Pairwise Local Alignment of Protein Interaction Networks Guided by Models of Evolution

Mehmet Koyutürk, Ananth Grama and Wojciech Szpankowski
Dept. of Computer Sciences, Purdue University, West Lafayette, IN 47907.
{koyuturk, ayg, spa}@cs.purdue.edu
Phone: 765 494 6010 Fax: 765 494 0739

Abstract— **With ever increasing amount of available data on protein interaction networks (PINs) and research revealing that these networks evolve at a modular level, discovery of conserved patterns in these networks becomes an important problem. Recent algorithms on aligning PINs target simplified structures such as conserved pathways to render these problems computationally tractable. However, since conserved structures that are parts of functional modules and protein complexes generally correspond to dense subnets of the network, algorithms that are able to extract conserved patterns in terms of general graphs are necessary. With this motivation, we focus here on discovering protein sets that induce subnets that are highly conserved in the interactome of a pair of species. For this purpose, we develop a framework that formally defines the pairwise local alignment problem for PINs, models the problem as a graph optimization problem, and presents fast algorithms for this problem.**

In order to capture the underlying biological processes correctly, we base our framework on duplication/divergence models that focus on understanding the evolution of PINs. Experimental results from an implementation of the proposed framework show that our algorithm is able to discover conserved interaction patterns very effectively (in terms of accuracies and computational cost). While we focus on pairwise local alignment of PINs in this paper, the proposed algorithm can be easily adapted to finding matches for a subnet query in a database of PINs.

Primary keyphrase: **Protein interaction.**

Secondary keyphrases: **Models of evolution, Pattern and motif discovery.**

I. Introduction

Increasing availability of experimental data relating to biological sequences, coupled with efficient tools such as BLAST and CLUSTAL have contributed to fundamental understanding of a variety of biological processes [1], [2]. These tools are used for discovering common subsequences and motifs, which convey functional, structural, and evolutionary information. Recent developments in molecular biology have resulted in a new generation of experimental data that bear relationships and interactions between biomolecules [3], [4]. An important class of molecular interaction data is in the form of protein interaction networks (PINs), which provide the experimental basis for understanding modular organization of cells, as well as useful information for predicting the biological function of individual proteins [5]. Common methods for obtaining protein interaction data include two-hybrid experiments [6], mass spectrometry experiments [7], and structural analysis such as phage-display [8].

As revealed by recent studies, PINs evolve at a modular level [9], [10] and consequently, understanding of conserved substructures through alignment of these networks can provide basic insights into a variety of biochemical processes. However, although vast amounts of high-quality data is becoming available, efficient network analysis counterparts to BLAST and CLUSTAL are not readily available for such abstractions. As is the case with sequences, key problems on graphs derived from biomolecular interactions include aligning multiple graphs [11], finding frequently occurring subgraphs in a collection of graphs [12], discovering highly conserved subgraphs in a pair of graphs and finding good matches for a subgraph in a database of graphs [13]. In this paper, we specifically focus on discovering highly conserved subnets in a pair of PINs. With the expectation that conserved subnets will be parts of complexes and modules, we base our model on the discovery of two subsets of proteins from each PIN such that the induced subnets are highly conserved.

Based on the understanding of the structure of PINs that are available for several species, theoretical models that focus on understanding the evolution of protein interactions have been developed. Among these, the duplication/divergence model has been shown to be successful in explaining the power-law nature of PINs [14]. According to this model, PINs evolve based on duplication of proteins together with their interactions followed by divergence of these proteins through elimination and emergence of interactions. In order to capture the underlying biological processes correctly, we base our framework on duplication/divergence models through definition of duplications, matches, and mismatches in a graph-theoretic framework. We then reduce the resulting alignment problem to a graph optimization problem and propose efficient heuristics to solve this problem. Experimental results based on an implementation of the proposed framework show that the proposed algorithm is able to discover conserved interaction patterns very effectively. Note that the proposed algorithm can be also adapted to finding matches for a subnet query in a database of PINs.

The rest of this paper is organized as follows: we start by outlining related work on alignment of PINs in the next section. Then, we provide a brief overview of duplication/divergence models for the evolution of PINs in Section III. In Section IV, we define the alignment problem based on these models of evolution, formulate the problem as a graph optimization problem, and propose efficient heuristics for the solution of the problem. We illustrate the effectiveness of the proposed framework with a selection of conserved subnets discovered by our technique in Section V. We conclude our discussion and outline avenues for future research in Section VI.

II. Related Work

As the amount of cell signaling data increases rapidly over the past few years, there have been various efforts aimed at developing graph counterparts to sequence analysis tools such as BLAST [1] and CLUSTAL [2]. In a relatively early effort, Dandekar et al. [15] comprehensively align glycolysis metabolic pathways through comparison of biochemical data, analysis of elementary modes, and comparative genome analysis, identifying iso-enzymes, several potential pharmacological targets, and organism-specific adaptations. While such efforts demonstrate the potential of interaction alignment in understanding cellular processes, these analyses are largely manual, motivating the need for automated alignment tools.

As partially complete interactomes of several species become available, researchers have explored the problem of identifying conserved topological motifs in different species [16], [10]. These studies reveal that many topological motifs are significantly conserved within and across species and proteins that are organized in cohesive patterns tend to be conserved to a higher degree. In a related effort, a publicly available tool, TopNet, is designed to compare topological characteristics of different sub-networks in the same network [17]. Another publicly available tool, PathBLAST [1], adopts the ideas in sequence alignment to protein interaction networks (PINs) to discover conserved protein pathways across species [18], [13]. By restricting the alignment to pathways, *i.e.*, linear chains of interacting proteins, this algorithm renders the alignment problem tractable, while preserving the biological implication of discovered patterns. PathBLAST accounts for gaps and mismatches by allowing unrepeated jumps and matching of non-orthologous proteins, based on the notion that the orthologous counterpart of a pair of interacting proteins in one species will, likely, be connected via a short path in the other.

Since the local alignment of PINs for patterns in the form of general graphs leads to computationally intractable problems, tools based on simplified models such as PathBLAST are generally useful. However, as functional modules and protein complexes are likely to be conserved across species [9], [10], algorithms for aligning general graphs are required for understanding conservation of such functional units. To address this, we develop a framework for aligning PINs to discover subsets of proteins in each species such that the subgraphs induced by these sets are highly conserved. The proposed framework relies on theoretical models that focus on understanding the evolution of protein interaction networks.

[1]PathBLAST is publicly available at `http://www.pathblast.org/`.

III. Theoretical Models for Evolution of Protein Interaction Networks

There have been a number of studies aimed at understanding the general structure of PINs. PINs are composed of a few hub proteins that interact with many other proteins, while a majority of proteins have only a few interaction partners. It has been shown that lethal proteins tend to be highly connected in the interactome [19], [20] and their removal causes a disruption in network structure [21]. This property of PINs ensures robustness of the network to random attacks.

There have also been efforts to model PINs using power-law graphs in which the relative frequency of nodes that interact with k proteins is proportional to $k^{-\gamma}$, where γ is a network-specific parameter [19]. These models led to efforts targeted on explaining the power-law nature of PINs. Barábasi and Albert's [22] network growth model based on preferential attachment is able to generate power-law graphs that resemble real-world power law graphs such as world wide web, scientific citation graphs, and biological networks. According to this model, networks expand continuously by addition of new nodes and these new nodes prefer to attach to well-connected nodes when joining the network. Observing that older proteins are better connected and the number of interactions gained by a protein is proportional to its connectivity, Eisenberg and Levanon [23] explain the evolutionary mechanisms behind such preference by the strength of selective pressure on maintaining connectivity of strongly connected proteins and creating proteins to interact with them. Furthermore, in a relevant study, it is observed that the interactions between groups of proteins that are temporally close in the course of evolution are likely to be conserved, suggesting synergistic selection during network evolution [24]. Note, however, that a protein might be lethal without being highly connected or that a protein of older origin may display low levels of connectivity, which can be explained by functional heterogeneity [25].

