BIOINFORMATICS

Vol. 00 no. 00 2010
Pages 1–7

Integrative Network Alignment Reveals Large Regions of Global Network Similarity in Yeast and Human

Oleksii Kuchaiev[2] and Nataša Pržulj[1]

[1]Department of Computing, Imperial College London; London, SW7 2AZ, UK
[2]Department of Computer Science, University of California, Irvine; Irvine, CA 92617, USA

ABSTRACT

Motivation: High-throughput methods for detecting molecular interactions have produced large sets of biological network data with much more yet to come. Analogous to sequence alignment, efficient and reliable network alignment methods are expected to improve our understanding of biological systems. Unlike sequence alignment, network alignment is computationally intractable. Hence, devising efficient network alignment heuristics is currently a foremost challenge in computational biology.

Results: We introduce a novel network alignment algorithm, called Matching-based Integrative GRAph ALigner (MI-GRAAL), which can integrate *any number and type* of similarity measures between network nodes (e.g., proteins), including, but not limited to, any topological network similarity measure, sequence similarity, functional similarity, and structural similarity. Hence, we resolve the ties in similarity measures and find a combination of similarity measures yielding the largest contiguous (i.e. connected) and biologically sound alignments. MI-GRAAL exposes the largest functional, connected regions of protein-protein interaction (PPI) network similarity to date: surprisingly, it reveals that 77.7% of proteins in the baker's yeast high-confidence PPI network participate in such a subnetwork that is fully contained in the human high-confidence PPI network. This is the first demonstration that species as diverse as yeast and human contain so large, continuous regions of *global* network similarity. We apply MI-GRAAL's alignments to predict functions of un-annotated proteins in yeast, human and bacteria validating our predictions in the literature. Furthermore, using network alignment scores for PPI networks of different herpes viruses, we reconstruct their phylogenetic relationship. This is the first time that phylogeny is exactly reconstructed from purely topological alignments of PPI networks.

Availability: Supplementary files and MI-GRAAL executables: http://bio-nets.doc.ic.ac.uk/MI-GRAAL/

Contact: natasha@imperial.ac.uk

1 INTRODUCTION

1.1 Background

Large amounts of biological network data of different types are increasingly becoming available, e.g., protein-protein interaction (PPI) networks, transcriptional regulation networks, signal transduction networks, and metabolic networks. PPI networks are of particular importance because proteins are crucial for almost all functions in the cell. Proteins almost never perform their functions alone, but they cooperate with other proteins by forming physical bonds, hence creating large, complex networks. Understanding these networks is a foremost challenge of the post-genomic era.

A PPI network is conveniently modeled as an undirected unweighted *graph*, denoted by $G(V, E)$, where V is the set of proteins, or *nodes*, and E is the set of PPIs, or *edges*. The interactions are usually obtained by high-throughput experimental bio-techniques, such as yeast-2-hybrid assays (e.g., Stelzl *et al.* (2005); Simonis *et al.* (2009); Fossum *et al.* (2009); Parrish *et al.* (2007)) and affinity purification coupled to mass spectrometry (Ho *et al.*, 2002; Krogan *et al.*, 2006). It has been shown that the topology of biological networks is not random (in the Erdös-Rényi random graph sense) and that it is linked to biological function (Milo *et al.*, 2004; Sharan *et al.*, 2005; Milenkovic and Pržulj, 2008).

Analogous to sequence alignment, *network alignment* can be vital for understanding how cells work. It tries to find the best way to fit one network into another (see Section 2.1). As for sequence alignment, there exist *local* and *global* network alignments. Local alignments aim to find small subnetworks corresponding to pathways or protein complexes conserved in PPI networks of different species. Such alignments can be ambiguous, since a node from one network can be mapped to many nodes in another network. In contrast, a global network alignment provides a unique alignment from every node in the smaller network to exactly one node in the larger network, even though this may lead to inoptimal matchings in some local regions.

The earliest local network alignment algorithm is PathBLAST (Kelley, B *et al.*, 2004). It searches for high-scoring alignments of pathways from two networks by taking into account both the probabilities that PPIs in a pathway are true PPIs rather than false-positives and the homology between the aligned proteins. A modification of PathBLAST, called NetworkBLAST-M (Sharan *et al.*, 2005), was developed to identify conserved protein complexes in multiple species. MaWISh local alignment algorithm is based on the duplication/divergence models that focus on understanding the evolution of protein interactions; it constructs a weighted global alignment graph and tries to find a maximum induced subgraph in it (Koyuturk *et al.*, 2006). Graemlin algorithm scores a possibly conserved module between different networks by computing the log-ratio of the probability that the module is subject to evolutionary constraints and the probability that it is under no constraints, taking into account phylogenetic relationships between species whose networks are being aligned (Flannick *et al.*, 2006).

The first global network alignment algorithm, IsoRank, is guided by the intuition that two nodes should be matched only if their neighbors can also be matched, formulated as an eigenvector

