ModuleAlign: module-based global alignment of protein–protein interaction networks

Somaye Hashemifar, Jianzhu Ma, Hammad Naveed, Stefan Canzar* and Jinbo Xu*

Toyota Technological Institute at Chicago, Chicago, IL 60637, USA

*To whom correspondence should be addressed.

Abstract

**Motivation:** As an increasing amount of protein–protein interaction (PPI) data becomes available, their computational interpretation has become an important problem in bioinformatics. The alignment of PPI networks from different species provides valuable information about conserved subnetworks, evolutionary pathways and functional orthologs. Although several methods have been proposed for global network alignment, there is a pressing need for methods that produce more accurate alignments in terms of both topological and functional consistency.

**Results:** In this work, we present a novel global network alignment algorithm, named ModuleAlign, which makes use of local topology information to define a module-based homology score. Based on a hierarchical clustering of functionally coherent proteins involved in the same module, ModuleAlign employs a novel iterative scheme to find the alignment between two networks. Evaluated on a diverse set of benchmarks, ModuleAlign outperforms state-of-the-art methods in producing functionally consistent alignments. By aligning Pathogen–Human PPI networks, ModuleAlign also detects a novel set of conserved human genes that pathogens preferentially target to cause pathogenesis.

**Availability:** http://ttic.uchicago.edu/~hashemifar/ModuleAlign.html

**Contact:** canzar@ttic.edu or j3xu.ttic.edu

**Supplementary information:** Supplementary data are available at Bioinformatics online.

1 Introduction

Protein–protein interaction (PPI) networks provide insight into the complex organization of biological processes in a cell at the system level. They are an effective tool for understanding the comprehensive map of functional interactions, and for identifying functional modules and pathways. The alignment of PPI networks facilitates the detection of evolutionary and functionally conserved pathways or complexes and the prediction of protein function.

PPI network alignments can be either local or global. The aim of local network alignment methods such as NetworkBlast (Sharan et al., 2005) and AlignNemo (Ciriello et al., 2012) is to find small and dense subnetworks corresponding to a pathway or motif. In contrast, global network alignment (GNA) aims to maximize the overall match between the input networks. GNA methods such as IsoRank (Singh et al., 2007, 2008), GRAAL family aligners (Kuchaiev et al., 2010; Kuchaiev and Pržulj, 2011; Malod-Dognin and Pržulj, 2012; Milenković et al., 2010), NETAL[10], GHOST[11], HubAlign (Hashemifar and Xu, 2014), MAGNA++ (Vijayan et al., 2015) and L-GRAAL[9] are designed for pairwise alignments while others such as ConvexAlign (Hashemifar et al., 2016) and IsoRankN[15] are able to align multiple networks.

IsoRank is the pioneer algorithm for aligning PPI networks globally. It is based on the eigenvalue concept of local neighborhood alignments. MI-GRAAL[5] and L-GRAAL[9] integrate network topology information such as graphlet signature and sequence similarity to align two proteins. L-GRAAL is unique in that it takes into account the mapped interactions and uses an integer programming method to compute the alignment. GHOST uses graph spectrum to assess the neighborhood similarity of proteins and finds an alignment through quadratic programming. NETAL is the first global aligner that iteratively updates the topological information during the alignment. HubAlign uses a minimum-degree heuristic to estimate the functional and topological importance of a protein from the global network topology. It aligns the topologically important proteins first and iteratively extends the alignment to the whole network. MAGNA++ uses a genetic algorithm to search for the optimal alignment.

Although several GNA methods have been developed, there is still a need to improve the scoring function, alignment strategy and
computational efficiency. Furthermore, current GNA methods do not fare well in detecting proteins that participate in the same pathways, i.e. are produced from orthologous genes.

This work introduces ModuleAlign, a novel pairwise global network alignment approach. Its novel scoring scheme integrates sequence information and both local and global network topology. Based on a hierarchical clustering of the input networks, we compute a homology score between proteins. We propose an iterative approach to find an alignment that scores high in our model while trying to preserve interactions. In our experiments on a diverse set of benchmarks, ModuleAlign outperforms state-of-the-art methods such as GHOST, MAGNA++, NETAL, HubAlign and LGRAAL in terms of both alignment accuracy and functional consistency.

