Pairwise Alignment of Protein Interaction Networks *

Mehmet Koyutürk[1,†], Yohan Kim[2], Umut Topkara[1],

Shankar Subramaniam[2,3], Wojciech Szpankowski[1], and Ananth Grama[1]

[1] Department of Computer Sciences, Purdue University, West Lafayette, IN 47906.

[2] Department of Chemistry and Biochemistry, University of California at San Diego, La Jolla, CA 92093.

[3] Department of Bioengineering, University of California at San Diego, La Jolla, CA 92093.

Author e-mail addresses:

koyuturk@cs.purdue.edu, ykim@ucsd.edu,utopkara@cs.purdue.edu,

shankar@sdsc.edu, spa@cs.purdue.edu, ayg@cs.purdue.edu

*A preliminary version of this manuscript was presented at RECOMB 2005. In view of the significant demand for this software following the meeting, we have made the source freely available at http://www.cs.purdue.edu/homes/koyuturk/mawish/

†Corresponding author

Abstract

With ever increasing amount of available data on protein-protein interaction (PPI) networks and research revealing that these networks evolve at a modular level, discovery of conserved patterns in these networks becomes an important problem. Recently proposed algorithms for aligning PPI networks 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, algorithms that are able to extract conserved patterns in terms of general graphs are necessary. With this motivation, we focus 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 PPI networks, model the problem as a graph optimization problem, and present fast algorithms for this problem. In order to capture the underlying biological processes accurately, we base our framework on duplication/divergence models that focus on understanding the evolution of PPI networks. Detailed experimental results from an implementation of the proposed framework show that our algorithm is able to discover conserved interaction patterns very effectively, both in terms of accuracies and computational cost.

1 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, 32]. 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 [16]. An important class of molecular interaction data is in the form of protein-protein interaction (PPI) networks. These networks provide the experimental basis for understanding modular organization of cells, as well as useful information for predicting the biological function of individual proteins [33]. High throughput screening methods such as two-hybrid analysis [18], mass spectrometry [13], and TAP [9] provide large amounts of data on these networks.

As revealed by recent studies, PPI networks evolve at a modular level [39] and consequently, understanding 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 [34], finding frequently occurring subgraphs in a collection of graphs [22], discovering highly conserved subgraphs in a pair of graphs, and finding good matches for a subgraph in a database of graphs [20]. In this paper, we specifically focus on discovering highly conserved subnets in a pair of PPI networks. With the expectation that conserved subnets will be parts of pathways, complexes, or modules, we base our model on the discovery of two subsets of proteins from each PPI network such that the induced subnets are highly conserved.

Based on the understanding of the structure of PPI networks 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 PPI networks [36]. In order to capture the underlying biological processes accurately, we base our framework on duplication/divergence models by defining 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 our framework show that the proposed algorithm is able to discover conserved interaction patterns very effectively. The proposed algorithm can be also adapted to finding matches for a subnet query in a database of PPI networks.

The rest of this paper is organized as follows: we start with a brief overview of duplication/divergence models for the evolution of PPI networks in Section 2. In Section 3, 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 on comprehensive pairwise alignment of the PPI networks for three eukaryotic species in Section 4. We then discuss existing literature on network alignment and compare the proposed framework with existing methods in Section 5. We conclude our discussion in Section 6.

2 Theoretical Models for Evolution of PPI Networks

There have been a number of studies aimed at understanding the general structure of PPI networks. These studies suggest that PPI networks can generally be modeled by power-law graphs, *i.e.*, the relative frequency of proteins that interact with k proteins is roughly proportional to $k^{-\gamma}$, where γ is a network-specific parameter [5]. In order to explain this power-law nature, Barábasi and Albert have proposed [5] a network growth model based on preferential attachment, which is able to generate networks with degree distribution similar to PPI 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, Eisenberg and Levanon [8] 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 [27].

A common model of evolution that explains preferential attachment is the duplication/divergence model, which is based on gene duplications [25, 36, 37, 38]. 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 protein duplication is shown in Figure 1. A protein loses many aspects of its functions rapidly after being duplicated. This translates to divergence of duplicated (paralogous) proteins in the interactome through elimination and emergence of interactions. Elimination of an interaction in a PPI network implies the loss of an interaction between two proteins due to structural and/or functional changes. Similarly, emergence of an interaction in a PPI network implies the introduction of a new interaction between two non-interacting proteins, caused by mutations that change protein surfaces. Examples of elimination and emergence of interactions are also illustrated in Figure 1. If an elimination or emergence is related to a recently duplicated protein, it is said to be correlated; otherwise, it is uncorrelated [25]. Since newly duplicated proteins are more tolerant to interaction loss because of redundancy, correlated elimination is generally more probable than emergence and uncorrelated elimination [36]. It is also theoretically shown that network growth models based on node duplications generate power-law distributions [6].

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 [37]. Indeed, 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 back [37].

Consequently, when we consider the PPI networks that belong to two separate species, the in-paralogs will be likely to have more common interactions than out-paralogs. Here, we use the terms in-paralog and out-paralog for proteins that are duplicated before and after speciation, respectively. While comparatively analyzing the proteome and interactome, it is important to distinguish in-paralogs from out-paralogs since the former are more likely to be functionally related. This, however, is a difficult task since out-paralogs also show sequence similarity.

In order to accurately identify and interpret conservation of interactions, complexes, and modules across species, we base our framework for the local alignment of PPI networks on duplication/divergence models. While searching for highly conserved groups of interactions, we evaluate mismatched interactions and paralogous proteins in light of the duplication/divergence model. Introducing the concepts of match (conservation), mismatch (emergence or elimination) and duplication, which are in accordance with widely accepted models of evolution, we are able to discover alignments that also allow speculation about the structure of the network in the common ancestor.

3 Pairwise Local Alignment of PPI Networks

In light of the theoretical models of evolution of PPI networks, we develop a framework for the comparison of PPI networks in two different species. We formally define a computational problem that captures the underlying biological phenomena using matches, mismatches, and duplications. We then formulate PPI network alignment as a graph optimization problem and propose efficient heuristics to effectively solve this problem.

3.1 The PPI Network Alignment Problem

A PPI network 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 PPI networks, we are given two PPI networks belonging to two different species, denoted by $G(U, E)$ 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, where $0 \leq S(u, v) \leq 1$. If u and v belong to the same species, then $S(u, v)$ quantifies the likelihood that the two proteins are in-paralogs. S is expected to be sparse, *i.e.,* each protein is expected to have only a few potential orthologs. We discuss the methodology for deriving similarity scores from sequence alignments in Section 3.1.3.

For PPI networks $G(U, E)$ 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 analog of a *duplication* is the duplication of a gene in the course of evolution. Each duplication is associated with a score that reflects the divergence of function between the two proteins, estimated using their similarity. 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 PPI network of one organism between a pair of proteins whose orthologs interact in the other organism. A mismatch may correspond to the emergence of a new interaction or the elimination of a previously existing interaction in one of the species after the split, or an experimental error. Thus, mismatches are 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 align-

ment as an optimization problem. Note that although PPI networks are undirected graphs, interactions are regarded as ordered pairs in the following 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 PPI networks.

