Topological network alignment uncovers biological function and phylogeny

Oleksii Kuchaiev[1,‡], Tijana Milenković[1,‡], Vesna Memišević[1], Wayne Hayes[1,3], and Nataša Pržulj[2,*]

[1] Department of Computer Science, University of California, Irvine, CA 92697-3435, USA
[2] Department of Computing, Imperial College London SW7 2AZ, UK
[3] Department of Mathematics, Imperial College London SW7 2AZ, UK
[‡]These authors contributed equally to this work

Sequence comparison and alignment has had an enormous impact on our understanding of evolution, biology, and disease. Comparison and alignment of biological networks will likely have a similar impact. Existing network alignments use information external to the networks, such as sequence, because no good algorithm for purely topological alignment has yet been devised. In this paper, we present a novel algorithm based solely on network topology, that can be used to align any two networks. We apply it to biological networks to produce by far the most complete topological alignments of biological networks to date. We demonstrate that both species phylogeny and detailed biological function of individual proteins can be extracted from our alignments. Topology-based alignments have the potential to provide a completely new, independent source of phylogenetic information. Our alignment of the protein-protein interaction networks of two very different species—yeast and human—indicate that even distant species share a surprising amount of network topology with each other, suggesting broad similarities in internal cellular wiring across all life on Earth.

1 Introduction and Motivation

Advances in high throughput experimental methods have yielded large amounts of biological network data, such as protein-protein interaction (PPI) networks. The two most commonly used high-throughput methods are yeast two-hybrid screening, resulting in binary interaction data,[1–6] and protein complex purification methods using mass-spectrometry, resulting in co-complex data.[7–12] Just as comparative genomics has led to an explosion of knowledge about evolution, biology, and disease, so will comparative proteomics. As more biological network data is becoming available, comparative analyses of these networks across species are proving to be valuable, since such systems biology types of comparisons may lead to transfer of knowledge between species as well as to exciting discoveries in evolutionary biology. The most common methods for such network comparisons are network alignments.

Network alignment is the problem of finding similarities between the structure or topology of two or more networks. In the biological context, comparing networks of different organisms in a meaningful manner is arguably one of the most important problems in evolutionary and systems biology.[13] Exactly analogous to sequence alignments between genomes, alignments of biological networks can be useful because we may know a lot about some of the nodes in one network and almost nothing about topologically similar nodes in the other network; then, specialized knowledge about one may tell us something new about the other. Network alignments can also be used to measure the global similarity between complete networks of different species. Given a group of such biological networks, the matrix of pairwise global network similarities can be used to infer phylogenetic relationships.

[*]To whom correspondence should be addressed; E-mail: natasha@imperial.ac.uk .

1.1 Theoretical Background

A network (or graph) is a collection of nodes (or vertices), and connections between them called edges. Graphs are used to describe, model, and analyze an enormous array of phenomena,[14, 15] including physical systems such as electrical power grids and communication networks, social systems such as networks of friendships or corporate and political hierarchies, physical relationships such as residue interactions in a folded protein or software systems such as call graphs or expression and syntax trees.

A graph $G(V, E)$, or G for brevity, has node set V and edge set E. The sheer number and diversity of possible graphs (about $2^{(n^2)}$ of them exist given n nodes) makes graph classification and comparison problems difficult. One particular comparison problem is called *subgraph isomorphism*, which asks if one graph G exists as an exact subgraph of another graph $H(U, F)$. This problem is *NP-complete*, which means that no efficient algorithm is known for solving it.[16] Network alignment[13] is the more general problem of finding the best way to "fit" G into H even if G does not exist as an exact subgraph of H. Some networks, such as the biological ones that we consider below, may also contain noise, i.e. missing edges, false edges, or both.[17] In these cases, and also due to biological variation, it is not even obvious how to measure the "goodness" of an inexact fit. One measure could be to assess the number of aligned edges—that is, the percentage of edges in E that are aligned to edges in F. We call this the "edge correctness" (EC).[18, 19] However, it is possible for two alignments to have similar ECs, one of which exposes large, dense, contiguous, and topologically complex regions that are similar in G and H, while the other fails to expose such regions of similarity. Additionally, although EC can easily be used to measure the quality of an alignment after the fact, it is not clear how to use it to *direct* an alignment algorithm; in fact, maximizing EC is an NP-hard problem since it implies solving the subgraph isomorphism problem. Thus, other strategies must be sought to guide the alignment process.