problem and using a greedy strategy (Singh *et al.*, 2007). It uses BLAST scores (Altschul *et al.*, 1990) for sequence similarity between nodes (proteins) into the node alignment cost function by having a user-defined weight λ that controls for the relative contribution of topological similarity. IsoRank has been extended to perform local and global alignments between multiple networks (Singh *et al.*, 2008; Liao *et al.*, 2009). Also, Graemlin has been extended to construct global networks alignments between multiple networks given their phylogenetic relationships (Flannick *et al.* , 2008). PISwap algorithm begins with a sequence-based network alignment and then iteratively adjusts the alignment by incorporating topological information (Chindelevitch *et al.*, 2010). PATH and GA algorithms use the same objective function which balances (using parameter λ) between matching similar pairs and increasing the number of aligned interactions (Zaslavskiy *et al.*, 2009). They are based on relaxations of the cost function over the set of doubly stochastic matrices. Natalie and HopeMap algorithms require homology information between proteins in the networks to create alignments (Klau, 2009; Tian and Samatova, 2009). Given such information, Natalie uses a Lagrangian relaxation approach, while HopeMap uses an iterative connected-components-based algorithm. Global network alignment algorithms based purely on network topology, called GRAAL (Kuchaiev *et al.*, 2010) and H-GRAAL (Milenkovic *et al.*, 2010), have also been designed. They can align networks of any type, not only biological ones, since they do not rely on sequence similarity between nodes. Instead, both algorithms use *graphlet degrees*, which give a highly constraining quantification of topological similarity between nodes (described below). GRAAL is a seed-and-extend approach, while H-GRAAL is based on the Hungarian algorithm for solving the assignment problem.

1.2 Our contribution

It could be argued that local network alignments are of more value, since distant species should not have large regions of global network similarity. Surprisingly, we demonstrate that even species as distant as yeast and human have large and contiguous (i.e., connected) regions of PPI network similarity. In particular, we show that 77.7% of the proteins in the baker's yeast high-confidence PPI network participate in the contiguous subnetwork that simultaneously exists both in the yeast and in the human high-confidence PPI networks.

We demonstrate this by presenting a novel algorithm for global network alignment, called Matching-based Integrative GRAph ALigner (MI-GRAAL), that outperforms all previous approaches. Its unique feature is the ability to integrate and *automatically*, *without any user specified parameters*, use several different sources of node similarity information to construct the alignment (see Section 2.2). Hence, it resolves ties in different node similarity measures and produces more *stable* alignments (see Supplementary Information), i.e., alignments that are almost the same for all runs of the algorithm (note that if ties are broken randomly, as is the case in other network aligners, different runs of the algorithm can produce quite different alignments). Also, it allows for exploration of the effects of many different node similarity measures on the quality of the alignments. We show that MI-GRAAL's alignments have better topological and biological quality over other approaches.

Furthermore, we perform an all-to-all solely topological alignment of five different herpesviral PPI networks and use the network alignment similarity scores to exactly reconstruct the phylogenetic relationship between these species. To our knowledge, this is the first time that phylogeny is exactly reconstructed from purely topological alignments of PPI networks.

2 ALGORITHM

2.1 Global Network Alignment

Several different formulations of the global network alignment problem have been proposed (Flannick *et al.* , 2008; Liao *et al.*, 2009; Zaslavskiy *et al.*, 2009). Unfortunately, unlike with the sequence alignment, any reasonable formulation of this problem makes it computationally hard. The reason for this is the underlying *subgraph isomorphism* problem: given two graphs, subgraph isomorphism asks if one graph exists as an exact subgraph of the other. This problem is NP-complete, meaning that no efficient algorithm for it is likely to be found (Cook, 1971).

We use the standard definition of the *global alignment* between two networks $G_1(V_1, E_1)$ and $G_2(V_2, E_2)$, where $|V_1| \leq |V_2|$, as a total injective function $f : V_1 \rightarrow V_2$ (Singh *et al.*, 2007; Zaslavskiy *et al.*, 2009; Kuchaiev *et al.*, 2010; Milenkovic *et al.*, 2010). Function f is *total* if it maps all elements of V_1 to some elements of V_2 and it is *injective* if it never maps different elements from V_1 to the same element of V_2. Hence, the alignment is *global* in the sense that each node in the smaller network is aligned to some node in the larger network. Also, no two nodes from the smaller network can be aligned to the same node in the larger network. To measure the topological quality of the alignment f, we use the *edge correctness (EC)* measure, which is defined as the percentage of correctly aligned edges (Singh *et al.*, 2007; Zaslavskiy *et al.*, 2009; Kuchaiev *et al.*, 2010; Milenkovic *et al.*, 2010):

$$EC = \frac{|\{(u,v) \in E_1 \wedge (f(u), f(v)) \in E_2\}|}{|E_1|} \times 100\% \qquad (1)$$

Hence, EC quantifies how topologically similar two networks are. Naturally, when aligning two networks, we want to achieve as high EC as possible, hence maximizing the number of aligned edges between two networks. Note that EC is equal to 100% if and only if G_2 contains an isomorphic copy of G_1, which implies the solution to the corresponding subgraph isomorphism problem. Therefore, maximizing edge correctness is an NP-hard problem and heuristic approaches must be devised. We introduce MI-GRAAL as one such heuristic. To our knowledge, along with EC, the size of the Largest Common Connected Subgraph (LCCS) (defined in Section 3.1) is the only other measure of topological quality of an alignment that has been used. To measure the biological quality of an alignment, other methods are used in conjunction with EC and LCCS (see Section 3 for details).

