NAIGO: An Improved Method to Align PPI Networks Based on Gene Ontology and Graphlets

Lijuan Zhu[1†], Ju Zhang[2†], Yi Zhang[3], Jidong Lang[4], Ju Xiang[5,6], Xiaogang Bai[3], Na Yan[4], Geng Tian[4], Huajun Zhang[1*] and Jialiang Yang[4*]

[1] College of Mathematics and Computer Science, Zhejiang Normal University, Jinhua, China, [2] Institute of Infectious Diseases, Beijing Ditan Hospital, Capital Medical University, and Beijing Key Laboratory of Emerging Infectious Diseases, Beijing, China, [3] Department of Mathematics, Hebei University of Science & Technology, Shijiazhuang, China, [4] Geneis Beijing Co., Ltd., Beijing, China, [5] Neuroscience Research Center & Department of Basic Medical Sciences, Changsha Medical University, Changsha, China, [6] School of Computer Science and Engineering, Central South University, Changsha, China

With the development of high throughput technologies, there are more and more protein–protein interaction (PPI) networks available, which provide a need for efficient computational tools for network alignment. Network alignment is widely used to predict functions of certain proteins, identify conserved network modules, and study the evolutionary relationship across species or biological entities. However, network alignment is an NP-complete problem, and previous algorithms are usually slow or less accurate in aligning big networks like human vs. yeast. In this study, we proposed a fast yet accurate algorithm called Network Alignment by Integrating Biological Process (NAIGO). Specifically, we first divided the networks into subnets taking the advantage of known prior knowledge, such as gene ontology. For each subnet pair, we then developed a novel method to align them by considering both protein orthologous information and their local structural information. After that, we expanded the obtained local network alignments in a greedy manner. Taking the aligned pairs as seeds, we formulated the global network alignment problem as an assignment problem based on similarity matrix, which was solved by the Hungarian method. We applied NAIGO to align human and *Saccharomyces cerevisiae* S288c PPI network and compared the results with other popular methods like IsoRank, GRAAL, SANA, and NABEECO. As a result, our method outperformed the competitors by aligning more orthologous proteins or matched interactions. In addition, we found a few potential functional orthologous proteins such as RRM2B in human and DNA2 in *S. cerevisiae* S288c, which are related to DNA repair. We also identified a conserved subnet with six orthologous proteins EXO1, MSH3, MSH2, MLH1, MLH3, and MSH6, and six aligned interactions. All these proteins are associated with mismatch repair. Finally, we predicted a few proteins of *S. cerevisiae* S288c potentially involving in certain biological processes like autophagosome assembly.

Keywords: PPI network, network alignment, gene ontology, graphlets, functional ortholog

INTRODUCTION

With the development of high-throughput techniques, such as yeast two-hybrid system (Uetz et al., 2000; Ito et al., 2001) and coimmunoprecipitation coupled to mass spectrometry (Krogan et al., 2006), a large amount of protein–protein interaction (PPI) networks in many species (e.g., human, yeast, and mouse) have been reported. Their nodes vary from hundreds to tens of thousands. Within a PPI network, each node denotes a protein, and each edge denotes an interaction between the two proteins connected. Interspecies network alignment is important for predicting protein functions (Dutkowski and Tiuryn, 2007; Singh et al., 2008) and understanding the PPI network evolutions (Flannick et al., 2006; Kuchaiev and PrŽulj, 2011). Many network alignment tools have been proposed. According to their comparison range, they could be categorized into local alignment and global alignment. The goal of local alignment is to find subnets conserved in two or more species, which are usually limited to a relative small range and involve only proteins with specific functions (Berg and Lässig, 2004, 2006; Brian et al., 2004; Sharan et al., 2005; Liang et al., 2006; Ciriello et al., 2012; Mina and Guzzi, 2014). On the contrary, the global alignment aims to find mappings traversing all nodes (Klau, 2009; Liao et al., 2009; Zaslavskiy et al., 2009; Milenković et al., 2010; Patro and Kingsford, 2012; Neyshabur et al., 2013; Faisal et al., 2015).

Since PPI network alignment is usually a generalized subgraph isomorphism problem, that is NP-hard (Lathrop, 1994), developing heuristic algorithms with good practical efficiency has become one of the foremost challenges. The existing network alignment algorithms could be classified into three categories: heuristic search method based on graph model, constraint optimization method based on objective function, and modular method based on the divide and conquer strategy. The heuristic search methods could establish interspecies alignment graphs with the orthologous proteins as nodes. They evaluate the similarity between PPI networks and design heuristic search algorithms generally adopting the greedy strategy of seed growth. Following similar ideas, different alignment tools also use different analytic strategies and algorithms. For example, MaWISH (Koyuturk et al., 2006) converts the local alignment into a maximum weight induced subgraph problem. Græmlin (Flannick et al., 2006) determines the initial matched proteins based on the log-likelihood ratio probability model and then gradually searches other similar protein nodes to expand the alignment graphs. The optimization-based methods could convert the alignment problem into an optimization problem. For instance, IsoRank (Singh et al., 2008) calculates the PPI network similarity with an eigenvalue matrix and extracts the global alignment from it. MNAligner (Li et al., 2007) converts the network alignment into an integer quadratic programming problem, while GRAAL (Kuchaiev et al., 2010) converts it into a topological structure problem. The other alignment tool, BinAligner (Yang et al., 2013), constructs a similarity matrix based on node similarity and "n-neighborhood" ($n \geq 1$) subnet similarity and implements alignment by integer programming. The modular-based methods are also widely used for alignments, considering the large size of PPI networks and their modular structures (Hartwell et al., 1999; Silva and Stumpf, 2005; Sharan and Ideker, 2006; Almaas, 2007; Srinivasan et al., 2007). For example, Match-and-Split alignment (Narayanan and Karp, 2007) divides the modules by matching and splitting, while BiNA (Towfic et al., 2009) divides the original network into multiple subnets and aligns them by a kernel function subsequently.

Although many algorithms for PPI network alignment have been developed, most current tools either have lower accuracy or fail to align large networks. There is still a need for a fast and accurate alignment algorithm. In this study, we propose an improved alignment method, NAIGO, which integrates divide-and-conquer strategy, optimization modeling, and graph-based features. It can align PPI networks locally and globally based on the node similarity, edge similarity, and topological similarity of the networks. To improve the calculation efficiency, NAIGO achieves the global alignment between large networks by expanding prealigned subnets in a greedy manner. In contrast to other alignment algorithms, NAIGO could also expand the smaller subnets by referring to the matched bigger ones and thus predict the unknown biological process (BP) of proteins. We applied NAIGO to align the PPI networks of human and *Saccharomyces cerevisiae* S288c. Compared with other popular methods such as GRAAL (Kuchaiev et al., 2010), IsoRank (Singh et al., 2008), SANA (Mamano and Hayes, 2017), and NABEECO (Ibragimov et al., 2003), NAIGO has a better alignment performance.

METHODS

Our algorithm consists of three steps: (1) divide the large networks into multiple small subnets, (2) align the corresponding subnets based on the similarity matrix, (3) expand the interspecies alignment graphs based on the heuristic search idea, with the best aligned subnets as nodes.

