Kazemi *et al. BMC Bioinformatics* (2016) 17:527
DOI 10.1186/s12859-016-1395-9

BMC Bioinformatics

RESEARCH ARTICLE Open Access

PROPER: global protein interaction network alignment through percolation matching

Ehsan Kazemi[1*], Hamed Hassani[2], Matthias Grossglauser[1] and Hassan Pezeshgi Modarres[3]

Abstract

Background: The alignment of protein-protein interaction (PPI) networks enables us to uncover the relationships between different species, which leads to a deeper understanding of biological systems. Network alignment can be used to transfer biological knowledge between species. Although different PPI-network alignment algorithms were introduced during the last decade, developing an accurate and scalable algorithm that can find alignments with high biological and structural similarities among PPI networks is still challenging.

Results: In this paper, we introduce a new global network alignment algorithm for PPI networks called PROPER. Compared to other global network alignment methods, our algorithm shows higher accuracy and speed over real PPI datasets and synthetic networks. We show that the PROPER algorithm can detect large portions of conserved biological pathways between species. Also, using a simple parsimonious evolutionary model, we explain why PROPER performs well based on several different comparison criteria.

Conclusions: We highlight that PROPER has high potential in further applications such as detecting biological pathways, finding protein complexes and PPI prediction. The PROPER algorithm is available at http://proper.epfl.ch.

Keywords: Global network alignment, Protein-protein interaction, Percolation graph matching, Biological network

Background

Proteins are large biomolecules that carry out vital functions in living cells. They typically carry out these functions in concert with other biomolecules, especially other proteins, which enables their diverse functionality, such as forming signaling networks and metabolic pathways, and regulating enzymatic activities [1]. In this context, the term protein-protein interaction (PPI) stands for the mutual interactions between pairs of proteins.

PPI data are obtained by high-throughput experimental techniques such as yeast 2-hybrid [2], synthetic lethality [3] and co-immunoprecipitation coupled mass spectrometry [4]. The data are deposited in more than 100 PPI databases [5] such as BioGRID [6], the Molecular Interaction Database (MINT) [7], the Human Protein Reference Database (HPRD) [8], and IntAct [9]. Despite the large amount of PPI data, the detection of the protein pathways and protein complexes is challenging because many of the PPIs are noisy and non-reproducible.

A comparative analysis of PPI networks provides insight into species evolution and information about evolutionarily conserved biological interactions, such as pathways across multiple species [1, 10–12]. Network alignment (also known as graph matching or network reconciliation in the computer science literature) algorithms were introduced to compare PPI networks between two or more species.

The comparison of PPI networks, by network alignment, shows that there are identical interaction patterns between proteins with high sequence similarity across different species [13]. For example, there are many common protein interactions between proteins in yeast networks and their corresponding protein orthologs in PPI networks of worms [14]. Because functional interactions are often conserved across species and false positives are unlikely to occur in multiple species, network alignment

*Correspondence: ehsan.kazemi@epfl.ch
[1]School of Computer and Communication Sciences, EPFL, Lausanne, Switzerland
Full list of author information is available at the end of the article

can increase the confidence level of an observed interaction in a database [15].

PPI-network alignment has many applications in areas such as detection of new pathways and of conserved motifs, prediction of the functions of proteins, orthology detection, drug design, protein-protein interaction prediction and phylogenetic tree reconstruction [16, 17].

Generally, PPI-network alignment methods assume that two functional ortholog proteins on two different PPI networks are likely to interact with proteins in the corresponding networks that are functionally orthologs themselves [1, 18, 19]. Following this line of thought, local network alignment (LNA) and global network alignment (GNA) methods are the main approaches for aligning PPI networks [1, 20, 21]. The LNA algorithms search for small but highly conserved subnetworks (e.g., homologous regions of biological pathways or protein complexes) between species, whereas GNA algorithms try to align all (or most of) the proteins to find large subgraphs that are functionally and structurally conserved over all the nodes in the two networks [1, 20, 21].

PPI-network alignment algorithms use topological (e.g., local and global network structures) and biological (e.g., amino acid sequences of proteins) information to align two networks. The topological information is more important than sequence information for aligning functionally conserved interactions [22, 23], hence the focus of network alignment algorithms shifted from using only biological information towards using topological information. Most of the early works on PPI-network alignment, such as PathBLAST [24], NetworkBLAST [10], NetAlign [25], MaWISh [26] and Grælin [27], study the LNA problem. More recent methods, such as IsoRank [28, 29], the GRAAL family [16, 23, 30–32], MAGNA and its successor MAGNA++ [33, 34], SPINAL [35], PINALOG [36], Netcoffee [37] and BEAMS [38], are examples of GNA algorithms.

In this paper, we consider the GNA of only two networks. Singh et al. [28] introduced IsoRank as the first GNA algorithm for PPI networks. The IsoRank algorithm is formulated as an eigenvalue problem, where it first computes a pairwise protein similarity metric (as a convex combination of protein sequence similarities and a structural similarity score), and then generates the final global alignment between the two networks based on this metric. The authors of [39, 40] developed approximation algorithms for efficient computation of the IsoRank similarities. GHOST [41] aligns two networks according to the similarity of spectral signatures of node couples. PINALOG [36] finds the final alignment by matching the communities of the two networks first. The GRAAL (GRAph ALigner) family is a group of GNA methods that use the graphlet-degree signature similarity to align two networks. GRAAL [30] is the first GNA algorithm that uses only structure of the two networks for alignment. It first selects a couple of nodes with high graphlet-degree signature similarity, and then by a seed-and-extend matching procedure it tries to expand the alignment around this couple in a greedy way. In general, a seed-and-extend algorithm starts the alignment procedure from a set of highly similar couples called seed pairs. Then, it proceeds to align iteratively similar couples among neighbors of the seed pairs. H-GRAAL [31] uses the Hungarian algorithm for improving the quality of alignments produced by GRAAL, at the cost of increased computational complexity. To align two networks, MI-GRAAL [16] integrates several metrics such as graphlet-degree signature similarity, local clustering coefficient differences, degree differences and protein sequence similarity. L-GRAAL [23] is the latest algorithm from the GRAAL family; it directly optimizes both the structural and sequence similarities with a heuristic seed-and-extend strategy based on a Lagrangian relaxation. The SPINAL algorithm [35] iteratively grows an alignment based on an a priori computed coarse-grained node similarity score. By using a genetic algorithm, MAGNA [33] tries to optimize the edge conservation between two networks.

In this paper, we design a new global pairwise-network alignment algorithm for PPI networks; it is built upon previous results for graph matching in computer science. We show the excellent performance of our algorithm (in terms of both accuracy and speed) compared to several state-of-the-art algorithms. We also introduce a new measure for evaluating the performance of algorithms in aligning biological pathways between species. We argue the suitability of our algorithm by analyzing its performance in a bigraph-sampling model of network evolution [42–44]. For this random-bigraph model, we use the results of [43, 45] to guarantee the performance of our algorithm.

Methods

GNA algorithms, by finding a one-to-one mapping of proteins, try to find large conserved sub-networks (as they are indicative of a common ancestor) and network motif[1] between several species [46]. Global pairwise-network alignment algorithms align proteins of only two species in order to maximize the biological and topological similarities (these concepts are defined precisely later in the text) between aligned proteins; they have been extensively studied in the literature [20, 21, 28, 35, 36].

A PPI network can be represented by a graph $G(V, E)$, where V is the set of proteins and each edge (u, v) in E is an indicator of interaction between the two proteins u and v. Formally speaking, given two networks $G_1(V_1, E_1)$ and $G_2(V_2, E_2)$, the purpose of global network alignment is to identify a bijection between the full (or partial) vertex

sets of two networks. The network alignment algorithms use the protein similarities and the network topology. In this study, the pairwise similarities between proteins are computed by the well-known basic local-alignment search tool (BLAST) [47] that considers the alignment of amino acid sequences of those proteins.

The PROPER algorithm

Network alignment has been an active area of research in the computer science literature. In the context of network analysis, the main goal is to align social graphs from different domains [48, 49]. In computer vision and pattern recognition, graph matching is used to find similar images in a database [50–52].

In many scenarios, there is some side information that could be used in the process of network alignment. For example, it is possible to use information from a small fraction of individuals who elect to reveal their identities in two social networks. Alternatively, some users link their accounts across multiple networks. This set of revealed identities or linked nodes (henceforth called a *seed set*) enables us to initiate an iterative procedure for matching the two graphs. In this regard, one main category of network alignment algorithms, known as *percolation graph matching* (PGM), assumes that there is side information in the form of a seed set of pre-matched node couples. In this class of algorithms, the alignment starts with a small set $\mathcal{A}$ of initial seed couples and percolates to other node couples, i.e., they build the alignment incrementally between nodes of the two networks [45, 48, 53–55].

We use the ideas from PGM algorithms (mainly [45]) to design our PROPER (PROtein-protein interaction network alignment based on PERcolatin) algorithm.

PROPER: two steps

In the process of PPI-network alignment by PROPER, initially we have as inputs two PPI networks $G_1(V_1, E_1)$ and $G_2(V_2, E_2)$, the set of pairwise BLAST bit-score similarities (call it $\mathcal{S}$) for couples of proteins in $V_1 \times V_2$, and fixed thresholds $\ell, r > 0$, where ℓ and r are the sequence similarity and the local topological similarity thresholds, respectively.

The PROPER algorithm uses the sequence similarities and network structures in a two-stage procedure: (i) At the first step, it uses the sequence similarities to generate a seed set for a PGM algorithm; and (ii) at the second step, to align remaining couples, it uses only the network structure and the seeds generated from the first step as inputs to the PGM algorithm. This is in contrast with many other pairwise-alignment algorithms, where they try to simultaneously maximize a function of both sequence and structural similarities. In this section, we first explain the process of generating seed set $\mathcal{A}$ from $\mathcal{S}$ (the SeedGeneration algorithm). Then, we explain how to align new couples, starting from the set $\mathcal{A}$ (the MapPercolation algorithm).

