BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 18 2014, pages 2619–2626
doi:10.1093/bioinformatics/btu358

Systems biology

Advance Access publication May 28, 2014

DualAligner: a dual alignment-based strategy to align protein interaction networks

Boon-Siew Seah[1,2,*], Sourav S. Bhowmick[1,2,*] and C. Forbes Dewey, Jr[2,3]

[1]Division of Software and Information Systems, School of Computer Engineering, Nanyang Technological University, [2]Singapore-MIT Alliance, Nanyang Technological University, Singapore 639798 and [3]Biological Engineering Department, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

Associate Editor: Igor Jurisica

ABSTRACT

Motivation: Given the growth of large-scale protein–protein interaction (PPI) networks obtained across multiple species and conditions, network alignment is now an important research problem. Network alignment performs comparative analysis across multiple PPI networks to understand their connections and relationships. However, PPI data in high-throughput experiments still suffer from significant false-positive and false-negatives rates. Consequently, high-confidence network alignment across entire PPI networks is not possible. At best, local network alignment attempts to alleviate this problem by completely ignoring low-confidence mappings; global network alignment, on the other hand, pairs all proteins regardless. To this end, we propose an alternative strategy: instead of full alignment across the entire network or completely ignoring low-confidence regions, we aim to perform highly specific protein-to-protein alignments where data confidence is high, and fall back on broader functional region-to-region alignment where detailed protein–protein alignment cannot be ascertained. The basic idea is to provide an alignment of multiple granularities to allow biological predictions at varying specificity.

Results: DualAligner performs *dual network alignment*, in which both region-to-region alignment, where whole subgraph of one network is aligned to subgraph of another, and protein-to-protein alignment, where individual proteins in networks are aligned to one another, are performed to achieve higher accuracy network alignments. Dual network alignment is achieved in DualAligner via background information provided by a combination of Gene Ontology annotation information and protein interaction network data. We tested DualAligner on the global networks from *IntAct* and demonstrated the superiority of our approach compared with state-of-the-art network alignment methods. We studied the effects of parameters in DualAligner in controlling the quality of the alignment. We also performed a case study that illustrates the utility of our approach.

Availability and implementation: http://www.cais.ntu.edu.sg/~assourav/DualAligner/

Contact: seah0097@ntu.edu.sg or assourav@ntu.edu.sg

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

Received on July 13, 2013; revised on March 18, 2014; accepted on May 21, 2014

1 INTRODUCTION

Protein–protein interaction (PPI) network models the cooperation between proteins to drive processes within a biological system. Advancement in high-throughput technologies has led to rapid growth of large scale PPI networks obtained across multiple species or conditions, and therefore, comparative analysis to understand their connections and relationships is now an important research problem (Sharan and Ideker, 2006). To this end, the *network alignment* problem aims to align a set of PPI networks belonging to different species to find *conserved* subnetworks within them. These conserved regions could explain mechanisms of function conservation and evolution beyond what is available to individual gene studies.

The network alignment problem consists of two key steps: a *scoring framework* that captures the knowledge about module evolution and *identification* of high scoring alignments (conserved subnetworks or modules) from among exponentially large set of possible alignments. Unfortunately, finding conserved subnetworks in a group of networks is NP-Hard (Sharan and Ideker, 2006). Consequently, a range of heuristic-based network alignment methods have been proposed that fall into two categories, namely, *local* (Kalaev *et al*., 2008; Kelley *et al*., 2004; Koyutürk *et al*., 2006) and *global* network alignment (Kuchaiev and Przulj, 2011; Liao *et al*., 2009; Singh *et al*., 2008; Zaslavskiy *et al*., 2009).

The local network alignment approach identifies multiple unrelated conserved regions between the input networks, each region implying a mapping independent of others. It is particularly useful in finding known functional components (pathways, complexes) in a new species. For instance, PathBLAST (Kelley *et al*., 2004) aligns linear pathways based on homology and interaction confidence. NetworkBLAST-M (Kalaev *et al*., 2009) finds highly conserved local regions greedily using inferred phylogeny. Another local alignment method is MaWish (Koyutürk *et al*., 2006), which is modeled on evolution of protein interactions.

On the other hand, the global network alignment approach aligns every node in the smaller network to the larger network to find 'best' overall alignment between the input networks. Particularly, it enables species-level comparisons and discovery of functional orthologs. For instance, IsoRank (Singh *et al*., 2008) and IsoRankN (Liao *et al*., 2009) identify a stationary random walk distribution to perform global network alignment. Græmlin 2.0 (Flannick *et al*., 2006) uses a training set of alignments to learn phylogeny relationships before performing an

*To whom correspondence should be addressed.

alignment. MI-GRAAL (Kuchaiev and Przulj, 2011) is an alignment framework that could integrate multiple similarity measures to construct a global alignment. Zaslavskiy *et al.* (2009) model global network alignment as a graph matching problem with hard constraints given by a set of *must-link* pairs of proteins. These *must-link* pairs are identified by sequence orthologs.

More recently, the ILP-based Natalie proposes a *Lagrangian relaxation approach* to solve the problem approximately (Pache and Aloy, 2012). PINALOG combines both the similarities of protein sequence and protein function to compute an alignment between two PPI networks (Phan and Sternberg, 2012). To this end, the Gene Ontology (GO) function similarity and sequence similarity are obtained independently and then combined using a linear combination of the independent similarities weighted by a factor θ.

Despite considerable progress made by the bioinformatics community in devising high-quality network alignment strategies, state-of-the-art network alignment techniques suffer from a key drawback. Specifically, they depend on protein sequence similarity to facilitate network alignment. Unfortunately, sequence similarity is only relevant to a subset of highly conserved proteins, leaving significant network regions poorly specified by sequence homology. Furthermore, it is well known that structural information of PPI networks suffers either from high false-positive rate of current high-throughput experiments or from false negatives because of incomplete data (Huang and Bader, 2009). *Consequently, in significant regions of a network, high-confidence alignment of proteins is not possible*. At best, local network alignment attempts to alleviate this problem by completely ignoring low-confidence mappings; global network alignment, on the other hand, pairs all proteins regardless.