Network Division

Considering that similar proteins participate in the same biological process in different species, we use the BP information of Gene Ontology (GO) as the criteria to divide the network. In our study, BP data was extracted by loading the GO.db package, and it contained a total of 14,291 GO terms. Based on the BP terms, we divided the network as follows: if two interacted proteins both involved in the same term, they will be included in the subnet (**Figure 1A**). The division method could avoid isolated vertices. According to the criteria, the PPI network of human could be divided into 6,781 non-empty subnets, and the PPI network of *S. cerevisiae* S288c could be divided into 1,836 non-empty subnets. Among them, there are 1,771 subnet pairs under the same terms of the two species. In the corresponding 1,771 terms, except for the biological process term, only one term pair had containment relationship, and the included term was removed. Thus, we consider aligning the 1,770 BP subnet pairs (i.e., align human subnet i to *S. cerevisiae* S288c subnet i, $i = 1, 2, \ldots, 1,770$) (**Figure 1B**).
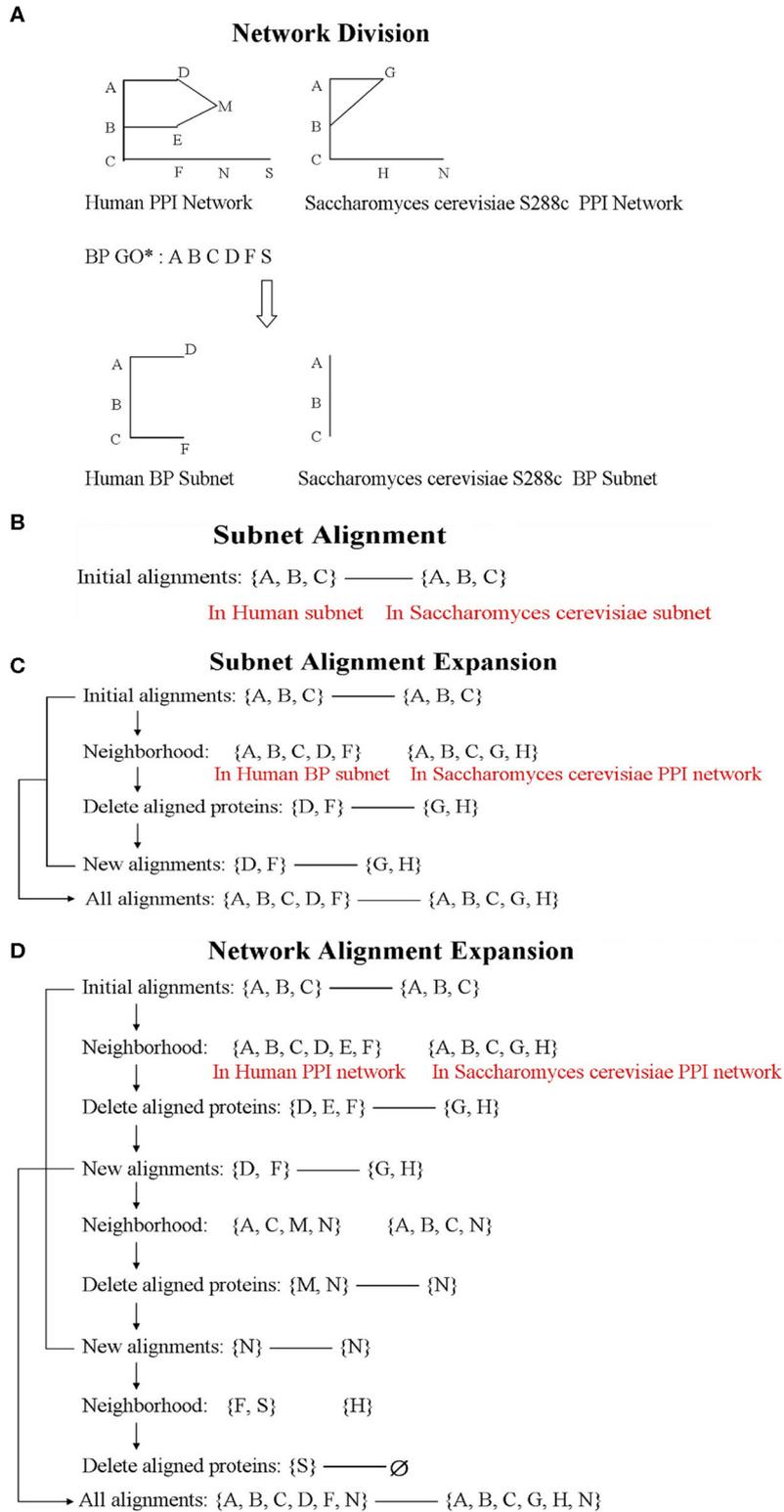
FIGURE 1 | The flowchart of Network Alignment by Integrating Biological Process (NAIGO). The letters, such as **(A–C)**, represent proteins in the figure. **(D)** Network alignment expansion.

Subnet Alignment Based on Similarity Matrix

A PPI network is an undirected graph $G = (V, E)$, where the node set V represents the proteins and the edge set E represents interactions between proteins. Given two PPI subnets, $G = (V, E)$ and $H = (U, F)$, the subnet alignment is defined as a one-to-one mapping (π) between node set V and U.

$$\pi : \{i \in V\} \mapsto \{j \in U\} \tag{1}$$

where π reflects the alignment of proteins, and we define the mapping value as Equation (2). If we get the π_{ij} value of each protein pair in the two PPI subnets ($i \in V, j \in U$) and confirm the correspondence of each interaction pair based on π_{ij}, we will achieve a subnet alignment.

$$\pi_{ij} = \begin{cases} 1, & if\, j = \pi\,(i) \\ 0, & otherwise \end{cases} \tag{2}$$

To get the optimal alignment efficiently, we constructed a similarity matrix (C) for each PPI subnet pair. We need to solve the integer linear programming problem to maximize the alignment score S in Equation (3), which is subject to restrictions as Equation (4). Hungarian algorithm was adopted to solve the problem.

$$S = \max_{\pi} \sum_{i \in V} \sum_{j \in U} C_{ij}\pi_{ij} \tag{3}$$

Subject to

$$\begin{cases} \sum_{i \in V} \pi_{ij} = 0 \text{ or } 1, & \forall j \in U \\ \sum_{j \in U} \pi_{ij} = 0 \text{ or } 1, & \forall i \in V \\ \pi_{ij} = 0 \text{ or } 1, & \forall i \in V \text{ and } j \in U \end{cases} \tag{4}$$

In the above equation, V and U represent the node set of subnet G and H, respectively.

Construction of Similarity Matrix

Obviously, the PPI subnet alignments were decided by their topological structure, as well as the conservation of the proteins (nodes) and interactions (edges) that mainly influenced by the protein orthology. Therefore, the similarity matrix C considers both the protein orthology and network topology as follows.

$$C = \theta_1 * (A + A^*) + \theta_2 * B (0 \le \theta_1, \theta_2 \le 1 \text{ and } \theta_1 + \theta_2 = 1) \tag{5}$$

On the one hand, we define the corresponding protein orthology matrix as A and satisfies Equation (6). The orthologous file between *S. cerevisiae* S288c and human was downloaded from Ensembl (http://www.ensembl.org/biomart/martview). Besides, we also lead A^* into the similarity matrix and define it as Equation (7), which reflects the contribution of the orthologous protein pairs to the edge matching. Within it, $N(i)$ is the set of

neighbors of node i, $|N(i)|$ is the size of this set, and $N(i) \times N(j)$ is the Cartesian product of sets $N(i)$ and $N(j)$.

$$A(i, j) = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are orthologs} \\ 0 & otherwise \end{cases} \tag{6}$$

$$A^*(i, j) = \frac{sum\{A[N(i) \times N(j)]\}}{|N(i)| * |N(j)|} \tag{7}$$

On the other hand, we also compare the topological structure around the nodes during the subnet alignment, which is based on the graphlet degree (PrŽulj, 2007). A graphlet is a connected non-isomorphic subgraph of a large network, in which non-automorphism positions are called orbits. We construct graphlet matrix $G_{n \times 73}$ based on 30 (2-, 3-, 4-, and 5-node) graphlets with 73 orbits, where n is the number of nodes in the subnet and $G(i, j)$ is the number of graphlets touched by node i at orbit $j - 1$ ($j = 1, 2, ..., 73$). Furthermore, the graphlet similarity matrix is defined as B and satisfies

$$B(i, j) = \frac{1}{1 + \sqrt{\sum_{k=1}^{73} \left[G_1(i, k) - G_2(j, k)\right]^2}} \tag{8}$$

G_1 and G_2 represent the graphlet matrices of two species. The number of rows is the number of proteins in the subnet, and the number of columns is the number of orbits.

Parameter Setting of Similarity Matrices

As displayed in Equation (5), how to weight the protein orthology (θ_1) and network topology (θ_2) is essential for constructing the similarity matrices. We consider the protein orthology as the major determinant when there are many orthologous protein pairs between corresponding subnets. Otherwise, we consider the topology as the major determinant. We thus calculate the weighting coefficient as follows (let m be the number of subnets).

$$\theta_1 = \begin{cases} p_i & \text{if } p_i < 1, \\ \underset{j=1,2,...,m}{mean}(p_j|p_j < 1) & \text{else.} \end{cases} \tag{9}$$

$$\theta_2 = 1 - \theta_1 \tag{10}$$

where p represents the orthologous protein ratio and is calculated with Equation (11), where n, n_1, and n_2 represents the number of orthologous protein pairs and the number of proteins in two PPI subnets, respectively.

$$p = \frac{n}{\min(n_1, n_2)} \tag{11}$$

When $p_i < 1$ in the subnet, the weighting coefficient of the protein orthology (θ_1) is equal to p_i; otherwise, θ_1 is the mean value of p_i of the other subnets with $p_i < 1$.