A common model of evolution that explains preferential attachment and power-law nature of PINs is the duplication/divergence model that is based on gene duplications [26], [14], [27], [28]. According to this model, when a gene is duplicated in the genome, the node corresponding to the product of this gene is also duplicated together with its interactions. An example of duplication of a protein is shown in Figure 1. The effective duplication rate is estimated to be around 8.3×10^{-4} per gene per million years for the yeast

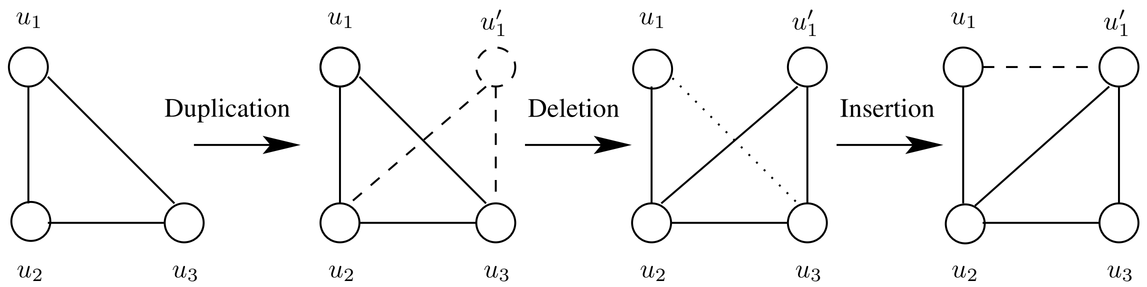


Fig. 1. A general duplication/divergence model for evolution of PINs. Starting with three interactions between three proteins, protein u_1 is duplicated to add u_1' into the network together with its interactions, (shown using dashed circle and lines). Then, u_1 loses its interaction with u_3 (shown using a dotted line). Finally, an interaction between u_1 and u_1' is added to the network (also shown using a dashed line).

genome [28]. A protein loses many aspects of its functions rapidly after being duplicated. This translates into divergence of duplicated proteins in the interactome through deletion and insertion of interactions. Deletion of an interaction in a PIN implies the elimination of an existing interaction between two proteins due to structural and/or functional changes. Similarly, insertion of a interaction into a PIN implies the emergence of a new interaction between two non-interacting proteins, caused by mutations that change protein surfaces. Examples of insertion and deletion of interactions are also illustrated in Figure 1. If a deletion or insertion is related to a recently duplicated protein, it is said to be correlated [26]; otherwise, it is uncorrelated. Since newly duplicated proteins are more tolerant to interaction loss because of redundancy, correlated deletions are generally more probable than insertions and uncorrelated deletions. For example, the rate of interaction loss after a gene is duplicated is over 2.2×10^{-3} deletions per million years [27], [28]. Indeed, while some models allow all kinds of insertions and deletions [28], other duplication/divergence models only allow correlated deletions to account for divergence of duplicated proteins [26], [14]. It is also theoretically shown that network growth models based on node duplications generate power-law distributions [29].

The interaction profiles of duplicated proteins tend to almost totally diverge in about 200 million years, as estimated on the yeast interactome. On the other hand, the correlation between interaction profiles of duplicated proteins is significant for up to 150 million years after duplication, with more than half of interactions being conserved for proteins that are duplicated less than 50 million years ago [27]. Since the elimination of interactions is related to sequence-level mutations, one can expect a positive correlation between similarity of interaction profiles and sequence similarity for paralogous proteins.

In order to accurately identify and interpret conservation of interactions, complexes, and modules across species, we base our framework for the local alignment of PINs on duplication/divergence models. Since the conservation of interactions is also related to protein function, interactions within a protein complex or functional module are expected to be conserved together [9]. Therefore, we define the local alignment problem as one of finding a set of proteins in each PIN such that the subset of interactions induced by these proteins is highly conserved. We do this with the expectation that these sets of proteins are parts of a conserved protein complex or more generally a functional module. While searching for highly conserved groups of interactions, we account for mismatches in light of the duplication/divergence model. Each duplication, insertion or deletion is included in the alignment with a certain penalty so that the trade-off between the cost of duplication and gain of matched interactions will be able to capture the extent of divergence between paralogous proteins. As the concepts of match, mismatch (insertion or deletion) and duplication are in accordance with widely accepted models of evolution, discovered alignments also allow speculation about the structure of the network in the common ancestor. Since the differences between the two PINs result not only from the evolutionary process but are also caused by experimental errors, we allow both correlated and uncorrelated insertions and deletions in our model. This facilitates detection of experimental errors through comparison of aligned networks as well.

IV. Pairwise Local Alignment of Protein Interaction Networks

In light of the theoretical models of evolution of PINs, we develop a generic framework for the comparison of PINs in two different species. We formally define a computational problem that captures the underlying biological phenomena through exact matches, mismatches, and duplications. We then formulate local alignment as a graph optimization problem and propose greedy algorithms to effectively solve this problem.

A. *The Pairwise Local Alignment Problem*

A PIN is conveniently modeled by an undirected graph $G(U, E)$, where U denotes the set of proteins and $uu' \in E$ denotes an interaction between proteins $u \in U$ and $u' \in U$. For pairwise alignment of PINs, we are given two PINs belonging to two different species, denoted $G(U, E)$ and $H(V, F)$. The homology between a pair of proteins is quantified by a similarity measure that is defined as a function

$S : (U \cup V) \times (U \cup V) \rightarrow \Re$. For any $u, v \in U \cup V$, $S(u, v)$ measures the degree of confidence in u and v being orthologous if they belong to different species and paralogous if they belong to the same species. A convenient source for $S(u, v)$ is the well-known BLAST E value measuring significance of sequence similarity [1], which is also used in PathBLAST with appropriate normalization [18]. We assume that similarity scores are non-negative, where $S(u, v) = 0$ indicates that u and v cannot be considered as potential orthologs or paralogs. In this respect, S is expected to be highly sparse, *i.e.*, the similarity between most of the protein pairs is zero since each protein is supposed to have only a few potential orthologs or paralogs. We discuss the reliability of possible choices for assessing protein similarity in detail in Section IV-D.1.

For PINs $G(U, E)$ and $H(V, F)$, a *protein subset pair* $P = \{\tilde{U}, \tilde{V}\}$ is defined as a pair of protein subsets $\tilde{U} \subseteq U$ and $\tilde{V} \subseteq V$. Any protein subset pair P induces a local alignment $\mathcal{A}(G, H, S, P) = \{\mathcal{M}, \mathcal{N}, \mathcal{D}\}$ of G and H with respect to S, characterized by a set of duplications $\mathcal{D}$, a set of matches $\mathcal{M}$, and a set of mismatches $\mathcal{N}$. The biological correspondent of a *duplication* is the duplication of a gene in the course of evolution. Each duplication is associated with a penalty, since duplicated proteins tend to diverge in terms of their interaction profiles in the long term [27]. A *match* corresponds to a conserved interaction between two orthologous protein pairs, which is rewarded by a match score that reflects our confidence in both protein pairs being orthologous. A *mismatch*, on the other hand, is the lack of an interaction in the PIN of one of the species between a pair of proteins whose orthologs interact in the other species. A mismatch may correspond to the emergence (insertion) of a new interaction or the elimination (deletion) of a previously existing interaction in one of the species after the split, or an experimental error. Thus, mismatches are also penalized to account for the divergence from the common ancestor. We provide formal definitions for these three concepts to construct a basis for the formulation of local alignment as an optimization problem. Note that although PINs are undirected graphs, interactions are regarded as ordered pairs in these definitions for convenience, *i.e.*, for an interaction $uu' \in E$, there is also an interaction $u'u \in E$, which is essentially the same interaction.