2 Method

Main idea: Proteins with similar sequences are more likely to have a similar function (Jaroszewski et al., 1998). Most of the network alignment methods thus consider sequence similarity (e.g. blast bit score) to detect homology between proteins. However, this score is not sensitive enough for remote homologs such that they may miss functionally similar proteins. In addition to the sequence of a protein, its interaction partners within a module or cluster indicate its function. Therefore, the homology of the partners of a gene in its corresponding modules can help to assess the homology of two proteins. In this work, we introduce a homology score that takes into account the local neighborhood within hierarchical clusters of a PPI net-works. This score is based on the observation that proteins with similar functions tend to form densely connected subnetworks (Park and Bader, 2011). We combine the homology score and a previously introduced topology score (Hashemifar and Xu, 2014) into an alignment score and propose a novel iterative strategy to optimally match proteins. Starting from an optimal bipartite matching of proteins, ModuleAlign iteratively selects highest-scoring protein pairs and adjusts alignment scores in their neighborhood to promote conservation of interactions.

Definitions: We represent a protein–protein interaction network by an undirected graph \( G = (V, E) \) where nodes in \( V \) represent proteins and edges in \( E \) their interactions. Let \( N(u) \) represent the neighborhood of node \( u \) in \( G \) and \( \deg(u) \) the degree of node \( u \), i.e. \( \deg(u) = |N(u)| \). A global alignment between two networks \( G_1(V_1, E_1) \) and \( G_2(V_2, E_2) \) is an injective function \( g : V_1 \rightarrow V_2 \) that maps node set \( V_1 \) to \( V_2 \). Let \( M \) be the set of aligned proteins, i.e. \( M = \{ (u, g(u)) : u \in V_1, g(u) \in V_2 \} \). The hierarchical structure of network \( G \) is represented by a set of clusters \( C = \{ c_1, c_2, \ldots, c_m \} \). Each cluster \( c_l \) consists of a subset of proteins in the corresponding network.

2.1 Hierarchical clustering of the network

It is well known (Dutkowski et al., 2013) that PPI networks have a hierarchical structure, which can be represented by a binary tree with leaves corresponding to the proteins and each internal non-leaf node to the clusters (see Fig. 1a). Each cluster contains the proteins at the leaves of the subtree rooted at the corresponding node, split between its left and right child. We use HAC-ML (Park and Bader, 2011) to infer the binary tree underlying a given network in which clusters have two main properties: (i) proteins within a cluster are relatively more densely connected than proteins in different clusters. (ii) Every cluster, except for those close to the root, corresponds to a specific network motif with its proteins performing similar functions (Park and Bader, 2011).

2.2 Novel cluster-based homology score

After determining the clusters in each input network, we calculate the similarity between each pair of clusters \( c_l \) and \( c'_l \) of \( G_1 \) and \( G_2 \), respectively. Knowing that proteins within each cluster have similar functions and thus their sequences are expected to show a certain degree of similarity, we compute the similarity \( \text{clusterSim}(c_l, c'_l) \) between a pair of clusters as the average blast score between contained proteins:

\[
\text{clusterSim}(c_l, c'_l) = \frac{\sum_{u \in c_l} \sum_{v \in c'_l} \text{blastSim}(u, v)}{|c_l| \times |c'_l|},
\]

where \( \text{blastSim}(u, v) \) is the blast score between proteins \( u \) and \( v \). We remove very high-level clusters (i.e. those on level higher than 3) as they contain a large number of genes. Note that \( \text{clusterSim} \) measures sequence similarity of proteins within functional modules of the network (see (Park and Bader, 2011)). This strategy implicitly takes into account network structure and avoids the blurring of the homology signal by comparing unrelated (or weakly related) proteins. At the same time, clusters that are similar according to \( \text{clusterSim} \) contain proteins with both similar sequences (definition of \( \text{clusterSim} \)) and similar interaction neighborhoods (definition of clusters by (Park and Bader, 2011)).