1.2 Previous Approaches

Analogous to sequence alignments, there exist *local* and *global* network alignments. Thus far, the majority of methods used for alignment of biological networks have focused on local alignments.[20–24] With local alignments, mappings are chosen independently for each local region of similarity. Many algorithms for local alignment have been developed. *PathBLAST* searches for high-scoring pathway alignments between two networks, by taking into account both the homology between the aligned proteins and the probabilities that PPIs in the path are true PPIs and not false-positives.[20] *NetworkBLAST* detects conserved protein clusters rather than paths, by deploying a likelihood-based scoring scheme that weighs the denseness of a subnetwork versus the chance of observing such network substructure at random.[25] *MaWISh* defines network alignment as a maximum weight induced subgraph problem and implements an evolution-based scoring scheme to detect conserved clusters; it extends the concepts of evolutionary events in sequence alignments to that of duplication, match, and mismatch in network alignments and evaluates the similarity between network structures through a scoring function that accounts for these evolutionary events.[26] *Graemlin*, the first method capable of identifying *dense* conserved subnetworks of *arbitrary* structure, scores a module by computing the log-ratio of the probability that the module is subject to evolutionary constraints and the probability that the module is under no constraints, while taking into account phylogenetic relationships between species whose networks are being aligned.[22]

Local alignments can be ambiguous, with one node having different pairings in different local alignments. 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. Previous local network alignment algorithms have not generally been able to identify large subgraphs that have been conserved during evolution.[20, 21]

Global network alignment has been studied previously in the context of biological networks.[19,27,28] Unlike the above algorithms that primarily aim to detect conserved subnetworks, *IsoRank*[27] aims to maximize the *overall* match between the two networks. It relies on spectral graph theory to compute scores of aligning pairs of nodes from different networks; it does so by using the heuristic that two nodes are a good match if their respective neighbors also match well. Thus, the score of a protein pair depends on the score of their neighbors, that, in turn, depend on the neighbors of their neighbors, and so on. Once these "topological" scores are computed for all node pairs, sequence-based BLAST scores are included in the pairwise alignment scores. IsoRank then constructs the node alignment with the repetitive greedy strategy of identifying among all protein pairs the highest scoring pair, outputting that pair, and removing all scores involving any of the two identified nodes.[27] The more recent *IsoRankN* relies on the notion of node-specific rankings and uses a method similar to *PageRank-Nibble* algorithm.[29] *Graemlin* has been extended to allow global network alignment by relying on a learning algorithm that uses a training set of known network alignments and their phylogenetic relationships to learn parameters for its scoring function, and by automatically adapting the learned objective function to any set of networks.[28]

Hence, most existing local and global network alignment methods incorporate some *a priori* information about nodes such as sequence similarities of proteins in PPI networks,[24,27] or they require a variety of biological information as the input such as phylogenetic relationships between species whose networks are being aligned or use some form of learning on a set of "true" alignments.[28]

1.3 Our Contribution

At best, all previous algorithms depend only implicitly or indirectly on the topology of the network, whereas we believe that there is important information encoded directly into the topology of biological networks. Furthermore, we believe that much of this topology-encoded information is not easily extracted through any means other than explicitly measuring network topology. For example, there exist identical protein sequences that can fold in different ways in different environments, leading to different functions and thus very different network topologies in a PPI network.[30–33] In such cases, homology information is more correctly encoded in the topology of network neighborhoods than in sequence similarity.[34] For this and other reasons, we believe that network topology, in and of itself, provides valuable biological information that is largely independent of other currently available information. Thus, we propose a network alignment algorithm whose cost function is based solely and explicitly on a strong, theoretically-grounded, direct measure of network topological similarity.

We introduce a novel method for aligning a pair of networks that is based *solely* on network topology. As such, this algorithm could be applied to *any* two networks, not just biological ones. For example, our algorithm can be applied to road maps or social networks, which obviously have no genetic or protein sequence associated with them. We apply our method to align two protein-protein interaction (PPI) networks and demonstrate that our alignment exposes far more topologically complex regions of similarity than existing methods. Also, we use our method to compute pairwise all-to-all network similarity matrix between a group of species, and then build a phylogenetic tree that bears a striking resemblance to the one based on sequence comparison. The significance of these results are that they extract statistically significant meaning from a new source of information—pure network topology—that is independent of sequence or any other commonly used biological information. We believe that the results in this paper just barely scratch the surface of the information that can be extracted from network topology.

2 Results and Discussion

We focus on topology instead of protein sequence because we aim to discover biological knowledge that is encoded in the PPI network topology. Since proteins aggregate to perform a function instead of acting in isolation, analyzing complex wirings around a protein in a PPI network could give deeper insights into inner working of cells than analyzing sequences of individual genes. Furthermore, network topology and protein sequences might give insights into different slices of biological information and thus, one could loose much information by focusing on sequence alone. Although protein sequence similarity correlates with functional similarity, there exist proteins with 100% sequence identity that have different functional roles.[30–32] Thus, restricting analysis to sequences might give incorrect functional assignments. Similarly, although high protein sequence similarity correlates with similarity in 3-dimensional structure, sequence-similar proteins can have structures that differ significantly from one another.[33] Thus, sequence-based homology analyses may mask important structural and functional information. On the other hand, since the structure of a protein is expected to define the number and type of its potential interacting parters in the PPI network, sequence-similar but structurally-dissimilar proteins are expected to have different PPI network topological characteristics. Moreover, entirely different sequences can produce identical structures.[32,35] In cases where such proteins are expected to share a common function, sequence-based function prediction would fail, where network topology-based one would not. Finally, we show that both sequence and topology have similar predictive power with respect to Gene Ontology (GO) terms[36] (Supplementary Figure 1), demonstrating that network topology can provide as much functional information as protein sequences. Since our goal is to uncover biological knowledge encoded in the topology of PPI networks, our alignments do not use protein sequence information. Thus, our method can align *any* type of network, not just biological ones. Note, however, that inclusion of sequence component into the cost function of our method is trivial (see Section 4), but this is out of the scope of the manuscript.