2.2 MI-GRAAL algorithm

MI-GRAAL is substantially different from GRAAL (Kuchaiev *et al.*, 2010) and H-GRAAL (Milenkovic *et al.*, 2010) algorithms (see below). The only conceptual similarity between GRAAL and MI-GRAAL is that both of them are, analogous to BLAST, seed-and-extend approaches. H-GRAAL is not a seed-and-extend approach. MI-GRAAL works as follows. During its first step, MI-GRAAL builds the matrix of confidence scores, C, of size $|V_1| \times |V_2|$. Entry $C(i, j)$ in this matrix reflects the confidence with which the algorithm can align node i from G_1 to node j from G_2. Matrix C can be built based on *any number and type* of similarity measures between nodes, including, but not limited to, any topological network similarity measure, sequence similarity, functional similarity, and structural similarity. Since the number of similarity measures can be very large, as a proof of concept, we implement MI-GRAAL to use the following four topological similarities between nodes in two networks (the definitions are given below), along with the fifth measure being the sequence similarity given by the BLAST (Altschul *et al.*, 1990) *E-value* score between protein sequences:

1. Graphlet degree signature distance (SD) (Milenkovic and Pržulj, 2008)

2. Relative degree difference (DD)
3. Relative clustering coefficient difference (CD)
4. Relative eccentricity difference (ED)
5. BLAST *E-value* for protein sequence similarity ($SeqD$)

For definitions on these measures see Supplementary Information. We choose these particular measures for the following reasons. The graphlet degree signature distance has already been shown to work very well for aligning biological networks (Kuchaiev *et al.*, 2010; Milenkovic *et al.*, 2010). Degrees, clustering coefficients and eccentricities are the most common simpler node measures. BLAST *E-values* are a standard measure for deciding whether two proteins are homologous. We emphasize again that MI-GRAAL can easily be extended to use any other topological distance measure between nodes (e.g., proteins) in a network and any number of topological distances, not only the five chosen in the current implementation described above. Also, it can use any protein distance information, such as sequence, structural, and ontology information, and any number of such distance measures as well, not just BLAST E-values chosen in the current implementation described above. Its flexibility is further enhanced by allowing the user to give different weights to different measures (see Supplementary).

We compute the four topological distance measures between all pairs of nodes from networks G_1 and G_2, as well as the sequence alignment cost matrix, $SeqD$ (and others in other possible implementations of MI-GRAAL), to obtain five $V_1 \times V_2$-sized matrices, SD, DD, CD, ED and $SeqD$. To compute confidence scores of aligning nodes from G_1 to nodes from G_2, MI-GRAAL treats each of these five cost matrices as independent *agents* that tell how confident they are about aligning node $i \in V_1$ with node $j \in V_2$. Note that the perfect alignment should minimize signature, degree, clustering, eccentricity and sequence differences between nodes. Hence, the confidence score between nodes i and j, $C(i,j)$, is computed as $C(i,j) = \sum_X conf_X(i,j)$, where $conf_X(i,j)$ is a fraction of elements in the i^{th} row of matrix X that are strictly greater than $X(i,j)$; here, X stands for SD, DD, CC, ED, or $SeqD$ matrix. Hence, if for some pair of nodes i and j, $X(i,j)$ is the smallest element in row i of matrix X, this means that matrix X is 100% confident that node i should be aligned with node j. This definition of the matrix of confidence scores, C, allows us to handle ties in the scores of any individual agent, as well as contradictions between different agents (e.g., i, j may be the best pair to align according to degrees, but not according to sequences), without any a priori, user input parameter adjustments, simply by taking the majority vote (see Supplementary). Note also that such approach makes our algorithm more robust to minor error in individual cost matrices. While building the matrix of confidence scores, MI-GRAAL simultaneously constructs a priority queue of node pairs in decreasing order of their confidence scores. The priority queue is used to quickly identify seed node pairs when necessary (the details are below). It is possible that several seed node pairs can have the same confidence scores, in which case the ties are broken randomly.

Algorithm 1 and Algorithm 2 present the pseudocode for MI-GRAAL algorithm and its subroutine $align\_neighborhoods$, respectively. Below, we define the specific concepts used in them.

ALGORITHM 1. *MI-GRAAL*(G_1, G_2)

Construct, or read in the cost matrices and build the matrix of confidence scores, C, as well as the priority queue of node pairs ordered by their confidence scores.
Initialize alignment A to an empty set.
while there are unaligned nodes in G_1 **do**
  Use the priority queue to find a seed pair of nodes, (u, v), $u \in G_1, v \in G_2$, i.e., the pair of nodes that can be aligned with the highest confidence, $C(u,v)$. Brake ties randomly.
  Add (u, v) to alignment A.
  for all $k \in \{1, ..., \min\{eccen(u), eccen(v)\}\}$ **do**
    Construct the k^{th} neighborhood of u in G_1, $N^k_{G_1}(u)$, and the k^{th}neighborhood of v in G_2, $N^k_{G_2}(v)$.
    $align\_neighborhoods(N^k_{G_1}(u), N^k_{G_2}(v), C, A)$
  end for
  If there are still unaligned nodes in G_1, raise both graphs to the next power (up to the 3rd power).
end while
return alignment A.
end

Graph G raised to power p is defined as $G^p = (V(G), E^p)$, where $E^p = \{(u_1, u_2) : dist_G(u_1, u_2) \leq p\}$ and the *distance* between u_1 and u_2, $dist_G(u_1, u_2)$, is the length of the shortest path between u_1 and u_2 in G. This allows us to model insertions and deletions of nodes in the paths conserved between two networks. We use up to the 3rd power because PPI networks have a small-world nature, i.e., they have small diameters. The k^{th} neighborhood of node u in network G_1, $N^k_{G_1}(u)$, is defined to be the set of nodes of G_1 that are at distance $\leq k$ from u. Hence, $N^k_{G_1}(u)$ can be thought of as the "ball" of nodes around u up to and including nodes at distance k.