Initial seeds play an important role in the alignment process. In the PPI setting, the BLAST bit-score is often a good indicator of functional similarities between proteins [56]. In other words, at high levels of sequence similarity it is possible to make a functional inference with an acceptable accuracy [57]. This means that, for couples of proteins with a high sequence similarity it is very likely that they have similar functions. The main approach in this paper is to use such couples as a starting point to find a global alignment. Indeed, the seeds to the PROPER algorithm are those couples of proteins with high sequence similarities. Also, a protein can be aligned with at most one protein from the other species. The degree of similarity between the couples in the seed set $\mathcal{A}$ is controlled by the threshold ℓ.

The seed set $\mathcal{A}$ is generated from the pairwise similarities (the set $\mathcal{S}$) in the following manner: Among all the couples of proteins with BLAST bit-score similarity above ℓ, couples $[i, j]$ are matched in a descending order of sequence similarity, unless i or j is matched already. More precisely, (i) we add the couple $[i, j] \in \mathcal{S}$ with the highest similarity to the seed set and match i to j; (ii) all the couples $[i, j']$ and $[i', j]$ are now forbidden and we remove them from $\mathcal{S}$. We repeat the steps (i) and (ii) until there is no remaining couple in the set $\mathcal{S}$ with BLAST bit-score similarity at least ℓ. Note that, in the process of seed generation, when there are several couples with the same sequence similarity, we randomly pick one of them.

Algorithm 1 describes the SeedGeneration algorithm in detail. In this algorithm, for a set of couples $\mathcal{A}$, $V_1(\mathcal{A})$ defines the set of nodes from network G_1 in $\mathcal{A}$, i.e., $V_1(\mathcal{A}) = \{i | \exists j \text{ s.t. } [i, j] \in \mathcal{A} \text{ for some} j\}$. We define $V_2(\mathcal{A})$ similarly. Also, $BlastBit(i, j)$ denotes the BLAST bit-score similarity between two proteins i and j.

A priori, the probability of biological similarity of a protein couple decreases with a decrease in the sequence similarity. Therefore, there is a trade-off between the number of protein couples with the same biological functions and the accuracy (i.e., the ratio of couples with the same functions over the size of seed set) based on ℓ. Clearly, choosing a high value for ℓ aligns proteins that, with a high probability, have similar functions. However, this can result in removing couples with lower sequence similarities, but the same functions from the initial seed-set.

The second step of PROPER (the MapPercolation algorithm) starts the alignment process from the seed couples (set $\mathcal{A}$) obtained from the set of pairwise similarities $\mathcal{S}$ (see the SeedGeneration algorithm). It then incrementally generates the set π of matched couples among $V_1 \times V_2 \setminus \mathcal{A}$. In the MapPercolation step, the PROPER algorithm relies only on the structure of $G_{1,2}$, and it does not use the

Algorithm 1: The SeedGeneration Algorithm

Input: BLAST bit-score similarities $\mathcal{S}$ and ℓ
Output: The seed set $\mathcal{A}$

$\mathcal{A} \leftarrow \emptyset$;
for *all couples* $[i,j] \in \mathcal{S}$ *from the highest similarity to the lowest* **do**
  if $i \notin V_1(\mathcal{A})$, $j \notin V_2(\mathcal{A})$ *and* $BlastBit(i,j) \geq \ell$ **then**
    add the couple $[i,j]$ to $\mathcal{A}$;
  end
end
return $\mathcal{A}$;

sequence similarities. In this regard, the seed couples are added to the set of aligned couples π. Then, at each time-step, the goal of the PGM algorithm is to add a new couple to the set π so that structural similarity is maximized.

In the process of the MapPercolation algorithm, we look at the neighboring couples of the previously matched couples. We say a couple of proteins $[i',j'] \in V_1 \times V_2$ is a neighbor of another couple $[i,j]$ if and only if $(i,i') \in E_1$ and $(j,j') \in E_2$. Indeed, the evidence for deciding which couple to match (called the score of a couple) is the number of common neighbors each couple has in the set of currently aligned couples. To achieve the maximum structural similarity, our algorithm chooses the next couple in a greedy way: it chooses the couple with the maximum number of common neighbors $score_{max}$ (provided there are at least r) in π and permanently aligns them. Therefore the number of conserved interactions by adding the couple with the highest score is $score_{max}$. Note that a new couple of proteins can be matched if its score is at least r.

When there are several couples with the maximum score, we tie-break by the minimum degree-difference in the two networks, i.e., we choose the couple $[i,j]$ with the minimum $|d_{1,i} - d_{2,j}|$, where $d_{1,i}$ and $d_{2,j}$ denote the degrees of nodes i and j in the networks G_1 and G_2, respectively. The proteins with closer number of interactions (i.e., closer degrees) have more structural similarity. If there is more than one couple with the minimum degree difference, we choose the couple with the minimum $d_{1,i} + d_{2,j}$. The couple with the minimum $d_{1,i} + d_{2,j}$ minimizes the number of mismatched interactions. Finally, if there are still several candidate couples, we randomly pick one of them. The process of alignment continues to the point where there is no remaining unmatched couple of proteins (we say a couple $[i,j]$ is unmatched if $i \notin V_1(\pi)$ and $j \notin V_2(\pi)$) with at least r common neighbors, in the current set of aligned proteins. Note that for a given value of r, only nodes with degree at least r can get enough score to be matched. More precisely, MapPercolation is not able to align: (i) unmatched nodes with a degree less than r, and (ii) couples that have not gained enough scores. Figure 1 presents an example of the second step of PROPER (the MapPercolation algorithm). Algorithm 2 describes this algorithm.

Algorithm 2: The MapPercolation algorithm

Input: $G_1(V_1,E_1)$, $G_2(V_2,E_2)$, seed set $\mathcal{A}$ and threshold r
Output: The set of aligned couples π

$\pi \leftarrow \mathcal{A}$;
while *there exists an unmatched couple with score at least* r **do**
  among all the couples with the highest score select the unmatched couple $[i,j]$ with the minimum $|d_{1,i} - d_{2,j}|$. If there is more than one couple with the minimum $|d_{1,i} - d_{2,j}|$, select the couple with the minimum $d_{1,i} + d_{2,j}$. Finally, if there are still several candidates, randomly pick one of them;
  add $[i,j]$ to the set π;
end
return π;

Performance measures

In this section, we explain the measures used for comparing alignment algorithms. As there is no single standard measure for evaluating the quality of alignments, we use several existing measures [20, 21, 46]. In addition, we introduce a new measure for comparison based

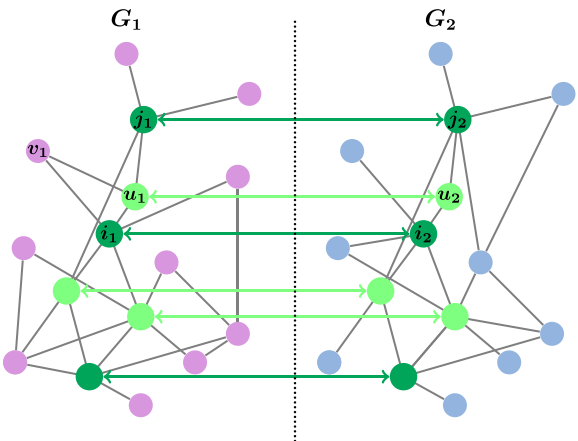

Fig. 1 *Dark-green nodes* correspond the initial seed-set. Couples $[i_1,i_2]$, $[i_1,j_2]$, $[j_1,j_2]$, $[j_1,i_2]$, $[v_1,i_2]$, $[v_1,j_2]$ are neighboring couples of the couple $[u_1,u_2]$. The couples $[i_1,i_2]$ and $[j_1,j_2]$ are the common neighbors of the couple $[u_1,u_1]$ in the set of already matched couples π, i.e., the score of couple $[u_1,u_2]$ is two. *Light-green nodes* are the nodes that are matched after the first three steps of the MapPercolation algorithm

on the performance of algorithms in aligning biological pathways.

For better illustration, in this section we assume that, without loss of generality, G_2 has at least as many nodes as G_1, i.e., $|V_1| \leq |V_2|$. Let π denote the mapping produced by an alignment algorithm. Also, let $G[U]$ denote the induced subgraph of G on a set of vertices U. Assume π maps the nodes $V_1' \subset V_1$ to the nodes $V_2' \subset V_2$, i.e., $V_2' = \pi\left(V_1'\right)$. Note that many global alignment algorithms do not match all the nodes from graph G_1 to a node from graph G_2, i.e., they align a large fraction of the nodes but not all of them. We define graph $G_0(V_0, E_0)$ as the intersection of the two graphs G_1 and G_2 under the alignment π, i.e., V_0 is the set of proteins in graph G_1 aligned by π to a protein in graph G_2; and E_0 is the set of interactions in G_1, conserved under the alignment π in the graph G_2. Formally, we have $V_0 = V_1'$ and $E_0 = E_{G_1[V_1']} \cap \pi^{-1}(E_{G_2[V_2']})$.

Structural and functional similarity measures

In this section, we review the measures that are used widely to evaluate the performance of network alignment algorithms.

(i) Node correctness (NC) of an alignment is defined as the ratio of the number of correctly aligned couples to the number of nodes in the smaller network (i.e., $|V_1|$) [30]. The precision is defined as the ratio of number of correctly aligned couples to the total number of couples $|\pi|$ in the alignment π. These measures are applicable only to synthetic networks, because they can be used only for alignments with known ground-truth [21].

As the true alignment between the proteins of two species is not known completely for real networks, it is not possible to directly calculate the NC and precision of an alignment [20, 21, 46]. To compare the performance of algorithms over real datasets, two different types of measures were introduced in the literature. The first group of measures uses the topological similarity of aligned networks. The second group measures the quality of an alignment by using other biological information.