Definition 1: **Local Alignment of PINs.**
Given protein interaction networks $G(U, E)$, $H(V, F)$ and a pairwise similarity function S defined over the union of their protein sets $U \cup V$, any protein subset pair $P = (\tilde{U}, \tilde{V})$ induces a local alignment $\mathcal{A}(G, V, S, P) = \{\mathcal{M}, \mathcal{N}, \mathcal{D}\}$, where

$$\begin{aligned} \mathcal{M} = \quad & \{u, u' \in \tilde{U}, v, v' \in \tilde{V} : S(u, v) > 0, \\ & S(u', v') > 0, uu' \in E, vv' \in F\} \end{aligned} \qquad (1)$$

$$\begin{aligned} \mathcal{N} = \quad & \{u, u' \in \tilde{U}, v, v' \in \tilde{V} : S(u, v) > 0, \\ & S(u', v') > 0, uu' \in E, vv' \notin F\} \\ & \cup \{u, u' \in \tilde{U}, v, v' \in \tilde{V} : S(u, v) > 0, \\ & S(u', v') > 0, uu' \notin E, vv' \in F\} \end{aligned} \qquad (2)$$

$$\begin{aligned} \mathcal{D} = \quad & \{u, u' \in \tilde{U} : S(u, u') > 0\} \\ & \cup \{v, v' \in \tilde{V} : S(v, v') > 0\} \end{aligned} \qquad (3)$$

Each match $M \in \mathcal{M}$ is associated with a score $\mu(M)$. Each mismatch $N \in \mathcal{N}$ and each duplication $D \in \mathcal{D}$ are associated with penalties $\nu(N)$ and $\delta(D)$, respectively.

The score of alignment $\mathcal{A}(G, H, S, P) = \{\mathcal{M}, \mathcal{N}, \mathcal{D}\}$ is defined as:

$$\sigma(\mathcal{A}) = \sum_{M \in \mathcal{M}} \mu(M) - \sum_{N \in \mathcal{N}} \nu(N) - \sum_{D \in \mathcal{D}} \delta(D). \qquad (4)$$

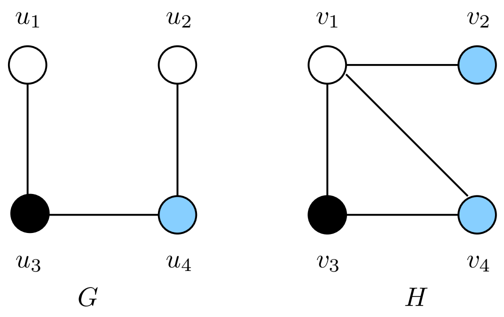
We aim to find all local alignments with score above a specified threshold. A protein subset pair with a high alignment score reveals a set of proteins in each species that are likely to have evolved from a common set of proteins in the common ancestor, with a relatively large number of interactions that are conserved during evolution, characterized by the match set.

The above definition for the local alignment problem provides a general framework for the comparison of PINs, without explicitly formulating match scores, mismatch, and duplication penalties. These functions can be selected and their relative contributions can be tuned based on theoretical models and experimental observations to effectively synchronize with the underlying evolutionary process. Clearly, an appropriate basis for deriving these functions is the similarity score function S. We discuss possible choices for scoring functions in detail in Section IV-D.2.

A sample instance of the pairwise local alignment problem is shown in Figure 2. Orthologous and paralogous proteins ($S > 0$) are shown by identically-colored nodes in the figure. Consider the alignment induced by the protein subset pair $\tilde{U} = \{u_1, u_2, u_3, u_4\}$ and $\tilde{V} = \{v_1, v_2, v_3\}$, which is shown in Figure 3. The only duplication in this alignment is (u_1, u_2), as $S(u_1, u_2) = 0.9 > 0$. All other intra-species similarities in the selected subsets are equal to zero. If this alignment is chosen to be a "good" one, then based on the existence of this duplication in the alignment, taking into account that $S(u_2, v_1) < S(u_1, v_1)$, we can speculate that u_1 and v_1 have evolved from the same gene in the common ancestor, while u_2 is an in-paralog that emerged from duplication of u_1 after split.

The match set consists of interaction pairs (u_1u_3, v_1v_3) and (u_2u_4, v_1v_2). Observe that v_1 is matched with both u_1 and u_2 in the context of two different interactions. This can be accounted for using the functional divergence of u_1 and u_2 after the duplication.

The mismatch set is composed of (u_1u_4, v_1v_2), (u_2u_3, v_1v_3), and (u_3u_4, v_3v_2). Note that the interaction u_3u_4 in G is left unmatched by this alignment, as the only possible pair of proteins in $\tilde{V}$ that are orthologous to these



(a)

S	u_1	u_2	u_3	u_4	v_1	v_2	v_3	v_4
u_1	1.0							
u_2	0.9	1.0						
u_3			1.0					
u_4				1.0				
v_1	1.0	0.9			1.0			
v_2				0.8		1.0		
v_3			1.0				1.0	
v_4				1.0		0.9		1.0

(b)

Fig. 2. An instance of the pairwise local alignment problem. (a) Two PINs G and H, containing 4 proteins each, consisting of 3 and 4 interactions, respectively. The proteins that have non-zero similarity scores (*i.e.*, are potentially orthologous), are colored the same. (b) Lower-diagonal part of the matrix that defines a similarity function over the protein. Empty cells show zero similarity between the two corresponding proteins.

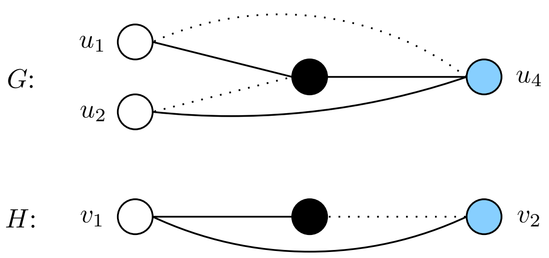


Fig. 3. An alignment of G and H in Figure 2 induced by the protein subset pair $\{u_1, u_2, u_3, u_4\}$ and $\{v_1, v_2, v_3\}$. Ortholog and paralog proteins are vertically aligned. Existing interactions are shown by solid lines, missing interactions that have an existing ortholog counterpart are shown by dotted lines. Solid interactions between two aligned proteins in separate species correspond to a match, one solid one dotted interaction between two aligned proteins in separate species correspond to a mismatch. Proteins in the same species that are on the same vertical line correspond to duplications.

two proteins are v_3 and v_2, which do not interact in H. One conclusion we can derive from this alignment is the insertion or deletion of this interaction in one of the species after the split. The indirect path between v_3 and v_2 through v_1 may also serve as a basis for the tolerability of the loss of this interaction. We can also simply attribute this observation to experimental noise. However, if we include v_4 in $\tilde{V}$ as well, then the induced alignment will be able to match u_3u_4 and v_3v_4. This will strengthen the probability that this interaction existed in the common ancestor. However, v_4 comes at the price of another duplication since it is paralogous to v_2. This example illustrates the challenge for correct matching of proteins with their orthologs in order to reveal the maximum amount of reliable information about the conservation of interaction patterns. Our model translates this problem into a trade-off between mismatches and duplications, favoring selection of duplicate proteins that have not quite diverged in the alignment.

In the following sections, we reduce the local alignment problem into a graph-theoretic optimization problem and propose efficient heuristics that are well-suited for the discovery of conserved modules in PINs.

B. Alignment Graphs and the Maximum-Weight Induced Subgraph Problem

It is possible to collect information on matches and mismatches between two PINs into a single alignment graph by computing a modified version of graph cartesian product [30] that takes orthology into account. Assigning appropriate weights to the edges of the alignment graph, the local alignment problem defined in the previous section can be reduced to an optimization problem on this alignment graph. We define an alignment graph for this purpose.

Definition 2: **Alignment Graph.**
For a pair of PINs $G(U, E)$, $H(V, F)$, and protein similarity function S, the corresponding weighted alignment graph $\mathbf{G}(\mathbf{V}, \mathbf{E})$ is computed as follows:

$$\mathbf{V} = \{\mathbf{v} = \{u, v\} : u \in U, v \in V \text{ and } S(u, v) > 0\} \qquad (5)$$

In other words, we have a node in the alignment graph for each pair of ortholog proteins that are in different PINs. Set $\mathbf{E}$ contains three types of edges. Let $\mathbf{v} = \{u, v\}$ and $\mathbf{v}' = \{u', v'\}$ be two nodes in $\mathbf{V}$. Edge $\mathbf{vv}' \in \mathbf{E}$ is a *match edge* if $uu' \in E$ and $vv' \in F$. Each match edge is associated with weight $w(\mathbf{vv}') = \mu(uu', vv')$. Edge $\mathbf{vv}' \in \mathbf{E}$ is a mismatch edge if $uu' \in E$ but $vv' \notin F$ or $vv' \in F$ but $uu' \notin E$. The weight of a *mismatch edge* corresponds to the mismatch penalty of its incident nodes, *i.e.*, $w(\mathbf{vv}') = -\nu(uu', vv')$. Edge $\mathbf{vv}' \in \mathbf{E}$ is a duplication edge if either $s(u, u') > 0$ or $s(v, v') > 0$. Each *duplication edge* is assigned weight $w(\mathbf{vv}') = -\delta(u, u')$ or $w(\mathbf{vv}') = -\delta(v, v')$, depending on which pair of proteins are paralogous. For any non-edge in the alignment graph $w(\mathbf{vv}') = 0$.