ALGORITHM 2. $align\_neighborhoods(N^k_{G_1}(u), N^k_{G_2}(v), C, A)$

1. Construct a bipartite graph $BP(N^k_{G_1}(u), N^k_{G_2}(v), E)$ with node partitions being $N^k_{G_1}(u)$ and $N^k_{G_2}(v)$ as follows:
   - Check the current alignment A and add an edge (u', v') to E, $u' \in N^k_{G_1}(u), v' \in N^k_{G_2}(v)$, if and only if nodes u' and v' have at least one pair of aligned neighbors. Hence, aligning them will increase the numerator of EC by at least 1.
   - To each edge (n, m) in E, assign the weight $C(n, m)$, the confidence with which we can align n and m.
2. Solve the *Maximum Weight Bipartite Matching Problem* for bipartite graph BP constructed above.
3. Add the optimal matching found in Step 2 above to the current alignment A.

end

A *bipartite graph*, $BP(V_1, V_2, E)$, is a graph with a node set V consisting of two partitions, $V = V_1 \cup V_2$, so that every edge $e \in E$ connects a node from V_1 with a node from V_2; that is, there are no edges between nodes of V_1 and there are no edges between nodes of V_2 – all the edges go across the node partition. A *matching* in a graph G is a set of edges such that no two edges from this set share a common endpoint. In a weighted bipartite graph, the *Maximum Weight Bipartite Matching Problem* is a problem of finding a matching of maximum weight. It can be solved in $O(|V|^2 log(|V|) + |V||E|)$ time using a modified shortest path search in the augmenting path algorithm (West, 2001). The total time complexity of MI-GRAAL algorithm for aligning networks $G_1(V_1, E_1)$ and $G_2(V_2, E_2)$ is $O(|V_1| \times (E_1 + |V_1| \times log(|V_1|))$ (see Supplementary Information).

3 RESULTS AND DISCUSSION

In this section, we use MI-GRAAL to do comparative analyses of biological networks and demonstrate the potential for its biological application. In Section 3.1, we align PPI networks of eukaryotes baker's yeast and human, while in Sections 3.2 and 3.3 we align PPI networks of bacteria and viruses, respectively. We demonstrate that MI-GRAAL exposes the largest PPI subnetworks common across species to date. It uncovers a surprising amount of common PPI network topology between yeast and human: 77.7% of the proteins in the yeast high-confidence PPI network comprise a

connected subnetwork that is fully contained within the human high-confidence PPI network. While previously local network alignments across species may have been considered to be of more value than global alignments due to believed conservation of smaller subnetworks only, that correspond to small functional subunits, this is the first demonstration of topological conservation of huge network regions. Biological implications of topological conservation of such large network regions are subject of future research. For now, we verify that MI-GRAAL aligns proteins of the same function, hence enabling function prediction of unannotated proteins. Furthermore, its edge correctness scores (defined in Section 2.1) can be used for successful reconstruction of phylogenetic relationships between species based purely on their PPI network topology, validating the quality of our alignments. Hence, global network alignment could be used as an alternative method for reconstructing phylogeny from a new source of biological information, PPI network topology.

3.1 Yeast-human PPI network alignment

We align with MI-GRAAL the high-confidence yeast *S. cerevisiae* PPI network (Collins *et al.*, 2007) with the high-confidence human PPI network (Radivojac *et al.*, 2008), henceforth denoted as "yeast" and "human," respectively. The former has 16,127 interactions amongst 2,390 proteins and the latter has 41,456 interactions amongst 9,141 proteins. To construct alignments, we explore *all* possible $2^5 - 1 = 32 - 1$ combinations of the four topological and sequence measures discussed in Section 2.2. To account for a possible randomness in the algorithm caused by randomly breaking ties, we run each of the 31 tests 30 times and compute the statistics (see bellow).

The highest edge correctness of 23.26%, comprised of 3,751 aligned interactions amongst 2,255 proteins, is obtained by an alignment that uses only signatures to score node pairs. We call this particular alignment *Alignment 1*. However, using only signatures does not resolve all possible ties and leads to different alignments for different runs with the average EC of 19.73% and the standard deviation of 1.39% over the 30 runs. That is, such an alignment is not stable (as defined in Section 1.2). Using only BLAST *E-values* does not resolve all possible ties either and also leads to different alignments for different runs with the average edge correctness of 13.30% and the standard deviation of 0.23% over the 30 runs. The best alignment obtained using only BLAST *E-values* has the EC of 13.73% and it consists of 2,215 aligned interactions amongst 2,208 proteins. We call this particular alignment *Alignment 2*. Note that Alignment 2 does use topology because of topological nature of MI-GRAAL. When we use signatures, degrees, clustering coefficients and BLAST scores, we obtain alignments that are 99.95% identical over the 30 runs and that always have the edge correctness of 18.68%, consisting of 3,012 aligned interactions amongst 2,280 proteins. Therefore, using these four cost functions resolves almost all ties and leads to almost stable alignments differing only in one or two aligned pairs. We pick one of them at random and call it *Alignment 3*. Experiments with all other possible combinations of node distance measures either result in smaller edge correctness scores, or lead to very different alignments across different runs.

Topological quality. We further analyze the topological quality of *Alignments 1, 2* and *3* by examining the size of their *largest common connected subgraphs (LCCSs)*. The LCCS is the largest connected subgraph that each of the aligned networks have as an exact copy. We examine this, since we prefer to align large and contiguous subgraphs rather than a number of small disconnected network regions (e.g., aligning only isolated edges would not give much insight into common topology of two networks). The size of the LCCS in *Alignment 1* is 1,858 nodes and 3,467 edges, which is about 77.7% and 21.5% of the yeast's nodes and edges, respectively. The LCCS uncovered by *Alignment 2* has 1,659 nodes and 1,837 edges. *Alignment 3*, that uses both sequence and topology, has the LCCS with 1,853 nodes and 2,490 edges (see Supplementary Figure 1). Thus, all of these alignments expose large contiguous common network regions (for comparison with other methods, see Section 3.4). None of these edge correctness scores are likely to be obtained at random (p-values are $\leq 10^{-9}$).