The following measures are used for evaluating the topological similarity of aligned networks.

(ii) The number of conserved interactions under the alignment π (call it Δ_π) is one of the measures used to evaluate the quality of algorithms based on the topological similarity [1]. Formally,

$$\Delta_\pi = |\pi(E_1) \cap E_2|.$$

(iii) Edge correctness (EC) is a measure of topological similarity among the aligned networks [30]. EC computes the ratio of edges from graph G_1, i.e., all the edges in the smaller network, which are conserved under the alignment π. Formally,

$$\mathrm{EC} = \frac{|\pi(E_1) \cap E_2|}{|E_1|}.$$

(iv) Recall that the numbers of proteins (nodes) in the two networks are not equal. Therefore, one drawback of the EC measure is that aligning sparse regions of G_1 with dense regions of G_2 can result in high values of EC. The induced conserved-structure score (*ICS*) measures the structural similarity of aligned networks by penalizing dense regions of G_2 [41]. The ICS score for an alignment π from graph G_1 with graph G_2 is

$$ICS = \frac{|\pi(E_1) \cap E_2|}{|E_{G_2[\pi(V_1)]}|}.$$

(v) The symmetric substructure score (S^3) is defined with respect to both $G_{1,2}$ networks [33]. The S^3 measure penalizes the alignments that map sparse regions of one network to denser regions of the other network. Formally, S^3 is defined as follows.

$$S^3 = \frac{|\pi(E_1) \cap E_2|}{|E_1| + |E_{G_2[\pi(V_1)]}| - |\pi(E_1) \cap E_2|}.$$

Note that $|E_1|$ refers to all the edges in the smaller network.

(vi) The largest connected shared-component (LCSC) is the largest connected subgraph of G_1, which is found to also exist in G_2, i.e., the largest connected component in graph G_0 [46]. Let $|LCSC|$ denote the number of nodes in LCSC. Also, the share of nodes in LCSC is defined as $\frac{|LCSC|}{|V_1|}$ [16].

We now introduce the second group of measures that are used for evaluating the biological quality of alignments.

(vii) The gene-ontology consistency (GOC) score measures the functional similarity of aligned proteins. Note that usually more than one gene ontology (GO) terms are assigned to a protein [58]. Also, as the GO datasets are noisy and proteins have diverse functions, it is possible that true ortholog proteins do not have exactly the same set of GO terms. GOC for an aligned couple of proteins $u \in V_1$ and $v \in V_2$ is defined as the Jaccard similarity coefficient between the GO terms of the two proteins [35]. Formally, it is defined as

$$GOC(u, v) = \frac{|GO(u) \cap GO(v)|}{|GO(u) \cup GO(v)|},$$

where $GO(u)$ denotes the set of GO terms associated with the protein u. $GOC(\pi)$ is calculated by summation over the GOC terms of all the aligned couples in π:

$$GOC(\pi) = \sum_{u \in V_1} GOC(u, \pi(u)). \quad (1)$$

For ease of notation we refer to $GOC(\pi)$ as GOC score.

(viii) To compare algorithms based on the sequence similarities of aligned proteins, we use a slightly modified version of the average normalized bit–score (ANBS) measure

proposed in [59]. ANBS for two graphs $G_1(V_1, E_1)$ and $G_2(V_2, E_2)$ under the alignment π is defined as follows.

$$ANBS(\pi) = |V_1|^{-1} \sum_{i \in V_1(\pi)} \frac{BlastBit(i, \pi(i))}{\sqrt{BlastBit(i,i)BlastBit(\pi(i),\pi(i))}}.$$

Pathway comparison measures

In order to evaluate the performance of algorithms in aligning biological pathways, we introduce a new measure in this section. This measure captures the quality of alignments based on a higher level of functional and structural similarities (beyond the introduced measures such as the similarity of GO terms and the number of conserved interactions).

It is known that there are many biological pathways with similar functions in different species [12]. The KEGG PATHWAY database [60] provides a set of experimentally found biological pathways. In this database, a pathway is called by the name of a species (e.g., hsa for Homo sapiens), followed by a number. The pathways with the same number have the same function in different species. For example, hsa03040, mmu03040, dme03040 and sce03040 are in Homo sapiens (human), Mus musculus (mouse), Drosophila melanogaster (fruit fly) and Saccharomyces cerevisiae (budding yeast), respectively. These pathways have the same functions.[2] Assume $PW_{i,1}$ denotes the set of proteins from a pathway with number i in the PPI network of the first species (i.e., G_1). Similarly, we define $PW_{i,2}$. For pathway i, $\Delta_{\pi,i}$ denotes the number of conserved interactions between the proteins in this pathway under the alignment π, i.e., $\Delta_{\pi,i} = E_{G_1[PW_{i,1}]} \cap \pi^{-1}(E_{G_2[PW_{i,2}]})$. Note that we are looking for pathways that are present in both aligned species.

We say a protein u from a pathway is aligned correctly, if it is mapped to a protein v from a pathway with the same function. For pathway i, we define the number of correctly mapped proteins as $|PW_{i,1} \cap \pi^{-1}(PW_{i,2})|$. This measure corresponds to the number of proteins that, from pathway i in the first species, are mapped to a protein from the same pathway in the second species. For pathway i, we define the accuracy as

$$acc_{\pi,i} = \frac{2|PW_{i,1} \cap \pi^{-1}(PW_{i,2})|}{|PW_{i,1}| + |PW_{i,2}|}. \quad (2)$$

This measure corresponds to the fraction of correctly mapped proteins in pathway i.

We conjecture that a good alignment algorithm should align proteins from pathways with the same functions across species, and many interactions among these proteins are conserved. To quantify this expectation, we set a threshold over the structural similarity of aligned pathways to consider them as a correct alignment. We say that an alignment π successfully aligns a pathway i, if there are at least δ conserved interactions under the alignment π for proteins in that pathway, i.e., if $\Delta_{\pi,i} \geq \delta$. This thresholding guarantees that the structural similarity of aligned pathways are more than a minimum value (here, δ conserved interactions). To evaluate the performance of an algorithm based on this thresholding criterion, we define a set of measures as follows.

1. We consider pathways with at least δ (say $\delta \geq 2$) interactions in each of the species. Let "#PW$_\delta$" denote the number of such pathways.
2. Alignment π successfully aligns pathway i, if $\Delta_{\pi,i} \geq \delta$. The variable "#FPW$_\delta$" refers to the number of successfully aligned pathways. We define the recall as
$$recall_{\pi,\delta} = \frac{\#\text{FPW}_\delta}{\#\text{PW}_\delta}. \quad (3)$$
3. Again, for a correctly aligned pathway i, we define $acc_{\pi,\delta,i}$ similar to (2).

The averages over all i of all the $acc_{\pi,i}$ and $acc_{\pi,\delta,i}$ values are represented by $\overline{acc_\pi}$ and $\overline{acc_{\pi,\delta}}$, respectively. Figure 2 provides a toy example of how to calculate the pathway alignment measures.

Results

In this section, we compare PROPER with the main state-of-the-art network alignment algorithms, specifically (i) with L-GRAAL as the most recent member of GRAAL family that takes into account both sequence and structural similarities [23]; (ii) with MAGNA++ that tries to maximize one of the EC, ICS or S^3 measures [33, 34] (In our experiments we run MAGNA++ in two different

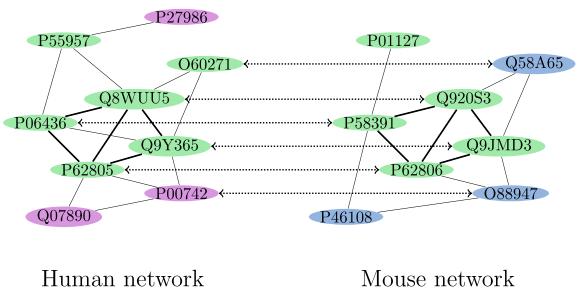

Fig. 2 In this figure, two example PPI networks are given. *Green nodes* are proteins which are in the same pathway (i.e., a pathway with the same number in both species). *Dotted lines* represent the alignment π between these two networks. Under this alignment, there are five conserved interactions between proteins in this pathway (shown by *thick black edges* in each network). Also, the number of correctly mapped proteins is four. Therefore, the accuracy of aligning this pathway is $acc_{\pi,i} = \frac{2\times4}{6+5}$, where there are six and five proteins from this pathway in each species, respectively

modes of maximizing S^3, which is the superior mode for MAGNA++ [33], and EC); (iii) with IsoRank [28] as one of the first global PPI-network alignment algorithms; (iv) with PINALOG [36]; and (v) with SPINAL I and II [35] as their performances are reported to be among the best alignment-algorithms [46]. Table 1 provides an overview of the arguments and parameters of the algorithms used in our comparisons. Note that it is recommended to use SPINAL and MAGNA++ in modes I and S^3, respectively. Also, the recommended settings for IsoRank is $\alpha = 0.6$. For the other algorithms, no default setting is provided. We evaluate the performance of PROPER with $r = 1$ and different values of ℓ.

All the algorithms use two sets of data as input: (i) the PPI networks of two species, and (ii) the pairwise BLAST similarities (in form of BLAST bit-score) between proteins from the first species and proteins from the second species. We use two different PPI-network databases for our comparisons. The first one is from IntAct molecular interaction database [9, 61]. This database enables us to compare algorithms based on large and more recent PPI networks. The GO annotation terms are extracted from the Gene Ontology Annotation (UniProt-GOA) Database [62, 63]. For pathway comparisons over these networks we can use data from [60]. The second database is Isobase [64], a common dataset used in comparison of recent algorithms [20, 46]. The results for experiments over Isobase dataset are provided in Additional file 1. For further evaluations, we use synthetic networks with a known ground-truth.