We use the hierarchical clusters to define a new homology score between proteins that does not solely rely on sequence information. We define the homology score between proteins \( u \) and \( v \) based on the similarity (as measured by \( \text{clusterSim} \)) between all clusters that contain \( u \) and \( v \):

\[
\text{HS}(u, v) = \frac{\sum_{c_l \in C_1 : u \in c_l} \sum_{c'_l \in C_2 : v \in c'_l} \text{clusterSim}(c_l, c'_l)}{|C_1| \times |C_2|},
\]

Intuitively, two proteins that belong to many similar clusters are expected to have a similar function.

2.3 Novel alignment strategy

We propose an alignment score between proteins that combines our homology score with a score based on global topological similarity. The topology scores are calculated in the same way as in HubAlign.
(Patro and Kingsford, 2012). For each node and every edge in the network, HubAlign calculates weights that it uses to infer the topological importance $S(v)$ for all nodes $v \in V$. Topological scores are normalized by $\max_{v \in V} S(v)$. Finally, HubAlign defines the topological similarity score $T_S(u, v)$ between two proteins $u \in V_1$ and $v \in V_2$ as the minimum of $S(u)$ and $S(v)$. We define the alignment score as follows:

$$A(u, v) = x \times H(u, v) + (1 - x)TS(u, v),$$

where $0 \leq x \leq 1$ is a tradeoff parameter that controls the contribution of global topological similarity relative to the homology score. In our implementation, we empirically set $x$ to 0.4 by default.

Our alignment strategy relies on a two-step process (see Fig. 1): In the first step, we apply the Hungarian method to compute an optimal matching $M_0$ of proteins in the two networks with respect to alignment scores. The Hungarian method is a primal–dual algorithm that starts with an empty matching and iteratively increases the size of the matching using maximum-weight augmenting paths. We refer the interested to read (Schrijver, 2003) for details on the Hungarian algorithm. The goal of the first step is to align proteins that are topologically and functionally consistent. In the second step, we additionally try to maximize the number of evolutionary conserved interactions using the initial alignment as our guidance. Starting from $M_0$, we iteratively fix the heaviest edge $(u_0, v_0)$ in the current alignment, i.e. the pair of proteins with largest alignment score. For each neighbor $u$ of $u_0$, we remove $(u, g(u))$ from the current alignment and upweight the alignment score between $u$ and all neighbors $v$ of $v_0$:

$$\forall v \in N(u_0), A(u, v) = A(u, v) + 1/\max_{v' \in V_1 \setminus V_2} S(u')$$

where $\max_{v' \in V_1 \setminus V_2} S(u')$ is the same normalization factor as used in the topology score. Intuitively, we increase the alignment score by one normalized unit since aligning neighbors $u$ of $u_0$ and $v$ of $v_0$ would yield one additional conserved interaction. Then, one primal–dual iteration of the Hungarian algorithm is performed to re-optimize w.r.t. the updated alignment scores. This procedure is repeated until all proteins of the smaller network are matched by the final alignment $M'$. See Algorithm 1 for more details. Without loss of generality, we assume $|V_1| < |V_2|$.}

### 2.4 Time complexity

Let $n = \max(|V_1|, |V_2|)$. It takes $O(n^3 \log(n))$ to compute the hierarchical clustering of the input networks. Moreover, calculating the topological scores takes $O(n^2)$. Thus, the time complexity for finding the alignment scores is $O(n^3 \log(n))$. The first step of the alignment strategy takes $O(n^3)$. In the second step, finding the best scoring pair takes $O(n)$ and updating the alignment scores can be done in $O(n^2)$. Finally, running one iteration of the Hungarian method takes $O(n)$. Therefore, the time complexity of Algorithm 1 is $O(n^3)$, yielding an overall time complexity of ModuleAlign of $O(n^3 \log(n))$.

### 3 Result

We compare ModuleAlign with several popular and publically available global network alignment methods NETAL[10], GHOST[11], HubAlign (Hashemifar and Xu, 2014), MAGNA++ (Vijayan et al., 2015), and L-GRAAL[5]. These methods have been shown to outperform other methods such as IsoRank and MI-GRAAL on several datasets (Hashemifar and Xu, 2014; Malod-Dognin and Pržulj, 2015). As recommended by the authors, we configured the genetic algorithm parameter of MAGNA++ to optimize the $S^3$ score (see definition below), running over 15 000 generations with a population size of 2000. Parameters of other methods are set to their default values. We evaluate the network alignment quality by several functional consistency and topological measures proposed in different studies (Hashemifar and Xu, 2014; Kuchaiev et al., 2010; Saraph and Milenković 2014). Functional consistency metrics are particularly important in the context of network alignments since one of their main purposes is the transfer of functional annotations and modules between networks.