Biological quality. To measure the biological quality of the *Alignments 1, 2* and *3*, we count the fraction of aligned proteins pairs that have at least 1, 2, 3, or more Gene Ontology (GO) terms (The Gene Ontology Consortium, 2000) in common (we exclude root GO terms from the analysis). The statistics and p-values are presented in Table 3.1. The fractions are with respect to the size of the yeast PPI network, since all yeast proteins are aligned to human proteins, but not vice versa, as the yeast PPI network is smaller than the human (see Section 2.1). *Alignment 1* has the highest EC; however it consists of aligned pairs statistically significant fraction of which share at least 1, 2, or 3, but no more GO terms. As expected, using BLAST E-value information (*Alignments 2* and *3*) improves GO term enrichment, since much of GO annotation is derived from sequence alignments. For this reason, *Alignment 2*, that uses only sequence information to score node similarities, has the highest enrichment in GO terms. To account for this, we repeat the same experiments using only experimental GO terms annotations (i.e., GO term evidence codes: IPI, IGI, IMP, IDA, IEP, TAS, and IC). In this case, Alignments 1, 2 and 3 have 12.4% (p-value$\leq 1.2 \times 10^{-2}$), 14.43% (p-value$\leq 1.3 \times 10^{-6}$) and 13.37% (p-value$\leq 3.4 \times 10^{-4}$) of protein pars with at least one common GO term, respectively. Alignments 2 and 3 also contain a significant fractions of pairs with more than 2 common GO terms: 1.95% (p-value$\leq 4 \times 10^{-2}$) and 2.19% (p-value$\leq 5 \times 10^{-3}$), respectively. Hence, *Alignment 3*, that uses both topology and sequence to score node similarities, is not much behind *Alignment 2*. Also, *Alignment 3* has a much higher EC of 18.68% than *Alignment 2*, which has the average EC of 13.30% and is not stable. Since *Alignment 3* is the most stable alignment, being 99.95% identical across different runs of the algorithm, and since it has high GO term enrichment, we choose this alignment to make protein function predictions for unannotated yeast and human proteins.

Protein function prediction. We make predictions for all three GO ontology types, Biological Process (BP), Molecular Function (MF) and Cellular Component (CC). To make predictions, for each of the three ontology types we simply scan our yeast-human alignment for protein pairs in which one protein is annotated and the other one is not and transfer the annotations to the unannotated protein. For human proteins, we make 45 BP predictions, 32 MF predictions and 64 CC predictions. For yeast proteins, we make 169 BP predictions, 446 MF predictions and 54 CC predictions. *Alignment 3* and these predictions are presented in Supplementary File 1. For validating

Table 1. Fraction of protein pairs in the yeast-human alignment that share GO terms. Numbers in brackets are p-values.

#terms	Alignment 1	Alignment 2	Alignment 3
≥ 1	46.67% (10^{-9})	50.58% (3.6×10^{-8})	47.84% (10^{-9})
≥ 2	14% (3.5×10^{-4})	20.52% (4×10^{-8})	16.67% (10^{-9})
≥ 3	3.58% (8.4×10^{-2})	8.19% (10^{-9})	6.08% (10^{-9})
≥ 4	1.01% (0.36)	4.10% (5×10^{-8})	2.81% (10^{-9})
≥ 5	0.32% (0.49)	1.89% (1.8×10^{-8})	1.61% (10^{-9})
≥ 6	0.05% (0.36)	0.97% (1.4×10^{-8})	0.97% (10^{-9})

our predictions, we use the literature search and text-mining web-service CiteXplorer (Labarga *et al.*, 2007) to perform automatic search of all published articles indexed in MEDLINE. For human proteins, this tool finds at least one article mentioning the protein of interest in the context of our predicted BP for 42.22% of our predictions. Similarly, we validate 50% and 53.13% of our MF and CC human predictions, respectively. For yeast, we validate 10.06% of our BP predictions, as well as 45.41% and 11.11% of our MF and CC predictions, respectively.

3.2 Aligning Bacterial PPI networks

3.2.1 Campylobacter jejuni vs Escherichia coli. We choose to align PPI networks of these two species since they are currently the most complete and well-studied bacterial PPI networks. The high-confidence functional interaction network of *E. coli* integrates high quality experimental PPI and computational data (Peregrin-Alvarez *et al.*, 2009). It consists of 3,989 interactions amongst 1,941 proteins. The high confidence *C. jejuni* PPI network consists of 2,988 interactions amongst 1,111 proteins; it is produced by yeast-2-hybrid experiments (Parrish *et al.*, 2007). Similar to our yeast-human alignments, we use MI-GRAAL to perform alignments using all possible combinations of costs functions (see Section 2.2 for details). We obtained protein sequences and GO annotation data for these bacteria from the European Bioinformatics Institute (EMBL-EBI) website in March 2010.

The best edge correctness of 26.14% (or 781 aligned interactions) is achieved when we use only topological parameters, signatures and degrees, to score node pairs. The LCCS for this alignment has 629 nodes and 693 edges. However, this alignment does not contain statistically significant fraction of protein pairs sharing GO terms. Moreover, using only signatures and degrees does not resolve all possible ties and thus leads to different alignments for different MI-GRAAL runs, with the average EC of 24.44% and the standard deviation of 0.61%. Hence, using only these two topological scores is not enough to construct unambiguous high-quality alignments.