Simplification of Similarity Matrices

When the dimension of matrix C is large, we simplify it to improve the calculation speed without affecting results. Let the matrix dimension be $m \times n$. If $m = n$, C could not be simplified.

Otherwise, the simplification principles are as follows, taking $m > n$ as an example.

(1) Extract the rows with orthologous proteins from C. Let the number of rows be r_1. The extracted matrix and remaining matrix are denoted by Z_1 and C_1, respectively.

(2) Extract the rows from C_1 which include the largest element of each column. Let the number of rows be r_2. The extracted matrix and remaining matrix are denoted by Z_2 and C_2, respectively.

(3) If $r_1 + r_2 \geq n$, $[Z_1; Z_2]$ is the simplified matrix C. If $r_1 + r_2 < n$, extract the rows from C_2 according to the step (2). Repeat the matrix extraction and test until $\sum_{q=1}^{Q} r_q \geq n$.

The Whole Network Alignment Expansion

After the subnet alignments, we treat the largest aligned subnet as a seed and expand it with the heuristic search method until aligning all the nodes. Then, we could achieve the global alignment between large PPI networks. The expansion steps are as follows (**Figure 1D**):

(1) Let the seed of Human be V and its neighbor (i.e., the node set that interacts with V) be V_1; the seed of *S. cerevisiae* S288c be U and its neighbor be U_1. Remove the repeatedly occurring proteins in their corresponding seed from V_1/U_1 (due to the interactions between the proteins in V/U, there are overlapping proteins in V/U and V_1/U_1).

(2) Find the alignment between V_1 and U_1 by resolving the optimization problem and maximizing the alignment score S as Equation (12), and thus expand the aligned seed.

(3) Let the neighbor of V_1 be V_2 and the neighbor of U_1 be U_2. Repeat steps (1) and (2) until the seed network expands to n layers and U_{n+1} or V_{n+1} is empty. Merge all results to achieve the global alignment between two species.

The alignment score S is also the criterion for assessing alignment. However, we use the scoring scheme of graph alignment (Berg and Lässig, 2006; Kolár et al., 2008) in the whole network alignment expansion, comparing to the similarity matrix based strategy in the subnet alignment. It is achievable when the subnet alignment is implemented as a seed and is more accurate. Briefly, we define S as the sum of the nodes score S_V and edge score S_E. In order to maximize S, we resolve the optimization problem with the Hungarian algorithm in Equation (12), which is subject to restrictions as Equation (13). Taking π_{ij} as the aligned seed, we could obtain the π_{kl} value for any k in V and l in U by resolving assignment problem.

$$S = \max \sum_{k \in V} \sum_{l \in U} \alpha_{kl} \pi_{kl} + \sum_{i,k \in V} \sum_{j,l \in U} \beta_{ijkl} \pi_{ij} \pi_{kl} \quad (12)$$

Subject to

$$\begin{cases} \sum_{i \in V} \sum_{j \in U} C_{ij} \pi_{ij} = \hat{s}, & \\ \sum_{k \in V} \pi_{kl} = 0 \text{ or } 1, & \forall l \in U \\ \sum_{l \in U} \pi_{kl} = 0 \text{ or } 1, & \forall k \in V \\ \pi_{kl} = 0 \text{ or } 1, & \forall k \in V \text{ and } l \in U \end{cases} \quad (13)$$

In the above equation, V and U represent the node set of human and *S. cerevisiae* S288c PPI network, respectively. S_V scores protein pairs, in which the sequence similarity score (α_{kl}) is used to determine the most likely orthologous proteins [$l = \pi(k)$]. S_E scores PPI pairs, in which β_{ijkl} is used to reward each matched PPI pair and punish each mismatched PPI pair. Given the nodes $i, k \in V$ and $j, l \in U$, as well as the PPIs $ik \in E$ and $jl \in F$, if $j = \pi(i)$ and $l = \pi(k)$, the PPIs ik and jl are matched edges; otherwise, they are mismatched edges. The parameter values are settled by Equations (14, 15), according to the reference (Yang et al., 2013).

$$\alpha_{kl} = \begin{cases} 4.4 & k \text{ and } l \text{ are orthologs} \\ -1.6 & \text{else} \end{cases} \quad (14)$$

$$\beta_{ijkl} = \begin{cases} 1.6 & \text{if } ik \in E \text{ and } jl \in F \\ -0.3 & \text{else} \end{cases} \quad (15)$$

The Subnet Alignment Expansion

The biological information of human PPI network is more complete, and the number of proteins in each subnet is greater than that of *S. cerevisiae* S288c. Therefore, the potential functions of many *S. cerevisiae* proteins could be further annotated by referring to the human subnets and only expanding the corresponding *S. cerevisiae* ones. The expansion steps are as follows (similar to the whole network expansion) (**Figure 1C**):

(1) Let the aligned protein set of *S. cerevisiae* subnet be X and its neighbor in whole PPI network be X_1. If X_1 contains the proteins of X, remove them from X_1.

(2) Let the aligned protein set of human subnet be Y and its neighbor in BP subnet be Y_1. Compare X_1 and Y_1 adopting the same solution method as the whole network expansion.

(3) According to the alignment results, the new aligned proteins of X_1 are predicted to participate in the corresponding BP.

Performance Assessment of Network Alignment

Node coverage (NC) has been widely used (Milenković et al., 2010) to evaluate how well an alignment reconstructs the true node mapping, and it is defined as the percentage of aligned orthologs in the node set of the smaller network (U). In addition, global network aligners often fail to align all the nodes of the smaller network to the larger one. Thus, we also use global node coverage (GNC) to evaluate global alignments, which measures the number of mapped nodes normalized by the number of nodes in the smaller network (Malod-Dognin et al., 2017). To measure how well edges are conserved under an alignment, three measures have been used to date: edge correctness (EC) (Kuchaiev et al., 2010), induced conserved structure (ICS) (Patro and Kingsford, 2012), and symmetric substructure score (S^3) (Saraph and Milenković, 2014). S^3 has been shown to be superior to EC and ICS, since, intuitively, not only it penalizes alignments from sparse graph regions to dense graph regions (as EC does) but also it penalizes alignments from dense graph regions to sparse graph regions (as ICS does). Hence, we only focus on S^3 ($S^3 = \frac{|E_1^*|}{|E_1| + |E_2'| - |E_1^*|}$, where $|E_1^*|$ is the number of edges from

smaller network G_1 that are conserved by alignment, $|E_2'|$ is the number of edges from the induced subnet of G_2 with aligned node set). Besides, the alignment score S is also the criterion for assessing alignment comprehensively. We use GO correctness (GC) to evaluate the biological quality of an alignment. GC is defined as the percentage of aligned protein pairs that share at least k GO terms ($k = 1, 2, \ldots$) (Kuchaiev et al., 2010). In our study, we choose $k = 1$. Since the subnet pairs are divided by the BP terms, the GCs of all subnet alignments are 1.

In addition, we also assess the alignment performance by exploring the functions of aligned proteins. The aligned proteins with similar functions will get higher evaluation. The other indicator for alignment quality is the number of found common subnets representing the conserved functional clusters.

RESULTS

NAIGO Algorithm and Benchmark Datasets

The NAIGO algorithm has integrated divide-and-conquer strategy, optimization modeling, and graph-based features. When aligning large networks, it first divides them into multiple subnets based on the BPs of proteins and then locally aligns the corresponding subnets by constructing similarity matrices. The alignment problem is thus formulated as an assignment problem and could be solved by a polynomial time algorithm called the Hungarian method. The similarity matrices integrate the orthologous information and topological similarity information of the networks. It is worth mentioning that we added the edge matching information to the orthologous information. After all the subnet comparisons, we consider the largest interspecies aligned subgraphs as seeds and use the scoring scheme of graph alignment to expand them in a greedy manner. Then, we finally achieve the global alignment of large networks.

