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

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. Recent algorithms on 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 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 PPI networks, 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 PPI networks. 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 PPI networks in this paper, the proposed algorithm can be easily adapted to finding matches for a subnet query in a database of PPI networks.

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, 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]. An important class of molecular interaction data is in the form of protein-protein interaction (PPI) networks, which provide the experimental basis for

* This paper was published in S. Miyano (Eds.): RECOMB 2005, Lecture Notes in Bioinformatics 3500, pp. 48-65, 2005

understanding modular organization of cells, as well as useful information for predicting the biological function of individual proteins [4]. High throughput screening methods such as two-hybrid analysis [5], mass spectrometry [6], and TAP [7] provide large amounts of data on these networks.

As revealed by recent studies, PPI networks evolve at a modular level [8] 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 [9], finding frequently occurring subgraphs in a collection of graphs [10], discovering highly conserved subgraphs in a pair of graphs, and finding good matches for a subgraph in a database of graphs [11]. 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 complexes and 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 [12]. 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 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.

2 Related Work

As the amount of cell signaling data increases rapidly, there have been various efforts aimed at developing methods for comparative network analysis. In a relatively early study, Dandekar et al. [13] 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 [8, 14]. 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 [11]. 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.

Since the local alignment of PPI networks for patterns in the form of general graphs leads to computationally intractable problems, tools based on simplified models are generally useful. However, as functional modules and protein complexes are likely to be conserved across species [8], algorithms for aligning general graphs are required for understanding conservation of such functional units. In a recent study, Sharan et al. [15] have proposed probabilistic models and algorithms for identifying conserved modules and complexes through cross-species network comparison. Similar to their approach, we develop a framework for aligning PPI networks to discover subsets of proteins in each species such that the subgraphs induced by these sets are highly conserved. In contrast to existing methods, our framework relies on theoretical models that focus on understanding the evolution of protein interaction networks.

3 Theoretical Models for Evolution of PPI Networks

There have been a number of studies aimed at understanding the general structure of PPI networks. It has been shown that these networks are power-law graphs, *i.e.*, the relative frequency of proteins that interact with k proteins is proportional to $k^{-\gamma}$, where γ is a network-specific parameter [16]. In order to explain this power-law nature, Barábasi and Albert have proposed [16] 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 [17] 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 [18].

A common model of evolution that explains preferential attachment and power-law nature of PPI networks is the duplication/divergence model that is based on gene duplications [12, 19–21]. 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 into divergence of duplicated (paralogous) proteins in the interactome through deletion and insertion of interactions. Deletion of an interaction in a PPI network implies the elimination of an existing interaction between two proteins due to structural and/or functional changes. Similarly, insertion of an interaction into a PPI network 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,

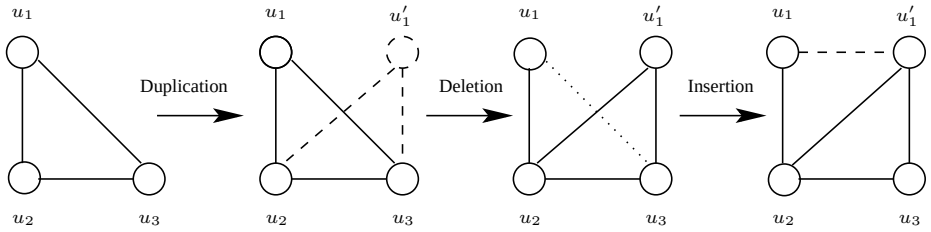

Fig. 1. Duplication/divergence model for evolution of PPI networks. Starting with three interactions between three proteins, protein u_1 is duplicated to add u_1' into the network together with its interactions (dashed circle and lines). Then, u_1 loses its interaction with u_3 (dotted line). Finally,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

it is said to be correlated; otherwise, it is uncorrelated [19]. Since newly duplicated proteins are more tolerant to interaction loss because of redundancy, correlated deletions are generally more probable than insertions and uncorrelated deletions [12]. 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 [20]. It is also theoretically shown that network growth models based on node duplications generate power-law distributions [22].