In this article, we address the aforementioned limitation by taking a GO annotation-driven 'dual alignment' (Here 'dual' refers to the two-step alignment process and not duality in optimization) strategy to align a pair of PPI networks. Instead of *only* performing protein-to-protein alignments, we perform high granularity protein-to-protein alignment where confidence is high and more general functional *region-to-region* alignment where detailed protein–protein alignment cannot be ascertained, which may still yield biologically informative conclusions. Consequently, our method not only aligns highly conserved protein pairs but also aligns *functional subgraphs* of one network to functional subgraphs of another. Informally, a *functional subgraph* is a connected component of the network whose nodes share a particular biological role or function (annotated using GO).

The region-to-region and protein-to-protein alignments in our dual alignment strategy are not independent of each other, but inextricably linked. Specifically, region-to-region alignment sets the foundation for broad associations between functional regions. On the other hand, protein-to-protein alignment specifies detailed associations between nodes within each associated regions. Figure 1 illustrates with an example on how functionally conserved subgraphs are first aligned, followed by alignment of the underlying interaction structure. First, functional subgraphs of the networks are identified. For instance, a subgraph that shares the transport role is identified in the human network. Alignment between pairs of functional subgraphs is carried out and high-confidence pairs are identified based on the structural and sequence similarities of their underlying subgraphs. For instance, f_2 in Figure 1(c) depicts an alignment between the transport-associated subgraph and microtubule binding-associated subgraph, highlighting the connection between the regions. Using region-to-region alignments, we model each region-to-region pair as *soft constraints* that *probabilistically restrict* protein-to-protein alignment within the regions. Following that, protein-to-protein alignments are computed using these soft constraints as priors.

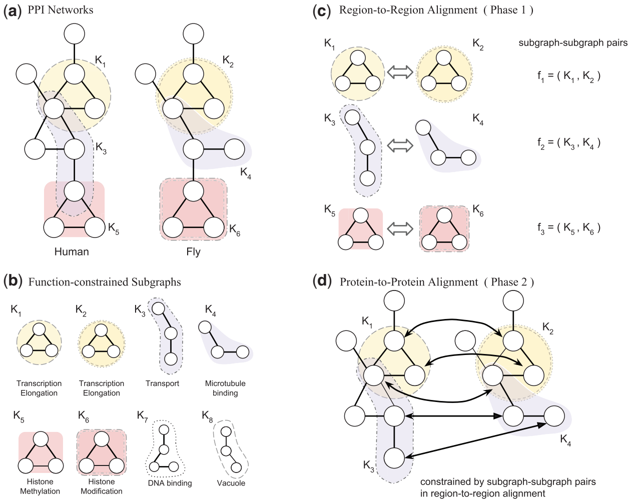

Fig. 1. Illustration of DualAligner. (**a**) GO annotated PPI networks are used as input networks. (**b**) Function-constrained subgraphs (connected subgraphs sharing a GO function) are constructed, representing functionally coherent regions in each network. (**c**) DualAligner computes a region-to-region alignment that best matches function-constrained subgraphs of the human network to the fly network. This involves pairwise subgraph-to-subgraph subalignment to identify optimally conserved subgraph pairs. Here, three pairwise alignments (f_1, f_2 and f_3) are shown. Note that K_7 and K_8 are not chosen in the alignment—only best conserved regions are aligned. (**d**) DualAligner computes fine-grained protein-to-protein alignment using the aligned regions in (c) as background information. Protein-to-protein alignments are *soft constrained* within subgraph–subgraph pairs, that is, any alignment between two proteins is *probabilistically* restricted within one or more subgraph–subgraph pairs in (c)

We refer to the aforementioned network alignment model as *function-constrained network alignment* (The first step, i.e. region-to-region alignment, represents function constraints that restricts the second step, i.e. protein-to-protein alignment; detailed in Section 2). It is worth mentioning the salient feature of our model. By integrating the rich set of functional information as soft constraints with the 'hard' constraints (Zaslavskiy *et al.*, 2009) (e.g. known orthologs) in the alignment framework, our dual alignment strategy not only limits the search space of 'unconstrained' alignment to low knowledge regions but also guides alignment of functionally informative regions under a reduced search space. Consequently, our approach identifies a suitable alignment from a smaller search space compared with state-of-the-art network alignment techniques, an important feature because of intractable nature of the network alignment problem. Additionally, by leveraging soft constraints, we have much lesser likelihood of generating alignments that conflict with known biological functions.

Compared with network-based approach, our approach considers both structural and functional information in defining subgraphs. Network-based methods rely purely on the assumption that densely connected proteins are likely to share similar function, and vice versa. However, there are cases where this assumption does not hold. In regulatory and signaling pathways, for example, proteins share similar function without being densely connected. An example is the MAPKKK cascade pathway. Such subgraphs can be uncovered via GO annotations but not topology alone because it is sparsely connected compared with protein complexes.

To realize the function-constrained network alignment problem, we propose an algorithm called DualAligner that uses the aforementioned dual alignment strategy, where region-to-region alignments are first made followed by detailed protein-to-protein alignment. Region-to-region alignment establishes high-level functional connections between the networks, while protein-to-protein alignment specifies the detailed connections within them. We demonstrate the utility and superiority of DualAligner over the state-of-the-art global network alignment techniques using real-world PPI networks in Section 3.