We employ KEGG Orthology (KO) annotations to measure the functional consistency of an alignment and examine the conservation of pathways by an alignment. KO annotations integrate pathway and genomic information in KEGG which is well known for its comprehensive pathway database (Kanehisa et al., 2011). About 68% of the genes in our benchmark are annotated by at least one KO. KO represents a group of orthologous genes and its direct link to KEGG allows for the identification of pathways that might provide therapeutic targets. The following functional consistency measures are based on the observation that orthologous genes often share identical KO annotations. In the following, a ‘class’ consists of a pair of aligned proteins.

**Precision:** We say a class is annotated if both of its proteins have KO annotations assigned. An annotated class is consistent if both of its proteins share at least one common KO annotation. Precision is defined as the fraction of proteins in consistent classes among proteins in annotated classes.

**Recall:** Recall is defined as the total number of proteins in consistent classes divided by the total number of proteins assigned at least one KO annotation.

Additionally, we employ a measure based on gene ontology (GO) terms. GO terms describe biological properties of proteins such as biological process (BP), molecular function (MF) and...
cellular component (CC). We exclude root GO terms from the analysis. About 75% of genes in our benchmark are annotated with at least one GO term. Based on the observation that proteins with similar GO terms are more likely to be functionally related, we apply the following quality measure.

**Average of functional similarity (AFS):** AFS is based on the semantic similarity of GO terms, which depends on their distance in the ontology. We use Schlicker’s similarity based on the Resnik ontological similarity to calculate the functional similarity (see Supplementary Section A). Let \( s_c(u, v) \) denote the GO functional similarity of proteins \( u \) and \( v \) in category \( c \) (i.e., BP, MF or CC). AFS of an alignment in category \( c \) is defined as follows (Hashemifar and Xu, 2014):

\[
AFS_c = \frac{1}{|V_1|} \sum_{u \in V_1} s_c(u, g(u)).
\]

We also evaluate the topological quality of an alignment by the following measures.

**Edge correctness (EC):** EC is defined as the percentage of edges in the first network that are preserved by the alignment (Kuchaiev et al., 2010). Let \( g : V_1 \rightarrow V_2 \) be an alignment between networks \((V_1, E_1)\) and \((V_2, E_2)\). Then,

\[
EC = \frac{|\{(u, v) \in E_1 : (g(u), g(v)) \in E_2\}|}{\min(|E_1|, |E_2|)} \times 100
\]

**Symmetric substructure score \((S^3)\):** \(S^3\) penalizes the alignment of sparse regions to denser ones and vice versa. A larger \(S^3\) indicates the ability of an aligner to map sparse regions onto each other. Let \( G[V] \) denote the subnetwork of \( G \) induced by node set \( V \) and let \( E(G) \) represent the edge set of network \( G \). Let \( f(V_1) = \{g(v) \in V_2 : v \in V_1\} \) and \( f(E_1) = \{(g(u), g(v)) \in E_2 : (u, v) \in E_1\} \). \( S^3\) is defined as follows (Vijayan et al., 2015):

\[
S^3 = \frac{|f(E_1)|}{|E_1| + |E(G_2[f(V_1)])|} + \frac{|f(E_1)|}{|E_1|} \times 100
\]

**Largest common connected subgraph (LCCS):** LCCS is calculated as the number of edges in the largest common connected subgraph that is preserved under the alignment (Kuchaiev et al., 2010). Larger and denser subgraphs provide more insights into the common topology between networks and are biologically more meaningful (Hu et al., 2005).