Consider the PINs in Figure 2. To construct the corresponding alignment graph, we first compute the product of these two PINs, shown in Figure 4(a). The node set of this graph contains all five ortholog protein pairs that are in different PINs. These nodes constitute the nodes of the alignment graph as well. In the figure, matching interactions are shown by solid lines, which correspond to the match edges of the alignment graph. For instance $\{u_1, v_1\}\{u_3, v_3\}$ is a match edge since $u_1u_3 \in E$ and $v_1v_3 \in F$. Interactions that exist only in G or H

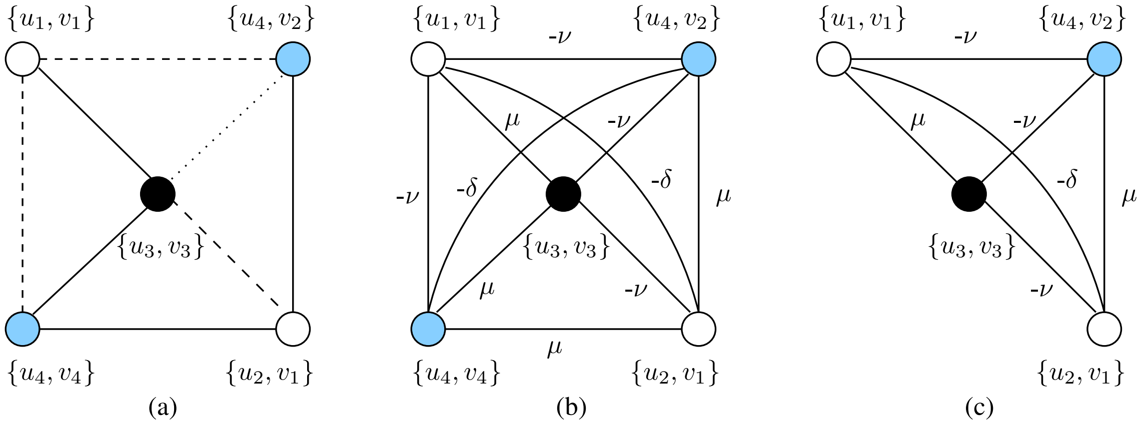


Fig. 4. (a) Product of PINs G and H in Figure 2. Each node is a pair of orthologous nodes, one from each species. Matching interactions are shown by solid edges, interactions that exist only in G and H are shown by dotted and dashed edges, respectively. (b) Corresponding alignment graph. Edges corresponding to matching interactions are weighted by positive match scores while those representing unmatched interactions are weighted by negative mismatch penalties. Two nodes that contain paralogous proteins in the same species are connected by an edge, weighted by a negative duplication penalty. (c) Subgraph induced by node set $\tilde{\mathbf{V}} = \{\{u_1, v_1\}, \{u_2, v_1\}, \{u_3, v_3\}, \{u_4, v_2\}\}$. This subgraph corresponds to the alignment shown in Figure 3.

are shown using dashed and dotted lines, respectively. These interactions correspond to mismatch edges in the alignment graph. For example, since $u_2u_3 \notin E$ although $v_1v_3 \in F$, the edge $\{u_2, v_1\}\{u_3, v_3\}$ is a mismatch edge, shown by a dashed line in the figure.

Adding duplication edges for each pair of nodes that contain paralogs and assigning appropriate weights to each edge, we obtain the alignment graph shown in Figure 4(b). This graph contains two duplication edges that are caused by paralogous proteins u_1,u_2 in G and v_1, v_2 in H. Note that the weights shown on the edges of the graph are not constants, instead they are functions of their incident nodes.

This construction of the alignment graph allows us to formulate the alignment problem as a graph optimization problem, which is defined as follows:

Definition 3: **Maximum Weight Induced Subgraph Problem.**
Given graph $\mathbf{G}(\mathbf{V}, \mathbf{E})$ and a constant ϵ, find a subset of nodes, $\tilde{\mathbf{V}} \in \mathbf{V}$ such that the sum of the weights of the edges in the subgraph induced by $\tilde{\mathbf{V}}$ is at least ϵ, *i.e.*,

$$\sum_{\mathbf{v}, \mathbf{v}' \in \tilde{\mathbf{V}}} w(\mathbf{v}\mathbf{v}') \geq \epsilon. \tag{6}$$

Not surprisingly, this problem is equivalent to the local alignment of PINs defined in the previous section, as stated by the following theorem:

Theorem 1: Given PINs G, H, and a protein similarity function S, let $\mathbf{G}(\mathbf{V}, \mathbf{E})$ be the corresponding alignment graph. If $\tilde{V}$ is a solution for the maximum weight induced subgraph problem on $\mathbf{G}(\mathbf{V}, \mathbf{E})$ with parameter ϵ, then $P = \{\tilde{U}, \tilde{V}\}$ induces an alignment $\mathcal{A}(G, H, S, P)$ with $\sigma(\mathcal{A}) \geq \epsilon$, where

$$\tilde{U} = \{u \in U : \exists v \in V \text{ s.t. } \{u, v\} \in \tilde{\mathbf{V}}\} \tag{7}$$

and

$$\tilde{V} = \{v \in V : \exists u \in U \text{ s.t. } \{u, v\} \in \tilde{\mathbf{V}}\}. \tag{8}$$

Proof. Follows directly from the construction of alignment graph.

The induced subgraph that corresponds to the local alignment in Figure 3 is shown in Figure 4(c).

It can be easily shown that the maximum-weight induced subgraph problem is NP-complete by reduction from maximum clique, by assigning unit weight to edges and $-\infty$ to non-edges. This problem is closely related to the maximum edge subgraph and maximum dispersion problems that are also NP-complete. Maximum edge subgraph problem, also known as heaviest subgraph or k-dense subgraph, restricts the edge weights to positive values and allows at most k nodes in the solution [31]. Maximum dispersion is also defined on graphs with positive edge weights, however, it asks for pairwise disjoint node subsets, each containing at most k nodes, such that the total edge weight of all subgraphs induced by these node subsets is maximized [32]. This is in accordance with the local alignment of PINs to a certain extent, since we are also looking for all local alignments with sufficiently high scores. However, the positive weight restriction on these problems limits the application of existing algorithms to the maximum weight induced subgraph problem since these algorithms generally rely on this restriction. Nevertheless, the nature of the conservation of PINs makes a simple greedy heuristic quite effective for the local alignment of PINs, as we discuss in the next section.