2 MATERIALS AND METHODS

In this section, we formally introduce the function-constrained network alignment problem. We begin by defining some terminology that we shall be using in the sequel. The set of key notations used in this article is given in Supplementary Material S1.1.

2.1 Terminology

Let $G = (V, E, \omega)$ be a PPI network where an edge $e \in E$ has a positive real weight ω that represents its interaction strength. Given a GO-directed acyclic graph, denoted as D, the ordered set $\Delta = \langle \Delta_1, \Delta_2, \ldots, \Delta_n \rangle$ is a topological sort of D, where Δ_i represents a single GO term. The *term association vector* of $v \in V$, denoted by Δ_v, is defined as $\Delta_v = \langle I_v(\Delta_1), I_v(\Delta_2), \ldots, I_v(\Delta_n) \rangle$, $I_v(\Delta_i) \in \{0, 1\}$, such that $I_v(\Delta_i) = 1$ if and only if the term Δ_i or its descendants are associated with protein v. Otherwise, $I_v(\Delta_i) = 0$. Note that Δ_v indicates GO terms that are associated with v.

A *function-constrained subgraph* $K_j = (V_j^K, E_j^K, \Delta_i)$ of G is a subgraph with the following properties: (i) it is a connected subgraph; and (ii) every node $v \in V_j^K$ shares a GO annotation Δ_i, i.e. $\forall v \in V_j^K, I_v(\Delta_i) = 1$. A function-constrained subgraph (which we refer to as a *subgraph constraint* for brevity) represents a constraint on a connected region of G that shares at least one functional role, namely the role represented by the term Δ_i. For example, in Figure 1, the subgraph constraint K_1 represents a connected subgraph of the human network such that every protein in it shares the transcription elongation annotation.

A *subgraph set* $S_i = \{K_1, K_2, \ldots, K_m\}$ represents m subgraph constraints on G. Note that typically $\cap_{K_j \in S_i} V_j^K \neq \emptyset$ and $\cup_{K_j \in S_i} V_j^K \neq V$, i.e. subgraphs in S_i may not form a partition on G and are highly overlapping. The overlapping nature of subgraph constraints does not indicate conflicting functions of its proteins, but rather the multi-attribute and multi-role nature of proteins. Figure 1(b) shows a sample subgraph set of function-constrained subgraphs for the fly and human PPI networks. Observe that the subgraph constraints may overlap (e.g. K_1 and K_3).

2.2 Region-to-region alignment

Given two PPI networks G_1 and G_2 with $|V_1| \leq |V_2|$, a *region-to-region alignment* between G_1 and G_2 is an injective function $rm : \mathcal{F}_1 \rightarrow \mathcal{F}_2$ that maps $K_m \in S_1$ (a function-constrained subgraph in G_1) to $K_n \in S_2$ (another function-constrained subgraphs in G_2). Figure 1(c) shows a region-to-region alignment of three pairs of function-constrained subgraphs. The transport function-constrained subgraph (K_3) in the human network is aligned to the microtubule binding subgraph (K_4) in the fly network, indicating that these two regions are conserved.

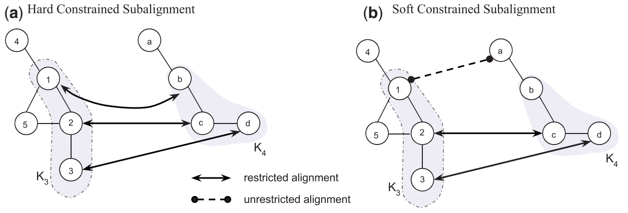

Fig. 2. Hard constraint versus soft constraint

To evaluate the conservation between function-constrained subgraphs during region-to-region alignment, it is necessary to consider protein-to-protein alignments within a pair of subgraphs. A protein-to-protein alignment between two subgraphs will reveal the extent of topological and functional similarities between them. To formalize this, we introduce the notion of *subalignment*. A subalignment is a subfunction $subm : V_i^s \rightarrow V_j^s$ of an alignment $Am : V_i \rightarrow V_j$ that aligns a subset $V_i^s \subset V_i$ of G_i to another subset $V_j^s \subset V_j$ of G_j. To illustrate this, consider in Figure 2 two toy PPI networks with nodes $V_i = \{1, 2, 3, 4, 5\}$ and $V_j = \{a, b, c, d\}$, respectively. Figure 2(a) shows a subalignment of nodes from $V_i^s = \{1, 2, 3\}$ to $V_j^s = \{b, c, d\}$. Notice that the subalignment aligns the function-constrained subgraph K_3 to K_4. On the other hand, Figure 2(b) shows a subalignment from $\{1, 2, 3\}$ to $\{a, c, d\}$, which is not an alignment between two function-constrained subgraphs.

Definition 1. (***Optimal Region-to-Region Alignment***). *Given two PPI networks G_i and G_j, let* **rm** *be a region-to-region alignment function that aligns function-constrained subgraphs $K_m \in S_i$ of G_i to function-constrained subgraphs $K_n \in S_j$ of G_j. Let $subm_{mn}$ be the best protein-to-protein subalignment between K_m and K_n and $S(subm_{mn})$ be the score of the $subm_{mn}$. The* ***optimal region-to-region alignment problem*** *is then defined as the problem of identifying the injective* **rm** *that maximizes:*

$$S(\mathbf{rm}) = \sum_{K_m \in S_i, rm(K_m) \in S_j} S(subm_{mn})$$

Here, protein-to-protein subalignment $subm_{mn}$ between two subgraphs can be performed using any existing network alignment algorithms. The score of the alignment is then $S(subm_{mn})$. Thus, the problem can be reformulated as the problem of identifying the one-to-one mapping between function-constrained subgraph pairs that maximizes the total subalignment scores.