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.

4 Pairwise Local Alignment of PPI Networks

In light of the theoretical models of evolution of PPI networks, we develop a generic framework for the comparison of PPI networks 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.

4.1 The Pairwise Local 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 if they belong to different species and paralogous if they belong to the same species. 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.*, each protein is expected to have only a few potential orthologs or paralogs. We discuss the reliability of possible choices for assessing protein similarity in detail in Section 4.4.

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 penalty, since duplicated proteins tend to diverge in terms of their interaction profiles in the long term [20]. 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 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.

Definition 1. Local Alignment of PPI networks. *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*

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

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

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

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:

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

We aim to find local alignments with locally maximal score (drawing an analogy to sequence alignment [23], *high-scoring subgraph pair*s). This definition of the local alignment problem provides a general framework for the comparison of PPI networks, 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 4.4.

A sample instance of the pairwise local alignment problem is shown in Figure 2(a). Consider the alignment induced by the protein subset pair $\tilde{U} = \{u_1, u_2, u_3, u_4\}$ and $\tilde{V} = \{v_1, v_2, v_3\}$, shown in Figure 2(b). 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. Moreover, 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 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 is 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 of correctly matching proteins to 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.

4.2 Alignment Graphs and the Maximum-Weight Induced Subgraph Problem

It is possible to collect information on matches and mismatches between two PPI networks into a single alignment graph by computing 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 an alignment graph for this purpose.

Definition 2. 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:*

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

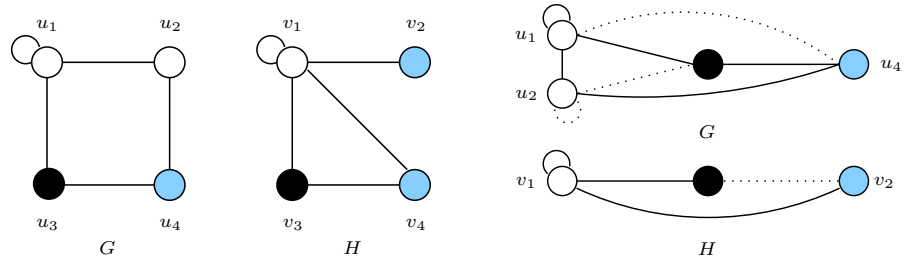

Fig. 2. (a) An instance of the pairwise local alignment problem. The proteins that have non-zero similarity scores (*i.e.,* are potentially orthologous), are colored the same. Note that S does not necessarily induce a disjoint grouping of proteins in practice. (b) A local alignment induced by the protein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

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*

$$w(\mathbf{vv'}) = \mu(uu', vv') - \nu(uu', vv') - \delta(u, u') - \delta(v, v'). \tag{6}$$

Here, $\mu(uu', vv') = 0$ *if* $(uu', vv') \notin \mathcal{M}$*, and similarly for mismatch and duplication penalties.*

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 put 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.

The weighted alignment graph is conceptually similar to the orthology graph of Sharan et al. [15]. However, instead of accounting for similarity of proteins through node weights, we encapsulate the orthology information in edge weights, which also allows consideration of duplications effectively. This construction of the alignment graph allows us to formulate the alignment problem as a graph optimization problem defined below.