Structural and functional based comparisons

Table 2 provides a brief description of the PPI networks for five major eukaryotic species, namely C. elegans (ce), D. melanogaster (dm), H. sapiens (hs), M. musculus (mm) and S. cerevisiae (sc); they are extracted from the IntAct database [9, 61]: The last column of Table 2 shows the number of pathways of each species from KEGG PATHWAY database [60]. The amino-acid sequences of proteins for each species are extracted in the FASTA format from UniProt database [65, 66]. The BLAST bit-score similarities [47] are calculated using these amino acid sequences.

Table 2 PPI networks of five major eukaryotic species from IntAct molecular interaction database [9, 61]

Species	#nodes	#edges	Avg. deg.	#pathways
C. elegans	4950	11550	4.67	117
D. melanogaster	8532	26289	6.16	127
H. sapiens	19141	83312	8.71	288
M. musculus	10765	22345	4.15	284
S. cerevisiae	6283	76497	24.35	98

Figure 3 compares algorithms based on the average ICS versus average GOC score for all the possible 10 pairwise alignments between the species. We observe that PROPER outperforms the other algorithms in both measures, i.e., the PROPER algorithm finds alignments with higher functional (GOC score) and structural (ICS) similarities. For the detailed comparisons of the algorithms refer to Figs. 6, 7, 8 and Additional files 1 and 2.

Note that many of the GO annotations are based on only sequence similarities, and these annotations could increase the GOC scores artificially. Clark and Kalita [46] (similar to [35]) propose to also compare algorithms by using only the experimentally verified GO terms (along the comparisons based on all the GO terms) to eliminate the effects of sequence similarities in the GOC evaluations. For this reason, in our next experiment, we consider only GO terms with codes "EXP", "IDA", "IMP", "IGI", "IEP" and "IPI" (the codes for experimental GO terms), and we exclude the annotations derived from computational methods. Figure 4 compares the GOC (based on experimentally verified GO terms) versus EC score. The result of this experiment confirms the superiority of PROPER over the other algorithms.

Figure 5 evaluates the performance of algorithms based on S^3 (for structural similarity) and ANBS (for functional similarity) measures. Again, the PROPER algorithm performs the best based on the two measures, simultaneously.

Table 3 reports the average number of aligned couples and the average of share of nodes in LCSC. We observe that MAGNA++ and IsoRank find, irrespective of the similarity of networks, alignments with full coverages, i.e., the

Table 1 Algorithms and their parameters

Algorithm	Commandline arguments	Parameters
IsoRank [28]	–K 50 –thresh 1e-5 –alpha α –maxveclen 1000000	$\alpha \in \{0.3, 0.5, 0.6, 0.7\}$
PINALOG [36, 77]	do not require arguments	none
L-GRAAL [23]	–a α –l 50	$\alpha \in \{0.3, 0.5, 0.7\}$
MAGNA++(S^3) [34]	–m S3 –p 1000 –n 15000 –f 5 –a α –t 16	$\alpha \in \{0.3, 0.5, 0.7\}$
MAGNA++(EC) [34]	–m EC –p 1000 –n 15000 –f 5 –a α –t 16	$\alpha \in \{0.3, 0.5, 0.7\}$
SPINAL I [35]	–mode -I –alpha α	$\alpha \in \{0.3, 0.5, 0.7\}$
SPINAL II [35]	–mode -II –alpha α	$\alpha \in \{0.3, 0.5, 0.7\}$

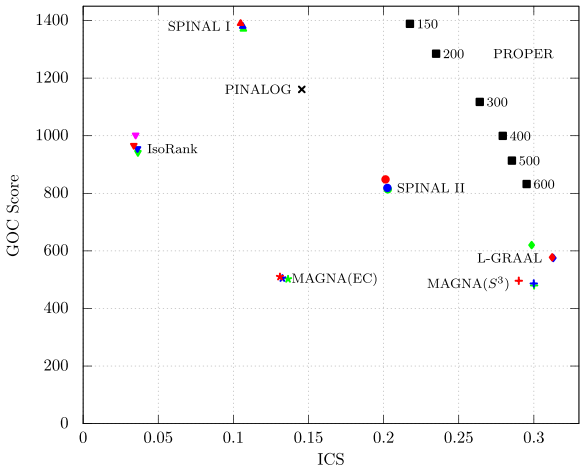

Fig. 3 Comparison of different global network aligners based on the average GO consistency vs. average integrated conserved structure score. For the PROPER algorithm, we set $r = 1$ and each point corresponds to a different value of ℓ. Also, the *red, blue, magenta and green points* correspond to the parameters $\alpha = 0.3, 0.5, 0.6$ and 0.7, respectively

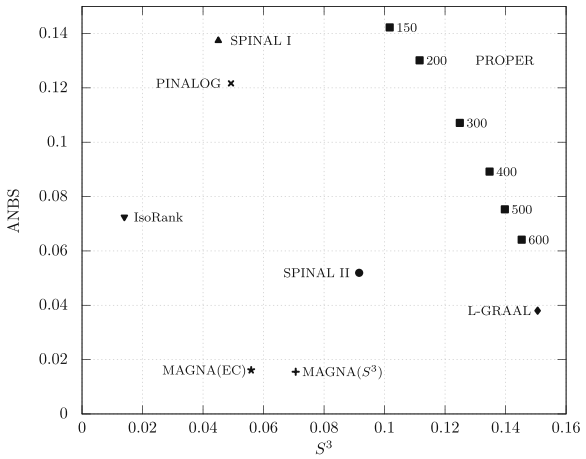

Fig. 5 Comparison of different global network aligners based on the average ANBS vs. average S^3 score. For the PROPER algorithm, we set $r = 1$ and each point corresponds to a different value of ℓ. The parameter α is 0.7

size of their alignments is equal to the number of nodes in the smaller network; and PINALOG has the lowest coverage among the algorithms. The size of an alignment alone is not a good indicator of its quality, because an algorithm with a large coverage might find alignments with low functional-similarities and structural-similarities. Instead, we can consider the sum of functional similarities of aligned proteins. To address this point, for example, GOC score (1) captures the total functional similarity, by summation over all the couples in π (see Figs. 3 and 4). We can also consider the size of shared structure between networks. To address this second point, we use LCSC. A larger LCSC implies that we have found a larger amount of shared structure between the two PPI networks [16]. From Table 3, we observe that PROPER, L-GRAAL and SPINAL II outperform the other algorithms (with huge margins), based on the share of nodes in LCSC.

Figure 6 provides a detailed comparison between the algorithms based on their performance in aligning H. sapiens with S. cerevisiae. Also, detailed comparisons between C. elegans and D. melanogaster, and M. musculus and S. cerevisiae are provided in Figs. 7 and 8, respectively. Note that in Figs. 6, 7 and 8, the values for each measure are normalized to the highest value, i.e., for each measure, in

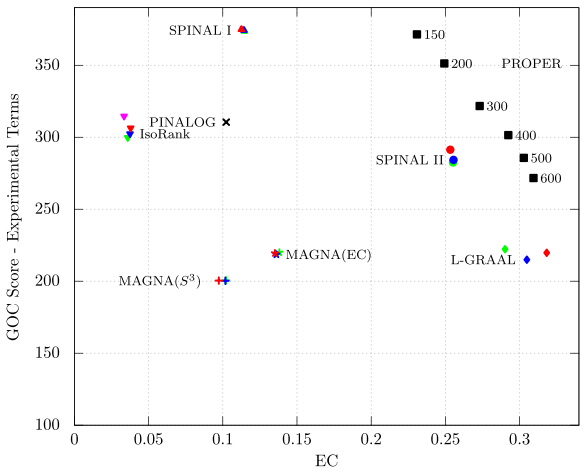

Fig. 4 Comparison of different global network aligners based on the average GO consistency (by considering only experimentally verified GO terms) vs. EC score. For the PROPER algorithm, we set $r = 1$ and each point corresponds to a different value of ℓ. Also, the *red, blue, magenta and green points* correspond to the parameters $\alpha = 0.3, 0.5, 0.6$ and 0.7, respectively

Table 3 The average number of aligned couples (i.e., $|\pi|$) and the average of share of nodes in LCSC (i.e., $|LCSC|/|V_1|$). We use $\alpha = 0.7$ for SPINAL, IsoRank, MAGNA and L-GRAAL, and $r = 1$ for PROPER

Algorithms	$\lvert\pi\rvert$	$\lvert LCSC\rvert/\lvert V_1\rvert$
PROPER ($\ell = 150$)	5521.2	0.528
PROPER ($\ell = 600$)	5320.5	**0.728**
SPINAL I	6364.3	0.219
SPINAL II	6433.4	0.720
PINALOG	3740.9	0.233
L-GRAAL	5616.4	0.726
MAGNA++(S^3)	**6647.8**	0.292
MAGNA++(EC)	**6647.8**	0.353
IsoRank	**6647.8**	0.051

The best value for each column is highlighted in boldface

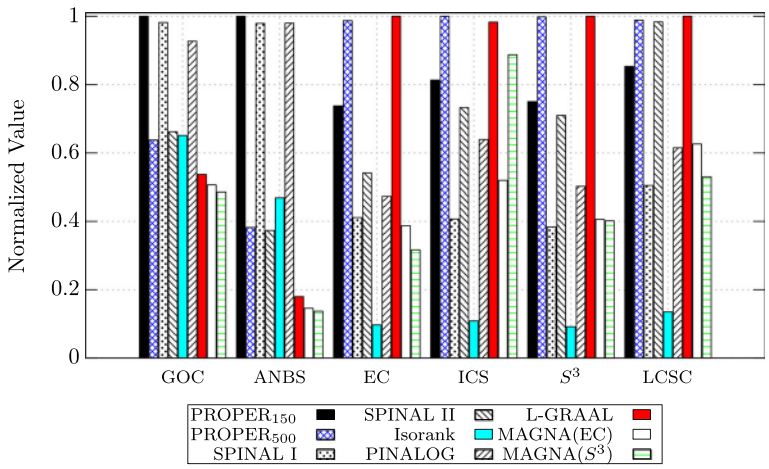

Fig. 6 Comparison of different global network aligners on aligning H. sapiens and S. cerevisiae based on six different measures. For the PROPER algorithm, we set $r = 1$ and $\ell \in \{150, 500\}$. The parameter α is 0.7

these figures, the maximum is 1 for the best algorithm and values for the other algorithms are normalized with respect to the maximum. We observe that PROPER outperforms the other algorithms in terms of most of GOC, ANBS, ICS, S^3, EC and LCSC measures.