Figure 1(c) depicts a region-to-region alignment of three pairs of function-constrained subgraphs and Figure 1(d) shows the potential subalignments between these pairs. Consider the subalignment between the transport function-constrained subgraph (K_3) in the human network and the microtubule binding subgraph (K_4) in the fly network. The subalignment K_3 to K_4 is shown by the arrowed lines within the circled subgraph in Figure 1(d).

2.3 Function-constrained network alignment problem

Given two PPI networks G_1 and G_2 with $|V_1| \leq |V_2|$, a *protein-to-protein alignment* between G_1 and G_2 is an injective function $Am : V_1 \rightarrow V_2$ that maps each protein in the smaller network to another protein in the larger network. For each $x \in V_1$, let $y_j = Am(x_i)$ be its corresponding aligned protein in V_2 given by Am.

Intuitively, a region-to-region alignment shows the functional conservation between the regions of both networks in a coarse-grained manner. Given such a region-to-region alignment, any detailed protein-to-protein alignment between these networks must ensure that it is consistent with the aligned functional subgraph regions (from region-to-region alignment). In other words, a protein-to-protein alignment should be guided by these region-to-region alignments. We shall now formalize concepts to let region-to-region alignment serve as constraint to protein-to-protein alignment.

A subalignment *subm* is said to be *hard constrained* to a pair of function-constrained subgraphs ($K_u \in S_i, K_v \in S_j$) iff $V_i^s \subseteq V_u^K$ and $V_j^s \subseteq V_v^K$ (recall that V_u^K and V_v^K are the vertices of subgraphs K_u and K_v, respectively). In this case, we refer to the pair of subgraphs as a *hard constraint*. Thus, the subalignment is strictly restricted within the subgraph regions specified by the hard constraint (K_u, K_v). The subalignment in Figure 2(a) is hard constrained to (K_3, K_4), while subalignment in Figure 2(b) is not.

Observe that it is computationally challenging to identify an alignment that satisfies a large number of (conflicting) hard constraints. Therefore, we introduce the notion of *probabilistic constraints* on subalignments. A *soft constraint* on a subalignment *subm* is a pair of function-constrained subgraphs (K_u, K_v) such that $\forall x \in V_i^s, P(subm(x) \in V_v^K) \geq p$ and $\forall x \in V_j^s, P(subm^{-1}(x) \in V_u^K) \geq p$. We refer to the subalignment as being *soft constrained* to the (K_u, K_v) pair. A soft constraint on a subalignment restricts the alignment to the region specified by (K_u, K_v) *probabilistically*. We shall later define a *score* that is associated with the probability parameter *p*. Suppose the subalignment in Figure 2(b) is a subalignment that is soft constrained by (K_3, K_4). For every protein in $x \in \{1, 2, 3\}$, it has probability of at least *p* to being mapped to a protein in $\{b, c, d\}$ (i.e. $\forall x \in V_i^s, P(subm(x) \in V_v^K) \geq p$). At the same time, for every protein in $x \in \{b, c, d\}$, it has probability of at least *p* to being mapped to a protein in $\{1, 2, 3\}$ (i.e. $\forall x \in V_j^s, P(subm^{-1}(x) \in V_u^K) \geq p$).

A *subgraph–subgraph pair* (or soft constraint) is a pair of subgraphs $f = (K_u, K_v)$, $K_u \in S_i$ and $K_v \in S_j$ such that there exists a subalignment *subm* of *m* that is soft constrained to *f* with probability *P*(*f*). This subgraph pair models a single region-to-region pairing. Let $\mathcal{F} = \{f_1, f_2, \ldots, f_K\}$ be the set of all subgraph–subgraph pairs, with each $f_k \in \mathcal{F}$ given a *certainty score* $s(f_k) \propto P(f_k)$. Here, $s(f_k)$ reflects the likelihood that subgraph K_u is aligned to subgraph K_v, and is associated with the network *subalignment score* between K_u and K_v in terms of both sequence and topological conservation.

It is reasonable to remove the top levels of GO terms as they are likely to contribute little to the alignment. However, choosing the number of top levels to prune is not a trivial task, as the specificity of the terms does not uniformly increase with level (e.g. some terms at level 3 may be more specific than others). Instead of a hard cutoff, `DualAligner` implicitly weighs the effects of GO terms via subalignment scores *s*(*f*). It relies on computing subalignment scores between GO term-associated subgraphs to determine its significance in functional conservation. Consider, for example, subalignment between GO term subgraphs associated with 'Biological Process to Cellular Component' and 'Golgi vesicle transport to vacuolar transport'. It is likely that the former subgraph pairs yield a poor subalignment score *s*(*f*) compared with the latter, which is likely to have higher percentage of conserved interactions and sequences. Because the alignment contribution of the subgraphs is weighted by *s*(*f*), `DualAligner` only weakly consider the effect of the former pair and strongly consider the latter. Thus, it can effectively discard the contribution of non-specific terms in a data-driven manner.

We propose, with simplifying assumptions, a *constraint mixture model*, where the overall distribution is modeled by a weighted superposition of distributions under each constraint $f_k \in \mathcal{F}$. The *mixing weight* for a soft constraint f_k is given by $P(f_k) = \frac{s(f_k)}{\sum_f s(f)}$, that is, the probability that constraint f_k is restricting a subalignment of *m*. The probability of $x \in V_1$ assigned to $y \in V_2$ under the constraint mixture model is as follows: $P(y|x) = \sum_{f_k} P(y|x, f_k)P(f_k)$. The conditional independence assumption is made such that conditioned on f_k, the matches are independent. $P(y|x, f_k)$ is the probability that $x \in V_1$ is aligned to $y \in V_2$ given that f_k constrains alignment of this pair of protein. Let $subm_{ij}$ be a subalignment that is soft constrained to $f = (K_i, K_j)$. Let $subm_{ij}^{max}$ be the best one-to-one alignment obtained. $subm_{ij}^{max}$ can be attained using any existing network alignment algorithms. Suppose an alignment generates a scoring function $\sigma(x, y)$ for each pair of aligned proteins. Then, we let the certainty score of *f* be $s(f_k) = \sum_{x, subm(x)} 1 + \sigma(x, subm(x))$. Under this formulation, we note that $P(y|x, f_k) = 0$ in most matches compared with sequence similarity matrix.