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.,* $W(\tilde{\mathbf{V}}) = \sum_{\mathbf{v}, \mathbf{v'} \in \tilde{\mathbf{V}}} w(\mathbf{vv'}) \geq \epsilon$*.*

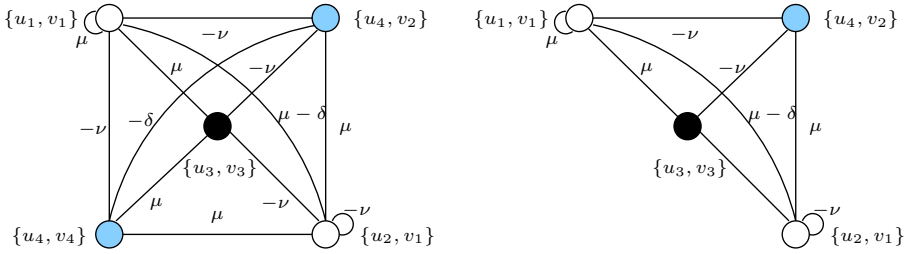

Fig. 3. (a) Alignment graph corresponding to the instance of Fig. 2(a). Note that match scores, mismatch and duplication penalties are functions of incident nodes, which is not explicitly shown in the figure for simplicity. (b) Subgraph induced by node set $\tilde{\mathbf{V}} = \{\{u_1, v_1\}, \{u_2,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

Not surprisingly, this problem is equivalent to the local alignment of PPI networks 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 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 [24] and maximum dispersion problems [25] that are also NP-complete. Although the positive weight restriction on these problems limits the application of existing algorithms to the maximum weight induced subgraph problem, the nature of the conservation of PPI networks makes a simple greedy heuristic quite effective for the local alignment of PPI networks.

4.3 A Greedy Heuristic for Local Alignment of Protein Interaction Networks

In terms of protein interactions, functional modules and protein complexes 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 either directly or through a common module hub, while it is only loosely connected to the rest of the network [26]. 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 [8], 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

procedure GREEDYMAWISH($\mathbf{G}$)
▷ **Input** $\mathbf{G}(\mathbf{V}, \mathbf{E}, w)$: Alignment graph
▷ **Input** ϵ: Threshold on subgraph weight
▷ **Output** $\tilde{\mathbf{V}}$: Subset of selected nodes
▷ $g(\mathbf{v})$: Gain of adding $\mathbf{v}$ into $\tilde{\mathbf{V}}$
▷ W: Total subgraph weight
1 **for** each $\mathbf{v} \in \mathbf{V}$ **do**
2 $g(\mathbf{v}) \leftarrow w(\mathbf{vv})$
3 $w(\mathbf{v}) = \sum_{\mathbf{vv}' \in \mathbf{E}} w(\mathbf{vv}')$
4 $\tilde{\mathbf{V}} \leftarrow \emptyset$, $W \leftarrow 0$
5 $\tilde{\mathbf{v}} \leftarrow \text{argmax}_{\mathbf{v} \in \mathbf{V}} w(\mathbf{v})$
6 **repeat**
7 $\tilde{\mathbf{V}} \leftarrow \tilde{\mathbf{V}} \cup \{\tilde{\mathbf{v}}\}$
8 $W \leftarrow W + g(\tilde{\mathbf{v}})$
9 **for** each $\mathbf{v} \in (\mathbf{V} \setminus \tilde{\mathbf{V}})$ s.t.$\tilde{\mathbf{v}}\mathbf{v} \in \mathbf{E}$ **do**
10 $g(\mathbf{v}) \leftarrow g(\mathbf{v}) + w(\tilde{\mathbf{v}}\mathbf{v})$
11 $\tilde{\mathbf{v}} \leftarrow \text{argmax}_{\mathbf{v} \in (\mathbf{V} \setminus \tilde{\mathbf{V}})} g(\mathbf{v})$
12 **until** $g(\mathbf{v}) \leq 0$
13 **if** $W > 0$
14 **then return** $\tilde{\mathbf{V}}$
15 **else return** $\emptyset$

Fig. 4. Greedy algorithm for finding a set of nodes that induces a subgraph of maximal total weight on the alignment graph.

conserved 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 PPI networks. Indeed, similar approaches are shown to perform well in discovering conserved or dense subnets in PPI networks [15, 27]. 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, likely being part of a functional module.