Given protein interaction networks $G(U, E)$, $H(V, F)$, let functions $\Delta_G(u, u')$ and $\Delta_H(v, v')$ denote the distance between two corresponding proteins in the interaction graphs G and H, respectively. Given a pairwise similarity function S defined over the union of their protein sets $U \cup V$, and a distance cutoff $\bar{\Delta}$, 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

$$\begin{aligned} \mathcal{M} = \ \{ \ & u, u' \in \tilde{U}, v, v' \in \tilde{V} : S(u, v) > 0, \ S(u', v') > 0, \\ & ((uu' \in E \wedge \Delta_H(v, v') \leq \bar{\Delta}) \vee (vv' \in F \wedge \Delta_G(u, u') \leq \bar{\Delta})) \ \ \} \end{aligned} \tag{1}$$

$$\begin{aligned} \mathcal{N} = \ \{ \ & u, u' \in \tilde{U}, v, v' \in \tilde{V} : S(u, v) > 0, \ S(u', v') > 0, \\ & ((uu' \in E \wedge \Delta_H(v, v') > \bar{\Delta}) \vee (vv' \in F \wedge \Delta_G(u, u') > \bar{\Delta})) \ \ \} \end{aligned} \tag{2}$$

$$\mathcal{D} = \{ \ u, u' \in \tilde{U} : S(u, u') > 0 \ \} \cup \{ \ v, v' \in \tilde{V} : S(v, v') > 0 \ \} \tag{3}$$

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

Following the definition of match and mismatch, while assessing the conservation of interactions, we take into account not only direct but also indirect interactions. If two proteins directly interact with each other in one organism, and their orthologs are reachable from each other via at most $\bar{\Delta}$ interactions in the other, we consider this a match. Conversely, a mismatch corresponds to the situation in which two proteins cannot reach each other via $\bar{\Delta}$ interactions in one network while their orthologs directly interact in the other. This approach is motivated by two observations. First, proteins that are linked by a short alternate path are more likely to tolerate losing their interaction because of relaxation of evolutionary

pressure. Second, high-throughput methods such as TAP [9] identify complexes that are associated with a single central protein and these complexes are recorded in the interaction database as star networks with the central protein serving as a hub. Therefore, all proteins that are part of a particular complex can be viewed as interacting by setting $\bar{\Delta} = 2$.

3.1.1 Scoring Match, Mismatch, and Duplications

For scoring matches and mismatches, we define the similarity between two protein pairs as follows:

$$S(uu', vv') = S(u, v)S(u', v') \tag{4}$$

$S(uu', vv')$ quantifies the likelihood that the interactions between u and v, and u' and v' are orthologous. Consequently, a match that corresponds to a conserved pair of orthologous interactions is rewarded as follows:

$$\mu(uu', vv') = \bar{\mu}S(uu', vv') \tag{5}$$

Here, $\bar{\mu}$ is the match coefficient that is used to tune the relative weight of matches against mismatches and duplications, based on the evolutionary distance between the species that are being compared.

A mismatch may correspond to the functional divergence of either interacting partner after speciation. It might also be due to a false positive or negative in one of the networks that is caused by incompleteness of data or experimental error [33]. However, considering indirect interactions as matches compensates for the second case to a certain extent. In most cases, interacting partners that are part of a common functional module are linked by short alternative paths. Therefore, even if an existing direct interaction is not observed, it is likely that a short alternate path linking them will exist in the data. Based on these observations, we penalize mismatches for possible divergence in function as follows:

$$\nu(uu', vv') = -\bar{\nu}S(uu', vv') \tag{6}$$

As for match score, mismatch penalty is also normalized by a coefficient $\bar{\nu}$ that determines the relative weight of mismatches w.r.t. matches and duplications.

While aligning PPI networks, the motivation is to identify conserved patterns of interactions between orthologous proteins. For assessing the likelihood of orthology between proteins, the similarity score defined above relies on sequence homology. However, out-paralogs, which are proteins that are duplicated before the species split hence cannot be considered orthologs, often show sequence similarities as well [28]. Since duplicated proteins rapidly lose their interactions, it is more likely that in-paralogs, *i.e.*, the proteins that are duplicated after a split, will share more interacting partners than out-paralogs do [37]. Therefore, penalizing mismatches implicitly favors *real* orthologs by penalizing the out-paralogs for each interaction that is lost after duplication. Furthermore, we employ sequence similarity as a means for distinguishing in-paralogs from out-paralogs. This is based on the observation that sequence similarity provides a crude approximation for the age of duplication [38]. With the expectation that recently duplicated proteins, which are more likely to be in-paralogs, show more significant sequence similarity than older paralogs, we define duplication score as follows:

$$\delta(u, u') = \bar{\delta}(S(u, u') - \bar{d}) \tag{7}$$

Here $\bar{d}$ is the cut-off for being considered in-paralogs. If $S(u, u') > \bar{d}$, suggesting that u and u' are likely to be in-paralogs, the duplication is rewarded by a positive score. If $S(u, u') < \bar{d}$, on the other hand, the proteins are considered out-paralogs, therefore the duplication is penalized.

3.1.2 Alignment Score and the Optimization Problem

The above formulation of match, mismatch, and duplication translates the problem of distinguishing orthologs and in-paralogs from out-paralogs to an optimization problem that accounts for the trade-off between conservation of sequences and interactions. This enables accurate identification of conserved

interactions between ortholog protein pairs, while allowing us to define the pairwise local alignment for inter-species comparison of PPI networks as an optimization problem.

Definition 2 Alignment Score and PPI Network Alignment Problem.

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

$$\sigma(\mathcal{A}) = \sum_{M \in \mathcal{M}} \mu(M) + \sum_{N \in \mathcal{N}} \nu(N) + \sum_{D \in \mathcal{D}} \delta(D). \tag{8}$$

The PPI network alignment problem is one of finding all maximal protein subset pairs P such that $\sigma(\mathcal{A}(G, H, S, P))$ is locally maximal, i.e. the alignment score cannot be improved by adding individual proteins to or removing proteins from P.

We aim to find local alignments with locally maximal score (drawing an analogy to sequence alignment [31], *high-scoring subgraph pair*s).

We illustrate the concepts of match, mismatch, and duplication using a simple example. Consider the two interaction networks G and H shown in Figure 2(a). The alignment induced by the protein subset pair $\tilde{U} = \{u_1, u_2, u_3, u_4\}$ and $\tilde{V} = \{v_1, v_2, v_3\}$ is shown in Figure 2(b), where we set $\bar{\Delta} = 1$. The only duplication in this alignment is (u_1, u_2). If this alignment is chosen to be a "good" one, then, based on the existence of this duplication in the alignment, if $S(u_2, v_1) < S(u_1, v_1)$, 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_1, v_1v_1), (u_1u_2, v_1v_1), (u_1u_3, v_1v_3), and (u_2u_4, v_1v_2). Observe that v_1 is mapped to both u_1 and u_2 in the context of different interactions. This is associated with the functional divergence of u_1 and u_2 after duplication. Furthermore, the self-interaction of v_2 in H is mapped to an interaction between paralogous proteins in G.