To test the NAIGO algorithm, human and *S. cerevisiae* S288c, two species separated by a long evolutionary distance, were picked out to perform the PPI network alignment. Although their PPIs have been widely explored, there is still a lack of studies on the PPI alignment between them, whether local or global. We downloaded the *S. cerevisiae* S288c data (BIOGRID-ORGANISM-Saccharomyces_cerevisiae_S288c-3.4.137.tab2.txt, see **Dataset S1**), human data (HPRD_PROTEIN_INTERACTIONS_Build9.txt, see **Dataset S1**), and the orthologous file between these two species from BioGRID (Andrew et al., 2015), Human Protein Reference Database (HPRD) (Keshava et al., 2009), and Ensembl (http://www.ensembl.org/biomart/martview), respectively.

The human PPI network consisted of 9,465 proteins and 39,240 interactions, among which there were a total of 2,201 self-interactions and repeated interactions. We denoted it by a graph $G = (V, E)$, in which $|V| = 9,465$ and $|E| = 37,039$ after removing the self-interactions and repeated interactions. In contrast, the *S. cerevisiae* S288c network consisted of 6,470 proteins and 342,123 interactions. We denoted it by $H = (U, F)$, in which $|U| = 6,470$ and $|F| = 228,703$. According to the BPs of proteins, the human and *S. cerevisiae* S288c network were divided into 6,781 and 1,836 non-empty subnets, respectively. Among them, there were

TABLE 1 | The influence of balancing parameters on the alignment.

θ_1	θ_2	Average NC	Average S^3
0	1	0.01	0.009
0.05	0.95	0.73	0.117
0.1	0.9	0.74	0.119
0.2	0.8	0.74	0.120
0.5	0.5	0.75	0.122
0.9	0.1	0.75	0.122
0.95	0.05	0.75	0.122
0.96	0.04	0.75	0.121
0.999	0.001	0.75	0.121
1	0	0.75	0.120
Defined θ_1	Defined θ_2	0.75	0.121

1,770 overlapping subnets in the two species after removing the included subnet.

According to the orthologous file, there were 3,871 ortholog pairs between V and U. Since many proteins have more than one ortholog in the other species, we finally confirmed 1,908 non-overlapping ortholog pairs by Hungarian algorithm. Each protein was involved in up to one non-overlapping ortholog pair, which guaranteed accurate and efficient alignments.

Effectiveness of Modified Algorithm and Defined Parameters

In NAIGO, the node and edge matching information were both introduced into the calculation of orthology (sequence-edge similarity), which was formulated into the similarity matrix together with the topological similarity. The parameters θ_1 and θ_2 were defined to balance contributions of nodes and edges.

This design ensured aligning more orthologous proteins on the matched edges and thus achieving better results. To evaluate the effectiveness of the integrative strategy and parameter setting, we randomly selected 30 subnet pairs. The interaction numbers of the smaller subnets ranged from 300 to 500. Adopting two common alignment scoring schemes, NC and S^3, we tested the performances of (1) setting one parameter to one and the other to zero (i.e., $\theta_1 = 0$ and $\theta_2 = 1$), (2) setting the continuously changing parameters ($0 < \theta_1 < 1$ and $\theta_2 = 1 - \theta_1$), and (3) defining parameters. The larger NC and S^3, the better the performance. As shown in **Table 1**, both the sequence-edge and topological similarities contributed to the node and edge alignments, and the former played more important roles in it. Meanwhile, the defining parameters got the better NC and S^3 performance. (For the defining parameters of 30 subnet pairs, see **Table S1**). The results demonstrated that the algorithm modification based on the sequence-edge similarity and defined parameters led to the good subnet alignment.

Performance Assessment of Local Alignment

We adopted the shared 1,770 BP subnets in the human and *S. cerevisiae* to evaluate the local alignment performance of the
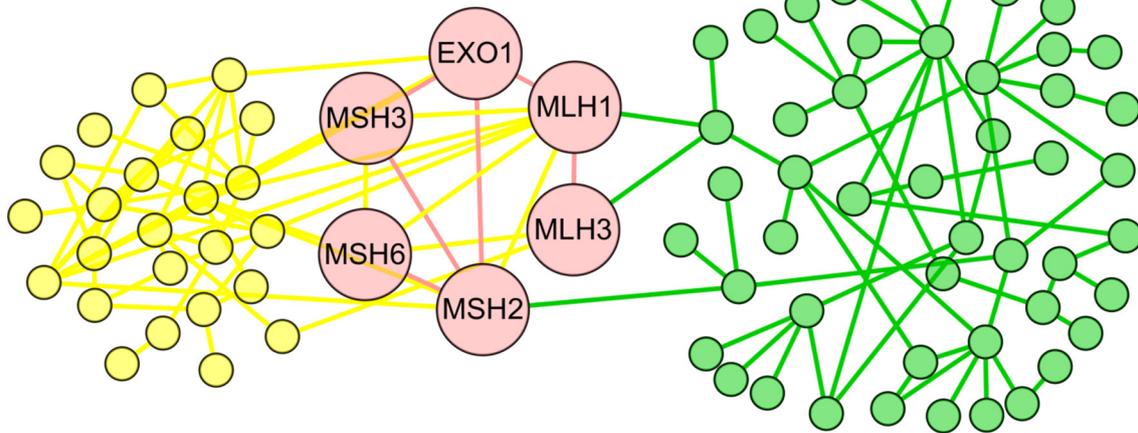
FIGURE 2 | The conserved subnet. The yellow nodes and edges represent the proteins and interactions only existing in the *Saccharomyces cerevisiae* S288c, and the green nodes and edges represent the proteins and interactions only existing in human. The red nodes and edges form the conserved subnet of the two species.

NAIGO algorithm. The mean value ($\pm$SD) of the NC is 0.57 ($\pm$0.35), and the mean value ($\pm$SD) of the S^3 is 0.30 ($\pm$0.33), the 1770 aligned subnets see **Dataset S2** - aligned subnets.

Within them, we achieved the conserved subnets in 1,240 aligned subnet pairs, which included at least two aligned edge pairs. We randomly selected several conserved subnets as examples (**Figure 2**, **Figures S1**, **S2**, and **Table S2**). As shown in **Figure 2**, the conserved subnet contained six orthologous proteins (EXO1, MSH3, MSH2, MLH1, MLH3, and MSH6) and six aligned interactions (EXO1-MSH3, EXO1-MSH2, EXO1-MLH1, MSH3-MSH2, MLH3-MLH1, and MSH6-MSH2). Refer to **Figure 3A** for the alignment between the subnets where the conserved subnets are located, and we know that both the human and *S. cerevisiae* S288c proteins in conserved subnets are related to mismatch repair.

We annotated the functions of the six orthologous proteins in the human and *S. cerevisiae* S288c (**Figures 3B,C**), displaying the similar patterns in the two species. Using ClueGO, a Cytoscape plugin, the orthologous proteins were annotated into 26 and 13 BP terms in the human and *S. cerevisiae*, respectively. Within them, we found that all the six orthologous proteins involved in the Mismatch repair in the two species. Furthermore, the BP terms could be classified into five clusters, including three consistent ones, such as mismatch repair (the enrichment *p*-value is 2.52E−16 in human and 5.44E−15 in *S. cerevisiae*, respectively), DNA replication (7.09E−10/1.21E−08), and DNA recombination (1.32E−09/1.19E−09). It should be noted that the biggest BP cluster in the human, B-cell-mediated immunity, was not identified in the *S. cerevisiae* S288c. However, it was due to the corresponding BP terms mainly correlating to the somatic diversification of immune receptors/immunoglobulins via recombination and mutation, which had evolved only in vertebrates. We also performed Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis on the six orthologous proteins in the human and *S. cerevisiae* S288c (**Figure S3**), and

the results also showed that they were enriched in mismatch repair pathway. We also performed KEGG enrichment analysis on conserved subnets in **Figures S1**, **S2** (see **Figures S4**, **S5**), and both subnet pairs had the same enrichment pathways. Therefore, these results indicated that the conserved subnets were functional conserved modules, and the NAIGO algorithm performed a correct local alignment.

In addition, the high performance of the NAIGO algorithm in the local alignment was also validated by annotating the functions of the non-orthologous protein pairs in the aligned subnets. We randomly selected five local alignments including 19 non-orthologous protein pairs and identified their functions by their BP terms and molecular function information on GeneCards. We totally found 10 non-orthologous protein pairs with similar functions (**Table 2**), indicating that these proteins were correctly aligned and should play similar roles in the two species.