A sketch of the greedy algorithm for finding a single conserved subgraph on the alignment graph is shown in Figure 4. This algorithm grows a subgraph, which is of locally maximal total weight. To find all non-redundant "good" alignments, we start with the entire alignment graph and find a maximal subgraph. If this subgraph is statistically significant according to the reference model described in Section 4.5, we record the alignment that corresponds to this subgraph and mark its nodes. We repeat this process by allowing only unmarked nodes to be chosen as seed until no subgraph with positive weight can be found. Restricting the seed to only non-aligned nodes avoids redundancy while allowing discovery of overlapping alignments. Finally, we rank all subgraphs based on their significance and report the corresponding alignments. A loose bound on the worst-case running time of this algorithm is $O(|\mathbf{V}||\mathbf{E}|)$, since each alignment takes $O(|\mathbf{E}|)$ time and each node can be the seed at most once. Assuming that the number of orthologs for each protein is bounded by a constant, the size of the alignment graph is linear in the total size of the input networks.

4.4 Selection of Model Components

In order for the discovered PPI network 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.

Similarity Function. Since proteins that are involved in a common functional module, or more generally, proteins that interact with each other, show local sequence similarities, care must be taken while employing pairwise sequence alignment as a measure of potential orthology between proteins. Furthermore, while aligning two PPI networks and interpreting the alignment, only duplications that correspond to 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 [28]. Unfortunately, distinguishing between in-paralogs and out-paralogs is not trivial. Therefore, we assign similarity scores to protein pairs conservatively by detecting orthologs and in-paralogs using a separate algorithm, INPARANOID [28], which is developed for finding disjoint ortholog clusters in two species. Each ortholog cluster discovered by this 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. We define the similarity between two proteins u and v as

$$S(u,v) = confidence(u) \times confidence(v). \tag{7}$$

This provides a normalized similarity function that takes values in the interval $[0,1]$ and quantifies the confidence in the two proteins being orthologous or paralogous.

Scores and Penalties. *Match score.* A match is scored positively in an alignment to reward a conserved interaction. Therefore, the score represents 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 interaction to obtain:

$$\mu(uu', vv') = \bar{\mu} S(uu', vv'), \tag{8}$$

where $S(uu', vv') = \min\{S(u,v), S(u',v')\}$ and $\bar{\mu}$ is a pre-determined parameter specifying the relative weight of a match in the total alignment score. While we use this definition of $S(uu', vv')$ in our implementation, $S(u,v) \times S(u',v')$ provides a reliable measure of similarity between the two protein pairs.

Mismatch penalty. Similar to match score, mismatch penalty is defined as:

$$\nu(uu', vv') = \bar{\nu} S(uu', vv'), \tag{9}$$

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.

Duplication penalty. Duplications are penalized to account for the divergence of the proteins after duplication. Sequence similarity provides a crude approximation to the age of duplication and likelihood of being paralogs [21]. Hence, duplication penalty is defined as:

$$\delta(u,u') = \bar{\delta}(d - S(u,u')), \tag{10}$$

where $\bar{\delta}$ is the relative weight of a duplication and $d \geq \max_{u,u' \in U} S(u,u')$ is a parameter that determines the extent of penalizing duplications. Considering the similarity function of (7), setting $d = 1.0$ results in no penalty for duplicates that are paralogous to the main ortholog with 100% confidence.

4.5 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 species 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. The probability $q_{uu'}$ of observing an interaction between two proteins $u, u' \in U$ for the degree sequence derived from G can be estimated by a Monte Carlo algorithm that repeatedly swaps the incident nodes of randomly chosen edges [15]. On the other hand, 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 of (7) 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.

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

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

is the expected weight of an edge in the alignment graph. Moreover, with the simplifying assumption of independence between 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 to hold.

4.6 Extensions to the Model

Accounting for Experimental Error. PPI networks obtained from high-throughput screening are prone to errors in terms of both false negatives and positives [4]. 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 [29–31]. Given the prior probability distribution for protein interactions and 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 4.1 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 indeed a special case of this model.