The MapPercolation algorithm and *r*

The PROPER algorithm has two main steps: (i) SeedGeneration and (ii) MapPercolation. The numbers of aligned couples in the first and second steps depend on ℓ and r, respectively. In Table 4, we report the average number of aligned couples (i.e., $|\pi|$) in the first and second steps of PROPER for different values of ℓ and $r \in \{1, 2\}$. We observe that by increasing the value of ℓ, the number of aligned couples in the first step decreases. This is because the number of couples with BLAST bit-score of at least ℓ has an inverse relationship with ℓ. In the second step, $|\pi|$ increases by a factor of 2.5 to 7.6 for $\ell \in \{150, 200, 300, 400, 500, 600\}$ with $r = 1$. For the detailed experimental result of PROPER with $r \in \{1, 2\}$ refer to Table S1 in Additional file 2.

Choosing smaller values of r reduces the required structural similarity for aligning a couple. This explains why the number of aligned couples for $r = 1$ is larger than for $r = 2$ in Table 4. Note that the MapPercolation algorithm, for a given value of r, cannot align nodes with degrees less than r. From Fig. 9, which reports the degree distribution of different networks, we observe that there are many nodes with degree one, e.g, almost half of nodes for C. elegans and M. musculus. These nodes of degree one

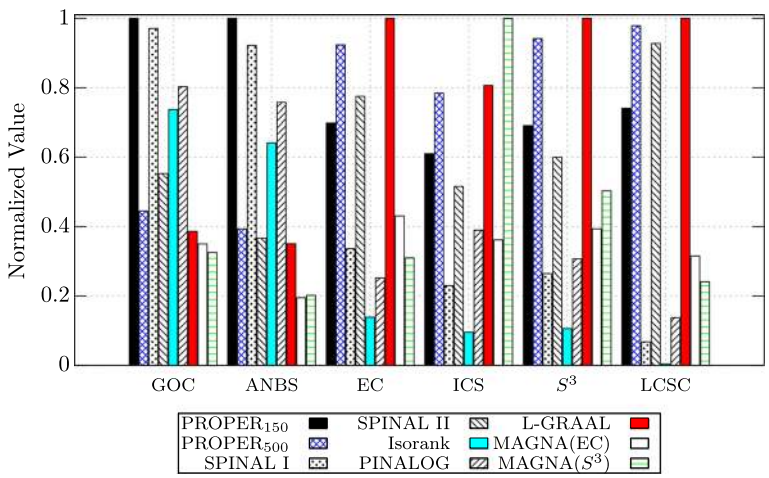

Fig. 7 Comparison of different global network aligners on aligning C. elegans and D. melanogaster based on six different measures. For the PROPER algorithm, we set $r = 1$ and $\ell \in \{150, 500\}$. The parameter α is 0.7

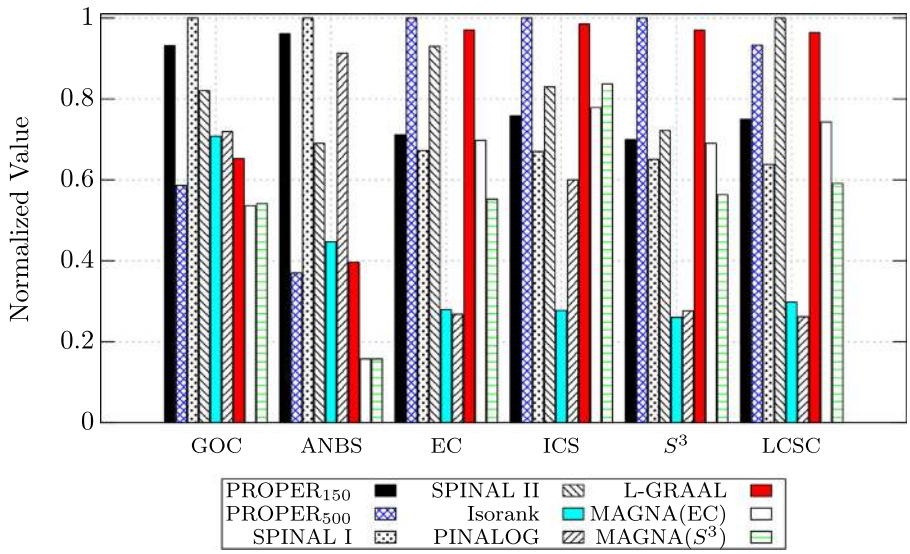

Fig. 8 Comparison of different global network aligners on aligning M. musculus and S. cerevisiae based on six different measures. For the PROPER algorithm, we set $r = 1$ and $\ell \in \{150, 500\}$. The parameter α is 0.7

can not be aligned with $r = 2$, and this is the reason we choose $r = 1$ for our experiments. In general, the value of r controls the strength of the structural evidence required before we decide to align a couple and a larger r makes errors less likely. We believe that by the increasing number of known PPI interactions over time, which consequently results in a decrease of the number of low-degree nodes, a larger value of r will generate better alignments.

Synthetic networks

In this section, we compare algorithms based on their performance over synthetic networks. For this, we consider the high-confidence yeast Saccharomyces cerevisiae PPI network with 1004 nodes and 8323 edges [33, 67]; this network serves as our "ground-truth". For this experiment, a noisy version of the yeast network is generated by sampling each of its nodes and interactions with a probability s. Here, s controls the similarity of a sampled network with the original network, and we take $1 - s$ as the "level of noise". Also, the sequence similarity for a subset of randomly chosen proteins is provided as a side information. In this experiment, the ground-truth node mapping is known by design, which enables us to calculate NC and precision. Note that in order to account for the randomness of our experiments, we provide the average of 50 different alignments for each level of noise and available sequence similarity.

In the first experiment, we align the original network with five networks that are generated by different levels of noise $1 - s \in \{5\%, 10\%, 15\%, 20\%, 25\%\}$. Also, the sequence similarity for 50% of randomly chosen proteins is provided. Figure 10 provides NC comparison over these synthetic networks for different levels of noise. From Fig. 10, for example, we observe that PROPER aligns networks which are sampled with the noise level $1 - s = 15\%$ with NC=0.86. Note that the average number of

Table 4 The average number of aligned couples when running (i) only the first step of PROPER (i.e., the SeedGeneration algorithm), and (ii,iii) PROPER with $r = \{1, 2\}$ with different values of ℓ

ℓ	SeedGeneration	$r = 2$	$r = 1$
150	2198.4	3116.1	5521.2
200	1875.6	2900.3	5471.4
300	1393.9	2618.4	5432.9
400	1083.1	2408.7	5416.4
500	861.0	2216.1	5347.4
600	696.4	2094.0	5320.5

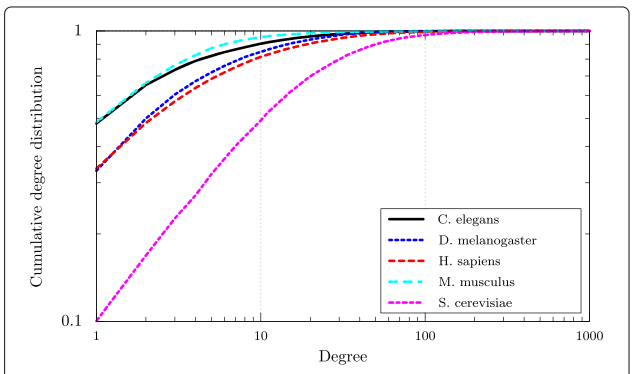

Fig. 9 Cumulative degree distribution for all the networks from Table 2

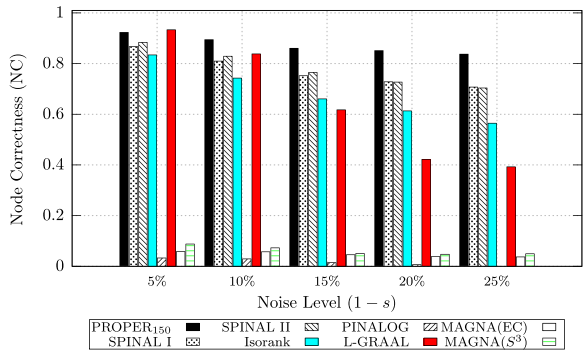

Fig. 10 Comparison of different global network aligners over synthetic networks based on node correctness (NC). The sequence similarity for 50% of randomly chosen proteins is provided. For the PROPER algorithm, we set $r = 1$ and $\ell = 150$. The parameter α is 0.7

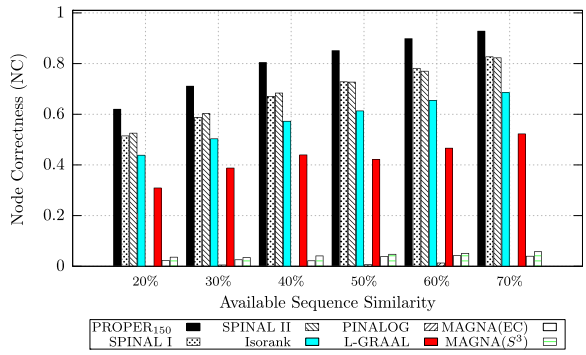

Fig. 11 Comparison of different global network aligners over synthetic networks based on node correctness (NC). The level of noise is set to $1 - s = 20\%$. For the PROPER algorithm, we set $r = 1$ and $\ell = 150$. The parameter α is 0.7

nodes for different noise levels (from 5 to 25%) is 946.48, 893.24, 832.54, 780.4 and 730.96, respectively. This means that PROPER correctly aligns $0.86 \times 832.54 \approx 716$ couples. Figure S1 in Additional file 2 compares algorithms based on precision. From the result of this experiment, we observe that for a low level of noise ($1 - s = 5\%$) L-GRAAL has the best performance and PROPER comes second. With increasing level of noise, the performance of PROPER remains almost unaffected, whereas the quality of the other alignments decreases quite markedly.