The mismatch set is composed of (u_1u_4, v_1v_2), (u_2u_2, v_1v_1), (u_2u_3, v_1v_3), and (u_3u_4, v_3v_2). The interaction u_3u_4 in G is left unmatched by this alignment, since the only possible pair of proteins in $\tilde{V}$ that are orthologous to these two proteins are v_3 and v_2, which do not interact in H. One conclusion that can be derived from this alignment is the elimination or emergence 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 tolerance to the loss of this interaction. Indeed, if we set $\bar{\Delta} = 2$, then this pair of a direct and an indirect interaction would be considered a match. However, if we include v_4 in $\tilde{V}$ as well, then the induced alignment is able to match u_3u_4 and v_3v_4. This strengthens the likelihood that this interaction existed in the common ancestor. However, v_4 comes with another duplication since it is paralogous to v_2. Hence, if $S(v_2, v_4) > \bar{d}$, the alignment that includes v_4 will be favored over the present one. However, if $S(v_2, v_4) < \bar{d}$, then v_4 must compensate for the duplication penalty with the strength of its matching interactions in order to be included in the alignment.

3.1.3 Estimation of Similarity Scores

The similarity score $S(u, v)$ quantifies the likelihood that proteins u and v are orthologous. We can approximate this likelihood using the BLAST [1] *E*-value for the alignment of u and v, $E(u, v)$. Given an *E*-value cutoff x and O_{uv} representing the event that u and v are orthologous, $P(E(u, v) > x|O_{uv})$ denotes the fraction of orthologs with *E*-values worse than (greater than) x. If we assume that the probability of a protein pair being orthologous is a monotonically decreasing function of the *E*-value, this quantity is a measure of the likelihood that two proteins with *E*-value x are orthologous. This monotonicity assumption is intuitive and we validate this using COG as well.

3.2 Alignment Graph and the Maximum-Weight Induced Subgraph Problem

It is possible to represent information regarding matches and mismatches between two PPI networks using a single alignment graph. This graph is a modified version of the graph Cartesian product 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 the following alignment graph:

Definition 3 Alignment Graph.

For a pair of PPI networks $G(U, E)$, $H(V, F)$, and protein similarity function S, the corresponding weighted alignment graph $\mathbf{G}(\mathbf{V}, \mathbf{E})$ is computed as follows:

$$\mathbf{V} = \{\mathbf{v} = \{u, v\} : u \in U, v \in V \text{ and } S(u, v) > 0\}. \tag{9}$$

In other words, we have a node in the alignment graph for each pair of ortholog proteins. Each edge $\mathbf{vv}' \in \mathbf{E}$, where $\mathbf{v} = \{u, v\}$ and $\mathbf{v}' = \{u', v'\}$, is assigned weight

$$w(\mathbf{vv}') = \mu(uu', vv') + \nu(uu', vv') + \delta(u, u') + \delta(v, v'). \tag{10}$$

Here, $\mu(uu', vv') = 0$ if $(uu', vv') \notin \mathcal{M}$, and similarly for mismatches and duplications.

Consider the PPI networks in Figure 2(a). To construct the corresponding alignment graph, we first compute the product of these two PPI networks to obtain five nodes that correspond to five ortholog protein pairs. We then insert an edge between two nodes of this graph if the corresponding proteins interact in both networks (*match edge*), interact in only one of the networks (*mismatch edge*), or at least one of them is paralogous (*duplication edge*), resulting in the alignment graph of Figure 3(a). Note that the weights assigned to these edges, which are shown in the figure, are not constant, but are functions of their incident nodes. Observe that the edge between $\{u_1, v_1\}$ and $\{u_2, v_1\}$ acts a match and

duplication edge at the same time, allowing analysis of the conservation of self-interactions of duplicated proteins. This construction of the alignment graph allows us to formulate the alignment problem as a graph optimization problem defined below.

Definition 4 Maximum Weight Induced Subgraph Problem (MAWISH**).** *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.,* $W(\tilde{\mathbf{V}}) = \sum_{\mathbf{v},\mathbf{v}' \in \tilde{\mathbf{V}}} w(\mathbf{v}\mathbf{v}') \geq \epsilon$*.*

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

Theorem 1 *Given PPI networks* G*,* H*, and a protein similarity function* S*, let* $\mathbf{G}(\mathbf{V}, \mathbf{E}, w)$ *be the corresponding alignment graph. If* $\tilde{\mathbf{V}}$ *is a solution to the maximum weight induced subgraph problem on* $\mathbf{G}(\mathbf{V}, \mathbf{E}, w)$*, then* $P = \{\tilde{U}, \tilde{V}\}$ *induces an alignment* $\mathcal{A}(G, H, S, P)$ *with* $\sigma(\mathcal{A}) = W(\tilde{\mathbf{V}})$*, where* $\tilde{U} = \{u \in U : \exists v \in V \text{ s.t. } \{u, v\} \in \tilde{\mathbf{V}}\}$ *and* $\tilde{V} = \{v \in V : \exists u \in U \text{ s.t. } \{u, v\} \in \tilde{\mathbf{V}}\}$*.*

Proof. Follows directly from the construction of alignment graph.

The induced subgraph that corresponds to the local alignment in Figure 2(b) is shown in Figure 3(b).

It can be shown that MAWISH 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 [14] and maximum dispersion problems [17], which are also NP-complete. However, the positive weight restriction on these problems limits the application of existing algorithms to the maximum weight induced subgraph problem. Nevertheless, the local PPI network alignment problem aims to find all locally maximal alignments, consequently, locally optimal solutions of MAWISH are sufficient. Observing the similarity between min-cut graph partitioning and MAWISH, we develop fast heuristics based

on common graph partitioning algorithms to identify locally maximal heavy subgraphs in the alignment graph.

3.3 Algorithms for Local Alignment of PPI Networks

In terms of protein-protein interactions, functional modules are likely to be densely connected while being separable from other modules, *i.e.*, a protein in a particular module interacts with most proteins in the same module either directly or through a common module hub, while it is only loosely connected to the rest of the network [35]. 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 [39], 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 conserved module will induce heavy subgraphs, while being loosely connected to other parts of the graph. This observation motivates the process of greedily growing a subgraph seeded at heavy nodes. This approach is shown to perform well in discovering conserved [29] or dense [4] subnets in PPI networks.

For min-cut graph partitioning, the most commonly applied heuristics are based on starting with a seed partition and repeatedly moving or swapping nodes with maximum gain on the objective function [21]. The key point here is that the move is performed even if it is associated with a negative gain in order to climb over poor local optima. Observe that minimizing the total weight of the cut edges (min-cut) in graph partitioning is equivalent maximizing the total weight of internal edges. This is very similar to the objective function of MAWISH. The difference is that the total weight of only one part is considered in MAWISH, and node balance is not an issue. Therefore, we apply this iterative improvement based heuristic to MAWISH in order to find locally maximal heavy subgraphs. The initial heavy subgraph is constructed by selecting the node with maximum number of matched interactions (*i.e.*, a

conserved hub) and adding all nodes that share a match edge with this node to the subgraph.

A sketch of this iterative improvement based algorithm for finding a single conserved subgraph on the alignment graph is shown in Figure 4. Each pass (*i.e.*, the loop between lines 3-13) of this algorithm works in linear time. In practice, we also limit the number of contiguous moves with negative gain. This allows us to tune the locality of identified patterns.

To find all non-redundant heavy subgraphs, we start with the entire alignment graph and find a maximally heavy subgraph. If this subgraph is statistically significant, we record the alignment that corresponds to this subgraph and mark its nodes. We repeat this process by considering only unmarked nodes. Once a new heavy subgraph is identified, we add the previously marked nodes that are positively connected to this subgraph. This approach allows identification of overlapping alignments while avoiding redundancy. Finally, we rank all subgraphs based on their significance and report the corresponding alignments.