C. A Greedy Heuristic for Local Alignment of Protein Interaction Networks

Proteins, along with other biomolecules, are organized into functional modules to perform specific cellular tasks in coordination [3]. Not surprisingly, in terms of protein interactions, these modules are densely connected while being separable from other modules, *i.e.*, a protein in a particular module interacts with most proteins in the same module, while it is

procedure GreedyMaWiSh($\mathbf{G}$, ϵ)
  ▷ **Input** $\mathbf{G}(\mathbf{V}, \mathbf{E})$: Weighted (w) alignment graph
  ▷ **Input** ϵ : Threshold on subgraph weight
  ▷ **Output** $\tilde{\mathbf{V}}$: Subset of selected nodes
  ▷ **Running** $g(\mathbf{v})$: Gain of adding $\mathbf{v}$ into $\tilde{\mathbf{V}}$
  ▷ **Running** W : Total subgraph weight
  for each $\mathbf{v} \in \mathbf{V}$ **do**
    $g(\mathbf{v}) \leftarrow w(\mathbf{vv})$
    $w(\mathbf{v}) = \sum_{\mathbf{vv'} \in \mathbf{E}} w(\mathbf{vv'})$
  $\tilde{\mathbf{V}} \leftarrow \emptyset$, $W \leftarrow 0$
  $\tilde{\mathbf{v}} \leftarrow \text{argmax}_{\mathbf{v} \in \mathbf{V}} w(\mathbf{v})$
  repeat
    $\tilde{\mathbf{V}} \leftarrow \tilde{\mathbf{V}} \cup \{\tilde{\mathbf{v}}\}$
    $W \leftarrow W + g(\tilde{\mathbf{v}})$
    for each $\mathbf{v} \in (\mathbf{V} \setminus \tilde{\mathbf{V}})$ s.t.$\tilde{\mathbf{v}}\mathbf{v} \in \mathbf{E}$ **do**
      $g(\mathbf{v}) \leftarrow g(\mathbf{v}) + w(\tilde{\mathbf{v}}\mathbf{v})$
    $\tilde{\mathbf{v}} \leftarrow \text{argmax}_{\mathbf{v} \in (\mathbf{V} \setminus \tilde{\mathbf{V}})} g(\mathbf{v})$
  until $g(\mathbf{v}) \leq 0$
  if $W > \epsilon$ **then return** $\tilde{\mathbf{V}}$, **else return** $\emptyset$

Fig. 5. Greedy algorithm for finding a set of nodes that induces a subgraph of maximal total weight on the alignment graph.

only loosely connected to the rest of the network [33]. Protein complexes, for example, form a single multimolecular machine by interacting with each other at the same time and place [34]. While duplicates of a protein tend to be in the same module (or more generally, subnet) immediately after duplication, they do not necessarily remain associated with the same group of proteins in the long term [27].

Since analysis of conserved motifs reveals that proteins in highly connected motifs are more likely to be conserved suggesting that such dense motifs are parts of functional modules [10], high-scoring local alignments are likely to correspond to functional modules. Therefore, in the alignment graph, we can expect that proteins that belong to a particular module will induce heavy subgraphs, while being loosely connected to other parts of the graph. This observation leads to a greedy algorithm that can be expected to work well for the solution of the maximum weight induced subgraph problem on the alignment graph of two PINs. Namely, by seeding a growing subgraph with a protein that has a large number of conserved interactions and small number of mismatched interactions (*i.e.*, a *conserved hub*) and adding proteins that share conserved interactions with this graph one by one, it is possible to discover a group of proteins with a set of dense interactions that are conserved, supposedly being part of some functional module. This algorithm is also able to account for divergence of duplicate proteins in terms of being connected to the same conserved module or not.

A sketch of the greedy algorithm for finding a single conserved subgraph on the alignment graph is shown in Figure IV-C. This algorithm greedily grows a subgraph. The first node in the subgraph is chosen to have maximum total edge weight, with the expectation that it will correspond to a conserved hub. Each node is associated with a running gain function that represents the increase in the total weight of the subgraph that will be gained by adding this node into the subgraph at that point. This gain is initialized to the match score or mismatch penalty of its self-interaction. Then, at each step of the algorithm, the node with largest gain is chosen, added to the growing subgraph and the gains of its neighbors are updated accordingly. The algorithm stops when no node with positive gain can be found. The resulting subgraph, which is of locally maximal total weight, is returned if its total weight is larger than the specified threshold. To find all disjoint "good" alignments, we start with the entire alignment graph and find a maximal subgraph. Then, we remove this subgraph from the alignment graph and run GreedyMaWiSh again. We repeat this process until no subgraph with total weight more than ϵ can be found. We then rank all subgraphs based on their score and report the corresponding alignments.

Note that the definition of match score and mismatch, duplication penalties may result in a modification of the maximum-weight induced subgraph problem to match the underlying local alignment problem. For instance, if we penalize duplications depending on existence of other duplications of proteins in the same paralogous group, this will translate into dynamic edge weights in the alignment graph, *i.e.*, the weight of an edge will be a function of the selected subgraph. The simple greedy algorithm presented in this section can be easily modified to handle such variations, allowing applicability of a wide range of models for scoring matches, mismatches, and duplications that are discussed in the next section.

D. Selection of Model Components

In order for the discovered PIN alignments to be biologically meaningful, selection of the underlying similarity function and the models for scoring and penalizing matches, mismatches, and duplications is crucial, as in the case of sequences. In this section, we discuss possible choices for these components taking into account the underlying biological processes and computational requirements.

1) Similarity Function: A good source for assessing similarity between proteins is clearly pairwise sequence alignment, which is also conveniently used by many studies that focus on comparing biological networks [15], [18]. For this purpose, the *E* value of BLAST pairwise local alignment [2] can be used with

[2]A web interface for `bl2seq`, a BLAST application for pairwise local alignment of nucleotide or protein sequences, can be found at `http://bioweb.pasteur.fr/seqanal/interfaces/bl2seq-simple.html`.

appropriate normalization. In PathBLAST [18], the similarity between two proteins u and v is defined as

$$S(u,v) = log_{10} \frac{p(u,v)}{p_{random}}, \tag{9}$$

where $p(u,v)$ is the probability of true homology between u and v given the BLAST E value of their alignment and p_{random} is the expected value of p over all proteins in the PINs being aligned.

Since proteins that are involved in a common functional module, or more generally, proteins that interact with each other, show local sequence similarities [35], care must be taken while employing significance of local alignment as a measure of potential orthology between proteins. One possible approach that can be employed for this purpose is setting cut-off values on the significance of alignments, setting similarity score to zero for protein pairs with significance below this threshold [18], [27]. Another method to assess sequence similarities conservatively is requiring the similarity to be as global as possible by setting cut-off values on the overlap between sequences as commonly applied in protein sequence clustering.

While aligning two PINs and interpreting this alignment, only duplications that correspond to paralogous proteins that are duplicated after the split of species are of interest. Such protein pairs are called in-paralogs, while the others are called out-paralogs [36]. Unfortunately, distinguishing between in-paralogs and out-paralogs is not trivial. Therefore, a more conservative approach to assigning similarity scores to protein pairs would be to detect orthologs and in-paralogs using a dedicated algorithm and computing similarity scores between potential orthologs and paralogs based on the confidence values as reported by the algorithm. INPARANOID [3] [36] is an algorithm developed for finding disjoint ortholog clusters in two species. Each ortholog cluster discovered by the algorithm is characterized by two *main orthologs*, one from each species, and possibly several other in-paralogs from both species. The main orthologs are assigned a confidence value of 1.0, while the in-paralogs are assigned confidence scores based on their relative similarity to the main ortholog in their own species. More specifically, the confidence for an in-paralog $u' \in U$ that is in the same ortholog cluster with main orthologs $u \in U$ and $v \in V$ is defined as

$$confidence(u') = \frac{score(u,u') - score(u,v)}{score(u,u) - score(u,v)}, \tag{10}$$

where $score(u,v)$ denotes the BLAST similarity score between proteins u and v in bits. In our experiments, we use the output of INPARANOID as the similarity function, where the similarity between two proteins u and v is defined as

$$S(u,v) = confidence(u) \times confidence(v). \tag{11}$$

[3] INPARANOID is publicly available at `http://inparanoid.cgb.ki.se/`.

This provides a normalized similarity function, where the similarity between two main orthologs of a cluster is 1.0, the similarity between two proteins that belong to separate clusters is zero and the similarity between two proteins that belong to the same cluster is scaled based on their similarity to the main orthologs.

2) Scores and Penalties: **Match Score.** A match is scored positively in an alignment to reward a conserved interaction. Therefore, the score is supposed to represent the similarity between the two interactions that are matched. Since the degree of conservation in the two ortholog protein pairs involved in the matched interactions need not be the same, it is appropriate to conservatively assign the minimum of the similarities at the two ends of the matching interactions to obtain:

$$\mu(uu', vv') = \bar{\mu} \min\{S(u,v), S(u',v')\}, \tag{12}$$

where $\bar{\mu}$ is a pre-determined parameter specifying the relative weight of a match in the total alignment score. An alternate approach that is less conservative would be to assign the average of the two similarity scores to the score of the match. However, this would cause assigning positive score to interactions involving totally irrelevant proteins at one end. In our implementation, we use the function given in Equation 12 to score matches. Match scores can also be scaled by the probability of experimental error as in PathBLAST to account for false positives in the discovery of interactions.