In the second experiment, we investigate the effect of available sequence similarity on the performance of algorithms. We consider different amounts of available sequence similarity and fix the level of noise to $1 - s = 20\%$. Figure 11 compares algorithms when the sequence similarities for 20%, 30%, 40%, 50%, 60% and 70% of randomly chosen proteins are provided. Figure S2 in Additional file 2 compares algorithms based on precision. We observe that PROPER outperforms the other algorithms over the entire range of available sequence similarities.

These two experiments confirm the success of the PROPER algorithm in aligning synthetic networks and its robustness to high levels of noise.

Aligning biological pathways

In this section, we compare algorithms based on their performance in aligning biological pathways. We use $\alpha = 0.7$ for SPINAL, IsoRank, MAGNA and L-GRAAL, and $r = 1, l = 150$ for PROPER. We use the measures introduced in the Pathway comparison measures section. For our comparisons, we consider the alignment of H. sapiens with the other four species from Table 2. We know that there are several proteins that belong to more than one pathway, because some proteins could be involved in different biological processes. For this reason, along the results for all the pathways, we consider a subset of non-overlapping pathways for each pair of species. Table 5 reports the number of common KEGG pathways between different pairs of species, where we consider (i) all the pathways, (ii) pathways with at least $\delta = 4$ interactions in each of the species, and (iii) a subset of non-overlapping pathways.

For the first experiment, we do not consider the topological similarities of aligned pathways. The result for alignments of pathways from different algorithms is provided in Table 6. We observe that PROPER outperforms the other algorithms in terms of accuracy. In the second experiment, for each algorithm we consider only the pathways with at least $\delta = 4$ conserved interactions across species (i.e., $\Delta_{\pi,i} \geq 4$). Table 7 provides the results for this case. Again, we observe that the PROPER algorithm outperforms the other algorithms, i.e., on average it aligns more pathways with a higher accuracy. MAGNA++ performs very poorly in this experiment and we omit it from Table 7.

For many pathways, the PROPER algorithm, compared to other algorithms, returns alignments with a larger portion of connected conserved subgraphs. For example, Fig. 12 shows the connected conserved subgraph of pathways hsa05200 and mmu05200 between human and mouse[3] The connected subgraph of this pathway has 37 nodes and 42 edges, which is larger than alignments by

Table 5 Number of common KEGG pathways between different pairs of species

Pair of species	#PW	#PW($\delta = 4$)	#PW (no-overlap)
hs-ce	116	19	37
hs-dm	122	31	40
hs-mm	283	152	49
hs-sc	98	32	34

Table 6 Comparison of algorithms based on aligning biological pathways. This table reports the average value of $\overline{acc_\pi}$ for pairwise alignments between Home sapiens and the four other species from Table 2

Algorithms	$\overline{acc_\pi}$	$\overline{acc_\pi}$ (no-overlap)
PROPER	**0.471**	**0.442**
SPINAL I	0.447	0.426
SPINAL II	0.115	0.134
PINALOG	0.409	0.397
L-GRAAL	0.232	0.218
MAGNA++(S^3)	0.016	0.020
MAGNA++(EC)	0.017	0.020
IsoRank	0.202	0.195

The best value for each column is highlighted in boldface

the other algorithms (see Additional file 3 for the detailed comparison results).

Execution time

A fast and scalable alignment algorithm is needed with the growing size of PPI networks. One of the key features of the PROPER algorithm is its low computational complexity and scalability. PROPER is able to align synthesis networks with millions of nodes in less than a hour. In fact, the complexity of our algorithm is $O((|E_1| + |E_2|)\min(D_1, D_2))$, where $D_{1,2}$ are the maximum degrees in the two networks. Table 8 provides the total execution time of algorithms for 10 pairwise alignments between the five species from Table 2. All computations are done on the same Linux machine with 16 GB of memory and 8 Intel Xeon E3-1270 CPUs working at clock speeds 3.50 GHz. We observe that PROPER runs much faster than the other algorithms.

Discussion

The purpose of network alignment algorithms is to find functional and structural similarities between PPI networks of different species [21]. Most of the works in the literature model global network alignment as an optimization problem over the convex combination of sequence and structural similarities between two networks [1, 28, 30]. This class of algorithms aims to maximize a cost function in order to increase the following two quantities simultaneously: (i) the pairwise similarities between aligned proteins (e.g., by maximizing the summation over all the BLAST similarities of aligned proteins), and (ii) the structural similarity between the two graphs, (e.g., by maximizing the conserved PPIs under the alignment) [46].

Table 7 Comparison of algorithms based on pathway alignment measures for $\delta = 4$ (i.e., $\Delta_{\pi,i} \geq 4$). This table reports the average value of measures for pairwise alignments between Home sapiens and the four other species from Table 2

Algorithms	#FPW	$\overline{acc_{\pi,\delta}}$	$recall_\pi$
PROPER	**42.5**	**0.585**	**0.584**
SPINAL I	38.75	0.554	0.536
SPINAL II	9.0	0.223	0.102
PINALOG	39.75	0.497	0.547
L-GRAAL	25.5	0.320	0.235
IsoRank	18.5	0.356	0.225

The best value for each column is highlighted in boldface

It appears that this particular formulation of the optimization problem precludes these algorithms from making good alignments by using both similarities jointly [46]. For example, the authors of [40] have shown that in the IsoRank algorithm for the structure-only ($\alpha = 1$) alignment, the similarity of two nodes is only a function of their degrees. Their results explicate the poor performance of IsoRank in finding alignments with good structural similarities. Also, our experimental results confirm the trade-off between structural and functional similarities in most of the state-of-the-art network alignment algorithms. We observe that each of the five algorithms evaluated here, namely L-GRAAL, MAGNA++, IsoRank, PINALOG and SPINAL, covers only a small portion of the trade-off frontier (see Figs. 3 and 4). In summary, we believe that these observations make it necessary to study the PPI network alignment problem under rigorous mathematical models.

The PROPER algorithm, in comparison, shows less compromise between the functional similarities among aligned proteins and the topological similarity. Figures 3, 4 and 5 show that our algorithm sweeps the frontier (i.e., has the best trade-off between both measures) more robustly than the other algorithms. In addition, large conserved subgraphs with the same function are aligned with PROPER. The PROPER algorithm not only aligns proteins and their corresponding interactions from two different species better than other algorithms, it also aligns the conserved pathways between the species with higher accuracy. This shows that instead of finding conserved single pairwise PPIs, PROPER represents a more biologically realistic performance by detecting sub-networks of conserved interactions from pathways with the same function among species.

In addition to its superior accuracy, PROPER performs better in terms of memory usage and speed, because the alignment process of PROPER is a very simple local propagation method.

Why the PROPER algorithm?

In the following, we provide two reasons why PROPER performs well in terms of the cost functions considered.

The first reason is that a high BLAST bit-score is a reliable indicator of a match, whereas a low BLAST bit-score is very unreliable for many functional characteristics

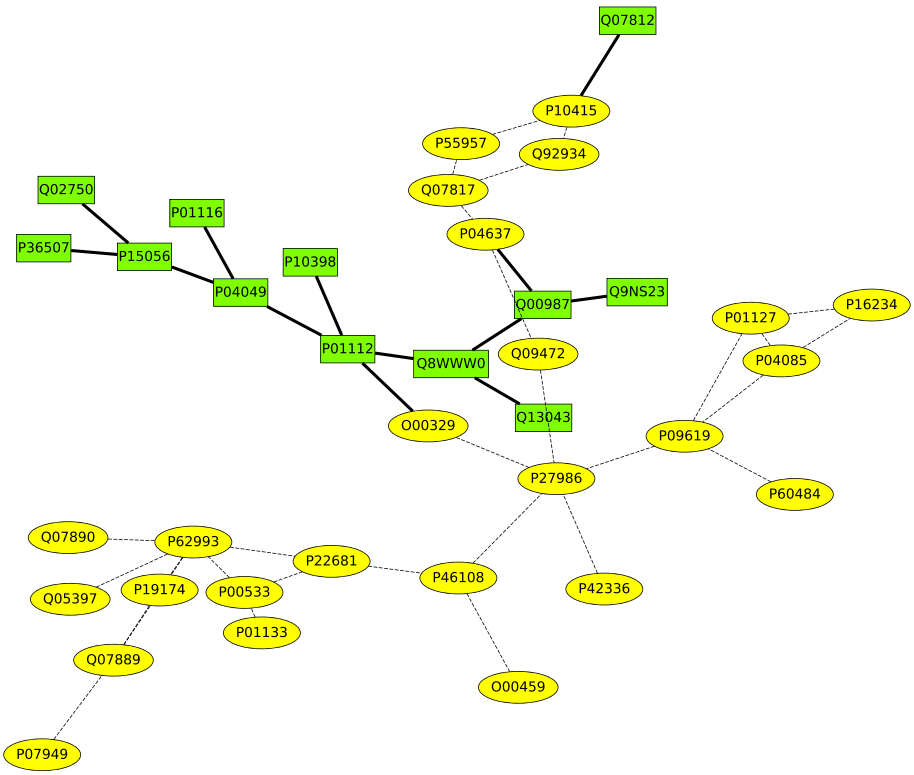

Fig. 12 The connected subgraph of hsa05200 and mmu05200 pathways in human and mouse from the PROPER algorithm, with conserved interactions in both species. This connected subgraph has 37 nodes and 42 edges. The PINALOG algorithm returns the second largest connected subgraph. The rectangular nodes and solid edges are the proteins and interactions among them that are found only by the PROPER algorithm

[68]. As a consequence, rather than optimizing a convex combination of functional similarity with structural similarity, it is advantageous to ascribe high confidence to the sparse set of high-BLAST couples, and to completely ignore low BLAST bit-scores. This is what PROPER does, by generating an initial seed-set of high BLAST couples, and then by propagating outwards from this seed set as a purely structure-driven process. Note that as the PGM class of algorithms are shown to be robust against noise in the seed set [45], PROPER is not too sensitive to the sequence similarity threshold ℓ for aligning new couples of proteins.