### 3.1 Alignment quality

The test data includes five PPI networks for *Homo sapiens* (human), *Saccharomyces cerevisiae* (yeast), *Drosophila melanogaster* (fly), *Caenorhabditis elegans* (worm) and *Mus musculus* (mouse) obtained from HINT[23], HINT integrates the interactions from several databases, including BioGRID[24], IntAct (Kerrien et al., 2011), and MINT[26], and manually removes erroneous interactions. The sizes of these networks are shown in Supplementary Table S1.

Figure 2 shows precision and recall of the alignments generated by the different methods. ModuleAlign significantly outperforms all other methods and predicts consistent classes with higher accuracy.

The consistent classes provide valuable information concerning the orthologous relationship of proteins from the two species. Taking this a step further, consistent classes allow the detection of evolutionary pathways conserved between species. In the yeast–human alignment produced by ModuleAlign, we find many conserved pathways such as RNA transport and RNA degradation. Tables 1 and 2 show the consistent classes whose KO is linked to the RNA transport pathway (i.e. ‘hsa03013’ in human and ‘sce03013’ in yeast) and to RNA degradation (i.e. ‘hsa03018’ in human and ‘sce03018’ in yeast), respectively. The definitions for all KO can be found at [27].

To measure the AFS of an alignment, we extract the GO terms from the Gene Ontology database (Botstein et al., 2000). We only consider aligned pairs in which both proteins have a GO term assigned. As shown in Figure 3 and Supplementary Figure S2, ModuleAlign yields better alignments than other methods in terms of AFS in categories BP, MF and CC. ModuleAlign finds more functionally consistent protein pairs for all pairs of networks which facilitate the identification of conserved functional modules (see Fig. 5).

Figure 4 shows that ModuleAlign produces alignments with larger EC and LCCS values than other methods except for NETAL. A larger EC value indicates that ModuleAlign is able to map densely connected proteins that potentially belong to similar structural or functional modules. While NETAL yields more aligned interactions than ModuleAlign, it is among the methods with lowest AFS indicating that many of the aligned proteins are not functionally similar (see Fig. 3). Similarly, Supplementary Figure S1 compares the \(S^3\) score achieved by the different methods. For the majority of the instances, ModuleAlign achieves the second best score after NETAL, demonstrating ModuleAlign’s ability to preserve sparse regions. Again, NETAL achieves a higher \(S^3\) score, but at the cost of a very low AFS (see Fig. 3). Overall, ModuleAlign is among the best aligners with respect to topological measures. More importantly, ModuleAlign aligns proteins with a substantially higher functional consistency than all competing methods and thus facilitates the transfer of functional annotations.

### 3.2 Finding conserved sub-networks

A major application of network alignments is the identification of conserved sub-networks across two species. Figure 5 shows two conserved sub-networks between yeast and human detected by
Table 1. Consistent classes in the yeast–human alignment that identify the RNA transport pathway

| Yeast  | Human    | KO ID  | KO title                                           |
|--------|----------|--------|---------------------------------------------------|
| SUB2   | DDX39B   | K12812 | ATP-dependent RNA helicase UAP56/SUB2             |
| THO2   | THOC2    | K12879 | THO complex subunit 2                             |
| STO1   | NCBP1    | K12882 | Nuclear cap-binding protein subunit 1             |
| CBC2   | NCBP2    | K12883 | Nuclear cap-binding protein subunit 2             |
| FAL1   | EIF4A3   | K13025 | ATP-dependent RNA helicase                        |
| SEC13  | SEC13    | K14004 | Protein transport protein SEC13                   |
| MLP1   | TPR      | K09291 | Nucleoprotein TPR                                  |
| TIF5   | EIF5     | K03262 | Translation initiation factor 5                   |
| RPF1   | EIF3A    | K03254 | Translation initiation factor 3 subunit A         |
| TIF34  | EIF3I    | K03246 | Translation initiation factor 3 subunit I         |
| GCD2   | EIF2B4   | K03680 | Translation initiation factor eIF-2B subunit delta |
| SU13   | EIF2S2   | K03238 | Translation initiation factor 2 subunit 2         |
| TFI1/TIF2 | EIF4A2  | K03257 | Translation initiation factor 4A                  |
| TRZ1   | ELAC2    | K00784 | Ribonuclease Z                                    |
| NAM7   | UPF1     | K14326 | Regulator of nonsense transcripts 1               |
| NMD2   | UPF2     | K14327 | Regulator of nonsense transcripts 2               |