Mismatch Penalty. A mismatch is penalized to account for the divergence of duplicate proteins, elimination of existing interactions in the common ancestor, insertion of new interactions after the split, or experimental error in terms of false positives or negatives in the discovery of interactions. A dual mismatch penalty for the match score of Equation 12 is defined as:

$$\nu(uu', vv') = \bar{\nu} \min\{S(u,v), S(u',v')\}, \tag{13}$$

where $\bar{\nu}$ is the relative weight of a mismatch. With this penalty function, each lost interaction of a duplicate protein is penalized to reflect the divergence of duplicate proteins. We use this function to penalize mismatches in our implementation.

As the insertion or deletion of interactions for proteins that are not duplicated are less probable, it may be appropriate to amplify the mismatch penalties associated with non-duplicated proteins. However, since observed interaction data is prone to experimental errors, care must be taken while employing such conservative approaches. An appropriate method for relaxing the model in order to reduce the effect of experimental errors would be to scale mismatch penalties by the probability of experimental error.

Mismatch penalty can be improved using a *shortest path* model based on the observation that proteins that are linked by a short alternative path is more likely to tolerate losing their interaction. The mismatch penalty according to this model can be defined as

$$\nu(uu', vv') = \bar{\nu}(\Delta(v, v') - 1) \min\{S(u, v), S(u', v')\}, \quad (14)$$

where $\Delta(v, v')$ is the length of the shortest path between proteins v and v', assuming that $vv' \notin F$ without loss of generality. While this model is likely to improve the insight of the alignment algorithm, it is computationally expensive since it requires solution of the all pairs shortest path problem on both PINs.

Duplication penalty. Duplications are penalized to account for the divergence of proteins after duplication. Although this is implicitly handled by mismatch penalties, it is also necessary to consider the likeliness of two proteins to be paralogs while scoring alignments involving duplicated proteins. Since sequence similarity provides a crude approximation to the age of duplication and likeliness of being paralogs, it is appropriate to penalize duplications based on their similarity score, leading to a duplication penalty of:

$$\delta(u, u') = \bar{\delta}(d - S(u, u')), \quad (15)$$

where $\bar{\delta}$ is the relative weight of a duplication and $d \geq \max_{u,u'} S(u, u')$ is a parameter that determines the extent of penalizing duplications. Considering the similarity function of Equation 11, setting $d = 1.0$ results in no penalty for duplicates that are paralogous to the main ortholog with 100% confidence. We use this function to penalize duplications in our implementation.

Note that the alignment graph model enforces each duplicate pair in an alignment to be penalized. For example, if an alignment contains n paralogous proteins in one species, $\binom{n}{2}$ duplications are penalized to account for each duplicate pair. However, in the evolutionary process, each paralogous protein is the result of a single duplication, *i.e.*, n paralogous proteins are created in only $n - 1$ duplications. Therefore, we refer to this as *quadratic duplication model*, since the number of penalties is a quadratic function of number of duplications. While this might be desirable as being more restrictive on duplications, to be more consistent with the underlying biological processes, it can be replaced by a *linear duplication model*, in which each duplicate protein is penalized only once, based on its similarity with the paralog that is most similar to itself. This model can be incorporated into the alignment graph model of Section IV-C with a simple modification in the greedy algorithm to dynamically reassign weights to edges that correspond to duplications.

E. Computational Complexity

The alignment algorithm proposed in Section IV-C is composed of three major computational components. The first component is a preprocessing step that involves computation of the similarity matrix, which is an outer component for the algorithm proposed here, since any PIN alignment algorithm must compute a similarity matrix. The computation of this matrix requires pairwise alignment of all pairs of proteins that are involved in the two PINs to be aligned. This computation is fairly expensive as it has a quadratic complexity in the total number of proteins and is likely to dominate the overall time complexity of the alignment process. For example, it takes more than five hours to align the 17036 proteins that take part in the PINs of the DIP interaction database on a 3 GHz Pentium IV processor with 512 MB RAM. However, since this is a preprocessing step that can be performed offline a-priori in most instances, we do not take the complexity of this computation into account while discussing the computational requirements of the alignment algorithm.

The two major inner components of the alignment algorithm are construction of the alignment graph and solution of the maximum weight induced subgraph problem. Since the complexity of both computations depend on the size of the alignment graph, we first analyze the size of the alignment graph. The number of nodes in this graph is at most $\sum_{u \in U} s(u)$, where $s(u) = |\{v \in U \cup V : S(u, v) > 0\}|$ for $u \in U \cup V$. Since the similarity function is highly sparse, with each protein having only a few potential orthologs independent of the size of the input PINs, we can safely assume that $s(u) \leq c$ for all $u \in U \cup V$ for some constant c. Then, the number of nodes in the alignment graph is at most $c(|U| + |V|)$. On the other hand, the number of edges in the alignment graph is at most $\sum_{u \in U} s(u) \times (d(u) + s(u)) + \sum_{v \in V} s(v) \times (d(v) + s(v))$, where $d(u) = |\{u' \in U : uu' \in E\}|$ for $u \in U$ and $d(v) = |\{v' \in V : vv' \in F\}|$ for $v \in V$. In this equation, $s(u)d(u)$ accounts for the replication of the interactions of u for match and mismatch edges, while $s(u)s(u)$ accounts for duplication edges of u. Hence, the number of edges in the alignment graph is at most $c(|E| + |F|) + c^2(|U| + |V|)$. Therefore, we conclude that $|\mathbf{V}| = \Theta(|U| + |V|)$ and $|\mathbf{E}| = \Theta(|E| + |F|)$.

During the construction of the alignment graph, with the help of sparse storage schemes supported by hash-based data structures that allow both sequential and random access to interactions and similarity scores in constant time, each edge in the alignment graph can be computed in $\Theta(1)$ time. Thus, construction of $\mathbf{G}$ requires $\Theta(|E| + |F|)$ time. In the greedy algorithm for computation of local alignments, nodes need to be organized in a priority queue for extraction of the node with maximum gain. Each node in $\mathbf{V}$ can be extracted from the priority queue and added to the growing subgraph at most once. On the other hand, during the update of node gains, each edge in $\mathbf{E}$ can be visited at most once, since its one end will be removed from the queue after duplication. Hence, if the priority queue is implemented as a heap, the overall complexity of the alignment algorithm is $O(|\mathbf{E}| + |\mathbf{V}| \log |\mathbf{V}|)$ $= O(|E| + |F| + (|U| + |V|) \log(|U| + |V|))$.

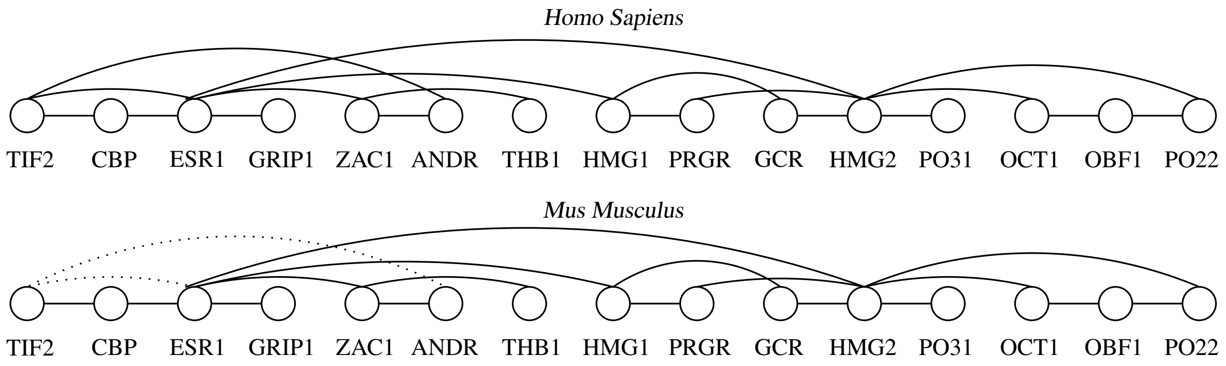


Fig. 6. A conserved subnet that is part of DNA-dependent transcription regulation in human and mouse PINs. Ortholog proteins are vertically aligned. Existing interactions are shown by solid edges, missing interactions that have an existing orthologous counterpart in the other species are shown by dotted edges.