The alignment constructed using signatures, clustering coefficients and BLAST *E-values* is the most stable, more than 99% identical across different runs, and it has a high edge correctness of 24.16% with the LCCS consisting of 630 interactions amongst 579 proteins. Interestingly, this combination of cost functions is very similar to the one found to be the best for the yeast-human alignment (see Section 3.1). Also, this alignment is of high biological quality: out of the aligned pairs with both proteins being annotated with GO terms, 43.01%, 21.24%, 11.40%, and 6.22% share at least 1, 2, 3 or 4 term, respectively, with the p-values for these GO terms enrichments of 4.45×10^{-6}, 4.86×10^{-9}, 6.9×10^{-9} and 1.49×10^{-8}, respectively. Since this alignment has statistically significant fraction of aligned protein pairs with 4 or more GO terms in common, we use it to predict function of unannotated proteins. As before, by using CiteXplorer (Labarga *et al.*, 2007), we perform automatic search of articles indexed in MEDLINE to validate our predictions. For *C. jejuni*, we predict BP for 219 proteins, 3.65% of which we validate. Also, we validate 20% of 30 and 10.88% of 377 of our predictions of MF and CC, respectively. For *E. coli*, we predict BP for 26 proteins, 38.46% of which we validate. Also, we validate 50% and 43.69% of our 48 MF and 103 CC predictions, respectively. Clearly, the validation rates for *E. coli* are much higher. This is due to the fact that *E. coli* is much more studied then *C. jejuni* and thus, there are more articles discussing the function of its proteins. The alignment and the predictions for these bacteria are presented in Supplementary File 2.

3.2.2 Mesorhizobium vs Synechocystis. The *E. coli* and *C. jejuni*, as well as the yeast and human PPI networks analyzed in the previous sections, are all produced by different research groups, sometimes using different experimental techniques. This implies different and hard to estimate levels of noise and incompleteness of the data. To estimate the highest level of edge correctness that can be achieved by MI-GRAAL for PPI networks, we align networks produced by the same lab and by the same experimental technique: PPI networks of *Mesorhizobium loti* and *Synechocystis sp. PCC6803*(Shimoda *et al.*, 2008; Sato *et al.*, 2007).These networks contain about 24% and 52% of the protein coding genes from these organisms, respectively. The interactions were identified using a modified high-throughput yeast two-hybrid system. The *Mesorhisobium* network contains 3,094 interactions amongst 1,804 proteins and *Synechocystis* network contains 3,102 interactions amongst 1,920 proteins.

Again, we use all possible combinations of cost functions described above. The protein sequences and GO annotations were downloaded from Kazusa DNA Research Institute[1] in March 2010. The largest EC of 41.79% was obtained when signatures, degrees, clustering coefficients, and eccentricities were used. This alignment has a tree-like LCCS with 1,142 nodes and 1,148 (37.10%) edges. Since interactions in these networks were detected by the same group and the same experimental technology, we obtain a substantially higher edge correctness than in our previous experiments in which we align networks published by different research groups. As before, to measure the biological quality of the alignments, we count how many of the aligned protein pairs share GO terms. The alignments based only on topological similarities between nodes do not have statistically significant enrichment in pairs that share GO terms. When we use all possible node scoring metrics described in Section 2.2, we obtain alignments with small drops in EC scores to the average of 39.75% that are almost stable, being 89% identical across different runs. In these alignments, a significant fraction of aligned pairs has at least 1, 2 or 3 GO terms in common, with p-values less than 10^{-5}. Hence, we use one of these alignments to predict functions of unannotated proteins (presented in Supplementary File 3). However, possibly due to different protein or gene naming schemes and also since these bacteria are not as well studied as *E. Coli* and *C. jejuni*, we were not able to validate these predictions in the literature.

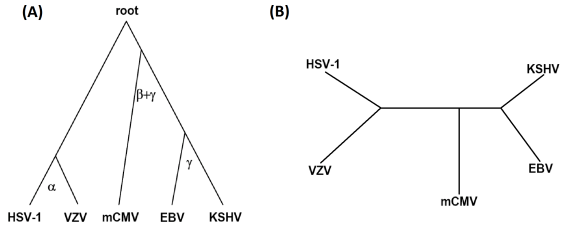

Fig. 1. Phylogeny of the five analyzed herpesviruses. **(A)** The gold standard tree (McGeoch and Gatherer, 2005; McGeoch *et al.*, 2006); **(B)** Unrooted phylogenetic tree reconstructed from edge correctness scores of topological alignments produced by MI-GRAAL.

3.3 Aligning viral PPI networks

All PPI networks discussed above contain only a fraction of proteins in each species and therefore, their alignment should not be used to reconstruct their phylogenetic relationship. The story is different for viral PPI networks described below.

We analyze PPI networks of five herpesviruses: varicella-zoster virus (VZV), Kaposis sarcoma-associated herpes virus (KSHV), herpes simplex virus 1 (HSV-1), murine cytomegalovirus (mCMV) and Epstein-Barr virus (EBV) (Fossum *et al.*, 2009). Although these networks contain false positives and negatives due to noise in experimental techniques, they contain genome-wide PPIs, meaning that all possible protein pairs in each virus were tested for interactions. It has been shown that meaningful inferences about the evolution of protein interaction networks require comparative analysis of reasonably closely related species (Agrafioti *et al.*, 2005). For these closely related herpesviruses, Fossum *et al.* (2009) have reconstructed phylogenetic relationships by counting the number of conserved interacting orthologous pairs in these species. Unlike Fossum *et al.* (2009), we do not use orthology or any sequence-based information. Instead, we use MI-GRAAL to perform all-to-all solely topological global alignment of these PPI networks (based on signatures, degrees, and clustering coefficients) and use the edge correctness scores as distances between species in the neighbor-joining algorithm of the PHYLIP package (Felsenstein, 1989) to exactly reconstruct the unrooted phylogenetic tree of these viruses (Figure 1). Also, we obtained the same results by using Fitch-Margoliash algorithm under additive tree model (Felsenstein, 1989). The phylogenetic tree does not change over different runs of MI-GRAAL. Hence, this is an evidence in support of our previous claim that purely topological network alignment may be used to reconstruct unrooted phylogenetic trees between closely related species (Kuchaiev *et al.*, 2010; Milenkovic *et al.*, 2010). Note that we were unable to exactly reproduce a rooted phylogenetic tree (which would assume evolutionary timescale) using either UPGMA, or Fitch-Margoliash under ultrametric model algorithms Felsenstein (1989). Hence, it remains an open question whether such information can be extracted from PPI network topology in principle and whether we were unable to do so because of the noise in the networks, or limitations of MI-GRAAL.