Performance Assessment of Global Alignment

We also evaluate the NAIGO performance in the global alignment of the human and *S. cerevisiae* PPI networks. A total of 6,440 protein pairs have been aligned with the NAIGO algorithm, including the orthologous and non-orthologous protein pairs. It accounted for 68.04 and 99.54% of all proteins in the PPI networks of human and *S. cerevisiae* S288c, respectively. The NC of the global alignment is 0.21, the S^3 is 0.02, and the GC is 0.15 (**Table 3**), the aligned network see **Dataset S2** - aligned network.

In addition, the NAIGO algorithm connected several fragmented subgraphs of the human PPI network by the global alignment. Within the HPRD database, the human PPI network consisted of 110 fragmented subgraphs, each of which had at least one edge, and had no interaction with each other. Correspondingly, there were two fragmented subgraphs in the PPI network of the *S. cerevisiae* S288c, one of which only had three vertices and three edges. After the global alignment, the
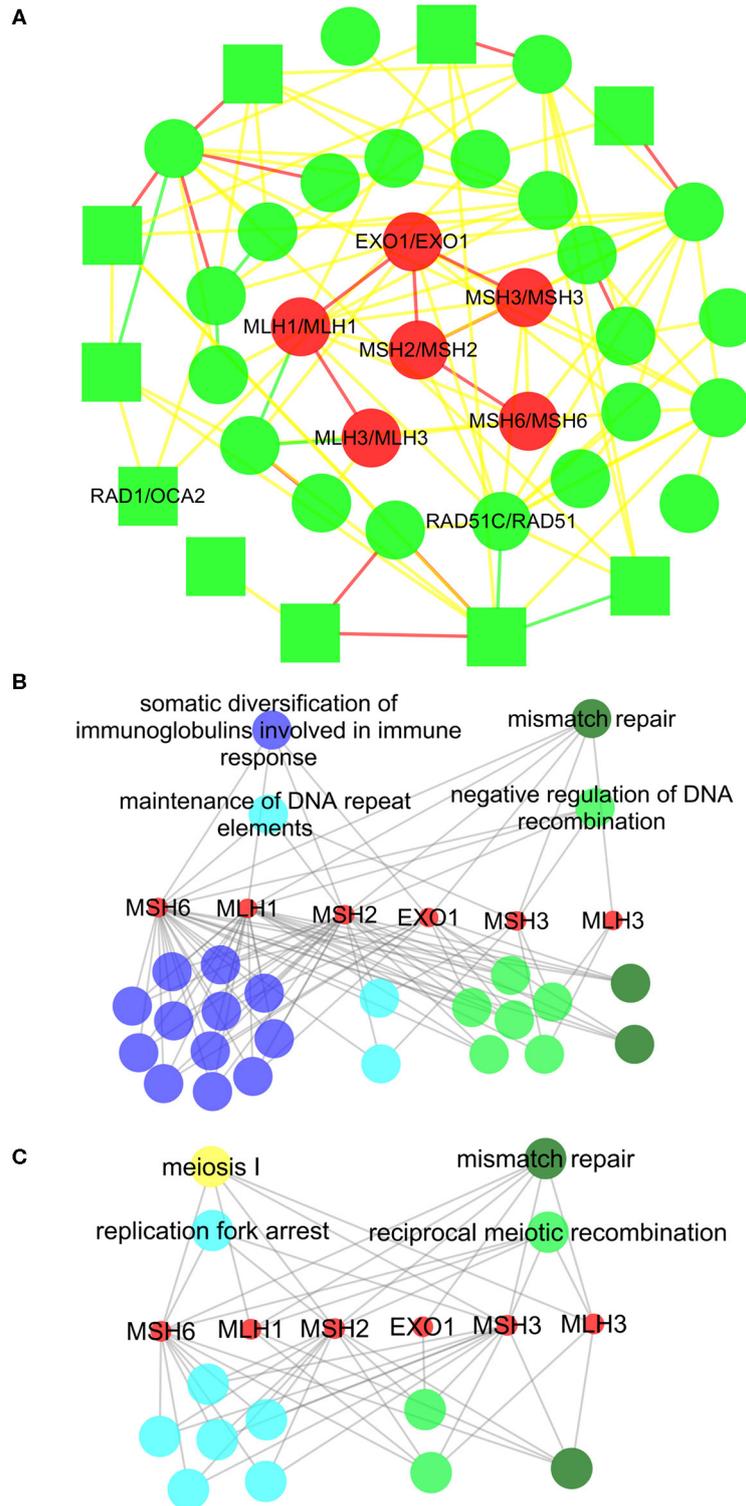
FIGURE 3 | The subnet alignment and the functional annotation of the six orthologous proteins in the conserved subnets. In the subnet alignment **(A)**, the round and square nodes represent the orthologous and non-orthologous protein pairs, respectively. If the human protein "A" aligned to the Saccharomyces cerevisiae S288c protein "B," then we represent the protein pair as "A/B." The yellow edges represent the interactions only existing in the Saccharomyces cerevisiae S288c, and the green edges represent those only existing in human. The red edges are the matched interactions. We colored the nodes in red based on broad GO term mismatch repair that had the most proteins. ClueGO is used to perform the gene enrichment analysis based on biological process (BP). The annotated BP groups and terms in
(Continued)

FIGURE 3 | the human **(B)** and Saccharomyces cerevisiae S288c **(C)**, with the statistical significance (P < 0.05), are displayed separately. Each node represents a BP term, and each color represents a BP group. The dark green group represents mismatch repair; the light blue group represents DNA replication; the light green group represents DNA recombination; the dark blue group represents B-cell-mediated immunity; and the yellow group represents meiosis. For simplicity, we only remark the most significant BP term of each group in the figures. The edges indicate the involved proteins in each functional item.

TABLE 2 | The functions of non-orthologous pairs.

Human/*Saccharomyces cerevisiae* proteins	BP terms	GeneCards functions
RRM2B/DNA2	DNA repair	Related to hereditary external ophthalmoplegia
LIG3/PIF1	NULL	DNA strand-break prevention and correction
RAD54L/CCT8	NULL	Telomere maintenance
RAD54B/REC114	Reciprocal meiotic recombination	Meiosis
AURKA/REC8	NULL	Structural maintenance of chromosome/spindle
RAD51/SPO11	Reciprocal meiotic recombination	NULL
TF/SCO1	NULL	Transportation
ATG16L1/ATG14	Protein transport Macroautophagy	Autophagy
GABARAP/ATG13	Autophagy Autophagosome assembly Macroautophagy	Neurotransmission
ARG2/CPS1	NULL	Urea metabolism

TABLE 3 | The predicted protein–protein interactions (PPIs) between subgraphs.

Predicted human interactions	Corresponding *Saccharomyces* interactions	Alignment type	Experimental evidence
VPS52-TSSK3	VPS52-AFT1	O-NO	*Science* paper (Menche et al., 2015)
VPS52-EPM2AIP1	VPS52-ERV41	O-NO	*Science* paper (Menche et al., 2015)
VPS52-COG3	VPS52-COG3	O-O	STRING database
VPS52-RAB6A	VPS52-YPT6	O-O	STRING database
VPS52-STX16	VPS52-TLG2	O-O	STRING database
VPS53-STX16	VPS53-TLG2	O-O	STRING database
VPS53-RAB6A	VPS53-YPT6	O-O	STRING database
USE1-STX5	USE1-SED5	O-O	STRING database
USE1-SCFD1	USE1-SLY1	O-O	STRING database
UBP1-ARHGAP21	UBP1-BEM2	O-O	STRING database

O-O represents the PPI alignment with orthologous protein pairs as both nodes. O-NO represents the PPI alignment with an orthologous protein pair as one node and a non-orthologous protein pair as the other node.

aligned PPI graph covered eight and one fragmented subgraphs of the human and *S. cerevisiae*, respectively. On the one hand, we deduced 10 PPIs among the eight human fragmented subgraphs by comparing with the aligned *S. cerevisiae* network. Interestingly, we found the experimental evidences for all the 10

predicted interactions by retrieving them in the String database (https://string-db.org/) and the PPI network data compiled by Menche et al. (2015), which contained much more PPIs (>140,000) than the HPRD database we used. The predicted interactions are listed in the **Table 3**, indicating that the PPI network alignment by NAIGO could correctly deduced unknown PPIs. On the other hand, the unmatched fragmented subgraphs in the human PPI network were small, with nodes less than five and edges less than four. As most of the proteins of the *S. cerevisiae* S288c had been aligned to the human ones, we deduced that those unmatched human subgraphs correlated with the multicellular organism associated functions that had not evolved in *S. cerevisiae* yet. Taken together, we concluded that NAIGO could achieve the global alignment with high quality and predict some unknown PPIs accordingly.