Definition 2 (***Function-constrained Network Alignment Problem***). *Given two PPI networks G_i and G_j and a region-to-region alignment* **rm**, *let* **Am** *be a protein-to-protein alignment function that aligns G_i to G_j constrained by soft constraints in* **rm**. *Let $P(y_j|x_i, f_k) = \sigma(x_i, y_j)$ if $f_k = (K_i, K_j) \in$* **rm** *and $(x_i, y_j) \in subm_{ij}^{max}$; $P(y_j|x_i, f_k) = 0$ if otherwise. The* ***function-constrained network alignment problem*** *is then defined as the problem of identifying the alignment function* **Am** *that maximizes*:

$$P(\mathbf{Am}(\mathbf{x})|\mathbf{x}) = \prod_{x_i} \sum_{f_k} P(Am(x_i)|x_i, f_k)P(f_k)$$

2.4 The `DualAligner` algorithm

We now present the `DualAligner` algorithm that identifies an alignment that maximizes $P(\mathbf{Am}(\mathbf{x})|\mathbf{x})$ via the aforementioned dual alignment strategy. The reader may refer to Supplementary Material S1.4 for formal description of the algorithm. `DualAligner` is composed of the following three phases.

Phase 1: Functional subgraph construction and ranking. Given two PPI networks G_1 and G_2, the first phase of `DualAligner` identifies and constructs the set of functional subgraphs S_1 and S_2. It exhaustively identifies for each GO term the induced subgraph of G_1 and G_2 sharing that term. Note that an induced subgraph of a GO term may contain multiple connected components. Our partitioning strategy transforms each connected component in the subgraph to a function-constrained subgraph of the GO term. For instance, if the induced subgraph of a GO term contains five connected components, then five function-constrained subgraphs of that term will be formed. Following that, pairs of subgraph–subgraph are constructed as soft constraints, and the certainty score *s*(*f*) is computed using the subalignment algorithm (discussed below). We introduce a pruning parameter β_s to remove near-duplicate functional subgraphs. While computing functional subgraphs, only non-redundant subgraphs are added to the set of functional subgraphs. In this study, we set $\beta_s = 0.9$ (remove subgraphs that share >90% of vertices). The user-defined parameter *k* is introduced to select only the top- *k* constraints for consideration in the next phase. These top- *k* constraints are selected based on their *s*(*f*) scores.

Phase 2: Region-to-region alignment. In the second phase, the soft constraint pairs are ranked by their certainty scores and subalignment is performed. Given a pair of subgraph–subgraph constraint $f = (K_i, K_j)$, a subalignment constrained by (K_i, K_j) is an injective mapping of vertices from V_i^K to V_j^K. Here, we propose a subalignment algorithm to achieve this. A pair of protein $u \in V_i^K, v \in V_j^K$ is scored using the following scoring function: $s(u, v) = b(u, v) + \lambda_s \sigma(u, v)$. The function $b(x_i, y_j)$ measures the sequence similarity score for aligning protein *u* to *v* and is defined as $b(x_i, y_j) = (1 + e^{-bscr(x_i, y_j)})^{-1}$, where $bscr(x_i, y_j)$ is the BLAST bit-score between the two proteins. The function $\sigma(u, v)$ measures the topological score for aligning protein *u* to *v* and is defined as follows:

$$\sigma(u, v) = |NE \cap ME|$$

$$NE = \{(x, y) : x \in N(u), y \in N(y)\}$$

$$ME = \{(Am(x), Am^{-1}(y)) : Am(x) \in N(x), Am^{-1}(y) \in N(v) \in E_2\}$$

where Am is function mapping currently aligned proteins and $N(x)$ are the neighbors of protein x. The parameter λ_s weighs the effect of sequence similarity versus structural similarity.

The subalignment method performs a greedy seed and extension strategy similar to several existing network alignment techniques such as MI-GRAAL, Græmlin. First, the best scoring pair of proteins (x, y) in V_i^K and V_j^K is identified as seed and added to alignment. Using this seed, an extension is performed by identifying the neighborhood of the seed, i.e. the vertices in V_i^K and V_j^K, which are adjacent to x and y, respectively. The pairs of proteins in this neighborhood sets are ranked according to their scoring function scores and added to the alignment. These steps are then repeated until the alignment is complete.

Observe that the pairwise subalignment step can be computationally expensive when subalignment is performed for every pair. Therefore, we introduce a pruning parameter γ_s to remove highly redundant functional subgraph–subgraph constraints. After ranking each constraint f by their confidence score $s(f)$, we greedily perform subalignment on the highest scoring constraint and update the conditional probabilities. Following that, we prune the remaining sets of constraints to remove near duplicates, such that any remaining constraint whose overlap ratio with f is greater or equal to γ_s is removed, i.e. given (K_i, K_j), we remove $\left\{(K_1, K_2) \in S : \frac{|V_1^K \cap V_i^K|}{|V_i^K|} \geq \gamma_s, \frac{|V_2^K \cap V_j^K|}{|V_j^K|} \geq \gamma_s\right\}$ from the set S. This is repeated until S is empty.