3.4 Comparison with other methods

The topological qualities of alignments produced by MI-GRAAL are impressive in comparison with alignments of the same networks with the three relevant global network alignment algorithms: IsoRank (Singh *et al.*, 2007), GRAAL (Kuchaiev *et al.*, 2010) and H-GRAAL (Milenkovic *et al.*, 2010) (see below for the discussion of other network alignment algorithms). We ran IsoRank, GRAAL and H-GRAAL on the same human, yeast, and bacterial PPI (*Mesorhizobium loti* and *Synechocystis sp. PCC6803*) networks that we aligned by MI-GRAAL (described above). Also, we ran IsoRank for all λ from 0 to 1 in increments of 0.1 (described in Section 1.1) using the same sequence similarity scores that we used in MI-GRAAL. Table 3.4 summarizes the comparisons: MI-GRAAL's alignments have more than twice as large EC and LCCS than the competing algorithms. Furthermore, unlike IsoRank, MI-GRAAL does not require the sequence score contribution to be adjusted manually by the user specified parameter λ. Instead, this is done automatically by using the confidence scores matrix (see Section 2.2 for details).

We do not compare MI-GRAAL to Graemlin 2 because it requires a variety of other input information, including phylogenetic relationships between the species being aligned (Flannick *et al.* , 2008). In contrast, we can use the output from MI-GRAAL to reconstruct the phylogenetic relationship between species (see Section 3.3). Later methods, such as HopeMap (Tian and Samatova, 2009) and Natalie (Klau, 2009), require homology information about proteins from both networks and therefore are not comparable to MI-GRAAL, since, in contract with them, by using MI-GRAAL, we can predict functional similarity between proteins of different species. Recently, a new algorithm IsoRankN was published (Liao *et al.*, 2009). However, its output is many-to-many mapping between nodes in the networks, whereas we define the global network alignment as a one-to-one node mapping (see Section 2.1). Therefore, strictly speaking, IsoRankN does not solve the global network alignment problem as we define it and its output can not be quantified using edge correctness scores. Hence, it is not comparable with MI-GRAAL. Current implementations of PATH and GA algorithms (Zaslavskiy *et al.*, 2009) cannot process networks of the sizes of yeast and human PPI networks that we aligned by MI-GRAAL[1]. PISwap algorithm has been shown to has the performance similar to that of IsoRank, GA and PATH

Table 2. Comparison of methods capable of solely topological network alignment

Algorithm	EC 1	EC 2	LCCS 1	LCCS 2	Other Sources
IsoRank	3.89%	5.33%	261	28	Yes (one)
GRAAL	11.72%	11.25%	900	46	No
H-GRAAL	10.92%	4.59%	1,290	22	No
MI-GRAAL	23.26%	41.79%	3,467	1,148	Yes (many)

Column "Algorithm" presents the name of the algorithm, columns "EC 1" and "EC 2" present the largest edge correctnesses achieved by the algorithms when aligning yeast and human (1), and *Mesorhizobium loti* and *Synechocystis sp. PCC6803* (2), respectively. Columns "LCCS 1," and "LCCS 2" present numbers of edges in the Largest Common Connected Subgraphs (LCCS) achieved by the algorithms when aligning yeast and human (1), and *Mesorhizobium loti* and *Synechocystis sp. PCC6803* (2), respectively. Column "Other Sources" indicates whether the algorithm can use any sources of information in addition to network topology (e.g., sequence) and how many.

[1] Personal communication with the authors of (Zaslavskiy *et al.*, 2009).

algorithms (Chindelevitch *et al.*, 2010) and therefore, we perform extensive comparison only with IsoRank (described above).

4 CONCLUDING REMARKS

We introduce a new global network alignment algorithm, MI-GRAAL, that is capable of integrating any number and type of node similarity measures, hence resolving ties in similarity measures and producing more stable alignments that almost do not change over different runs of the algorithm. We demonstrate that MI-GRAAL exposes a surprisingly large amount of common topology between PPI networks to date. In particular, it uncovers that 77.7% of the proteins in the yeast high-confidence PPI network are linked into a connected subnetwork that is fully contained in the human high-confidence PPI network. This is the first demonstration of the existence of such a surprisingly large amount of shared topology between species as distant as yeast and human. The biological reasons for sharing this much topology could only be speculated at the moment, including hypothesizing common structural principles across eukaryotic life. Further, we verify that the protein pairs aligned across species share biological function, which enables us to use the alignments to transfer function from annotated to unannotated parts of the aligned networks. In addition, we use our network alignment scores to successfully reconstruct phylogeny from PPI network topology only. This validates our alignment algorithm and implicates that it could be used to reconstruct phylogeny from the new source of biological information – PPI network topology.

Aligning biological networks of different species is expected to be a valuable tool in the future, since, as demonstrated above, such alignments may lead to transfer of knowledge across networks and potential discoveries in evolutionary biology. In the light of forthcoming accumulation of huge amounts of biochemical and other domain network data, network alignment methods are expected to become increasingly valuable in improving our understanding and control of not only biological, but also social and technological networks.