3.4 Statistical Significance

To evaluate the statistical significance of discovered high-scoring alignments, we compare them with a reference model generated by a random source. In the reference model, it is assumed that the interaction networks that belong to the two organisms are independent from each other as well as the protein sequences. To accurately capture the power-law nature of PPI networks, we assume that the interactions are generated randomly from a distribution characterized by a given degree sequence. (Note that the power law nature of the graphs is not critical to our algorithm. The degree distribution can be computed explicitly from the database of interactions). If u and u' are interacting with d_u and $d_{u'}$ proteins, respectively, then the probability $q_{uu'}$ of observing an interaction between u and u' can be estimated as $q_{uu'} = d_u d_{u'} / \sum_{v \in U} d_v$ [7]. We assume that the sequences are generated by a memoryless source, such

that $u \in U$ and $v \in V$ are orthologous with probability p. Similarly, $u, u' \in U$ and $v, v' \in V$ are paralogous with probability p_U and p_V, respectively. Since the similarity function provides a measure of the probability of true homology between a given pair of proteins, we estimate p by $\frac{\sum_{u \in U, v \in V} S(u,v)}{|U||V|}$. Hence, $E[S(u,v)] = p$ for $u \in U, v \in V$. The probabilities of paralogy are estimated similarly.

Recall that the weight of a subgraph of the alignment graph is equal to the score of the corresponding alignment. Hence, in the reference model, the expected value of the score of an alignment induced by $\tilde{\mathbf{V}} \subseteq \mathbf{V}$ is $E[W(\tilde{\mathbf{V}})] = \sum_{\mathbf{v},\mathbf{v}' \in \tilde{\mathbf{V}}} E[w(\mathbf{v}\mathbf{v}')]$, where

$$\begin{aligned} E[w(\mathbf{v}\mathbf{v}')] = \; & \bar{\mu} p^2 q_{uu'} q_{vv'} - \bar{\nu} p^2 (q_{uu'}(1 - q_{vv'}) + (1 - q_{uu'}) q_{vv'}) \\ & + \bar{\delta}(p_U(p_U - \bar{d}) + p_V(p_V - \bar{d})) \end{aligned} \tag{11}$$

is the expected weight of an edge in the alignment graph. With the simplifying assumption of independence of interactions, we have $Var[W(\tilde{\mathbf{V}})] = \sum_{\mathbf{v},\mathbf{v}' \in \tilde{\mathbf{V}}} Var[w(\mathbf{v}\mathbf{v}')]$, enabling us to compute the z-score to evaluate the statistical significance of each discovered high-scoring alignment, under the normal approximation that we assume.

While the approach described above enables quick calculation of significance without repeated simulations or extensive numerical computations, it has a few shortcomings. First, the significance of an identified pattern is estimated for the proteins involved in that conserved subgraph, rather than computing the probability of the existence of the pattern anywhere in the networks. Second, the model does not take into account the variability in the distribution of orthologs. These cause low variability of alignment score in the reference model, leading to overestimated z-scores, since the observed variances in alignment score are fairly high, which indeed is statistically significant. While these shortcomings can be addressed explicitly, the cost associated with the computation of significance scores also increases accordingly.

3.5 Extensions to the Model

The proposed model can be extended to account for data quality as well as algorithm parameters.

3.5.1 Accounting for Experimental Error.

PPI networks obtained from high-throughput screening are prone to errors in terms of both false negatives and positives [33]. While the proposed framework can be used to detect experimental errors through cross-species comparison to a certain extent, experimental noise can also degrade the performance of the alignment algorithm. In other words, mismatches should be penalized for lost interactions during evolution, not for experimental false negatives. To account for such errors while analyzing interaction networks, several methods have been developed to quantify the likelihood of an interaction or complex co-membership between proteins [2, 15, 19]. Given the prior probability distribution for protein interactions and a set of observed interactions, these methods compute the posterior probability of interactions based on Bayesian models. Hence, PPI networks can be modeled by weighted graphs to account for experimental error more accurately.

While the network alignment framework introduced in Section 3.1 assumes that interactions are represented by unweighted edges, it can be easily generalized to a weighted graph model as follows. Assuming that weight ϖ_{uv} represents the posterior probability of interaction between u and v, we can define match score and mismatch penalty in terms of their expected values derived from these posterior probabilities. Therefore, for any $u, u' \in U$ and $v, v' \in V$, we have

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

$$\nu(uu', vv') = \bar{\nu} S(uu', vv') (\varpi_{uu'}(1 - \varpi_{vv'}) + (1 - \varpi_{uu'}) \varpi_{vv'}). \tag{13}$$

Note that match and mismatch sets are not necessarily disjoint here in contrast to the unweighted graph

model, which is a special case of this model.

3.5.2 Tuning Model Components and Parameters.

Contracting Paralogs. An alternate approach for handling duplications is contracting the proteins in the same species that are likely to be in-paralogs. This approach fits into the alignment graph model since in-paralogs are expected to be consistently orthologous to the same set of proteins in the other organism. It also reduces the computational complexity since the number of nodes will be decreased by node contraction and the edges that correspond to duplications will be eliminated. Contraction of nodes is also shown to be effective for multiple alignment of metabolic pathways using graph mining [22]. However, clustering proteins in the same organism to identify in-paralogs requires preprocessing to solve a difficult problem. Clustering algorithms that are specifically designed for this purpose, such as INPARANOID [28] serve as a reliable tool. However, the resulting graphs may produce conservative alignments since the search space is narrowed down by the clustering of proteins [23]. In contrast, accounting for duplications using duplication edges provides more flexibility and uses conservation of interactions as additional information to distinguish in-paralogs from out-paralogs, as discussed above.

Shortest-path mismatch model. In the above discussion, while we consider proteins that are linked by at most $\bar{\Delta}$ interactions as interacting, we do not take into account the distance while penalizing mismatches. We can extend this to a shortest-path mismatch model, defined as follows:

$$\nu(uu', vv') = \bar{\nu} S(uu', vv')(\max\{\Delta_G(u, u'), \Delta_H(v, v')\} - \bar{\Delta}), \tag{14}$$

While this model may improve the alignment algorithm, it is computationally expensive since it requires solution of the all pairs shortest path problem on both PPI networks.

Linear duplication model. The alignment graph model forces each duplicate pair in an alignment to be scored. For example, if an alignment contains n paralogous proteins in one species, $\binom{n}{2}$ duplications are

scored 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 the current model as *quadratic duplication model*, since the number of scored duplications is a quadratic function of number of duplicates. 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 this model, 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 3.3 with a simple modification of the algorithm that dynamically reassigns weights to edges that correspond to duplications.

4 Experimental Results

4.1 Data & Implementation

We implement the proposed algorithms in the C programming language and test on PPI networks that belong to three commonly studied eukaryotic organisms. The source code of the software is available at `http://www.cs.purdue.edu/homes/koyuturk/mawish/` along with detailed alignment results. The interaction data are downloaded from BIND [3] and DIP [40] molecular interaction databases. The statistics for the PPI networks of S. Cerevisiae (yeast), C. Elegans (nematode), and D. Melanogaster (fruit fly) are shown in Table 1.

We align all pairs of these three organisms using a fixed set of parameters to be able to compare the results with each other. We set these parameters conservatively in order to obtain a compact set of illustrative results. For any pair of PPI networks, we set the *E*-value threshold adaptively based on the estimated similarity scores so that the minimum similarity score for any pair of potential orthologs is 0.6.