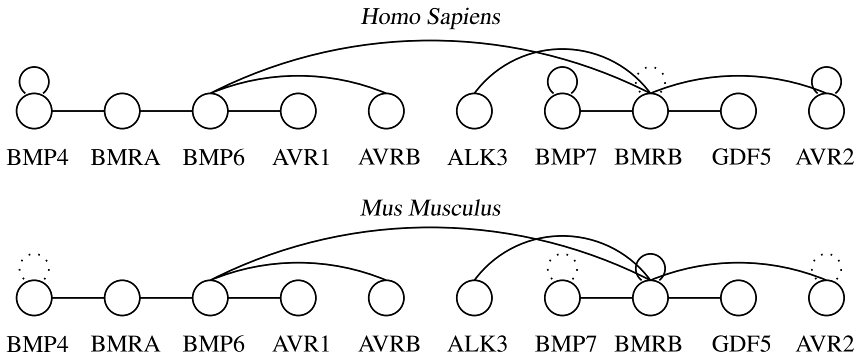


Fig. 7. A conserved subnet that is part of transforming growth factor beta receptor signaling pathway in human and mouse PINs.

V. Experimental Results

In this section, we present sample local alignment results to illustrate the effectiveness of the proposed framework and the underlying algorithm. For this purpose, we gather available protein interaction data from the DIP [4] protein interaction database. We first align the PINs of two mammalians that are available in the database; *Homo sapiens* (Hsapi) and *Mus musculus* (Mmusc). As of October 2004, the Hsapi PIN contains 1369 interactions among 1065 proteins while Mmusc PIN contains 286 interactions among 329 proteins. Running INPARANOID on this set of 1351 proteins, we discover 237 ortholog clusters. Based on the similarity function induced by these clusters, we construct an alignment graph that consists of 273 nodes and 1233 edges. We then compute local alignments using the greedy algorithm of Section IV-C on this graph. Note that construction of alignment graph and discovery of local alignments on this graph takes only a few milliseconds on all pairs of species that are currently available in DIP.

A subnet of DNA-dependent transcription regulation that is highly conserved in human and mouse PINs is shown in Figure 6. The subnet is composed of three major common functional groups, namely transcription factors and coactivators PO22, PO31, OCT1, TIF2, OBF1, steroid hormone receptors GCR, ANDR, ESR1, PRGR, GRIP1, THB1, and high mobility proteins HMG1 and HMG2. Indeed, it is known that HMG1 and HMG2 are co-regulatory proteins that increase the DNA binding and transcriptional activity of the steroid hormone class of receptors in mammalian cells [37]. All proteins in this subnet are localized in nucleus, with mobility proteins particularly localizing in condensed chromosome. This subnet contains 17 matching interactions between 15 proteins. Two interactions of TIF2 (transcriptional intermediary factor 2) that exist in human are missing in mouse. If we increase the relative weight of mismatch penalties in the alignment score, the alignment does not contain TIF2 any more, providing a perfect match of 16 interactions.

The subnet that is part of transforming growth factor beta receptor signaling pathway, which is highly conserved in human and mouse PINs is shown in Figure 7. This subnet contains 8 matching interactions among 10 proteins. It is composed of two separate subnets that are connected through the interaction of their hubs, namely BMP6 (bone morphogenetic protein 6 precursor) and BMRB (activin receptor-like kinase 6 precursor). All proteins in this subnet have the common function of transforming growth factor beta receptor activity and are localized in the membrane. Note that self-interactions of three proteins in this subnet that exist in human PIN are

[4]PIN datasets for seven species can be downloaded, browsed or searched on the DIP website at `http://dip.doe-mbi.ucla.edu/`.

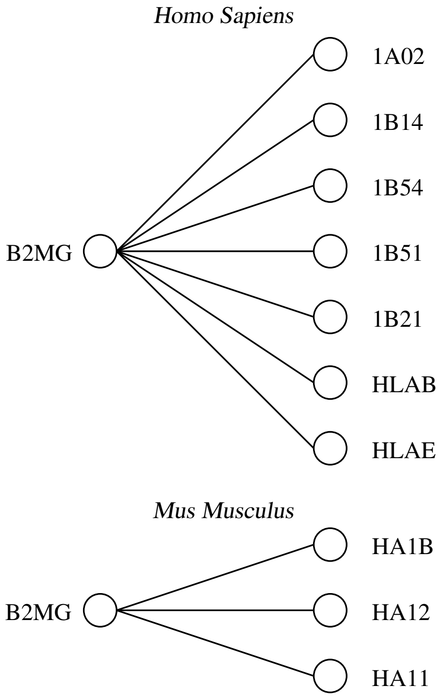


Fig. 8. A conserved subnet that is part of antigen presentation and antigen processing in human and mouse PINs. Paralogous proteins in a species are shown in order of confidence in being orthologous to the respective proteins in the other species.

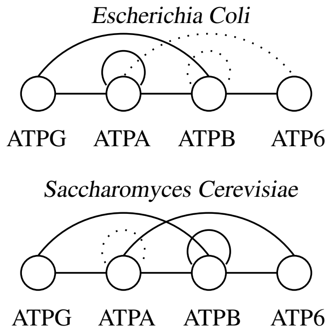


Fig. 9. A conserved subnet that is part of ATP-synthesis coupled proton transport in ecoli and yeast PINs.

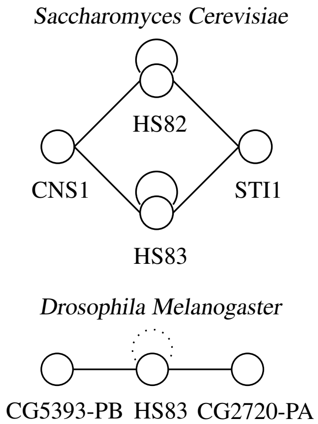


Fig. 10. A conserved subnet that is part of response to stress in yeast and fruit fly PINs.

missing in mouse and one self-interaction that exists in mouse is missing in human.

As an example of duplication, a subnet that is part of antigen presentation and antigen processing, which is conserved in human and mouse PINs is shown in Figure 8. This subnet is a star network of several paralogous class I histocompatibility antigens interacting with B2MG (beta-2 microglobulin precursor) in both species. In the figure, paralogous proteins are displayed in order of confidence in being orthologous to the corresponding proteins in the other species from top to bottom. This star network is associated with MHC class I receptor activity. Since all proteins that are involved in these interactions are homologous, we can speculate that all these interactions have evolved from a single common interaction. Note that such patterns are found only with the help of the duplication concept in the alignment model. Neither a pathway alignment algorithm, nor an algorithm that tries to match each protein with exactly one ortholog in the other species will be able to detect such conserved patterns. Indeed, if the weight of duplication penalty in the scoring of alignment is high, this subnet cannot be detected by the alignment algorithm.

We also perform experiments to align PINs of several pairs of species among the seven available species-specific PINs available in DIP. However, for space limitations, we do not discuss these results in detail. We conclude this section with two small conserved subnets from two different pairs of PINs. The first conserved subnet shown in Figure 9 is part of ATP-synthesis coupled proton transport [38] and is common to a pair of prokaryotic and eukaryotic PINs, namely ecoli and yeast. Another sample conserved subnet is shared by yeast and fruit fly, with a duplication of heat shock proteins in yeast as shown in Figure 10. This small subnet is composed of proteins that are localized in cytoplasm and is part of the response to stress.

VI. Concluding Remarks and Ongoing Work

This paper presents a framework for local alignment of protein interaction networks that is guided by theoretical models of the evolution of these networks. The model is based on discovering sets of proteins that induce conserved subnets with the expectation that these proteins will constitute a part of protein complexes or functional models, which are expected to be conserved together. A preliminary implementation of the proposed algorithm reveals that this framework is quite successful in uncovering conserved substructures in protein interaction data.