Obviously, if one is to build meaningful alignments based solely upon network topology, one must first have a highly constraining *measure* of topological similarity. The simplest (and weakest) description of the topology of a node is its *degree*, which is the number of edges that touch it. Our much more highly constraining measure is a generalization of the degree of a node. We define a *graphlet* as a small, connected, *induced* subgraph of a larger network.[37–39] An *induced* subgraph on a node set $X \subseteq V$ of G is obtained by taking X and *all* edges of G having both end-nodes in X. Figure 1 shows all the graphlets on 2, 3, 4, and 5 nodes. For a particular node v in a large network, we define a vector of “graphlet degrees”[40] that counts the number of each kind of graphlet that touch v (Figure 2). This vector, or *signature*, of v describes the topology of its neighborhood and captures its interconnectivities out to a distance of 4 (see Section 4.1 and Figure 2).[40] This measure is superior to all previous measures, since it is based on all up to 5-node graphlets, which is practically enough due to the small-world nature of many real-world networks.[41]

For our purposes, an alignment of two networks G and H consists of a set of ordered pairs (x, y), where x is a node in G and y is a node in H. Our algorithm, called GRAAL (GRAph ALigner), incorporates facets of both local and global alignment. We match pairs of nodes originating in different networks based on their *signature similarity*[40] (see Section 4.1), where a higher signature similarity between two nodes corresponds to a higher topological similarity between their extended neighborhoods (out to distance 4). The cost of aligning two nodes is modified to align the densest parts of the networks first; the cost is reduced as the degrees of both nodes increase, since higher degree nodes with similar signatures provide a tighter constraint than correspondingly similar low degree nodes (see Section 4.2 and the Supplementary Information); α is a parameter in [0,1] that controls the contribution of the node signature similarity to the cost function, the other contribution being simply the degree of the node (see Section 4.2). In the case of two node alignments comparing equally, the tie is broken randomly. Thus, different runs of the alignment algorithm can produce different results. However, we find that for PPI networks that we analyze below, a deterministic "core" alignment containing 60% of all aligned pairs remains across all runs (see Section 2.1).

We align each node in the smaller network to exactly one node in the larger network. The matching proceeds using a technique analogous to the "seed and extend" approach of the popular BLAST[42] algorithm for sequence alignment: we first choose a single "seed" pair of nodes (one node from each network) with high signature similarity. We then expand the alignment radially outward around the seed as far as practical using a greedy algorithm (see Section 4.2). Although local in nature, our algorithm produces large and dense global alignments. By "dense" we mean that the aligned subgraphs share many edges, which would not be the case in a low-quality or random alignment. We believe that the high quality of our alignments is based less on the details of the extension algorithm and more on having a good measure of pair-wise topological similarity between nodes.[40]

2.1 Pairwise Alignment of Yeast and Human PPI Networks

Using GRAAL, we align the human PPI network of Radivojac et al.[43] to the Collins et al. yeast PPI network,[12] which we call "human1" and "yeast2," respectively. We chose yeast as our second species because currently it has a high quality PPI network, with 16,127 interactions (edges) among 2,390 proteins (nodes). The "best" alignment (defined below) found by GRAAL aligns 1,890 of the edges in yeast2 to edges in human1. Thus, the edge correctness (EC) of our alignment is 11.72%. There are 970 nodes involved in these "correct" edge alignments, representing 40% of all yeast2 nodes. We obtained similar EC for aligning other yeast[12,44] and human[5,44,45] networks (Supplementary Figure 2). The best alignment is defined as follows. Due to the existence of the α parameter in the cost function (as explained above) and some randomness in the GRAAL algorithm (see Section 4.2 and the Supplementary Information for details), the actual alignments and ECs vary across different values of α, and across different runs of the algorithm for the same α. With this in mind, the best alignment is the alignment with the highest EC over all values of α, and over all runs for the given α. The highest EC is obtained for α of 0.8; the minimum EC over all runs for this α is higher than the maximum EC over all runs for any other α. Thus, we focus on alignments produced for α of 0.8. Variation of EC over different runs for this α is small, with minimum and maximum EC of 11.5% and 11.72%, respectively. Moreover, intersection of alignments from up to 40 different runs at α of 0.8 contains 1,433 pairs, i.e., about 60% of the entire alignment. We call this intersection the *core* alignment.

In addition to counting aligned edges, it is important that the aligned edges cluster together to form large and dense connected subgraphs, in order to uncover such regions of similar topology. We define a *common connected subgraph* (CCS) as a connected subgraph (not necessarily induced) that appears in both networks. The largest CCS in our best alignment (Figure 3A) has 900 interactions amongst 267 proteins, which comprises 11.2% of the proteins in the yeast2 network. Our second largest CCS has 286 interactions amongst 52 nodes, depicted in Figure 3B. The entire common subgraph is presented in Supplementary Figure 3.