Table 2. Consistent classes in the yeast–human alignment that identify the RNA degradation pathway

| Yeast  | Human    | KO ID  | KO title                                           |
|--------|----------|--------|---------------------------------------------------|
| CDC39  | CNOT1    | K12604 | CCR4-NOT transcription complex subunit 1           |
| POP2   | CNOT8    | K12581 | CCR4-NOT transcription complex subunit 7/8         |
| SSC1   | HSPA9    | K04043 | Molecular chaperone DnaK                           |
| XRN1   | XRN1     | K12618 | 5’-3’-Exoribonuclease 1                            |
| DIS3   | DIS3     | K12585 | Exosome complex exonuclease DIS3/RRP44             |
| CDC36  | CNOT2    | K12605 | CCR4-NOT transcription complex subunit 2           |
| RRP46  | EXOSC5   | K12590 | Exosome complex component RRP46                    |
| RRP45  | EXOSC9   | K03678 | Exosome complex component RRP45                    |
| RAT1   | XRN2     | K12619 | 5’-3’-Exoribonuclease 2                            |

Fig. 3. Performance of all methods in terms of AFS in categories BP and MF

Fig. 4. Performance of ModuleAlign and competing methods in terms of EC and LCCS
ModuleAlign is among the fastest global alignment methods. We report the running times of all methods for the alignment of the bacterial PPI networks.

Table 3. EC, S^3 and AFS of the alignments on the bacterial PPI networks

| Precision | Recall | AFS (BP) | AFS (MF) | EC   | S^3 |
|-----------|--------|----------|----------|------|-----|
| NETAL     | 0      | 0        | 0.15     | 0.09 | 32.30 | 19.54 |
| GHOST     | 0.05   | 0.04     | 0.19     | 0.14 | 22.79 | 15.14 |
| HubAlign  | 0.31   | 0.24     | 0.25     | 0.22 | 24.65 | 16.51 |
| MAGNA++   | 0      | 0        | 0.24     | 0.18 | 24.83 | 19.32 |
| L-GRAAL   | 0.11   | 0.09     | 0.12     | 0.10 | 24.61 | 17.83 |
| ModuleAlign | 0.37  | 0.30     | 0.31     | 0.28 | 25.95 | 16.92 |

Best results are indicated in bold.

代谢途径。我们使用ModuleAlign来对各病原体-人类PPI网络进行对齐，以发现现有PPI网络中的蛋白质。我们使用病原体-人类PPI网络来对齐，用于对齐中的蛋白质。PANTHER是一个全面的软件系统，用于根据基因与病原体的交互作用来发现功能相等的基因。PANTHER是一个全面的软件系统，用于对齐其中的蛋白质。

对于三个对齐，我们生成了六个集群集（每个对齐可以对应到各自的病原体-人类PPI网络）。所有六套簇都证实病原体目标了人类免疫响应和代谢途径。而且，对齐时我们发现病原体目标了与运输相关的蛋白质。具体来说，Sec31A和NPM1是出现在五个最大的集群在所有对齐生成的。

3.4 Detecting conserved clusters in pathogen–human PPI networks

*Bacillus anthracis*, *Franciscella tularensis* and *Yersinia pestis* can suppress or evade immune response to cause pathogenesis (*Dyer et al., 2010*). They are known to interact with human proteins involved in the immune response pathway and to a lesser extent with the metabolism pathway. We have used ModuleAlign to align the pathogen–human PPI networks to discover the function of human proteins present in the clusters conserved between these pathogen–human PPI networks. We used pathogen–human PPI networks for *B. anthracis*, *F. tularensis* and *Y. pestis* that were constructed using the yeast two-hybrid assay by *Dyer et al., 2010*.

We ran ModuleAlign on each pair of pathogen–human PPI networks. In each computed alignment, we select the five largest connected components and submit them to PANTHER[33] to assign GO biological functions to the human genes that interact with pathogen genes. PANTHER is a comprehensive software system for inferring the functions of genes based on their evolutionary relationships. The function of the human genes contained in our top ranked clusters is reported at https://uchicago.box.com/ModuleAlign. In the statistical analysis we apply Bonferroni correction for multiple hypothesis testing. Only a biological function that shows at least 1.5-fold enrichment is reported.