Comparison With Other Algorithms

To further assess the performance of the NAIGO algorithm, we compared it with four popular network alignment algorithms, IsoRank (Singh et al., 2008), GRAAL (Sharan et al., 2005), SANA (Mamano and Hayes, 2017), and NABEECO (Ibragimov et al., 2003). Within them, IsoRank calculates the network similarity by eigenvalue matrix and extracts the alignment, and GRAAL implements alignment only based on network topology structure. Both of the algorithms have a parameter to balance the contributions of the nodes and edges. Overall, NAIGO performed much better than the two algorithms. For IsoRank, we chose its default parameter to align the PPI networks of the human and *S. cerevisiae* S288c globally. As shown in **Table 4**, NAIGO achieved the larger NC/GNC, S^3/S and GC than IsoRank. For GRAAL, we first randomly selected 30 subnet pairs, since it is hard for GRAAL to align the large networks. These pairs were the same with those for assessing the effectiveness of the NAIGO algorithm, with the interaction numbers of the smaller subnets ranging from 300 to 500. The GRAAL's parameter was optimized in steps of 0.1 from zero to one, and the best performance is listed in **Table 4**. Comparing with NAIGO, GRAAL's average NC performance of the 30 subnet pairs was significantly inferior, but its average S^3 performance was better. Then, we compared the performances of the two methods only based on topology information (i.e., $\theta_1 = 0$ and $\theta_2 = 1$). The average NC of NAIGO was 0.01, which was better than that of GRAAL (0.002), but its average S^3 (0.009) was worse than GRAAL's (0.15). The results reflected the misses of many orthologous protein pairs due to the better topology performance of GRAAL algorithm, which should be more obvious in the network alignments with large size differences. To balance the NC and S^3 performances of GRAAL and thus compare with NAIGO better, we chose the other 15 subnet pairs. The protein numbers of the smaller subnets were >11,110, and the protein number differences of

TABLE 4 | Comparison of five methods on aligning the PPI networks of human and *Saccharomyces cerevisiae* S288c.

Method (the whole network pair)	NC/GNC	S^3/S	GC
IsoRank	0.19/0.762	0.004/−2.6393E + 09	0.13
SANA	0/1	0.07/−2.6392E + 09	0.14
NAIGO	0.21/0.995	0.02/−2.6392E + 09	0.15
Method (the first set of subnet pairs)			
GRAAL	0.002	0.15	1
NABEECO	0.36	0.17	1
NAIGO	0.75	0.12	1
Method (the second set of subnet pairs)			
GRAAL	0.05	0.22	1
NAIGO	0.64	0.24	1

TABLE 5 | The computational time of different methods.

Method	SANA	IsoRank	NABEECO	GRAAL	NAIGO
Time (h)	0.05	1.5	>72	>72	30

the protein pairs were <5. We also optimized the GRAAL's parameter in steps of 0.1 and found that GRAAL achieved much better NC performance without affecting the mean value of S^3 (**Table 4**). Nevertheless, the NAIGO's NC and S^3 performance in the alignment of these 15 subnet pairs were significantly better than GRAAL (**Table 4**). The performances of the two algorithms in global alignment and two sets of local alignment suggest that NAIGO performed substantially better than GRAAL.

SANA and NABEECO are recent algorithms. Among them, SANA is a Simulated Annealing Network Aligner that can produce the alignment in a short time. NABEECO is a novel and robust Network Alignment heuristic based on Bee Colony Optimization. For SANA, we align the PPI networks of the human and *S. cerevisiae* S288c globally. SANA's GNC and S^3 performance were better, but its NC and GC were inferior, especially that it missed all orthologous protein pairs. In addition, its largest connected component of aligned subgraph contained 6,118 nodes, while NAIGO contained 6,440 nodes, which indicated that the alignment result of NAIGO had stronger connectivity. For NABEECO, we choose the 30 subnet pairs to compare the performances because it needs to calculate the graphlet signature vector for long time when the network is large. Comparing with NAIGO, NABEECO's average NC performance of the 30 subnet pairs was significantly inferior, but its average S^3 performance was better. It reflected that NABEECO resulted in high topology but low functional quality.

We compare the computational time of different methods based on the PPI networks of the human and *S. cerevisiae* S288c. The comparison results are shown in **Table 5**.

Protein Function Deduction by Expanding Local Alignments

Due to the BP-based subnets, NAIGO could deduce the functions of certain proteins by expanding the local alignments. It was the

unique advantage of the NAIGO algorithm, which could not be achieved by most of the existing alignment tools.

For a subnet of *S. cerevisiae* S288c, we predict that its aligned neighbor proteins will participate in the corresponding BP; thus, BP and subnet expansion was achieved. In order to evaluate the performances of BP and subnet expansion, we randomly selected 100 BPs and performed David analysis on the original BPs and the corresponding expanded BPs, of which 13 BPs were more significant, and 10 BPs had lower *P*-values. This shows that the BP proteins of *S. cerevisiae* S288c that we predicted are of practical significance.

On the other hand, we randomly selected 1 of the 10 expanded subnets for gene enrichment analysis, the corresponding BP of which is autophagosome assembly. This subnet contains 7 nodes with 12 edges and 10 nodes with 29 edges after expansion, whose corresponding BP contains 74 proteins and 77 proteins after expansion. Refer to **Figure 4A** for the expanded subnet alignment, and there are three expanded protein pairs. We predicted that *S. cerevisiae* S288c proteins ATG8, ATG1, and ATG4 involved in the biological process autophagosome assembly, and this is indeed the case that the associated genes of this term in expanded subnet contain them (see **Tables S3, S4**). Based on the term group analysis in **Figures 4B,C**, we found that the term percentage of autophagosome assembly group increased but macroautophagy group reduced. From **Figure 5**, it can be seen that a new enrichment term organelle assembly appeared in the autophagosome assembly group; thereby, the term percentage of this group increased. This also shows that our expansion is practical. Besides, it also can be seen that the number of terms in the macroautophagy group is unchanged but the overall enriched terms increased, so the term percentage of this group reduced. Due to the addition of these three proteins, there are new groups such as late nucleophagy, C-terminal protein lipidation, and cellular response to nitrogen starvation that further illustrate that our subnet expansion algorithm is biologically significant.