In other words, two proteins that belong to two different species are considered potentially orthologous only if they have a BLAST *E*-value less than 60% of ortholog pairs in COG. On the other hand, we set $\bar{d} = 0.9$, *i.e.*, two proteins in the same organism are considered potential in-paralogs only if they have BLAST *E*-value less than 90% of protein pairs in this organism that are in the same COG. For potential out-paralogs, we consider protein pairs that have a BLAST *E*-value less than 0.1 but greater than 10% of the ortholog pairs in COG. By setting these cut-off values on similarity score, we only consider the homologous protein pairs that have the highest positive or negative contribution on the alignment score. This eliminates noise to a certain extent while improving the computational efficiency. However, for more detailed analysis and discovery of loosely visible patterns, it may be necessary to relax and set these parameters based on the evolutionary distance between the two organisms being compared.

4.2 Results & Discussion

We perform pairwise alignment of the three PPI networks by tuning the alignment parameters to $\bar{\mu} = 1.0$, $\bar{\nu} = 1.0$, and $\bar{\delta} = 0.1$. Detailed statistics on alignment of the three pairs of eukaryotic PPI networks are shown in Table 2. In this table, we list the number of nodes in the alignment graph, nodes with at least one matched edge, matches, mismatches and duplications in both organisms. The number of matches and the number matched nodes are shown for two values of $\bar{\Delta}$, where only direct interactions $\bar{\Delta} = 1$ and indirect interactions through a single protein $\bar{\Delta} = 2$ are considered as matches. In practice, we eliminate all nodes that do not have any matching interactions from the alignment graph. As evident in the table, this improves the computational performance of the algorithm significantly.

Alignment of S. Cerevisiae PPI network with D. Melanogaster PPI network results in identification of 412 conserved subnets. Eight of the conserved subnets with highest alignment scores are shown in Table 3. Similarly, sample high-scoring conserved subnets identified by the alignment of S. Cerevisiae

vs C. Elegans and C.Elegans vs D. Melanogaster PPI networks are shown in Tables 4 and 5, respectively. In total, 83 conserved subnets are identified on S. Cerevisiae and C. Elegans, and 146 are identified on C. Elegans and D. Melanogaster. For each conserved subnet, we count the biological processes that the proteins in the subnet take part in, according to GO annotations. We identify the biological process that is represented by the largest number of proteins in an organism as the dominant biological process for that organism. The dominant biological processes for the conserved subnets are also shown in the tables. While most of the conserved subnets are dominated by one particular processes and the dominant processes are generally consistent across species, there also exist different processes in different organisms that are mapped to each other by the discovered alignments. This illustrates that the comparative analysis of PPI networks is effective in not only identifying particular functional modules, pathways, and complexes, but also in discovering relationships between different processes in separate organisms and crosstalk between known functional modules and pathways.

A selection of interesting conserved subnets is shown in Figure 5. The alignments in the figure illustrate that the alignment algorithm takes into account the conservation of interactions in addition to sequence similarity while mapping orthologous proteins to each other. In all of the alignments shown in the figure, the interactions of proteins that belong to the same orthologous group are highly conserved, suggesting relatively recent duplications.

Detailed examination of the conserved subnets in S. Cerevisiae and D. Melanogaster shows that many of them do correspond to some functional modules. There are multiple instances of 20S proteosome (10,11). All seven of the alpha subunits in the 20S proteosome, a subcomplex of the 26S proteosome involved in protein degradation, are present in the alignment #10 [11]. In addition, there is a subnet for the proteosome regulatory particle (6,9) as well as one for calcium induced pathways (2). Interestingly, proteins that make up the regulatory particle of the 26S proteosome (Rpt1, Rpt2, Rpt3, Rpt4, Rpt5 and

Rpt6) are also present in the alignment #9 [10]. The method also detected a number of components involved in calcium-dependent stress-activated signaling pathways (Cmd1, Cna1, Cna2 and Cnb1) as well as those associated with budgrowth of yeast (Cmd1, Myo2 and Myo4) in alignment #2 [12]. Many of the subnets found for yeast are overlapping, possibly reflecting the fact that drosophila uses a functional module in various contexts.

In some cases, the self-interaction of a single protein in one organism is aligned with a clique of interactions between its orthologs that are part of a particular module. For example, in alignment #7, five proteosome regulatory particle proteins (Rpt1, Rpt3, Rpt4, Rpt5, Rpt6) are mapped to one protein (Rpt4) in drosophila.

Based on these results, we establish pairwise alignment of PPI networks as a tool for not only identifying conserved modules, but also assessing functional differences and similarities of homologous proteins based on shared and missing interactions. Moreover, alignment results provide a means for discovery of new functional modules in relatively less studied organisms through mapping of functions at a modular level rather than at the level of single protein homologies.

5 Related Work

As partially complete interactomes of several species become available, researchers have explored the problem of identifying conserved topological motifs in different species [24, 39]. 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. A publicly available tool, PathBLAST, adopts the ideas in sequence alignment to PPI networks to discover conserved protein pathways across species [20]. 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. In [26], Pinter et al. align metabolic pathways based on subtree homeomorphism, observing that this model not only leads to tractable solutions, but also can describe the variations in metabolic pathways effectively.

In a recent study, Sharan et al. [29] have proposed probabilistic models and algorithms for identifying conserved complexes in bacteria and yeast through cross-species network comparison. Their approach is similar to the framework proposed here in that they construct an orthology graph with nodes that correspond to pairs of ortholog proteins. The edges of the orthology graph are weighted according to a probabilistic framework that compares null and conserved complex models based on log-likelihood. In contrast to their model, our framework is based on concepts of matches, mismatches and duplications and the edges are weighted in order to reward or penalize these evolutionary events. This allows tuning of the parameters based on relative divergence of the species being compared and interpretation of discovered alignments in terms of evolutionary models. One may therefore conclude that their model is designed to identify conserved complexes while our framework is designed for comparative analysis of PPI networks that belong to two different species. The idea of constructing product graphs by joining orthologous nodes is also applied to the comparative analysis of PPI networks that belong to multiple species [30].

6 Conclusion

This paper presents a framework for local alignment of protein interaction networks. The framework is guided by theoretical models of evolution of these networks. The model is based on discovering sets

of proteins that induce conserved subnets based on scoring match and mismatch of interactions, and duplication of proteins. An implementation of the proposed algorithm reveals that this framework is successful in uncovering conserved substructures in protein interaction data.

References

[1] S. F. Altschul, T. L. Madden, A. A. Schffer, Z. Zhang J. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nuc. Acids Res.*, 25(17):3389–3402, 1997.

[2] S. Ashtana, O. D. King, F. D. Gibbons, and F. P. Roth. Predicting protein complex membership using probabilistic network reliability. *Genome Research*, 14:1170–1175, 2004.

[3] G. D. Bader, I. Donalson, C. Wolting, B. F. Quellette, T. Pawson, and C. W. Hogure. BIND-the Biomolecular Interaction Network Database. *Nuc. Acids Res.*, 29(1):242–245, 2001.

[4] J. S. Bader. Greedily building protein networks with confidence. *Bioinformatics*, 19:1869–1874, 2003.

[5] A. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[6] F. Chung, L. Lu, T. G. Dewey, and D. J. Galas. Duplication models for biological networks. *J Comp. Bio.*, 10(5):677–687, 2003.