We are currently working on a comprehensive implementation of the proposed framework that allows adaptation of several models for assessing protein similarities and scoring/penalizing matches, mismatches and duplications. Another important feature that needs to be added to the framework is a means of evaluating significance. Once these enhancements are completed, the proposed framework will be established as a tool for pairwise alignment of PINs, that will be publicly available through a web interface. The framework will also be generalized to the search of input queries in the form of subnets in a database of PINs. Using this tool researchers will be able to find conserved counterparts of newly discovered complexes or modules in several species.

REFERENCES

[1] S. F. Altschul, T. L. Madden, A. A. Schffer, Z. Z. J. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nuc. Acids Res.*, vol. 25, no. 17, pp. 3389–3402, 1997.

[2] J. D. Thompson, D. G. Higgins, and T. J. Gibson, "CLUSTAL-W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," *Nuc. Acids Res.*, vol. 22, pp. 4673–4680, 1994.

[3] L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray, "From molecular to modular cell biology," *Nature*, vol. 402, pp. C47–C51, 1999.

[4] Z. N. Oltvai and A. L. Barabási, "Life's complexity pyramid," *Science*, vol. 298, pp. 763–764, 2002.

[5] B. Titz, M. Schlesner, and P. Uetz, "What do we learn from high-throughput protein interaction data?" *Expert Rev. Proteomics*, vol. 1, no. 1, pp. 111–121, 2004.

[6] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki, "A comprehensive two-hybrid analysis to explore the yeast protein interactome," *PNAS*, vol. 98, no. 8, pp. 4569–4574, 2001.

[7] Y. H. et al., "Systematic identification of protein complexes in Saccharomyces cerevisae by mass spectrometry," *Nature*, vol. 415, pp. 180–183, 2002.

[8] A. H. T. et al., "A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules," *Science*, vol. 295, pp. 321–324, 2002.

[9] A. Vespignani, "Evolution thinks modular," *Nature Gen.*, vol. 35, no. 2, pp. 118–119, 2003.

[10] S. Wuchty, Z. N. Oltvai, and A. L. Barabási, "Evolutionary conservation of motif constituents in the yeast protein interaction network," *Nature Gen.*, vol. 35, no. 2, pp. 176–179, 2003.

[11] Y. Tohsato, H. Matsuda, and A. Hashimoto, "A multiple alignment algorithm for metabolic pathway analysis using enzyme hierarchy," in 8^{th} *Intl. Conf. Int. Sys. Mol. Bio. (ISMB'00)*, 2000, pp. 376–383.

[12] M. Koyutürk, A. Grama, and W. Szpankowski, "An efficient algorithm for detecting frequent subgraphs in biological networks," *BioInformatics (12th Intl. Conf. Int. Sys. Mol. Bio. (ISMB 2004))*, vol. 20, no. Suppl. 1, pp. i200–i207, 2004.

[13] B. P. Kelley, B. Yuan, F. Lewitter, R. Sharan, B. R. Stockwell, and T. Ideker, "PathBLAST: a tool for aligment of protein interaction networks," *Nuc. Acids Res.*, vol. 32, pp. W83–W88, 2004.

[14] A. Vázquez, A. Flammini, A. Maritan, and A. Vespignani, "Modeling of protein interaction netwokrs," *ComPlexUs*, vol. 1, pp. 38–44, 2003.

[15] T. Dandekar, S. Schuster, B. Snel, M. Huynen, and P. Bork, "Pathway alignment: application to the comparative analysis of glycolytic enzymes," *Biochemistry J.*, vol. 343, pp. 115–124, 1999.

[16] E. Y. .Lotem, S. Sattath, N. Kashtan, S. Itzkovitz, R. Milo, R. Y. Pinter, U. Alon, and H. Margalit, "Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction," *PNAS*, vol. 101, no. 16, pp. 5934–5939, 2004.

[17] H. Yu, X. Zhu, D. Greenbaum, J. Karro, and M. Gerstein, "TopNet: a tool for comparing biological sub-networks, correlating protein properties with topological statistics," *Nuc. Acids Res.*, vol. 32, no. 1, pp. 328–337, 2004.

[18] B. P. Kelley, R. Sharan, R. M. Karp, T. Sittler, D. E. Root, B. R. Stockwell, and T. Ideker, "Conserved pathways withing bacteria and yeast as revealed by global protein network alignment," *PNAS*, vol. 100, no. 20, pp. 11 394–11 399, 2003.

[19] H. Jeong, S. P. Mason, A. Barabási, and Z. N. Oltvai, "Lethality and centrality in protein networks," *Nature*, vol. 411, pp. 41–42, 2001.

[20] S. Wuchty, "Evolution and topology in the yeast protein interaction network," *Gen. Res.*, vol. 14, pp. 1310–1314, 2004.

[21] N. Pržulj, D. A. Wigle, and I. Jurisica, "Functional topology in a network of protein interactions," *Bioinformatics*, vol. 20, no. 3, pp. 340–348, 2004.

[22] A. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, pp. 509–512, 1999.

[23] E. Eisenberg and Y. Levanon, "Preferential attachment in the protein network evolution," *Phys. Rev. Letters*, vol. 91, no. 13, p. 138701, 2003.

[24] H. Qin, H. H. S. Lu, W. B. Wu, and W. Li, "Evolution of the yeast protein interaction network," *PNAS*, vol. 100, no. 22, pp. 12 820–12 824, 2003.

[25] V. Kunin, J. B. Pereira-Leal, and C. A. Ouzounis, "Functional evolution of the yeast protein interaction network," *Mol. Bio. Evol.*, vol. 21, no. 7, pp. 1171–1176, 2004.

[26] R. Pastor-Satorras, E. Smith, and R. V. Solé, "Evolving protein interaction networks through gene duplication," *J. Theo. Bio.*, vol. 222, pp. 199–210, 2003.

[27] A. Wagner, "The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes," *Mol. Bio. Evol.*, vol. 18, no. 7, pp. 1283–1292, 2001.

[28] ——, "How the global structure of protein interaction networks evolves," *Proc R Soc Lond B Biol Sci.*, vol. 270, no. 1514, pp. 457–466, 2003.

[29] F. Chung, L. Lu, T. G. Dewey, and D. J. Galas, "Duplication models for biological networks," *J. Comp. Bio.*, vol. 10, no. 5, pp. 677–687, 2003.

[30] F. Harary, *Graph Theory*. MA: Addison-Wesley, 1994.

[31] U. Feige, D. Peleg, and G. Kortsarz, "The dense k-subgraph problem," *Algorithmica*, vol. 29, no. 3, pp. 410–421, 2001.

[32] R. Hassin, S. Rubinstein, and A. Tamir, "Approximation algorithms for maximum dispersion," *Operations Res. Letters*, vol. 21, pp. 133–137, 1997.

[33] S. Tornow and H. W. Mewes, "Functional modules by relating protein interaction networks and gene expression," *Nuc. Acids Res.*, vol. 31, no. 21, pp. 6283–6289, 2003.

[34] V. Spirin and L. A. Mirny, "Protein complexes and functional modules in molecular networks," *PNAS*, vol. 100, no. 21, pp. 12 123–12 128, 2003.

[35] Y. Bilu and M. Linial, "On the predictive power of sequence similarity in yeast," in *Proc. 5th Intl. Conf. Res. Comp. Mol. Bio. (RECOMB 2001)*, 2001, pp. 39–48.

[36] M. Remm, C. E. V. Storm, and E. L. L. Sonnhammer, "Automatic clustering of orthologs and in-paralogs from pairwise species comparisons," *J. Mol. Bio.*, vol. 314, pp. 1041–1052, 2001.

[37] V. B. et al., "High-mobility group chromatin proteins 1 and 2 functionally interact with steroid hormone receptors to enhance their DNA binding in vitro and transcriptional activity in mammalian cells," *Mol. Cell. Bio.*, vol. 18, no. 8, pp. 4471–4488, 1998.

[38] R. H. Fillingame, "Coupling h+ transport and atp synthesis in f1f0-atp synthases: glimpses of interacting parts in a dynamic molecular machine," *J. Exp. Biol.*, vol. 200, pp. 217–224, 1997.