	No. of complexes	No. of proteins		Specificity	Sensitivity(%)	MIPS purity	GO enrichment	
		Yeast	Worm				Yeast (%)	Worm (%)
DOMAIN	**21**	**84**	**63**	**36.4**	**3.3**	**75.0**	**90.5**	9.5
NetworkBLAST	19	82	51	7.7	0.8	60.0	89.5	**10.5**
MaWISh	11	42	32	11.1	1.6	42.8	63.6	9.1

The largest value in each column is indicated in bold.

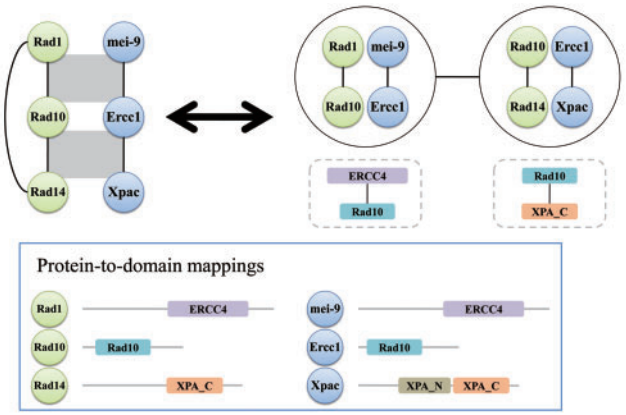

Fig. 3. DOMAIN reports a network alignment between the yeast NEF1 complex (MIPS category 510.180.10.10) and the fly proteins mei-9, Ercc1 and Xpac. The object to the right of the double arrow depicts the corresponding subgraph of this alignment in the APE graph.

species (the DDI probabilities of `ERCC4-Rad10` are 1.00 in both species; the DDI probabilities of `Rad10-XPA_C` are 1.00 and 0.54 in yeast and fly, respectively).

Another advantage of DOMAIN is that often it provides a more comprehensive means of interpreting the identified network alignments, because protein domains are directly relevant to function in many cases. For instance, Rad14 and Xpac may play a similar role in the biological process of nucleotide excision repair, as they share a common `XPA_C` domain. Furthermore, although the `XPA_N` domain is not reported as a significant domain for Rad14 in Pfam (*E*-value = 0.023), the alignment of yeast Rad14 to fly Xpac suggests that `XPA_N` is potentially an important functional domain in Rad14.

Identifying conserved biological pathways across species is another important application of network alignment. Figure 4a demonstrates an example of alignment reported by DOMAIN

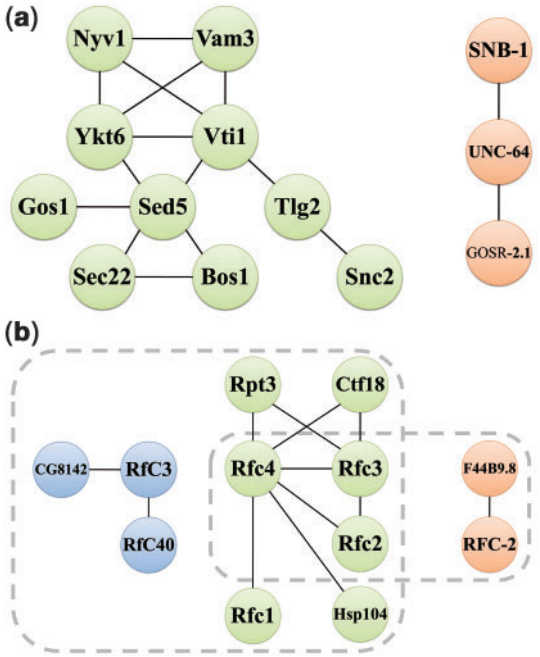

Fig. 4. (**a**) DOMAIN reports an alignment between 10 yeast proteins and 3 worm proteins that significantly matches the pathway of SNARE interactions in vesicular transport in KEGG. (**b**) An example of improving network alignment by combining several cross-species pairwise alignments. (Green, yeast proteins; blue, fly proteins; orange, worm proteins.)

between 10 yeast proteins and three worm proteins, in which nine of the yeast proteins (all except Nyv1) and all three worm proteins are known to be involved in the pathway of SNARE interactions in vesicular transport in KEGG (Kanehisa and Goto, 2000).

Alignment performance may further be improved by combining several cross-species pairwise network alignments. Figure 4b shows an example of combining three alignments taken from yeast–fly, yeast–worm and fly–worm network alignments, respectively. By aligning yeast and fly networks, DOMAIN detects an alignment between three fly proteins (CG8142, RfC3, and RfC40) and seven yeast proteins, and four of them (Rfc1-4) are involved in the replication factor C complex (MIPS: 410.40.30). As the yeast replication factor C complex contains five proteins (Rfc1-5), the F-score[1] is 0.67. Further, we see that two worm proteins (F44B9.8 and Rfc-2) are aligned to all these three fly proteins in fly–worm alignment and three of these seven yeast proteins (Rfc2-4) in yeast–worm alignment. This three-way alignment suggests that the alignment between fly proteins CG8142, RfC3 and RfC40 and yeast proteins Rfc2-4 are of high confidence, and the F-score is increased to 0.75.

4 CONCLUSIONS

In this study, we described DOMAIN, a domain-oriented pairwise network alignment framework. To our knowledge, DOMAIN is the first algorithm to introduce protein domains into the network alignment problem. Also, DOMAIN uses a novel *direct-edge-alignment* paradigm to directly detect equivalent PPI pairs across species and suggests a new graph representation to merge these equivalent PPI pairs and their network evolutionary-based relationships into one graph. We tested DOMAIN to identify conserved protein complexes in the yeast–fly and yeast–worm protein interaction networks, and the experimental results show that DOMAIN exhibits better performance than two recent pairwise network alignment methods in most performance metrics.