DISCUSSION

In the study, we used BP as the standard for PPI network division, which made a preliminary restriction on the local alignment. It was helpful for finding more functional orthologous proteins and conserved functional modules. The BP associated PPI subnets were aligned by constructing pairwise similarity matrices of nodes. Considering that the network differences were mainly reflected in the node itself and the edge structure, we constructed the matrices based on the sequence similarity and graphlets similarity. In the related protein orthology matrix, we also added the matched edge information, in order to assess the sequence-edge similarity, and thus find more orthologous proteins and functional modules. By tuning the weight coefficients of the sequence-edge similarity and structure similarity, we found that the sequence-edge similarity made the major contribution to the local alignment. Moreover, the aligned proteins were annotated with the similar functions in the two species. In other words, NAIGO
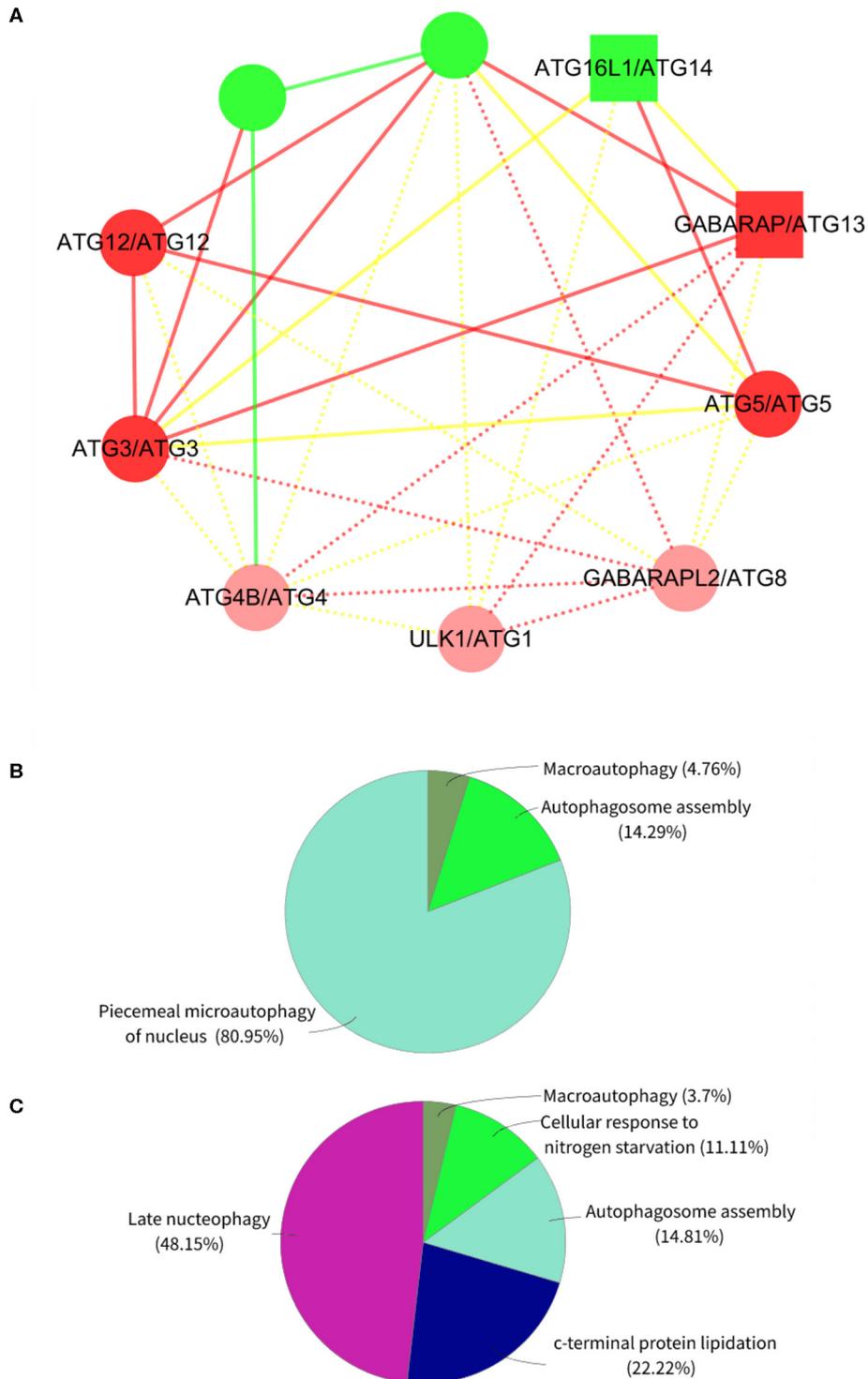
FIGURE 4 | The expanded subnet alignment and the term group analysis of subnet in the *Saccharomyces cerevisiae* S288c. In the expanded subnet alignment **(A)**, the round and square nodes represent the orthologous and non-orthologous protein pairs, respectively. If the human protein "A" aligned to the *Saccharomyces cerevisiae* S288c protein "B," then we represent the protein pair as "A/B." The yellow edges represent the interactions only existing in the *Saccharomyces cerevisiae* S288c, and the green edges represent those only existing in human. The red edges are the matched interactions. We colored the nodes in red based on broad GO term autophagosome assembly that had most proteins. In addition, we represent the expanded proteins and interactions in *Saccharomyces cerevisiae* as light red nodes and dotted lines, respectively. ClueGO is used to perform the gene enrichment analysis based on biological process (BP). Each annotated BP group and its corresponding term percentage (the term percentage of a group is the ratio of the number of terms in the group to the number of terms in all groups) of subnet **(B)** and expanded subnet **(C)** are displayed. Each color represents a group, and we remark the most significant BP term of each group in the figures.

FIGURE 5 | The gene enrichment of subnet in the *Saccharomyces cerevisiae* S288c. ClueGO is used to perform gene enrichment analysis on subnet **(A)** and expanded subnet **(B)** based on biological process (BP). The enrichment terms with statistical significance ($p < 0.05$) and the generated groups based on the relationships between terms are displayed separately. Each node represents a BP term, and each color represents a BP group. We remark the most significant BP term of each group in the figures. The edges indicate the relationships between the terms based on the similarity of their associated genes.

could achieve good alignments with biological significance by fully evaluating the orthology of the paired nodes and their 1-neighborhoods.

Based on the observations, we further expanded the BP-associated subnets with local alignments, and predicted the functions of newly aligned proteins accordingly. Because most

human subnets were larger than the corresponding ones in the *S. cerevisiae* S288c, we tried to map the 1-neighborhoods of the *S. cerevisiae* subnets to the unaligned proteins in the human subnets. The result showed that such expansion was functionally effective. There were two probable reasons: on the one hand, the bigger human subnets provide references; on the other hand, we only expanded one layer. Similarly, we expanded the local alignment of the largest subnets to the global alignment, mapping the 1-neighborhoods in two species to each other until 1-neighborhoods of one species was empty. It offered the possibility to connect the fragmented subgraphs of a PPI network.

Taken together, the NAIGO algorithm could achieve local alignment by similarity matrix integrating the sequence-edge similarity and graphlets similarity and achieve global alignment by expanding the local alignment of the largest subnets. The PPI network alignments in the human and *S. cerevisiae* S288c showed that NAIGO outperformed some popular algorithms by aligning more orthologous proteins or more protein interactions. Although we focused on the alignment of two PPI networks in this study, NAIGO could also compare any other types of biological networks, such as gene regulatory networks, metabolic networks, and so on. Our algorithm is suitable for networks with thousands of nodes and edges without affecting the alignment results, and the scalability is still relatively strong.

On the other hand, NAIGO take the advantage of known prior knowledge for alignment, such as gene ontology and protein orthologous information, which provide biological references for the network alignment to a certain extent (Seah et al., 2014). However, if the alignment method relies on the biological information very much, it may miss some functional orthologs for the network pairs with similar sizes and topologies. On the other hand, if the method only relies on the topology information for the network pairs with large size differences, it will also result in the alignments with low functional quality. Thus, finding the optimal combination of biological information and topology information is still what we need to explore in the future.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

AUTHOR CONTRIBUTIONS

JY and HZ conceived the project. LZ, JZ, and YZ implemented the experiments and analyzed the data. JL, JX, XB, NY, and GT prepared the data and performed literature search. LZ, JZ, and JY wrote the manuscript. All authors approved the final manuscript.

FUNDING

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00547/full#supplementary-material

NAIGO is available at https://github.com/xujunlin123/nacode and **Supplementary Material** - NAIGO-SW.

REFERENCES

Almaas, E. (2007). Biological impacts and context of network theory. *J. Exp. Biol.* 210, 1548–1558. doi: 10.1242/jeb.003731

Andrew, C. A., Bobby-Joe, B., Rose, O., Lorrie, B., Sven, H., Daici, C., et al. (2015). The BioGRID interaction database: 2015 update. *Nucleic Acids Res.* 43, D470–D478. doi: 10.1093/nar/gku1204

Berg, J., and Lässig, M. (2004). Local graph alignment and motif search in biological networks. *Proc. Natl. Acad. Sci. U.S.A.* 101, 14689–14694. doi: 10.1073/pnas.0305199101

Berg, J., and Lässig, M. (2006). Cross-species analysis of biological networks by bayesian alignment. *Proc. Natl. Acad. Sci. U.S.A.* 103, 10967–10972. doi: 10.1073/pnas.0602294103

Brian, P. K., Bingbing, Y., Fran, L., Roded, S., Brent, R. S., and Trey, I. (2004). PathBLAST: a tool for alignment of protein interaction networks. *Nucleic Acids Res.* 32, W83–W88. doi: 10.1093/nar/gkh411

Ciriello, G., Mina, M., Guzzi, P. H., Cannataro, M., and Guerra, C. (2012). AlignNemo: a local network alignment method to integrate homology and topology. *PLoS ONE* 7:e38107. doi: 10.1371/journal.pone.0038107