Table 8 The total execution time of algorithms for 10 pairwise alignments between the five species from Table 2

Aligner	Time
PROPER	317 s
L-GRAAL	4 h and 2 min
MAGNA++(S^3)	7 h and 47 min
MAGNA++(EC)	7 h and 41 min
PINALOG	2 days, 5 h and 26 min
SPINAL I	10 h and 51 min
SPINAL II	11 h and 56 min
IsoRank	12 h and 43 min

The second reason is more speculative and has to do with the statistical structure of the two networks being matched. Computational biology postulates evolutionary models to explain the difference between PPI networks. Studies have identified gene duplication and the gain or loss of genes and their interactions as the key evolutionary events in forming biological networks [69–71]. Several evolutionary models for regulatory networks and protein–interaction networks have been introduced based on these observed evolutionary processes [72–74].

Percolation-based methods for network alignment are well-suited for network pairs whose structural differences arise from the random deletions of nodes and edges. Specifically, in prior works, the authors of [42, 43] define the $G(n, p; t, s)$ random bigraph model for generating two correlated networks $G_{1,2}$ that rely on node and edge sampling processes. The two parameters t and s control the node and edge similarity of the generated graphs. Although the analysis in these prior works is for a different algorithm within the PGM class, we believe the main concepts carry over to PROPER.

More specifically, for the sake of simplicity, we assume that the evolutionary process can only delete proteins and interactions among proteins. We call this model $Evolve(G, t, s)$, where we postulate an ancestor network $G(V, E)$, from which both observable networks $G_{1,2}$ derive through independent evolutionary processes. The parameter t is the probability that a protein in G survives in $G_{1,2}$ (proteins are lost with probability $1 - t$); and parameter s is the probability that an interaction between proteins, i.e., an edge in G, survives in $G_{1,2}$ (interactions are lost with probability $1 - s$). With the additional assumption that the ancestor network G is an Erdős-Rényi [75] random graph (i.e., a $G(n, p)$ graph with n nodes, where each of the $\binom{n}{2}$ possible edges occurs independently with probability $0 < p < 1$) this evolutionary model is equivalent to the $G(n, p; t, s)$ model studied in the literature [42–44].

Under this model, conditions for the success of PGM-based network alignment have been established. In particular, a sharp phase transition in terms of the seed-set size have been shown: If the seed-set size is above some threshold (which depends on the network parameters n, p, t, and s), PGM-based alignment can correctly match, with high probability, almost all the node couples by using a purely structural process. Also, from the result of [43], we know that under a similar random bigraph model, the correct alignment maximizes the number of conserved interactions between the two networks. This simple parsimonious evolutionary model provides guarantees for the performance of the PROPER algorithm over random graphs similar to [45]. Note that, in practice, these algorithms are able to successfully align large real-networks, as well as many types of random graphs. In conclusion, it seems that mapping a (small) subset of nodes through a seed-generation step and matching the rest by using only structure of the two graphs works very well under an evolutionary model.

Conclusion

In this paper, we have introduced a new global pairwise-network alignment algorithm called PROPER. We have compared our algorithm with the state-of-the-art algorithms. We have shown that PROPER outperforms the other algorithms in both accuracy and speed. Also, we have shown that the PROPER algorithm can detect large conserved subnetworks between species.

Our results suggest that network-evolutionary models could be beneficial in designing network alignment algorithms. We believe that, for future work, considering a model that also takes into account gene duplication, network motifs, clustering within networks and modularity of biological networks (e.g., [76]) would increase the accuracy of global network alignments. Finally, to find biological pathways and protein complexes using the PROPER algorithm, the next step would be to design methods that can detect sub-networks as potential pathways or complexes (similar to the method used in [12, 24]).

Endnotes

[1] A network motif is a small recurrent connected-subgraph that occurs in PPI and other biological networks significantly more often than in random networks.

[2] These pathways are Spliceosome. Spliceosome removes introns from a transcribed pre-mRNA, a type of primary transcript.

[3] Pathways hsa05200 and mmu05200 are in the class cancer Homo sapiens (human).

Additional files

Additional file 1: IsoBase: experimental results. The results for experiments over Isobase dataset are provided in this appendix. (PDF 269 kb)

Additional file 2: The PROPER algorithm: detailed comparisons. The detailed experimental results for PROPER are provided in this appendix. (PDF 183 kb)

Additional file 3: Pathways: experimental results. The detailed experimental results for aligning biological pathways are provided in this appendix. (PDF 165 kb)

Abbreviations

BLAST: Basic local-alignment search tool; PGM: Percolation graph matching; PPI: Protein-protein interaction

Acknowledgements

We would like to thank Prof. Bernard Morêt for our fruitful discussions and his feedback on the first version of the manuscript. Also, we thank Frederic Morêt for the initial development of the http://proper.epfl.ch website.

Funding

Not applicable.

Availability of data and materials

The PROPER algorithm is publicly available at http://proper.epfl.ch. All the datasets we used in this research are collected from public databases (cited in the manuscript).

Authors' contributions

All the authors contributed to performing the research and writing the paper. MG designed the original research question. EK performed data collection, implementation of the algorithm and computational analysis. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Author details
[1]School of Computer and Communication Sciences, EPFL, Lausanne, Switzerland. [2]Department of Computer Science, ETHZ, Zurich, Switzerland. [3]School of Life Sciences, EPFL, Lausanne, Switzerland.

Received: 3 May 2016 Accepted: 29 November 2016
Published online: 12 December 2016