[7] F. Chung, L. Lu, and V. Vu. Spectra of random graphs with given expected degrees. *PNAS*, 100(11):6313–6318, 2003.

[8] E. Eisenberg and Y. Levanon. Preferential attachment in the protein network evolution. *Phys. Rev. Let.*, 91(13):138701, 2003.

[9] A. C. Gavin et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868):141–147, 2002.

[10] H. Fu et al. Subunit interaction maps for the regulatory particle of the 26s proteasome and the cop9 signalosome. *EMBO J*, 20(24):7096–7107, 2001.

[11] M. Groll et al. Structure of 20s proteasome from yeast at 2.4 a resolution. *Nature*, 386(6624):463–471, 1997.

[12] M. S. Cyert et al. Genetic analysis of calmodulin and its targets in saccharomyces cerevisiae. *Annu Rev Genet*, 35:647–672, 2001.

[13] Y. Ho et al. Systematic identification of protein complexes in Saccharomyces cerevisae by mass spectrometry. *Nature*, 415:180–183, 2002.

[14] U. Feige, D. Peleg, and G. Kortsarz. The dense k-subgraph problem. *Algorithmica*, 29(3):410–421, 2001.

[15] M. A. Gilchrist, L. A. Salter, and A. Wagner. A statistical framework for combining and interpreting proteomic datasets. *Bioinformatics*, 20(5):689–700, 2003.

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

[17] R. Hassin, S. Rubinstein, and A. Tamir. Approximation algorithms for maximum dispersion. *Oper. Res. Let.*, 21:133–137, 1997.

[18] 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*, 98(8):4569–4574, 2001.

[19] R. Jansen, H. Yu, and D. Greenbaum et al. A bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, 302:449–453, 2003.

[20] 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.*, 32:W83–W88, 2004.

[21] B. W. Kernighan and S. Lin. An efficient heuristic procedure for partitioning graphs. *The Bell System Technical Journal*, 49(2):291–307, 1970.

[22] M. Koyutürk, A. Grama, and W. Szpankowski. An efficient algorithm for detecting frequent subgraphs in biological networks. In *Bioinformatics Suppl. 12th Intl. Conf. Intel. Sys. Mol. Bio. (ISMB'04)*, pages i200–i207, 2004.

[23] M. Koyutürk, A. Grama, and W. Szpankowski. Pairwise local alignment of protein interaction networks guided by models of evolution. In *S. Miyano (Eds.): RECOMB 2005, Lecture Notes in Bioinformatics*, volume 3500, pages 48–65, 2005.

[24] 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*, 101(16):5934–5939, 2004.

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

[26] R. Y. Pinter, O. Rokhlenko, E. Yeger-Lotem, and M. Ziv-Ukelson. Alignment of metabolic pathways. *Bioinformatics*, in press, 2005.

[27] H. Qin, H. H. S. Lu, W. B. Wu, and W. Li. Evolution of the yeast protein interaction network. *PNAS*, 100(22):12820–12824, 2003.

[28] 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.*, 314:1041–1052, 2001.

[29] R. Sharan, T. Ideker, B. P. Kelley, R. Shamir, and R. M. Karp. Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data. In *8th Intl. Conf. Res. Comp. Mol. Bio. (RECOMB'04)*, pages 282–289, 2004.

[30] R. Sharan, S. Suthram, R. .M. Kelley, T. Kuhn, S. McCuine, P. Uetz, T. Sittler, R. M. Karp, and T. Ideker. Conserved patterns of protein interaction in multiple species. *PNAS*, 102(6):1974–1979, 2005.

[31] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *J Mol. Bio.*, 147(1):195–197, 1981.

[32] 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.*, 22:4673–4680, 1994.

[33] B. Titz, M. Schlesner, and P. Uetz. What do we learn from high-throughput protein interaction data? *Exp. Rev. Prot.*, 1(1):111–121, 2004.

[34] Y. Tohsato, H. Matsuda, and A. Hashimoto. A multiple alignment algorithm for metabolic pathway analysis using enzyme hierarchy. In *8th Intl. Conf. Intel. Sys. Mol. Bio. (ISMB'00)*, pages 376–383, 2000.

[35] S. Tornow and H. W. Mewes. Functional modules by relating protein interaction networks and gene expression. *Nuc. Acids Res.*, 31(21):6283–6289, 2003.

[36] A. Vázquez, A. Flammini, A. Maritan, and A. Vespignani. Modeling of protein interaction networkrs. *ComPlexUs*, 1:38–44, 2003.

[37] A. Wagner. The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol. Bio. Evol.*, 18(7):1283–1292, 2001.

[38] A. Wagner. How the global structure of protein interaction networks evolves. *Proc. R. Soc. Lond. Biol. Sci.*, 270(1514):457–466, 2003.

[39] S. Wuchty, Z. N. Oltvai, and A. L. Barabási. Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nature Gen.*, 35(2):176–179, 2003.

[40] I. Xenarios, L. Salwinski, X. J. Duan, P. Higney, S. Kim, and D. Eisenberg. DIP: The Database of Interacting Proteins. A research tool for studying cellular networks of protein interactions. *Nuc. Acids Res.*, 30:303–305, 2002.

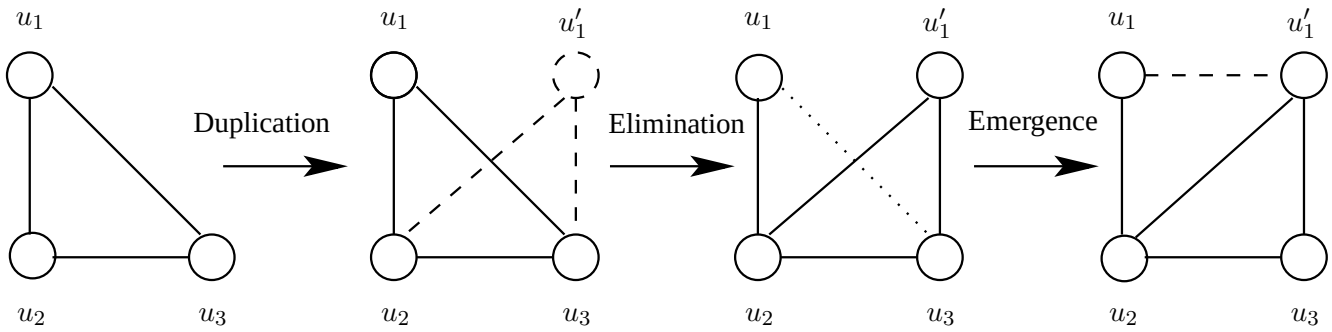

Figure 1: Duplication/divergence model for evolution of PPI networks. Starting with three interactions between three proteins, protein u_1 is duplicated to add u_1' into the network together with its interactions (dashed circle and lines). Then, u_1 loses its interaction with u_3 (dotted line). Finally, an interaction between u_1 and u_1' is added to the network (dashed line).