Dutkowski, J., and Tiuryn, J. (2007). Identification of functional modules from conserved ancestral protein-protein interactions. *Bioinformatics* 23, i149–i158. doi: 10.1093/bioinformatics/btm194

Faisal, F. E., Zhao, H., and Milenković, T. (2015). Global network alignment in the context of aging. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 12, 40–52. doi: 10.1109/TCBB.2014.2326862

Flannick, J., Novak, A., Srinivasan, B. S., McAdams, H. H., and Batzoglou, S. (2006). Graemlin: general and robust alignment of multiple large interaction networks. *Genome Res.* 16, 1169–1181. doi: 10.1101/gr.5235706

Hartwell, L. H., Hopfield, J. J., Leibler, S., and Murray, A. W. (1999). From molecular to modular cell biology. *Nature* 402, C47–C52. doi: 10.1038/35011540

Ibragimov, R., Martens, J., Guo, J., and Baumbach, J. (2003). "NABEECO: biological network alignment with bee colony optimization algorithm," in *GECCO 2013 Companion: Proceedings of the 15th Annual Conference Companion on Genetic and Evolutionary Computation*, ed C. Blum (New York, NY: Association for Computing Machinery), 43–44. doi: 10.1145/2464576.2464600

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. U.S.A.* 98, 4569–4574. doi: 10.1073/pnas.061034498

Keshava, P. T. S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., et al. (2009). Human protein reference database−2009 update. *Nucleic Acids Res.* 37, D767–D772. doi: 10.1093/nar/gkn892

Klau, G. W. (2009). A new graph-based method for pairwise global network alignment. *BMC Bioinformatics* 10:S59. doi: 10.1186/1471-2105-10-S1-S59

Kolár, M., Lässig, M., and Berg, J. (2008). From protein interactions to functional annotation: graph alignment in herpes. *BMC Syst. Biol.* 2:90. doi: 10.1186/1752-0509-2-90

Koyuturk, M., Kim, Y., Topkara, U., Subramaniam, S., Szpankowski, W., and Grama, A. (2006). Pairwise alignment of protein interaction networks. *J. Comput. Biol.* 13, 182–199. doi: 10.1089/cmb.2006.13.182

Krogan, N. J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., et al. (2006). Global landscape of protein complexes in the yeast saccharomyces cerevisiae. *Nature* 440, 637–643. doi: 10.1038/nature 04670

Kuchaiev, O., Milenkovic, T., Memisevic, V., Hayes, W. B., and Przulj, N. (2010). Topological network alignment uncovers biological function and phylogeny. *J. R Soc. Interface* 7, 1341–1354. doi: 10.1098/rsif.2010.0063

Kuchaiev, O., and PrŽulj, N. (2011). Integrative network alignment reveals large regions of global network similarity in yeast and human. *Bioinformatics* 27, 1390–1396. doi: 10.1093/bioinformatics/btr127

Lathrop, R. (1994). The protein threading problem with sequence amino acid interaction preferences is NP-complete. *Protein Eng.* 7, 1059–1068. doi: 10.1093/protein/7.9.1059

Li, Z. P., Zhang, S. H., Wang, Y., Zhang, X., and Chen, L. N. (2007). Alignment of molecular networks by integer quadratic programming. *Bioinformatics* 23, 1631–1639. doi: 10.1093/bioinformatics/btm156

Liang, Z., Xu, M., Teng, M., and Niu, L. (2006). NetAlign: a web-based tool for comparison of protein interaction networks. *Bioinformatics* 22, 2175–2177. doi: 10.1093/bioinformatics/btl287

Liao, C. S., Lu, K., Baym, M. H., Singh, R., and Berger, B. (2009). IsoRankN: spectral methods for global alignment of multiple protein networks. *Bioinformatics* 25, i253–i258. doi: 10.1093/bioinformatics/btp203

Malod-Dognin, N., Ban, K., and Przulj, N. (2017). Unified alignment of protein-protein interaction networks. *Sci. Rep.* 7:953. doi: 10.1038/s41598-017-01085-9

Mamano, N., and Hayes, W. (2017). SANA: simulated annealing far outperforms many other search algorithms for biological network alignment. *Bioinformatics* 33, 2156–2164. doi: 10.1093/bioinformatics/btx090

Menche, J., Sharma, A., Kitsak, M., Ghiassian, S. D., Vidal, M., Loscalzo, J., et al. (2015). Uncovering disease-disease relationships through the incomplete interactome. *Science* 347:1257601. doi: 10.1126/science.1257601

Milenković, T., Ng, W. L., Hayes, W. B., and PrŽulj, N. (2010). Optimal network alignment with graphlet degree vectors. *Cancer Inform.* 9, 121–137. doi: 10.4137/CIN.S4744

Mina, M., and Guzzi, P. H. (2014). Improving the robustness of local network alignment: design and extensive assessment of a markov clustering-based approach. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 11, 561–572. doi: 10.1109/TCBB.2014.2318707

Narayanan, M., and Karp, R. M. (2007). Comparing protein interaction networks via a graph match-and-split algorithm. *J. Comput. Biol.* 14, 892–907. doi: 10.1089/cmb.2007.0025

Neyshabur, B., Khadem, A., Hashemifar, S., and Arab, S. S. (2013). NETAL: a new graph-based method for global alignment of protein–protein interaction networks. *Bioinformatics* 29, 1654–1662. doi: 10.1093/bioinformatics/btt202

Patro, R., and Kingsford, C. (2012). Global network alignment using multiscale spectral signatures. *Bioinformatics* 28, 3105–3114. doi: 10.1093/bioinformatics/bts592

PrŽulj, N. (2007). Biological network comparison using graphlet degree distribution. *Bioinformatics* 23, e177–e183. doi: 10.1093/bioinformatics/btl301

Saraph, V., and Milenković, T. (2014). MAGNA: maximizing accuracy in global network alignment. *Bioinformatics* 30, 2931–2940. doi: 10.1093/bioinformatics/btu409

Seah, B. S., Bhowmick, S. S., and Dewey, C. F. (2014). DualAligner: a dual alignment-based strategy to align protein interaction networks. *Bioinformatics* 30, 2619–2626. doi: 10.1093/bioinformatics/btu358

Sharan, R., and Ideker, T. (2006). Modeling cellular machinery through biological network comparison. *Nat. Biotechnol.* 24, 427–433. doi: 10.1038/nbt1196

Sharan, R., Suthram, S., Kelley, R. M., Kuhn, T., McCuine, S., Uetz, P., et al. (2005). Conserved patterns of protein interaction in multiple species. *Proc. Natl. Acad. Sci. U.S.A.* 102, 1974–1979. doi: 10.1073/pnas.0409522102

Silva, E., and Stumpf, M. P. H. (2005). Complex networks and simple models in biology. *J. R. Soc. Interface* 2, 419–430. doi: 10.1098/rsif.2005.0067

Singh, R., Xu, J. B., and Berger, B. (2008). Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proc. Natl. Acad. Sci. U.S.A.* 105, 12763–12768. doi: 10.1073/pnas.0806627105

Srinivasan, B. S., Shah, N. H., Flannick, J. A., Abeliuk, E., Novak, A. F., and Batzoglou, S. (2007). Current progress in network research: toward reference networks for key model organisms. *Brief Bioinform.* 8, 318–332. doi: 10.1093/bib/bbm038

Towfic, F., Greenlee, M. H. W., and Honavar, V. (2009). "Aligning biomolecular networks using modular graph kernels," in *WABI 2009: Proceedings of the 9th International Conference on Algorithms in Bioinformatics*, ed S. L. Salzberg (Philadelphia, PA: Springer, University of Pennsylvania), 345–361. doi: 10.1007/978-3-642-04241-6_29

Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., et al. (2000). A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. *Nature* 403, 623–627. doi: 10.1038/35001009

Yang, J., Li, J., Grünewald, S., and Wan, X. F. (2013). BinAligner: a heuristic method to align biological networks. *BMC Bioinformatics* 14:S8. doi: 10.1186/1471-2105-14-S14-S8

Zaslavskiy, M., Bach, F., and Vert, J. P. (2009). Global alignment of protein–protein interaction networks by graph matching methods. *Bioinformatics* 25, i259–i267. doi: 10.1093/bioinformatics/btp196