References

1. Zaslavskiy M, Bach F, Vert JP. Global alignment of protein-protein interaction networks by graph matching methods. Bioinformatics. 2009;25(12):1259–67.
2. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Nat Acad Sci. 2001;98(8):4569–74.
3. Tong AHY, Evangelista M, Parsons AB, Xu H, Bader GD, Pagé N., Robinson M, Raghibizadeh S, Hogue CW, Bussey H, et al. Systematic genetic analysis with ordered arrays of yeast deletion mutants. Science. 2001;294(5550):2364–8.
4. Aebersold R, Mann M. Mass spectrometry-based proteomics. Nature. 2003;422(6928):198–207.
5. Orchard S, Kerrien S, Abbani S, Aranda B, Bhate J, Bidwell S, Bridge A, Briganti L, Brinkman FS, Cesareni G, et al. Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. Nat Methods. 2012;9(4):345–50.
6. Chatr-aryamontri A, Breitkreutz BJ, Heinicke S, Boucher L, Winter A, Stark C, Nixon J, Ramage L, Kolas N, O'Donnell L, et al. The BioGRID interaction database: 2013 update. Nucleic Acids Res. 2013;41(D1):816–23.
7. Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, Galeota E, Sacco F, Palma A, Nardozza AP, Santonico E, et al. MINT, the molecular interaction database: 2012 update. Nucleic Acids Res. 2012;40(D1):857–61.
8. Peri S, Navarro JD, Kristiansen TZ, Amanchy R, Surendranath V, Muthusamy B, Gandhi T, Chandrika K, Deshpande N, Suresh S, et al. Human protein reference database as a discovery resource for proteomics. Nucleic Acids Res. 2004;32(suppl 1):497–501.
9. Hermjakob H, Montecchi-Palazzi L, Lewington C, Mudali S, Kerrien S, Orchard S, Vingron M, Roechert B, Roepstorff P, Valencia A, et al. IntAct: an open source molecular interaction database. Nucleic Acids Res. 2004;32(suppl 1):452–5.
10. Sharan R, Suthram S, Kelley RM, Kuhn T, McCuine S, Uetz P, Sittler T, Karp RM, Ideker T. Conserved patterns of protein interaction in multiple species. Proc Nat Acad Sci. 2005;102(6):1974–9.
11. Suthram S, Sittler T, Ideker T. The Plasmodium protein network diverges from those of other eukaryotes. Nature. 2005;438(7064):108–12.
12. Kelley BP, Sharan R, Karp RM, Sittler T, Root DE, Stockwell BR, Ideker T. Conserved pathways within bacteria and yeast as revealed by global protein network alignment. Proc Nat Acad Sci. 2003;100(20):11394–9.
13. Yu H, Luscombe NM, Lu HX, Zhu X, Xia Y, Han JD, Bertin N, Chung S, Vidal M, Gerstein M. Annotation transfer between genomes: protein–protein interologs and protein–DNA regulogs. Genome Res. 2004;14(6):1107–18.
14. Matthews LR, Vaglio P, Reboul J, Ge H, Davis BP, Garrels J, Vincent S, Vidal M. Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". Genome Res. 2001;11(12):2120–6.
15. Sharan R, Ideker T. Modeling cellular machinery through biological network comparison. Nat Biotechnol. 2006;24(4):427–33.
16. Kuchaiev O, Pržulj N. Integrative network alignment reveals large regions of global network similarity in yeast and human. Bioinformatics. 2011;27(10):1390–6.
17. Singh R, Xu J, Berger B. Pairwise global alignment of protein interaction networks by matching neighborhood topology. In: Research in Computational Molecular Biology. Oakland: Springer; 2007. p. 16–31.
18. Sjölander K. Phylogenomic inference of protein molecular function: advances and challenges. Bioinformatics. 2004;20(2):170–9.
19. Remm M, Storm CE, Sonnhammer EL. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J Mol Biol. 2001;314(5):1041–52.
20. Elmsallati A, Clark C, Kalita J. Global Alignment of Protein-Protein Interaction Networks: A Survey. IEEE/ACM Trans Comput Biol Bioinformatics. 2016;13(4):689–705.
21. Faisal FE, Meng L, Crawford J, Milenković T. The post-genomic era of biological network alignment. EURASIP J Bioinformatics Syst Biol. 2015;2015(1):1–19.
22. Davis D, Yaveroğlu ÖN, Malod-Dognin N, Stojmirovic A, Pržulj N. Topology-function conservation in protein–protein interaction networks. Bioinformatics. 2015;31(10):1632–9.
23. Malod-Dognin N, Pržulj N. L-GRAAL: Lagrangian graphlet-based network aligner. Bioinformatics. 2015;31(13):2182–9.
24. Kelley BP, Yuan B, Lewitter F, Sharan R, Stockwell BR, Ideker T. PathBLAST: a tool for alignment of protein interaction networks. Nucleic Acids Res. 2004;32(suppl 2):83–8.
25. Liang Z, Xu M, Teng M, Niu L. NetAlign: a web-based tool for comparison of protein interaction networks. Bioinformatics. 2006;22(17):2175–7.
26. Koyutürk M, Kim Y, Topkara U, Subramaniam S, Szpankowski W, Grama A. Pairwise alignment of protein interaction networks. J Comput Biol. 2006;13(2):182–99.
27. Flannick J, Novak A, Srinivasan BS, McAdams HH, Batzoglou S. Graemlin: general and robust alignment of multiple large interaction networks. Genome Res. 2006;16(9):1169–81.
28. Singh R, Xu J, Berger B. Global alignment of multiple protein interaction networks with application to functional orthology detection. Proc Nat Acad Sci. 2008;105(35):12763–8.
29. Liao CS, Lu K, Baym M, Singh R, Berger B. IsoRankN: spectral methods for global alignment of multiple protein networks. Bioinformatics. 2009;25(12):253–8.
30. Kuchaiev O, Milenković T, Memišević V, Hayes W, Pržulj N. Topological network alignment uncovers biological function and phylogeny. Journal of the Royal Society Interface. 2010;7(50):1341–1354.
31. Milenković T, Ng WL, Hayes W, Pržulj N. Optimal network alignment with graphlet degree vectors. Cancer Inform. 2010;9:121.
32. Memišević V, Pržulj N. C-GRAAL: C ommon-neighbors-based global GRA ph AL ignment of biological networks. Integr Biol. 2012;4(7):734–43.
33. Saraph V, Milenković T. Magna: Maximizing accuracy in global network alignment. Bioinformatics. 2014;30(20):2931–40.
34. Vijayan V, Saraph V, Milenković T. MAGNA++: Maximizing Accuracy in Global Network Alignment via both node and edge conservation. Bioinformatics. 2015;31(14):2409–11.
35. Aladag AE, Erten C. SPINAL: scalable protein interaction network alignment. Bioinformatics. 2013;29(7):917–24.
36. Phan HTT, Sternberg MJE. PINALOG: a novel approach to align protein interactionnetworks–implications for complex detection and function prediction. Bioinformatics. 2012;28(9):1239–45.
37. Hu J, Kehr B, Reinert K. NetCoffee: a fast and accurate global alignment approach to identify functionally conserved proteins in multiple networks. Bioinformatics. 2014;30(4):540–548.
38. Alkan F, Erten C. BEAMS: backbone extraction and merge strategy for the global many-to-many alignment of multiple PPI networks. Bioinformatics. 2014;30(4):531–9.
39. Bayati M, Gleich DF, Saberi A, Wang Y. Message-passing algorithms for sparse network alignment. ACM Trans Knowl Discov Data (TKDD). 2013;7(1):3.
40. Kazemi E, Grossglauser M. On the Structure and Efficient Computation of IsoRank Node Similarities. 2016. arXiv:1602.00668v2.
41. Patro R, Kingsford C. Global network alignment using multiscale spectral signatures. Bioinformatics. 2012;28(23):3105–14.
42. Pedarsani P, Grossglauser M. On the privacy of anonymized networks. In: Proceedings of ACM SIGKDD 2011. San Diego; 2011.
43. Kazemi E, Yartseva L, Grossglauser M. When Can Two Unlabeled Networks Be Aligned Under Partial Overlap? In: Proceedings of the 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton). Monticello; 2015. p. 33–42.
44. Cullina D, Kiyavash N. Improved Achievability and Converse Bounds for Erdos-Renyi Graph Matching. In: Proceedings of the 2016 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Science. New York: ACM; 2016.
45. Kazemi E, Hassani SH, Grossglauser M. Growing a Graph Matching from a Handful of Seeds. Proc VLDB Endowment. 2015;8(10):1010–21.
46. Clark C, Kalita J. A comparison of algorithms for the pairwise alignment of biological networks. Bioinformatics. 2014;30(16):2351–9.
47. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10.

48. Narayanan A, Shmatikov V. De-anonymizing Social Networks. In: Proceedings of IEEE Symposium on Security and Privacy 2009. Oakland; 2009.
49. Melnik S, Garcia-Molina H, Rahm E. Similarity Flooding: A Versatile Graph Matching Algorithm and Its Application to Schema Matching. In: Proceedings of ICDE 2002. San Jose; 2002.
50. Conte D, Foggia P, Sansone C, Vento M. Thirty years of graph matching in pattern recognition. Int J Pattern Recognit Artif Intell. 2004;18(03):265–98.
51. Torresani L, Kolmogorov V, Rother C. In: Forsyth D, Torr P, Zisserman A, editors. Feature Correspondence Via Graph Matching: Models and Global Optimization. Berlin: Springer; 2008, pp. 596–609.
52. Egozi A, Keller Y, Guterman H. A Probabilistic Approach to Spectral Graph Matching. Pattern Anal Mach Intell IEEE Trans. 2013;35(1):18–27.
53. Yartseva L, Grossglauser M. On the performance of percolation graph matching. In: Proceedings of ACM COSN 2013. Boston; 2013.
54. Chiasserini CF, Garetto M, Leonardi E. De-anonymizing scale-free social networks by percolation graph matching. In: Proceedings of IEEE INFOCOM 2015. Hong Kong; 2015.
55. Korula N, Lattanzi S. An efficient reconciliation algorithm for social networks. Proc VLDB Endowment. 2014;7(5):377–88.
56. Joshi T, Xu D. Quantitative assessment of relationship between sequence similarity and function similarity. BMC Genomics. 2007;8(1):222.
57. Pesquita C, Faria D, Bastos H, Ferreira AE, Falcão AO, Couto FM. Metrics for GO based protein semantic similarity: a systematic evaluation. BMC Bioinformatics. 2008;9(Suppl 5):4.
58. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. Gene Ontology: tool for the unification of biology. Nat Genet. 2000;25(1):25–9.
59. Seah B, Bhowmick SS, Jr CFD. DualAligner: a dual alignment-based strategy to align protein interaction networks. Bioinformatics. 2014;30(18): 2619–26.
60. KEGG pathway database - Kyoto University Bioinformatics Centre. http://www.genome.jp/kegg/pathway.html. Data acquired on 04 April 2016.
61. IntAct: an open source molecular interaction database. http://www.ebi.ac.uk/intact/. Data acquired on 04 April 2016.
62. The GOA database. http://www.ebi.ac.uk/GOA/downloads. Data acquired on 04 April 2016.
63. Barrell D, Dimmer E, Huntley RP, Binns D, O'Donovan C, Apweiler R. The GOA database in 2009–an integrated Gene Ontology Annotation resource. Nucleic Acids Res. 2009;37(suppl 1):396–403.
64. Park D, Singh R, Baym M, Liao CS, Berger B. IsoBase: a database of functionally related proteins across PPI networks. Nucleic Acids Res. 2011;39(Database-Issue):295–300.
65. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, et al. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2004;32(suppl 1):115–9.
66. UniProt: the universal protein knowledgebase. http://www.uniprot.org/. Data acquired on 04 April 2016.
67. Collins SR, Kemmeren P, Zhao XC, Greenblatt JF, Spencer F, Holstege FC, Weissman JS, Krogan NJ. Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. Mol Cell Proteomics. 2007;6(3):439–50.
68. Devos D, Valencia A. Practical limits of function prediction. Proteins: Structure, Function, Bioinformatics. 2000;41(1):98–107.
69. Madan Babu M, Teichmann SA, Aravind L. Evolutionary dynamics of prokaryotic transcriptional regulatory networks. J Mol Biol. 2006;358(2): 614–33.
70. Roth C, Rastogi S, Arvestad L, Dittmar K, Light S, Ekman D, Liberles DA. Evolution after gene duplication: models, mechanisms, sequences, systems, and organisms. J Exp Zool Part B: Mol Dev Evol. 2007;308(1): 58–73.
71. Teichmann SA, Babu MM. Gene regulatory network growth by duplication. Nat Genet. 2004;36(5):492–6.
72. Zhang X, Moret BM. Refining regulatory networks through phylogenetic transfer of information. Comput Biol Bioinformatics, IEEE/ACM Trans. 2012;9(4):1032–45.
73. Berg J, Lässig M, Wagner A. Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications. BMC Evol Biol. 2004;4(1):51.
74. Sahraeian SME, Yoon BJ. A network synthesis model for generating protein interaction network families. PloS One. 2012;7(8):41474.
75. Erdös P, Rényi A. On random graphs I. Publ Math Debrecen. 1959;6:290–7.
76. Navlakha S, Kingsford C. Network archaeology: uncovering ancient networks from present-day interactions. PLoS Comput Biol. 2011;7(4): 1001119.
77. PINALOG web server for protein interaction network alignment. http://www.sbg.bio.ic.ac.uk/~pinalog/. Data acquired on 04 April 2016.