Phase 3: Expanded protein-to-protein alignment. In the final phase, we extend the coverage of protein–protein alignment beyond richly annotated regions. Phases 1 and 2 of the alignment align only proteins that have GO annotations. Thus, unannotated proteins could not be aligned yet. This phase enables alignment of such proteins by using Phase 2 aligned proteins as seeds, and then identifies all remaining unaligned proteins and ranks them pairwise by their scoring function $\sigma(u, v)$ score with reference to the seeds. Following that, each ranked pair, starting from the highest scoring pair, is added to the alignment. Finally, the remaining unaligned proteins are aligned based on the topological conservation with respect to the seeds.

The worst-case time complexity of DualAligner is $O(\alpha(|V|) + |S|^2 |V|^3)$ (Supplementary Material S1.5).

3 RESULTS

The DualAligner algorithm is implemented in Scala. We now present the experiments conducted to study the performance of DualAligner and report some of the results here (additional results are in Supplementary Material S1.6). The experiments were conducted on a 1.66 GHz Intel Core 2 Duo T5450 machine with 3 GB memory. We align the PPI networks of the global human, fly and yeast (Table 1). We use GO annotations that are part of the IntAct species dataset. No other annotations were used.

Table 1. Datasets

Dataset	Number of nodes	Number of edges	Source
Homo sapiens	9131	34 362	IntAct (Kerrien *et al.*, 2007)
Saccharomyces cerevisiae	4768	40 457	IntAct
Drosophila melanogaster	3114	6472	IntAct

Evaluation criteria. We use several criteria to evaluate the performance of DualAligner. We define the *coverage* of an alignment m between two networks G_1 and G_2, denoted as $cov(m)$, as the fraction of protein pairs aligned by an alignment method. That is, $cov(m) = |m|/min(|V_1|, |V_2|)$, where $|m|$ is the number of protein pairs aligned in m. Observe that $cov(m) = 1$ if the alignment m is a one-to-one mapping of V_1 to V_2.

To measure the structural similarity of an alignment m between two networks G_1 and G_2 with $cov(m) \leq 1$, we propose a modified version of the *edge correctness* (EC) measure (Kuchaiev and Przulj, 2011; Zaslavskiy *et al.*, 2009). Given G_1, G_2 and alignment m, the EC measure is defined as $\mathrm{EC} = \frac{\mathrm{E1SET} \cap \mathrm{E2SET}}{|\mathrm{E2SET}|}$ where

$$\mathrm{E1SET} = \{(x, y) : (x, y) \in E_1\}$$

$$\mathrm{E2SET} = \{(Am^{-1}(x), Am^{-1}(y)) : (x, y) \in E_2\}$$

Observe that EC indicates the fraction of correctly aligned edges among the proteins that are aligned (E2SET is restricted by domain of Am). Unlike the standard EC measure that is only useful for full coverage alignment, this modified measure can be used for not only global alignment results but also local alignments (i.e. alignments with $cov(m) < 1.0$).

To measure the sequence similarity of an alignment m between G_1 and G_2, we propose the *average normalized bit-score* measure as follows:

$$\mathrm{ANBS}(G_1, G_2, m) = |m|^{-1} \sum_{(x,y) \in m} \frac{bitscr(x, y)}{(bitscr(x, x) bitscr(y, y))^{-1/2}} \quad (1)$$

Observe that it is simply the average normalized BLAST bitscore of the paired proteins in m. A high ANBS score implies that sequence homologs are well matched, whereas low score implies that sequence homologs are not being matched. We do not report the variance of the ANBS scores because it does not indicate alignment quality (even in an optimal alignment, the ABNS variance can be high). Lastly, to measure the biological function quality of an alignment m, we use a modified *functional coherence* measure (Singh *et al.*, 2008) that incorporates the specificity of the GO terms:

$$\mathrm{FC}(G_1, G_2, m) = |m|^{-1} \sum_{(x,y) \in m} maxspecificity(\Delta_x \cap \Delta_y) \quad (2)$$

where Δ_x and Δ_y are the GO terms associated with proteins x and y, respectively, and $maxspecificity(\cdot)$ measures the most specific GO term shared by x and y:

$$maxspecificity(X) = \max_{\Delta_i \in X} 1 - \frac{|\{v \in V : \Delta_i \in \Delta_v\}|}{|V|} \quad (3)$$

Identification of conserved regions through region-to-region alignment. We first demonstrate the importance of region-to-region alignment for network alignment. Recall that in a region-to-region alignment, pairs of subgraphs from both networks indicate conservation at a broader granularity. That is, the subgraph as whole is deemed to be functionally conserved with allowance for protein and topological differences. Here, we evaluate several high-scoring region-to-region alignments. A few of these conserved subgraphs are depicted in Figure 3 (see also region-to-region alignment statistics in Supplementary Material S1.3). Each pair of subgraphs separated by the dashed lines represents a functional subgraph of the human network matched to its yeast network counterpart. Observe that while several region-to-region alignments contain a high-confidence one-to-one protein alignments, such as the conserved NAT1-ARD1-NATS acetyltransferase subgraph, other region-to-region alignments do not admit a proper one-to-one protein alignment. Furthermore, the COG complex, which plays an important role in intra-Golgi trafficking, is well conserved between yeast and human network. However, COG1, COG2 and COG7 are structurally unique to mammalian cells and are not homologous to their yeast counterpart (Ungar *et al.*, 2002), which imply that no individual pairing between COG1, COG2 and COG7 could exist. This partially explains the minor differences between the subgraphs and also justifying the need for region-to-region alignment in addition to individual protein-to-protein alignment.

We also note that the proteosome core complex is conserved. This complex is known to be ubiquitous and highly conserved in eukaryotes. Other relationships can be inferred from well-aligned subgraphs. For instance, transport mechanisms are found to be highly conserved between human and yeast. In summary, well-aligned regions show important conservation relationships between complex–complex, complex–function and function–process.