For the three alignments, we generated six sets of clusters (each alignment can be mapped to either of the pathogen–human PPI networks). All six sets of clusters confirm that the pathogens target the human immune response and metabolism pathways. Moreover, for the first time we discovered that the pathogens target/interact with transport related proteins. Specifically, genes Sec31A and NPM1 that are present in the five largest clusters in all alignments generated by ModuleAlign are conserved across the three pathogen-human PPI. Sec31A is involved in ER to Golgi vesicle-mediated transport (GO:0006888) and protein transport (GO:0015031), and NPM1 is involved in intracellular protein transport (GO:0006886) and nucleocytoplasmic transport (GO:0006913). Similarly, genes MCM3AP and TINAGL1 are conserved across the three pathogen–human PPI networks and are present in the five largest clusters in all alignments.

3.3 Alignment of bacterial PPI networks

We also ran ModuleAlign to align the PPI networks of the two bacterial species *Escherichia coli* and *Campylobacter jejuni* for which the most complete PPI networks among all bacteria exist. The *E. coli* and *C. jejuni* PPI networks have 1941 nodes, 3989 edges, 1111 nodes and 2988 edges, respectively (*Peregrin-Alvarez et al., 2009; Parrish et al., 2007*). *Escherichia coli* is a model organism for studying fundamental cellular processes such as gene expression and signaling. Table 3 shows that ModuleAlign achieves a significantly higher precision and recall than all other methods. ModuleAlign also outperforms other methods in terms of AFS whose absolute values are small due to insufficient GO annotations of bacterial proteins. This again indicates that ModuleAlign is able to align proteins with consistent functions.
human with the yeast and fly networks, which are the largest networks in our benchmark. On a 1400 MHz Linux system with 2GB RAM, it takes NETAL, HubAlign, MAGNA++, L-GRAAL, GHOST and ModuleAlign 4, 15, 621, 64, 41 and 15 min, respectively, to align the yeast and human networks, and 7, 25, 757, 79, 50 and 30 min, respectively, to align the fly and human networks.

5 Discussion

We present ModuleAlign, a novel method for pairwise global alignment of PPI networks. ModuleAlign for the first time exploits module information to compute a homology score between proteins. It uses a novel strategy to find the best network alignment under a carefully designed scoring function. We have tested ModuleAlign on several PPI networks. We evaluated the output alignments using different performance measures and compared them with state-of-the-art global network alignments methods such as NETAL, GHOST, HubAlign, MAGNA++ and L-GRAAL. Experimental results show that ModuleAlign not only generates functionally consistent alignments efficiently but also is tolerant to partly missing interactions, a useful property in the context of incomplete PPI networks. ModuleAlign also finds several conserved and biologically important modules, which are missed by all other alignment methods. Moreover, network pharmacology is considered to be the next paradigm in drug discovery, where integrating network biology and poly-pharmacology is going to be the key to discovering drug targets (Hopkins, 2008). Our improved network alignment algorithm can help in transferring the results of network pharmacology studies from model organism to humans.

Funding

This work is supported in part by NSF CAREER award CCF-1149811.

Conflict of Interest: none declared.

References

Botstein, D. et al. (2000) Gene ontology: tool for the unification of biology. Nat. Genet., 25, 25–29.

Chatr-Aryamontri, A. et al. (2013) The BioGRID interaction database: 2013 update. Nucleic Acids Res., 41, D816–D823.

Ciriello, G. et al. (2012) Alignnemo: a local network alignment method to integrate homology and topology. PLoS One, 7, e38107.

Das, J. and Yu, H. (2012) HINT: high-quality protein interactomes and their applications in understanding human disease. BMC Syst. Biol., 6, 92.

Dutkowski, J. et al. (2013) A gene ontology inferred from molecular networks. Nat. Biotechnol., 31, 38–45.

Dyer, M.D. et al. (2010) The human–bacterial pathogen protein interaction networks of Bacillus anthracis, Francisella tularensis, and Yersinia pestis. PloS One, 5, e12089.