2.2 Statistical Significance of GRAAL's Yeast-Human Alignment

In the following three paragraphs, we look at three distinct ways in which to judge the statistical significance of our alignment: first, we judge the quality of our alignment compared to a random alignment of these two particular networks; second, we comment on the amount of similarity found between yeast and human in our alignment; and third, we interpret the biological significance of our alignment. Section 4 and Supplementary Information provide more details on all of the above.

Given a random alignment of yeast2 to human1, the probability of obtaining an edge correctness of 11.72% or better (p-value) is less than 7×10^{-8}. The probability of obtaining a large CCS would be significantly smaller, so this represents a weak upper bound on our p-value.

Judging the amount of similarity found between the yeast2 and human1 networks in our alignment requires us to state carefully what we are comparing against. If we align with GRAAL networks drawn from several different random graph models[46] that have the same number of nodes and edges as yeast2 and human1, we find that the edge correctness between random networks is significantly lower than the edge correctness of our yeast2-human1 alignment. For example, aligning two Erdös-Rényi random graphs with the same degree distribution as the data ("ER-DD") gives an edge correctness of only about $0.31 \pm 0.22\%$. Similar alignments of Barabási-Albert type scale-free networks ("SF-BA"),[47] stickiness model networks ("STICKY"),[48] or 3-dimensional geometric random graphs ("GEO-3D"),[37] give edge correctness scores of only $2.86 \pm 0.57\%$, $5.89 \pm 0.39\%$ and $8.8 \pm 0.39\%$, respectively. Accepting GEO-3D as the best available null model (see Section 4.3), the p-value of our yeast2-human1 alignment is at most 8.4×10^{-3}. This tells us that yeast and human, two very different species, enjoy more network similarity than chance would allow.

We measure the biological significance of our alignment by counting how many of our aligned pairs share common Gene Ontology (GO) terms.[36] GO terms succinctly describe the many biological properties that a given protein may have. For this analysis, we consider the "complete" GO annotation data set, containing all GO annotations, independent of GO evidence code. GO annotation data was downloaded in September 2009. Across our entire best yeast2-human1 alignment, 45.1%, 15.6%, 5.1%, and 2.0% of aligned protein pairs share at least one, two, three, and four GO terms, respectively. Compared to random alignments, the p-values for these percentages are all in the 10^{-6} to 10^{-8} range. Furthermore, the results improve across GRAAL's *core* yeast2-human1 alignment: 50.9%, 19.3%, 7.3%, and 3.0% of aligned protein pairs share at least one, two, three, and four GO terms, respectively; the p-values for these percentages are all in the 10^{-8} to 10^{-9} range.

2.3 Comparison with Other Methods

GRAAL produces by far the most complete topological alignments of biological networks to date and uncovers CCSs (Common Connected Subgraphs) that are substantially larger and denser than those produced by currently published algorithms, as demonstrated below. The best currently published global alignment of similar networks is the alignment of yeast and fly by IsoRank,[27] which uses sequence information in addition to topological information. It aligns 1,420 edges, but its largest CCS contains just 35 nodes and 35 edges. We applied IsoRank to our yeast2-human1 data using only topological information. We found that it aligns 628 interactions, giving an edge correctness of only 3.89%, compared to GRAAL's edge correctness of 11.72%. Hence, we align 3 times more edges than IsoRank does. IsoRank's largest CCS has just 261 interactions among 116 proteins, compared to GRAAL's largest CCS with 900 interactions amongst 267 proteins. Thus, GRAAL's largest CCS is 2.3 and 3.5 times larger than IsoRank's largest CCS in terms of the number of nodes and edges, respectively. Note that we do not include sequence information in IsoRank's alignment cost function, since Singh et al. (2007) have shown that the highest edge correctness is obtained when topology alone is used.[27]

Additionally, our results are better then those achieved by IsoRank with respect to the number of shared GO terms even though GRAAL does not use any protein sequence information. In the global alignment produced by IsoRank, 44.2%, 14.1%, 4.1%, and 1.5% of aligned protein pairs have at least one, two, three, and four GO terms in common, respectively, compared to GRAAL's percentages of 45.1%, 15.6%, 5.1%, and 2.0%, respectively. Furthermore, if we restrict our analysis only to the largest CCS, in IsoRanks's CCS, the percentages are 60.6%, 11.9%, and 0% for sharing at least 1, 2, and 3 common GO terms, respectively, while in GRAAL's CCS, these percentages are 67.2%, 22.0%, and 5.2%, respectively.

Recently, IsoRankN, an algorithm for global alignment of *multiple* networks, has been introduced.[29] However, a comparison with GRAAL is not feasible, since the output of the two algorithms is different. While GRAAL outputs a list of one-to-one node mappings between the networks being aligned, IsoRankN's alignment contains sets of aligned proteins, where no two sets overlap, but each set can contain more than one node (i.e., many-to-many node mapping) from each of the networks being aligned. Thus, IsoRankN's output can not be quantified topologically with EC, since one many-to-many node alignment can produce exponentially many one-to-one node alignments and enumerating all of them is computationally infeasible.