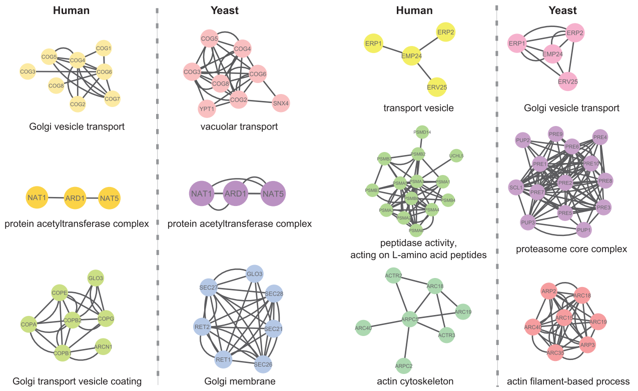

Fig. 3. Selected region-to-region alignments showing highly conserved subgraphs between human and yeast networks. Note that there may not necessarily be an optimal protein-to-protein alignment between the subgraphs

Global protein-to-protein alignment. Next, we compare the performance of DualAligner with the following state-of-the-art global network alignment methods: IsoRank (Singh *et al.*, 2008), IsoRankN (Liao *et al.*, 2009), MI-GRAAL (Kuchaiev and Przulj, 2011), Natalie (Klau, 2009), PINALOG (Phan and Sternberg, 2012), Shih2012 (Shih and Parthasarathy, 2012) and PISwap (Chindelevitch *et al.*, 2013). We ran alignments between the global yeast and human networks as well as between the global fly and yeast networks. We set *min* = 3, $\lambda_s = 0.8$ and $\beta_s = 0.9$. We vary the parameter k to obtain alignments of varying coverage. For IsoRank and IsoRankN, we used the standard settings recommended by the tool with $\alpha = 0.8$. We use the default parameters for PINALOG, and for Natalie we chose the setting with highest EC value. For MI-GRAAL, we enabled signatures, sequences and degrees cost matrices. To use PISwap as an independent method, we use the Hungarian method to obtain a best matching based on sequence homology as input. For Shih2012, we set trade-off parameter to 50. Note that PATH (Zaslavskiy *et al.*, 2009) suffers from scalability problem as highlighted in (Kuchaiev and Przulj, 2011), and as a result it failed to complete the alignment of the human and yeast global networks within 24 h.

Figure 4 plots the results of the alignment between the global human and yeast networks, comparing DualAligner to existing methods. In each figure, the scores obtained using DualAligner are indicated by a line because we obtain multiple instances of network alignments with different coverage. While no methods are superior on all measures, DualAligner is among the best performer in each measure (when compared with another method having the same coverage). It achieves the best balance of edge correctness, sequence similarity and functional coherence by being the best or close to the best performer in each measure. DualAligner outperforms all methods except MI-GRAAL based on EC. On the other hand, DualAligner significantly outperforms MI-GRAAL based on average normalized bit-score. It also outperforms all methods in functional coherence. Notice that at lower coverage, DualAligner even outperforms other methods in both measures. Finally, DualAligner outperforms Natalie, PINALOG, IsoRank and PISwap on all measures. Thus, our approach can be used to obtain high-quality alignments if maximizing coverage is not a requirement.

Figure 5 plots the results of alignment between the global fly and yeast networks. Similar to the preceding comparison, DualAligner is either the best performer or close to the best

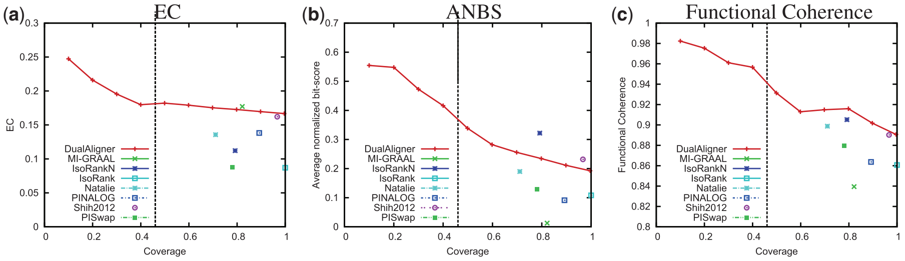

Fig. 4. Performance of DualAligner (human versus yeast) protein-to-protein alignment; alignment of each method may not have the same coverage. With DualAligner, one can adjust the trade-off between alignment quality and coverage. The dashed vertical line indicates the portion of the DualAligner alignment that is aligned from region-to-region alignment. (a) EC, (b) ANBS, (c) Functional Coherence

on all measures. It achieves the best score in FC except PINALOG. Nevertheless, we urge caution in interpreting the quality of the protein-to-protein alignments. The global fly network is highly incomplete and has large gaps in interaction data (Table 1). This could also explain the low coverage of PINALOG, which rely on topology-based clustering to align the networks. The rapid drop in EC and bit-score similarities at higher coverage is indicative of this issue. In this case, a broad region-to-region alignment would be more suitable than inferring detailed protein-to-protein alignment out of the incomplete interaction data.

Finally, compared with other methods, DualAligner is able to assign confidence scores to matched protein pairs that correlate with alignment quality (by EC, bit-score and functional coherence measures). This is demonstrated by its ability to allow users to control the trade-off between alignment coverage and quality. Overall, DualAligner is competitive with all protein-to-protein methods while being able to perform region-to-region alignment that is consistent with existing knowledge. Note that the fly network is sparsely connected with few GO annotations in comparison with the human network. Hence, the sequence homology of proteins may be crucial here for better alignment. Consequently, PISwap is advantageous here, as it relies on sequence-based Hungarian matching.

A unique advantage of DualAligner is its ability to have functionally interpretable subgraph-to-subgraph alignments. Methods that are based on clustering (e.g. PINALOG and Shih2012) may not yield functionally interpretable subgraph-to-subgraph alignment. DualAligner is especially advantageous when there is no clear one-to-one protein alignment between the subgraphs. In that case, functional interpretation of alignment between regions can be useful. On both alignments, DualAligner is able to achieve almost complete coverage. Thus, despite not considering isolated proteins, a significant portion of the networks can be aligned.