Tuning Model Components and Parameters. *Sequence similarity.* A more flexible approach for assessing similarity between proteins is direct employment of sequence alignment scores. In PathBLAST [32], the similarity between two proteins is defined as the log-likelihood ratio for homology, *i.e.*, $S(u, v) = \log(p(u, v)/\bar{p})$, where $p(u, v)$ is the probability of true homology between u and v given the BLAST E value of their alignment and $\bar{p}$ is the expected value of p over all proteins in the PPI networks being aligned. To avoid consideration of similarities that do not result from orthology, it is necessary to set cut-off values on the significance of alignments [32, 20].

Shortest-path mismatch model. Since proteins that are linked by a short alternative path are more likely to tolerate losing their interaction, mismatch penalty can be improved using a shortest-path mismatch model, defined as:

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

where $\Delta(u, u')$ is the length of the shortest path between proteins u and u'. While this model is likely to 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 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 the current model 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 this model, each duplicate protein is penalized only once, based on its similarity with the paralog that is most similar to itself.

5 Experimental Results

In this section, we present local alignment results to illustrate the effectiveness of the proposed framework and the underlying algorithm on interaction data retrieved from the

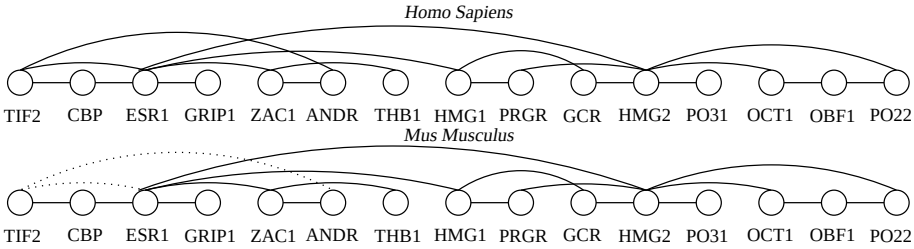

Fig. 5. A conserved subnet that is part of DNA-dependent transcription regulation in human and mouse PPI networks. Ortholog proteins are vertically aligned. Existing interactions are shown by solid edges, missing interactions that have an existing orthologous counterpart in the other species are shown by dotted edges.

DIP protein interaction database [33]. We align the PPI networks of two mammalians that are available in the database; *Homo sapiens* (Hsapi) and *Mus musculus* (Mmusc). As of October 2004, the Hsapi PPI network contains 1369 interactions among 1065 proteins while Mmusc PPI network 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. The alignment graph contains 305 matched interactions, 205 mismatched interactions in Hsapi, 149 mismatched interactions in Mmusc, 536 duplications in Hsapi, and 384 duplications in Mmusc. We then compute local alignments using the algorithm of Section 4.3 on this graph. By trying alternate settings for the relative weights of match score and mismatch, duplication penalties, we identify 54 non-redundant alignments, 15 of which contain at least 3 proteins on each network. Note that construction of alignment graph and discovery of local alignments on this graph takes only a few milliseconds.

A conserved subnet of DNA-dependent transcription regulation that is found to be statistically significant (z-score=18.1) is shown in Figure 5. 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 [34]. 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 significantly conserved (z-score=19.9) in human and mouse PPI networks is shown in Figure 6. This subnet contains 8 matching interactions among 10 proteins. It is composed of two separate subnets that are connected through the interaction of

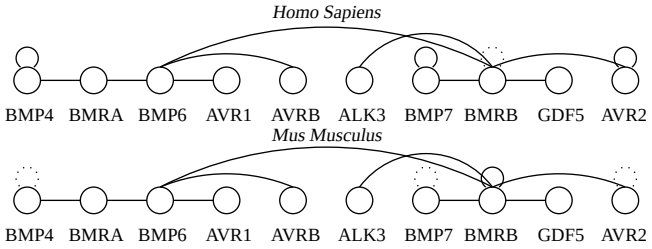

Fig. 6. A conserved subnet that is part of transforming growth factor beta receptor signaling pathway in human and mouse PPI networks.