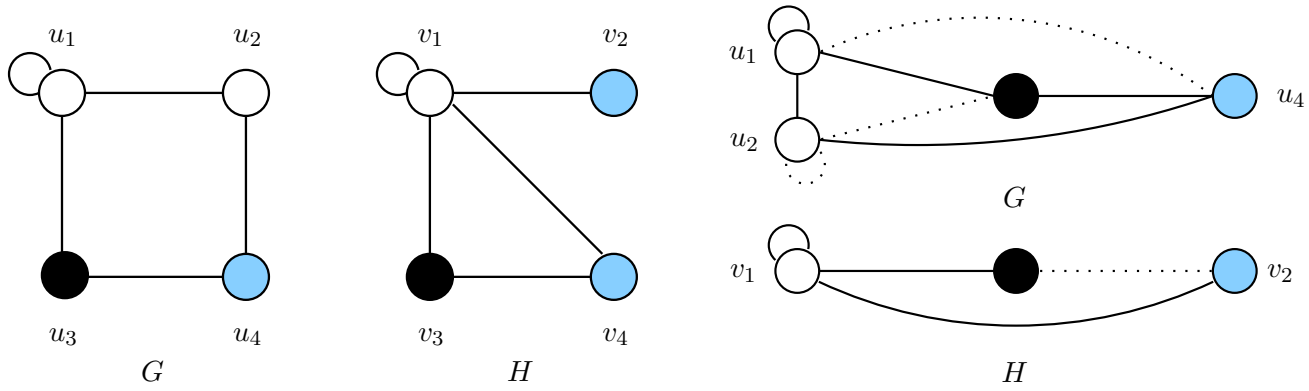

Figure 2: (a) An instance of the pairwise local alignment problem. The proteins that have non-zero similarity scores (*i.e.*, are potentially orthologous), are colored the same. Note that S does not necessarily induce a disjoint grouping of proteins in practice. (b) A local alignment induced by the protein 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.

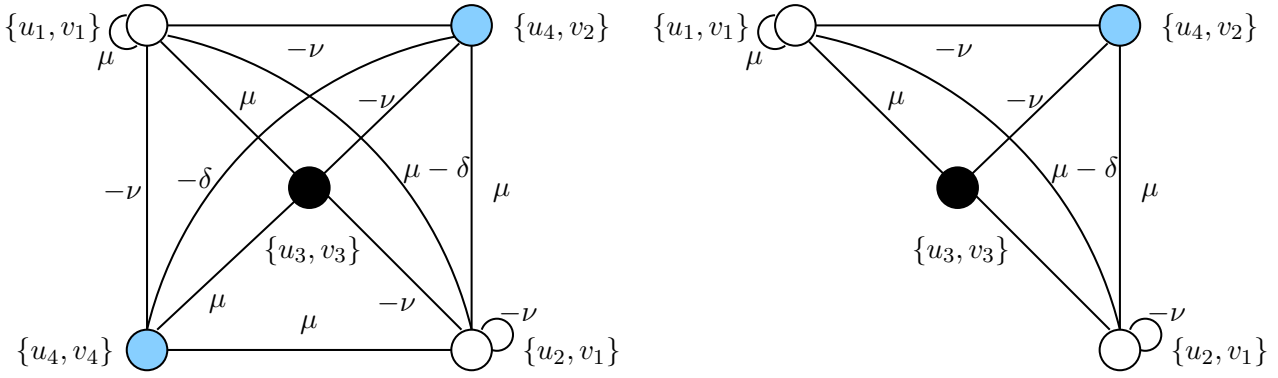

Figure 3: (a) Alignment graph corresponding to the instance of Fig. 2(a). Note that match scores, mismatch and duplication penalties are functions of incident nodes, which is not explicitly shown in the figure for simplicity. (b) Subgraph induced by node set $\tilde{\mathbf{V}} = \{\{u_1, v_1\}, \{u_2, v_1\}, \{u_3, v_3\}, \{u_4, v_2\}\}$, which corresponds to the alignment shown in Fig. 2(b).

procedure HEAVIESTSUBGRAPH($\mathbf{G}$)

▷ **Input** $\mathbf{G}(\mathbf{V}, \mathbf{E}, w)$: Alignment graph

▷ **Output** $\tilde{\mathbf{V}}$: Subset of nodes that induces a maximally heavy subgraph in $\mathbf{G}$

$\tilde{\mathbf{v}} \leftarrow \text{argmax}_{\mathbf{v} \in \mathbf{V}} |\{\mathbf{v}' \in \mathbf{V} : (\mathbf{v}, \mathbf{v}') \text{ is a match edge}\}|$

$\tilde{\mathbf{V}} \leftarrow \{\tilde{\mathbf{v}}\} \cup \{\mathbf{v} \in \mathbf{V} : (\tilde{\mathbf{v}}, \mathbf{v}) \text{ is a match edge}\}$

repeat

  $Q \leftarrow \{\mathbf{v} \in \mathbf{V} : key(\mathbf{v}) = -\sum_{\mathbf{v}' \in \tilde{\mathbf{V}}} w(\mathbf{v}, \mathbf{v}')$ **if** $\mathbf{v} \in \tilde{\mathbf{V}},\ key(\mathbf{v}) = \sum_{\mathbf{v}' \in \tilde{\mathbf{V}}} w(\mathbf{v}, \mathbf{v}')$ **else**$\}$

  $W_{max} \leftarrow W(\tilde{\mathbf{V}})$

  while $Q \neq \emptyset$

    $\mathbf{v} \leftarrow$ EXTRACTMAX(Q)

    if $\mathbf{v} \in \tilde{\mathbf{V}}$ **then** $\tilde{\mathbf{V}} \leftarrow \tilde{\mathbf{V}} \setminus \{\mathbf{v}\}$ **else** $\tilde{\mathbf{V}} \leftarrow \tilde{\mathbf{V}} \cup \{\mathbf{v}\}$

    if $W(\tilde{\mathbf{V}}) > W_{max}$ **then** $W_{max} \leftarrow W(\tilde{\mathbf{V}})$, $bestmove \leftarrow \mathbf{v}$

    for all $\mathbf{v}'$ such that $\mathbf{v}\mathbf{v}' \in \mathbf{E}$ update $key(\mathbf{v}')$

  endwhile

  roll back all moves after $bestmove$

until $bestmove = \texttt{NULL}$

return $\tilde{\mathbf{V}}$

Figure 4: Fast heuristic for finding a subset of nodes that induces a subgraph of maximal total weight on the alignment graph.