Another popular global network alignment method is Graemlin.[28] We do not compare our alignment to the one produced by Graemlin because Graemlin requires a variety of other input information, including phylogenetic relationships between the species being aligned. In contrast, GRAAL's *output* can be used to infer phylogenetic relationships.

Finally, other methods potentially better than IsoRank exist.[19] However, their current implementations failed to process networks of the size of yeast2 and human1[§]. Moreover, we do not benchmark these methods on the yeast and fly data analyzed by Zaslavskiy et al. (2009)[19] because they did not try to align the entire yeast and fly networks but they focused only on their smaller induced subgraphs defined on proteins covered by Inparanoid clusters.[§] Thus, although their "global" yeast-fly alignment aligns each node in the smaller subnetwork (defined above) to a node in the larger subnetwork, it is not truly global, as it aligned only parts of the original yeast and fly networks. Therefore, we found it inappropriate to evaluate GRAAL's global alignment of the entire yeast and fly networks with their "global" alignments of partial yeast and fly networks. Moreover, we believe that a good network alignment algorithm should both produce high-quality alignments and be capable of dealing with large data sets; this is especially true for biological networks, since their sizes will only continue to grow. Thus, the methods by Zaslavskiy et al. (2009)[19] that failed to process any larger data set are not relevant to the large networks we consider.

2.4 Application to Protein Function Prediction

With the above validations in hand (Section 2.2), we believe that GRAAL's alignments can be used to predict biological characteristics (i.e., GO molecular function (MF), biological process (BP), and cellular component (CC)) of un-annotated proteins based on their alignments with annotated ones.

Here, we distinguish between two different sets of GO annotation data: the complete set described above, containing all GO annotations, independent of GO evidence codes, and biologically-based set, containing GO annotations obtained by experimental evidence codes only (see[36] for details). Since in the complete GO annotation data set, many GO terms were assigned to proteins computationally (e.g., from sequence alignments), that set is biologically less confident than the biologically-based one. We make predictions with respect to both GO annotation data sets, as described below.

First, we analyze GRAAL's best yeast2-human1 alignment (i.e., the alignment with the highest EC over all runs for alpha of 0.8, as explained in Section 2.1) to identify aligned protein pairs where one of the proteins is annotated with a "root" GO term only: GO:0003674 for MF, GO:0008150 for BP, or GO:0005575 for CC; this means that one of the proteins in the pair has no known functional information.[36] Next, we check if aligned partners of such proteins with unknown function are annotated with a known MF, BP, or CC GO term, with respect to both the complete and biologically-based GO annotation data sets. If so, we assign all known MF, BP, or CC GO terms to the unannotated protein.

[§]Personal communication with the authors.

With respect to the complete GO data set, we predict MF for 44 human and 435 yeast proteins, BP for 53 human and 157 yeast proteins, and CC for 52 human and 54 yeast proteins. Since GO database offers a list with an explicit note that a protein is not associated with a given GO term, we were able to examine directly whether our predictions contradicted this list. We found no contradictions in GO database for any of the yeast or human proteins with respect to MF or BP; we found contradiction only for one of our human predictions with respect to CC. We also attempted to validate all of our predictions using the literature search and text mining tool CiteXplorer.[49] For 34.1%, 43.4%, and 46.2% of our MF, BP, and CC human predictions, respectively, this tool found at least one article mentioning the protein of interest in the context of at least one of our predictions for that protein. For yeast, these percentages are 42.07%, 3.18%, and 12.96%, respectively. Our human and yeast predictions made with respect to the complete GO data set are presented in Supplementary Tables 1 and 2, respectively.

With respect to biologically-based GO data set, we predict MF for 30 human and 214 yeast proteins, BP for 42 human and 41 yeast proteins, and CC for 45 human and 17 yeast proteins. None of these predictions were contradicted in the GO database. We validated with CiteXplorer 10%, 4.76%, and 20% of our biologically-based MF, BP, and CC human predictions, respectively. We also validated 48.1% of our biologically-based MF yeast predictions. Our human and yeast predictions made with respect to biologically-based GO data set are presented in Supplementary Tables 3 and 4, respectively.

2.5 Reconstruction of Phylogenetic Trees by Aligning Metabolic Pathways Across Species

Finally, we describe a completely different application: how purely topological alignment of metabolic networks obtained by GRAAL can be used to recover phylogenetic relationships.