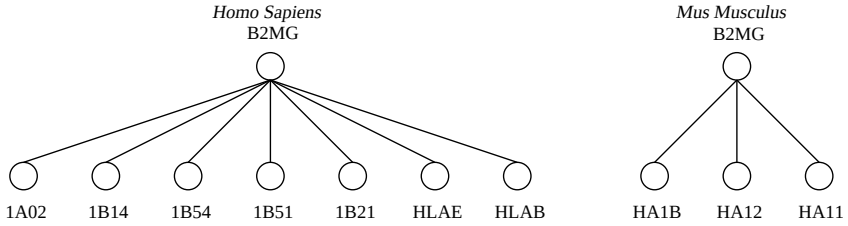

Fig. 7. A conserved subnet that is part of antigen presentation and antigen processing in human and mouse PPI networks. Homologous proteins are horizontally aligned. Paralogous proteins in a species are shown from left to right in the order of confidence in being orthologous to the respective proteins in the other species.

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

As an example for duplications, a subnet that is part of antigen presentation and antigen processing, which is significantly conserved (z-score=456.5) in human and mouse PPI networks is shown in Figure 7. 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, this subnet can only be discovered when the duplication coefficient is small ($\bar{\delta} \leq 0.12\bar{\mu}$).

6 Concluding Remarks and Ongoing Work

This paper presents a framework for local alignment of protein interaction networks that is guided by theoretical models of 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. Furthermore, we are working on a rigorous analysis of distribution of the alignment score, which will enable more reliable assessment of statistical significance. Once these enhancements are completed, the proposed framework will be established as a tool for pairwise alignment of PPI networks, 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 PPI networks. Using this tool researchers will be able to find conserved counterparts of newly discovered complexes or modules in several species.

Acknowledgments

This research was supported in part by NIH Grant R01 GM068959-01 and NSF Grant CCR-0208709. The authors would like to thank Prof. Shankar Subramaniam of UCSD for many valuable discussions.

References

1. Altschul, S.F., Madden, T.L., Schffer, A.A., J. Zhang, Z.Z., Miller, W., Lipman, D.J.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nuc. Acids Res. **25** (1997) 3389–3402
2. Thompson, J.D., Higgins, D.G., Gibson, T.J.: 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** (1994) 4673–4680
3. Hartwell, L.H., Hopfield, J.J., Leibler, S., Murray, A.W.: From molecular to modular cell biology. Nature **402** (1999) C47–C51
4. Titz, B., Schlesner, M., Uetz, P.: What do we learn from high-throughput protein interaction data? Exp. Rev. Prot. **1** (2004) 111–121
5. Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., Sakaki, Y.: A comprehensive two-hybrid analysis to explore the yeast protein interactome. PNAS **98** (2001) 4569–4574
6. Ho, Y. et al.: Systematic identification of protein complexes in Saccharomyces cerevisae by mass spectrometry. Nature **415** (2002) 180–183
7. Gavin, A.C. et al.: Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature **415** (2002) 141–147
8. Wuchty, S., Oltvai, Z.N., Barabási, A.L.: Evolutionary conservation of motif constituents in the yeast protein interaction network. Nature Gen. **35** (2003) 176–179