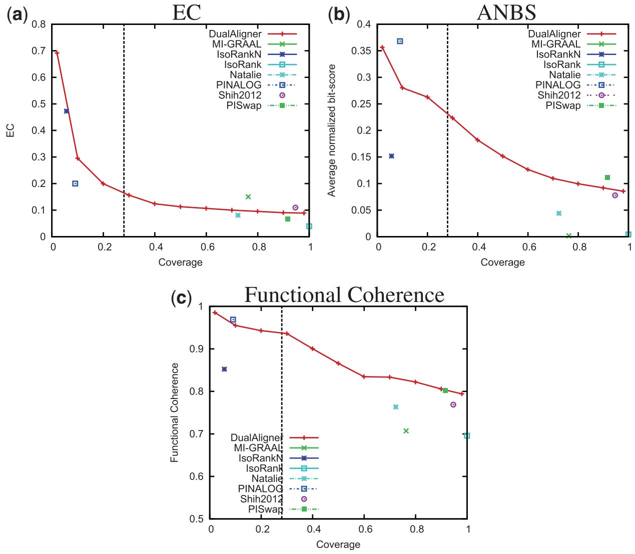

Fig. 5. Performance of DualAligner (fly versus yeast). (a) EC, (b) ANBS, (c) Functional Coherence

Effect of λs. The parameter λ_s controls the effect of sequence similarity versus topology similarity in the scoring model. We study the effect of λ_s on the alignments by varying λ_s from 0 to 400. Figure 6 depicts the results showing its effect on EC and ANBS scores. As λ_s increases, the EC value improves while ANBS decreases. Thus, λ_s allows a trade-off between sequence similarity and topology similarity. At low values of λ_s, more emphasis in preserving high average bit-score alignment is observed. The greater the value λ_s, the greater is the effect of topology similarity. This effect, however, converges to a steady state at subsequently higher values.

Running times. Table 2 reports the running times of DualAligner compared with other methods. Observe that DualAligner outperforms the tested approaches except PISwap. Although IsoRankN is designed to scale to multiple networks, IsoRank significantly outperforms IsoRankN in alignment of a pair of global PPI networks. However, the running times of IsoRank increase much more rapidly with the size of the networks. In comparison, DualAligner scales better with the size of the networks. The Natalie method is designed to terminate after 3600 s.

4 SUMMARY

We propose DualAligner, a network alignment algorithm that performs a dual alignment strategy, in which both region-to-region alignment (i.e. whole subgraph of one network is aligned to subgraph of another) and protein-to-protein alignment (i.e. individual proteins in networks are aligned to one another) are performed to achieve superior-quality network alignment. Specifically, global alignment is achieved in

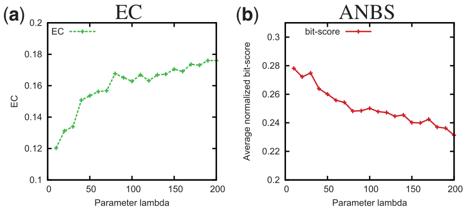

Fig. 6. Effect of λ_s showing its role in controlling the trade-off between topology and sequence conservation. (a) EC, (b) ANBS

Table 2. Running times (in seconds)

Networks	IsoRankN	IsoRank	MI-GRAAL	DualAligner	Natalie	PISwap
Human PPI–Yeast PPI	23 911	3950	38 849	1831	3600	803
Fly PPI–Yeast PPI	5725	1050	31 472	1192	3600	441

DualAligner via the background information provided by a combination of GO annotation information and protein interaction data. We empirically demonstrate that our proposed approach outperforms state-of-the-art global network alignment techniques and demonstrates its ability to rank regions of alignment by their alignment quality.

Conflicts of Interest: none declared.

REFERENCES

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

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

Huang,H. and Bader,J.S. (2009) Precision and recall estimates for two-hybrid screens. *Bioinformatics*, **25**, 372–378.

Kalaev,M. *et al.* (2008) NetworkBLAST: comparative analysis of protein networks. *Bioinformatics*, **24**, 594–596.

Kalaev,M. *et al.* (2009) Fast and accurate alignment of multiple protein networks. *J. Comput. Biol.*, **16**, 989–999.

Kelley,B.P. *et al.* (2004) PathBLAST: a tool for alignment of protein interaction networks. *Nucleic Acids Res.*, **32**, W83–W88.

Kerrien,S. *et al.* (2007) IntAct–open source resource for molecular interaction data. *Nucleic Acids Res.*, **35**, D561–D565.

Klau,G.W. (2009) A new graph-based method for pairwise global network alignment. *BMC Bioinformatics*, **10** (**Suppl. 1**), S59.

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

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.

Pache,R.A. and Aloy,P. (2012) A novel framework for the comparative analysis of biological networks. *PloS One*, **7**, e31220.

Phan,H.T.T. and Sternberg,M.J.E. (2012) PINALOG: a novel approach to align protein interaction networks–implications for complex detection and function prediction. *Bioinformatics*, **28**, 1239–1245.

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

Shih,Y.K. and Parthasarathy,S. (2012) Scalable global alignment for multiple biological networks. *BMC Bioinformatics*, **13** (**Suppl. 3**), S11.

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.

Ungar,D. *et al.* (2002) Characterization of a mammalian Golgi-localized protein complex, COG, that is required for normal Golgi morphology and function. *J. Cell Biol.*, **157**, 405–415.

Zaslavskiy,M. *et al.* (2009) Global alignment of protein-protein interaction networks by graph matching methods. *Bioinformatics*, **25**, i259–i267.