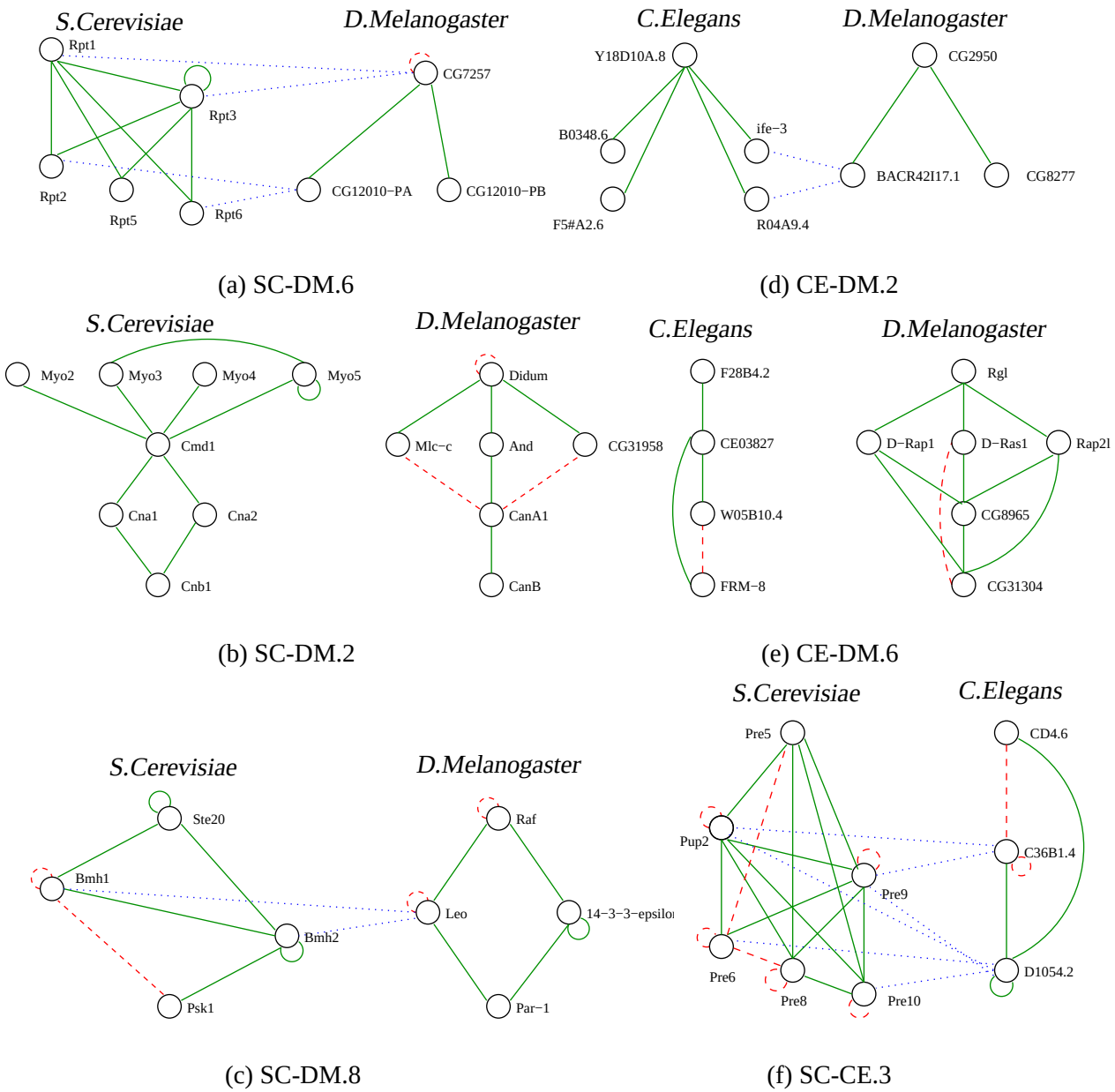

Figure 5: Sample conserved subnets identified by the alignment algorithm. Orthologous and paralogous proteins are either vertically aligned or connected by blue dotted lines. Existing interactions are shown by green solid lines, missing interactions that have an orthologous counterpart are shown by red dashed lines. The rank of each alignment in the set of alignments discovered for the respective pair of organisms is indicated in its label.

Table 1: Description of aligned PPI networks.

Organism	# Proteins	# Interactions
S. Cerevisiae	5157	18192
C. Elegans	3345	5988
D. Melanogaster	8577	28829

Table 2: Alignment statistics for the three pairs of eukaryotic organisms. For each alignment, the number of nodes in alignment graphs (# of orthologous pairs), number of nodes with at least one matched edge, number of matches, number of mismatches and number of duplications for both organisms are shown. Number of mismatches for $\bar{\Delta} = 2$ can be derived from other statistics.

Organism pair	# Nodes	# Matched nodes		# Matches		# Mismatches	# Duplications	
		$\bar{\Delta} = 1$	$\bar{\Delta} = 2$	$\bar{\Delta} = 1$	$\bar{\Delta} = 2$	$\bar{\Delta} = 1$	Org. 1	Org. 2
SC vs CE	2746	312	1230	412	3007	40262	6107	6886
SC vs DM	15884	1730	8622	2061	42781	1054241	6107	32670
CE vs DM	11805	491	3391	455	6626	205593	6886	32670

Table 3: Eight high-scoring conserved subnets identified by the alignment of S. Cerevisiae and D. Melanogaster. For each conserved subnet, its rank (R), score (S), number of nodes in alignment graph and corresponding number of proteins in each organism (#P), number of matches (#M), number of mismatches (#N), and number of duplications in each organism (#D) are shown in the corresponding row. The dominant biological process in which the majority of proteins in the conserved subnet participate is shown for each organism, in the first and second rows, respectively.

R	S	#P	#M	#N	#D	Dominant Process
1	15.97	18 (16, 5)	28	6	(4, 0)	protein amino acid phosphorylation (3)
						JAK-STAT cascade (2)
2	13.93	13 (8, 6)	16	6	(3, 1)	endocytosis (4)
						calcium-mediated signaling (3)
3	12.44	22 (14, 4)	32	10	(3, 0)	protein amino acid phosphorylation (4)
						protein amino acid phosphorylation (2)
6	8.05	8 (5, 3)	12	2	(0, 1)	ubiquitin-dependent protein catabolism (4)
						proteolysis and peptidolysis (1)
7	6.96	5 (5, 1)	10	5	(0, 0)	ubiquitin-dependent protein catabolism (5)
						ubiquitin-dependent protein catabolism (1)
8	6.83	6 (4, 4)	12	6	(0, 1)	pseudohyphal growth (3)
						polarity specification of anterior/posterior axis (1)
9	6.76	8 (6, 3)	16	9	(0, 1)	ubiquitin-dependent protein catabolism (5)
						proteolysis and peptidolysis (1)
10	6.75	10 (7, 3)	24	12	(0, 1)	ubiquitin-dependent protein catabolism (7)
						biological process unknown(2)

Table 4: Five high-scoring conserved subnets identified by the alignment of S. Cerevisiae and C. Elegans.

R	S	#P	#M	#N	#D	Dominant Process
1	36.14	13 (5, 3)	65	24	(0, 3)	ubiquitin-dependent protein catabolism
						protein catabolism
2	8.47	20 (11, 5)	19	4	(1, 1)	protein amino acid phosphorylation (2)
						protein amino acid phosphorylation (2)
3	6.28	8 (6, 3)	21	12	(0, 0)	ubiquitin-dependent protein catabolism (6)
						ubiquitin-dependent protein catabolism (3)
8	3.23	4 (3, 3)	4	1	(1, 1)	mismatch repair (2)
						mismatch repair (1)
15	1.70	3(3, 3)	2	0	(0, 0)	vesicle-mediated transport (2)
						physiological process (2)

Table 5: Five high-scoring conserved subnets identified by the alignment of C. Elegans and D. Melanogaster.

R	S	#P	#M	#N	#D	Dominant Process
1	26.75	17 (4, 9)	52	4	(0, 4)	thermosensory behavior (1)
						regulation of transcription from RNA polymerase II promoter (4)
2	4.65	9 (5, 3)	8	0	(2, 1)	translational initiation (2)
						translational initiation (1)
3	4.57	7 (5, 4)	9	2	(1, 0)	protein amino acid phosphorylation (2)
						protein amino acid phosphorylation (2)
6	4.00	6 (4, 6)	8	2	(0, 2)	signal transduction (2)
						signal transduction (1)
10	3.48	5 (4, 4)	6	3	(1, 0)	regulation of transcription, DNA-dependent (2)
						regulation of transcription from RNA polymerase II promoter (3)