Hashemifar, S. and Xu, J. (2014) HubAlign: an accurate and efficient method for global alignment of protein–protein interaction networks. Bioinformatics, 30, i438–i444.

Hashemifar, S.Q. et al. (2016) Joint alignment of multiple protein–protein interaction networks via convex optimization. arXiv preprint, arXiv:1604.03482.

Hopkins, A.L. (2008) Network pharmacology: the next paradigm in drug discovery. Nat. Chem. Biol., 4, 682–690.

http://www.genome.jp/kegg/ko.html.

Hu, H. et al. (2005) Mining coherent dense subgraphs across massive biological networks for functional discovery. Bioinformatics, 21, i213–i221.

Jaroszewski, L. et al. (1998) Fold prediction by a hierarchy of sequence, threading, and modeling methods. Protein Sci. (a publication of the Protein Society), 7, 1431.

Kanehisa, M. et al. (2011) KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res., gkr988.

Kerrien, S. et al. (2011) InrAct molecular interaction database in 2012. Nucleic Acids Res., gkr1088.

Kuchaiev, O. and Pržulj, N. (2011) Integrative network alignment reveals large regions of global network similarity in yeast and human. Bioinformatics, 27, 1390–1396.

Kuchaiev, O. et al. (2010) Topological network alignment uncovers biological function and phylogeny. J. R. Soc. Interface, 7, 1341–1354.

Liao, C.S. et al. (2009) IsoRankN: spectral methods for global alignment of multiple protein networks. Bioinformatics, 25, i253–i258.

Licata, L. et al. (2012) MINT, the molecular interaction database: 2012 update. Nucleic Acids Res., 40, D857–D861.

Malod-Dognin, N. and Pržulj, N. (2013) L-GRAAL: Lagrangian graphlet-based network aligner. Bioinformatics, 31, 2182–2189.

Memisević, V.N. and Pržulj, C. (2012) GRAAL: common-neighbors-based global graph alignment of biological networks. Integr. Biol., 4, 734–743.

Mi, H. et al. (2010) PANTHER version 7: improved phylogenetic trees, orthology and collaboration with the Gene Ontology Consortium. Nucleic Acids Res., 38, D204–D210.

Milenković, T. et al. (2010) Optimal network alignment with graphlet degree vectors. Cancer Informatics, 9, 121.

Neyshabur, B. et al. (2013) NETAL: a new graph-based method for global alignment of protein–protein interaction networks. Bioinformatics, 29, 1654–1662.

Park, Y. and Bader, J.S. (2011) Resolving the structure of interactomes with hierarchical agglomerative clustering. BMC Bioinformatics, 12, 544.

Parrish, J.R. et al. (2007) A proteome-wide protein interaction map for Campylobacter jejuni. Genome Biol., 8, R130.

Patro, R. and Kingsford, C. (2012) Global network alignment using multiscale spectral signatures. Bioinformatics, 28, 3105–3114.

Peregrín-Alvarez, J.M. et al. (2009) The modular organization of protein interactions in Escherichia coli. PloS Comput. Biol., 5, e1000523.

Saraph, V. and Milenković, T. (2014) MAGNA: maximizing accuracy in global network alignment. Bioinformatics, 30, 2931–2940.

Schrijver, A. (2003) Combinatorial Optimization: Polyhedra and Efficiency, vol. 24. Springer Science & Business Media, New York.

Sharan, R. et al. (2005) Conserved patterns of protein interaction in multiple species. Proc. Natl. Acad. Sci. USA, 102, 1974–1979.

Singh, R. et al. (2008) Global alignment of multiple protein interaction networks with application to functional orthology detection. Proc. Natl. Acad. Sci., 105, 12763–12768.

Singh, R.J. et al. (2007) Pairwise global alignment of protein interaction networks by matching neighborhood topology. In: Research in Computational Molecular Biology. Springer, Berlin.

Szklarczyk, D. et al. (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res., 39, D561–D568.

Vijayan, V. et al. (2015) MAGNA++: Maximizing Accuracy in Global Network Alignment via both node and edge conservation. Bioinformatics, 31, 2409–2411.