Several studies analyzing metabolic pathways in different species have aimed to find an evolutionary relationship between the species and construct their phylogenetic trees.[50–53] Different distance metrics have been used for constructing phylogenetic trees. For example, similarities between pathways have been computed from sequence similarities between corresponding substrates and enzymes from individual pathways[50] or as a combination of similarities of enzymes from individual metabolic networks and topologies of these networks.[51,53] The similarity of enzymes is based on the similarity of their sequences, structures, or Enzyme Commission numbers.[54] The topological similarity of two pathways has been based on the similarity between nodes (corresponding to enzymes) and the similarity of their neighborhoods, measuring whether a node influences similar nodes and whether it is influenced by similar nodes itself.[51] In addition, topological similarity of metabolic pathways combining global network properties, such as the diameter and clustering coefficient, and similarities of shared node (i.e., enzyme) neighborhoods has been used.[52]

Therefore, although related attempts exist,[53] they all still use some biological or functional information external to network topology, such as sequence similarities, to define node similarities and derive phylogenetic trees from pathways. Since we use only network topology to define protein similarity, our information source is fundamentally different. Thus, our algorithm recovers phylogenetic relationships (but not the evolutionary timescale of species divergence at this point) in a completely novel and independent way from all existing methods for phylogenetic recovery.

It has been shown that PPI network structure has subtle effects on the evolution of proteins and that reasonable phylogenetic inference can only be done between closely related species.[55] In the KEGG pathway database, there are 17 Eucaryotic organisms with fully sequenced genomes,[56] of which seven are protists, six are fungi, two are plants, and two are animals. Here we focus on protists (see the Supplementary Information for fungi). For each organism, we extract the union of all metabolic pathways from KEGG, and then find all-to-all pairwise network alignments between species using GRAAL. The edge correctness scores between pairs of protist networks range from 29.6% to 76.7%. We create phylogenetic trees using the average distance algorithm¶, with pairwise edge correctness as the distance measure. We compare our phylogenetic trees to the published ones‖ obtained from genetic or amino acid sequence alignments.[57,58] Figure 4 presents our phylogenetic tree for protists and shows that it is very similar to that found by sequence comparison.[57] We can estimate the statistical significance of our tree by measuring how it compares to trees built from random networks of the same size as the metabolic networks (see the Supplementary Information); we find that the p-value of our tree is less than 1.3×10^{-3}. Phylogenetic trees based on alignments made by IsoRank do not differ significantly from random ones (see the Supplementary Information). We also find that the topologies of the entire metabolic networks of Cryptosporidium parvum and Cryptosporidium hominis are very similar, having edge correctness of 75.72%. This result is encouraging since these organisms are two morphologically identical species of Apicomplexan protozoa with 97% genetic sequence identity, but with strikingly different hosts[59] that contribute to their divergence.[60]

Note that all of the metabolic networks that we align are derived from a mix of experimentally obtained data and network reconstructions based on orthology relationships between species. Hence, the fact that we largely recover the phylogenetic trees obtained from sequence alignments is a strong validation of our method. Moreover, the phylogenetic tree in the literature is obtained from sequence alignments of mitochondrial proteins or ribosomal RNA, whereas metabolic networks in KEGG are partially obtained by sequence alignments of protein sequences. Therefore, since different source of sequence data is used for reconstructing phylogenetic trees in the literature and for reconstructing metabolic networks, the phylogenetic trees obtained from our network alignments might already be viewed as new and independent sources of phylogenetic information. This will gain in biological significance when purely experimentally obtained networks become available further providing validation of sequence-based phylogeny.

Given that our phylogenetic tree is slightly different from that produced by sequence, there is no reason to believe that the sequence-based one should *a priori* be considered the correct one. Sequence-based phylogenetic trees are built based on multiple alignment of gene sequences and whole genome alignments. Multiple alignments can be misleading due to gene rearrangements, inversions, transpositions, and translocations that occur at the substring level. Furthermore, different species might have an unequal number of genes or genomes of vastly different lengths. Whole genome phylogenetic analyses can also be misleading due to non-contiguous copies of a gene or non-decisive gene order.[61] Finally, the trees are built incrementally from smaller pieces that are “patched” together probabilistically,[57] so probabilistic errors in the tree are expected. Our tree suffers from none of the above problems.

3 Conclusions

In summary, we present evidence that it is possible to extract biological knowledge from network topology only. We introduce a new global network alignment algorithm that is based solely on network topology. As such, it can be applied to any network type, not just biological ones. We apply our method to align PPI networks of yeast and human and demonstrate that it produces topologically statistically significant alignments in which many aligned proteins perform the same biological function. Given the high quality of our yeast-human alignment, we predict biological function of unnanotated proteins based on the function of their annotated aligned partners, validating a large number of our predictions in the literature. Additionally, we succesfully reconstruct phylogenetic trees from topological alignments of metabolic networks, demonstrating that network topology can be used as a novel and independent source of phylogenetic information.

¶http://www.mathworks.com/access/helpdesk/help/toolbox/bioinfo/index.html
‖http://fungal.genome.duke.edu/

Network alignment has applications across an enormous span of domains, from social networks to software call graphs. In the biological domain, the mass of currently available network data will only continue to increase and we believe that high-quality topological alignments can yield new and pivotal insights into function, evolution, and disease.