ACKNOWLEDGMENTS

We thank Prof. Wayne Hayes and Prof. Tijana Milenković for helpful discussions and suggestions. This project was supported by NSF CAREER IIS-0644424 and NSF CDI OIA-1028394 grants.

REFERENCES

Agrafioti *et al.* (2005). Comparative analysis of the saccharomyces cerevisiae and caenorhabditis elegans protein interaction networks. *BMC Evolutionary Biology*, **5**(1), 23.

Altschul, S. F., Gish, W., Miller, W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

Chindelevitch, L., Liao, C., and Berger, B. (2010). Local optimization for global alignment of protein interaction networks. *Pacific Symposium on Biocomputing*, pages 123–132.

Collins *et al.* (2007). Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. *Molecular and Cellular Proteomics*, **6**, 439–450.

Cook, S. (1971). The complexity of theorem-proving procedures. In *Proc. 3rd Ann. ACM Symp. on Theory of Computing: 1971; New York*, pages 151–158.

Felsenstein, J. (1989). Phylip-phylogeny inference package (version 3.2). *Cladistics*, **5**, 164–166.

Flannick *et al.* (2008). Automatic parameter learning for multiple network alignment. *RECOMB*, pages 214–231.

Flannick *et al.* (2006). Graemlin: General and robust alignment of multiple large interaction networks. *Genome Res.*, **16**, 1169–1181.

Fossum *et al.* (2009). Evolutionarily conserved herpesviral protein interaction networks. *PLoS Pathog*, **5**, e1000570.

Ho *et al.* (2002). Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*, **415**, 180–3.

Kelley, B *et al.* (2004). PathBLAST: a tool for alignment of protein interaction networks. *Nucl. Acids Res.*, **32**, 83–88.

Klau, G. (2009). A new graph-based method for pairwise global network alignment. *BMC Bioinformatics*, **10**, S59.

Koyuturk, M., Kim, Y., Topkara, U., Subramaniam, S., Szpankowski, W., and Grama, A. (2006). Pairwise alignment of protein interaction networks. *Journal of Computational Biology*, **13**, 182–199.

Krogan *et al.* (2006). Global landscape of protein complexes in the yeast saccharomyces cerevisiae. *Nature*, **440**, 637–643.

Kuchaiev, O., Milenkovic, T., Memisevic, V., Hayes, W., and Pržulj, N. (2010). Topological network alignment uncovers biological function and phylogeny. *Journal of the Royal Society Interface*, **7**, 1341–1354.

Labarga, A., Valentin, F., Andersson, M., and Lopez, R. (2007). Web services at the european bioinformatics institute. *Nucleic Acids Research*, **35**, W6–W11.

Liao, C.-S., Lu, K., Baym, M., Singh, R., and Berger, B. (2009). IsorankN: spectral methods for global alignment of multiple protein networks. *Bioinformatics*, **25**, 253–258.

McGeoch, D., Rixon, F., and Davison, A. (2006). Topics in herpesvirus genomics and evolution. *Virus Res*, **117**, 90–104.

McGeoch, D. J. and Gatherer, D. (2005). Integrating Reptilian Herpesviruses into the Family Herpesviridae. *J. Virol.*, **79**, 725–731.

Milenkovic, T. and Pržulj, N. (2008). Uncovering biological network function via graphlet degree signatures. *Cancer Informatics*, **6**, 257–273.

Milenkovic, T., Leong Ng, W., Hayes, W., and Pržulj, N. (2010). Optimal network alignment with graphlet degree vectors. *Cancer Informatics*, **9**, 121–137.

Milo *et al.* (2004). Superfamilies of evolved and designed networks. *Science*, **303**, 1538–1542.

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

Peregrin-Alvarez *et al.* (2009). The modular organization of protein interactions in Escherichia coli. *PLoS Comput Biol*, **5**, e1000523.

Radivojac *et al.* (2008). An integrated approach to inferring gene-disease associations in humans. *Proteins*, **72**, 1030–1037.

Sato, S., Shimoda, Y., Muraki, A., Kohara, M., Nakamura, Y., and Tabata, S. (2007). A Large-scale Protein protein Interaction Analysis in Synechocystis sp. PCC6803. *DNA Res*, **14**, 207–216.

Sharan *et al.* (2005). Conserved patterns of protein interaction in multiple species. *PNAS*, **102**(6), 1974–1979.

Shimoda *et al.* (2008). A large scale analysis of protein-protein interactions in the nitrogen-fixing bacterium Mesorhizobium loti. *DNA Res*, **15**, 13–23.

Simonis *et al.* (2009). Empirically controlled mapping of the Caenorhabditis elegans protein-protein interactome network. *Nature Methods*, **6**, 47–54.

Singh, R., Xu, J., and Berger, B. (2007). Pairwise global alignment of protein interaction networks by matching neighborhood topology. In *Research in Computational Molecular Biology*, pages 16–31. Springer.

Singh, R., Xu, J., and Berger, B. (2008). Global alignment of multiple protein interaction networks. *Proceedings of Pacific Symposium on Biocomputing*, pages 303–314.

Stelzl *et al.* (2005). A human protein-protein interaction network: A resource for annotating the proteome. *Cell*, **122**, 957–968.

The Gene Ontology Consortium (2000). Gene ontology: tool for the unification of biology. *Nature Genetics*, **25**, 25–29.

Tian, W. and Samatova, N. (2009). Pairwise alignment of interaction networks by fast identification of maximal conserved patterns. *Pacific Symposium on Biocomputing*, pages 99–110.

West, D. B. (2001). *Introduction to Graph Theory*. Prentice Hall, Upper Saddle River, NJ., 2nd edition.

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