9. Tohsato, Y., Matsuda, H., Hashimoto, A.: A multiple alignment algorithm for metabolic pathway analysis using enzyme hierarchy. In: 8th Intl. Conf. Intel. Sys. Mol. Bio. (ISMB'00). (2000) 376–383
10. Koyutürk, M., Grama, A., Szpankowski, W.: An efficient algorithm for detecting frequent subgraphs in biological networks. In: Bioinformatics Suppl. 12th Intl. Conf. Intel. Sys. Mol. Bio. (ISMB'04). (2004) i200–i207
11. Kelley, B.P., Yuan, B., Lewitter, F., Sharan, R., Stockwell, B.R., Ideker, T.: PathBLAST: a tool for aligment of protein interaction networks. Nuc. Acids Res. **32** (2004) W83–W88
12. Vázquez, A., Flammini, A., Maritan, A., Vespignani, A.: Modeling of protein interaction netwokrs. ComPlexUs **1** (2003) 38–44
13. Dandekar, T., Schuster, S., Snel, B., Huynen, M., Bork, P.: Pathway alignment: application to the comparative analysis of glycolytic enzymes. Biochem. J **343** (1999) 115–124
14. .Lotem, E.Y., Sattath, S., Kashtan, N., Itzkovitz, S., Milo, R., Pinter, R.Y., Alon, U., Margalit, H.: Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction. PNAS **101** (2004) 5934–5939
15. Sharan, R., Ideker, T., Kelley, B.P., Shamir, R., Karp, R.M.: Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data. In: 8th Intl. Conf. Res. Comp. Mol. Bio. (RECOMB'04). (2004) 282–289
16. Barabási, A., Albert, R.: Emergence of scaling in random networks. Science **286** (1999) 509–512
17. Eisenberg, E., Levanon, Y.: Preferential attachment in the protein network evolution. Phys. Rev. Let. **91** (2003) 138701
18. Qin, H., Lu, H.H.S., Wu, W.B., Li, W.: Evolution of the yeast protein interaction network. PNAS **100** (2003) 12820–12824
19. Pastor-Satorras, R., Smith, E., Solé, R.V.: Evolving protein interaction networks through gene duplication. J Theo. Bio. **222** (2003) 199–210
20. Wagner, A.: The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. Mol. Bio. Evol. **18** (2001) 1283–1292
21. Wagner, A.: How the global structure of protein interaction networks evolves. Proc. R. Soc. Lond. Biol. Sci. **270** (2003) 457–466
22. Chung, F., Lu, L., Dewey, T.G., Galas, D.J.: Duplication models for biological networks. J Comp. Bio. **10** (2003) 677–687
23. Smith, T.F., Waterman, M.S.: Identification of common molecular subsequences. J Mol. Bio. **147** (1981) 195–197
24. Feige, U., Peleg, D., Kortsarz, G.: The dense k-subgraph problem. Algorithmica **29** (2001) 410–421
25. Hassin, R., Rubinstein, S., Tamir, A.: Approximation algorithms for maximum dispersion. Oper. Res. Let. **21** (1997) 133–137
26. Tornow, S., Mewes, H.W.: Functional modules by relating protein interaction networks and gene expression. Nuc. Acids Res. **31** (2003) 6283–6289
27. Bader, J.S.: Greedily building protein networks with confidence. Bioinformatics **19** (2003) 1869–1874
28. Remm, M., Storm, C.E.V., Sonnhammer, E.L.L.: Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J Mol. Bio. **314** (2001) 1041–1052
29. Jansen, R., Yu, H., et al., D.G.: A bayesian networks approach for predicting protein-protein interactions from genomic data. Science **302** (2003) 449–453
30. Ashtana, S., King, O.D., Gibbons, F.D., Roth, F.P.: Predicting protein complex membership using probabilistic network reliability. Genome Research **14** (2004) 1170–1175
31. Gilchrist, M.A., Salter, L.A., Wagner, A.: A statistical framework for combining and interpreting proteomic datasets. Bioinformatics **20** (2003) 689–700

32. Kelley, B.P., Sharan, R., Karp, R.M., Sittler, T., Root, D.E., Stockwell, B.R., Ideker, T.: Conserved pathways withing bacteria and yeast as revealed by global protein network alignment. PNAS **100** (2003) 11394–11399
33. Xenarios, I., Salwinski, L., Duan, X.J., Higney, P., Kim, S., Eisenberg, D.: DIP: The Database of Interacting Proteins. A research tool for studying cellular networks of protein interactions. Nuc. Acids Res. **30** (2002) 303–305
34. Boonyaratanakornkit, V. 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. **18** (1998) 4471–4488