4 Methods

A graph $G(V, E)$, or G for brevity, has node set V and edge set E. Given $n = |V|$ nodes, the maximum number of undirected edges is $M = n(n-1)/2$, and the number of possible undirected graphs on n nodes is thus 2^M. The sheer number and diversity of possible graphs makes graph classification and comparison problems difficult. One of those problems is called *subgraph isomorphism*: given two arbitrary graphs $G(V, E)$ and $H(U, F)$ such that $|V| \leq |U|$, does G exist as a subgraph of H? That is, is there a discrete map $\sigma : V \rightarrow U$ defined $\forall v \in V$ such that $(x, y) \in E \Rightarrow (\sigma x, \sigma y) \in F$? This problem is *NP-complete*, which means that no efficient algorithm is known for finding the mapping σ—the only known generally applicable way is to search through all possible mappings from V to U.[16] Since the number of such mappings is exponential in both $|V|$ and $|U|$, this is considered an intractable problem.

4.1 Graphlet Degree Signatures and Signature Similarities

GRAAL aligns a pair of nodes originating in different networks based on a similarity measure of their local neighborhoods.[40] This measure generalizes the degree of a node, which counts the number of edges that the node touches, into the vector of *graphlet degrees*, counting the number of graphlets that the node touches, for all 2-5-node graphlets (see Figure 1). Note that the degree of a node is the first coordinate in this vector, since an edge (graphlet G_0 in Figure 1) is the only 2-node graphlet. Since it is topologically relevant to distinguish between, for example, nodes touching graphlet G_1 at an end or at the middle, the notion of *automorphism orbits* (or just *orbits*, for brevity) is used. By taking into account the "symmetries" between nodes of a graphlet, there are 73 different orbits across all 2- to 5-node graphlets. We number the orbits from 0 to 72.[39] The full vector of 73 coordinates is the *signature* of a node (Figure 2).

The signature of a node provides a novel and highly constraining measure of local topology in its vicinity and comparing the signatures of two nodes provides a highly constraining measure of local topological similarity between them. The *signature similarity*[40] is computed as follows. For a node $u \in G$, u_i denotes the i^{th} coordinate of its signature vector, i.e., u_i is the number of times node u is touched by an orbit i in G. The distance $D_i(u, v)$ between the i^{th} orbits of nodes u and v is defined as $D_i(u, v) = w_i \times \frac{|log(u_i+1) - log(v_i+1)|}{log(max\{u_i, v_i\}+2)}$, where w_i is a weight of orbit i that accounts for dependencies between orbits; for example, differences in counts of orbit 3 will imply differences in counts of all orbits that contain a triangle, such as orbits 10-14, 25, 26, etc. and thus, a higher weight is assigned to orbit 3, w_3, than to the orbits that contain it.[40] The total distance $D(u, v)$ between nodes u and v is defined as: $D(u, v) = \frac{\sum_{i=0}^{72} D_i}{\sum_{i=0}^{72} w_i}$. The distance $D(u, v)$ is in [0, 1), where distance 0 means that signatures of nodes u and v are identical. Finally, the signature similarity, $S(u, v)$, between nodes u and v is $S(u, v) = 1 - D(u, v)$.

4.2 GRAAL (GRAph ALigner) Algorithm

When aligning two graphs $G(V, E)$ and $H(U, F)$, GRAAL first computes costs of aligning each node in G with each node in H. The cost of aligning two nodes takes into account the signature similarity between them, modified to reduce the cost as the degrees of both nodes increase, since higher degree nodes with similar signatures provide a tighter constraint than correspondingly similar low degree nodes (see the Supplementary Information). α is the parameter in [0,1] that controls the contribution of the signature similarity to the cost function; that is, $1 - \alpha$ is the parameter that controls the contribution of node degrees to the cost function. In this way, we align the densest parts of the networks first.

It is also possible to add protein sequence component to the cost function, to balance between topological and sequence similarity of aligned nodes. This can be done trivially by adding another parameter β to the cost function that would control the contribution of the current topologically-derived costs, while $1 - \beta$ would control the contribution of node sequence similarities to the total cost function; similar has been done by other relevant studies.[19,27,29] However, as we aim to extract only biological information encoded in network topology, analyzing how balancing between the topological and sequence similarity affects the resulting alignments is out of the scope of our manuscript and is the subject of future work.

GRAAL chooses as the initial seed a pair of nodes (v, u), $v \in V$ and $u \in U$, that have the smallest cost. Ties are broken randomly, which results in slightly different results across different runs. Once the seed is found, GRAAL builds "spheres" of all possible radii around nodes v and u. A sphere of radius r around node v is the set of nodes $S_G(v, r) = \{x \in V : d(v, x) = r\}$ that are distance r from v where the distance $d(v, x)$ is the length of the shortest path from v to x. Spheres of the same radius in two networks are then greedily aligned together by searching for the pairs $(v', u') : v' \in S_G(v, r)$ and $u' \in S_H(u, r)$ that are not already aligned and that can be aligned with the minimal cost. When all spheres around the seed (v, u) have been aligned, some nodes in both networks may remain unaligned. For this reason, GRAAL repeats the same algorithm on a pair of networks (G^p, H^p) for $p = 1, 2$, and 3, and searches for the new seed again, if necessary. We define a network G^p as a new network $G^p = (V, E^p)$ with the same set of nodes as G and with $(v, x) \in E^p$ if and only if the distance between nodes v and x in G is less than or equal to p, i.e., $d_G(v, x) \leq p$. Note that $G^1 = G$. Using $G^p, p > 1$ allows us to align a path of length p in one network to a single edge in another network, which is analogous to allowing "insertions" or "deletions" in a sequence alignment. GRAAL stops when each node from G is aligned to exactly one node in H.