Although DOMAIN can be applied only to the subset of proteins with domain mappings, we notice that most functionally annotated proteins contain domain structures and remain in this subset. To overcome this restriction, we may employ a larger domain database, e.g. CDD (Marchler-Bauer *et al.*, 2007), or combine DOMAIN with other network aligners. In addition, as the set of defined domains expands and is refined over time, this will gradually become less of a restriction.

Further directions for research include extending this approach to multiple network alignment and to network querying. Since multiple network alignment requires more than two networks by definition, we would simply need to devise an appropriate scoring scheme that can handle more than a pair of alignable PPIs at once, and then extend the notion of the APE graph accordingly.

The goal of network querying is to identify subnetworks in a given network that are similar to the query. Typically, the query is a hypothetical or known functional module. We may simply treat the query as a small input network and apply our DOMAIN method directly on it. A more sophisticated approach would be to devise a sequence-profile-like structure to describe the DDI contents of the network query, as well as perhaps constructing such structures for the full network as a one-time expense for many successive queries.

[1] F-score is defined as $F = 2 \times (\text{precision} \times \text{recall})/(\text{precision} + \text{recall})$.

ACKNOWLEDGEMENTS

The authors would like to thank Bruce Donald and anonymous reviewers for helpful discussions and comments on this article, and Michael Mayhew for suggesting the name DOMAIN.

Funding: Duke Graduate School Fellowship (to X.G.); a National Science Foundation CAREER award (NSF 0347801 to A.J.H.); Alfred P. Sloan Research Fellowship (to A.J.H.); National Institutes of Health (P50-GM081883-01 and R01-ES015165-01 to A.J.H.); DARPA (HR0011-08-1-0023 to A.J.H.).

Conflict of Interest: none declared.

REFERENCES

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Royal Stat. Soc.,* **57**, 289–300.

Berg,J. *et al*. (2004) Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications. *BMC Evol. Biol.*, **4**, 51.

Bernard,A. *et al*. (2007) Reconstructing the topology of protein complexes. In *Proceedings of the 11th Annual International Conference on Research in Computational Molecular Biology (RECOMB 2007)*, LNBI 4453, pp. 32–46.

Boyle,E. *et al*. (2004) GO::TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics,* **20**, 3710–3715.

Deng,M. *et al*. (2002) Inferring domain-domain interactions from protein-protein interactions. *Genome Res.*, **12**, 1540–1548.

Dutkowski,J. and Tiuryn,J. (2007) Identification of functional modules from conserved ancestral protein-protein interactions. *Bioinformatics,* **23**, i149–i158.

Finn,R. *et al*. (2008) The Pfam protein families database. *Nucleic Acids Res.*, **36**, D281–D288.

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

Goh,C.S. *et al*. (2000) Co-evolution of proteins with their interaction partners. *J. Mol. Biol.*, **299**, 283–293.

Hirsh,E. and Sharan,R. (2007) Identification of conserved protein complexes based on a model of protein network evolution. *Bioinformatics*, **23**, e170–e176.

Itzhaki,Z. *et al*. (2006) Evolutionary conservation of domain-domain interactions. *Genome Biol.*, **7**, R125.

Jothi,R. *et al*. (2006) Co-evolutionary analysis of domains in interacting proteins reveals insights into domain-domain interactions mediating protein-protein interactions. *J. Mol. Biol.*, **362**, 861–875.

Kalaev,M. *et al*. (2008) Fast and accurate alignment of multiple protein networks. In *Proceedings of the 12th Annual International Conference on Research in Computational Molecular Biology (RECOMB 2008)*, LNBI 4955, pp. 246–256.

Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.

Kelley,B.P. *et al*. (2003) Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proc. Natl Acad. Sci. USA*, **100**, 11394–11399.

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

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

Marchler-Bauer,A. *et al*. (2007) CDD: a conserved domain database for interactive domain family analysis. *Nucleic Acids Res.*, **35**, D237–D240.

Mewes,H. *et al*. (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.*, **30**, 31–34.

Mintseris,J. and Weng,Z. (2005) Structure, function, and evolution of transient and obligate protein-protein interactions. *Proc. Natl Acad. Sci. USA*, **102**, 10930–10935.

Pazos,F. *et al*. (1997) Correlated mutations contain information about protein-protein interaction. *J. Mol. Biol.*, **271**, 511–523.

Riley,R. *et al*. (2005) Inferring protein domain interactions from databases of interacting proteins. *Genome Biol.*, **6**, R89.

Schuster-Böckler,B. and Bateman,A. (2007) Reuse of structural domain-domain interactions in protein networks. *BMC Bioinformatics*, **8**, 259.

Sharan,R. and Ideker,T. (2006) Modeling cellular machinery through biological network comparison. *Nat. Biotechnol.*, **24**, 427–433.

Sharan,R. *et al*. (2005a) Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data. *J. Comput. Biol.*, **12**, 835–846.

Sharan,R. *et al*. (2005b) 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 with application to functional orthology detection. *Proc. Natl Acad. Sci. USA*, **105**, 12763–12768.

Srinivasan,B. *et al*. (2006) Integrated protein interaction networks for 11 microbes. In *Proceedings of the 10th Annual International Conference on Research in Computational Molecular Biology (RECOMB 2006)*, LNBI 3909, pp. 1–14.

Srinivasan,B. *et al*. (2007) Current progress in network research: toward reference networks for key model organisms. *Brief. Bioinformatics*, **8**, 318–332.

Xenarios,I. *et al*. (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.*, **30**, 303–305.