GRAAL produces global alignments. We note that optimal global alignments are not necessarily unique. Given any particular cost function, there may be many distinct alignments that all share the optimal cost. In this paper, we analyze just one specific alignment that we believe is a good one, although it may not be optimal even according to our measure. Enumerating all optimal (or at least good) alignments requires extending our algorithm to allow many-to-many mappings between the nodes in the two networks, and is the subject of the future work. Thus, many more predictions of equal validity to those in this paper are likely to be possible. However, we empirically demonstrate that a large portion (about 60%) of the entire alignment is conserved across different runs of the algorithm; thus, this core alignment is independent of the randomness in the algorithm.

The algorithm's pseudo code and details about the complexity analysis are presented in the Supplementary Information. The software and data used in this paper are available upon request.

4.3 Statistical Significance of our Yeast-Human Alignment

Given a GRAAL alignment of two networks $G(V, E)$ and $H(U, F)$, we compute the probability of obtaining a given or better edge correctness score at random. For this purpose, an appropriate null model of random alignment is required. A random alignment is a random mapping f between nodes in two networks $G(V, E)$ and $H(U, F)$, $f : V \rightarrow U$. GRAAL produces *global* alignments, so that all nodes in the smaller network (smaller in terms of the number of nodes) are aligned with nodes in the larger network. In other words, f is defined $\forall v \in V$. This is equivalent to aligning each edge from $G(V, E)$ with a *pair of nodes* (not necessarily an edge) in $H(U, F)$. Thus, we define our null model of random alignment as a random mapping

$g : E \to U \times U$. We define $n_1 = |V|$, $n_2 = |U|$, $m_1 = |E|$, and $m_2 = |F|$. We also define the number of node pairs in H as $p = \frac{n_2(n_2-1)}{2}$, and let $EC = x\%$ be the edge correctness of the given alignment. We let $k = [m_1 \times EC] = [m_1 \times x]$ be the number of edges from G that are aligned to edges in H. Then, the probability P of successfully aligning k or more edges by chance is the tail of the hypergeometric distribution: $P = \sum_{i=k}^{m_2} \frac{\binom{m_2}{i}\binom{p-m_2}{m_1-i}}{\binom{p}{m_1}}$. For our yeast2–human1 alignment, we find $P \approx 7 \times 10^{-8}$.

Now we describe how to estimate the statistical significance of the amount of similarity we find between yeast2 and human1 in our alignment. To do that, we need to estimate how much similarity one would expect to find between two *random* networks and doing that, in turn, requires us to specify how we generate model random networks. Given two models that purport to fit a set of observations, we generally consider as superior the one that has fewer tunable parameters. For example, the STICKY and ER-DD models are constructed to preserve the degree distribution of the data. These and other data-driven models of random networks[62–64] are thus expected to model particular PPI networks better than theoretical network models. However, they are not an appropriate choice to judge whether the yeast2 and human1 networks share a significant amount of structural similarity; this is because these models are strongly conditioned on these particular networks and thus they might transfer onto the model networks the similarities between yeast2 and human1 that we aim to detect in the first place. Thus, we search for a well-fitting *theoretical* null model. Arguably the best currently known theoretical model for PPI networks, requiring the fewest tunable parameters, is the *geometric random graph* model ("GEO"),[37,39,65] in which proteins are modeled as existing in a metric space and are connected by an edge if they are within a fixed, specified distance of each other.

Although early, incomplete PPI datasets were modeled well by scale-free networks because of their power-law degree distributions,[47,66] it has been argued that such degree distributions were an artifact of noise.[67–69] In the light of new PPI network data, several studies[37,39,65] have presented compelling evidence that the structure of PPI networks is closer to geometric than to scale-free networks. This was done by comparing frequencies of graphlets in real-world and model networks[37] and by measuring a highly-constraining agreement between "graphlet degree distributions."[39] Finally, it has been shown that PPI networks can be successfully embedded into a low-dimensional Euclidean space, thus directly confirming that they have a geometric structure.[65] The superior fit of the GEO model to PPI networks over other models may not be surprising, since it can be biologically motivated. In particular, the currently accepted paradigm for evolution is based on a series of gene duplication and mutation events. We outline our crude *geometric gene duplication* model.[70] We model genes, and proteins as their products, as existing in some biochemical metric space. Although the dimension and axes of this space are not obvious, we assume that when a parent gene is duplicated, the child gene starts at a similar location in the metric space, since it is structurally identical to the parent and thus inherits interactions from the parent. As mutations and "evolutionary optimization" act on the child, it drifts away from the parent in the metric space. The child may preserve some of the parent's interacting partners, but it may also establish new interactions with other genes.[70] Similarly, in a geometric graph, the closer two nodes are to each other, the more interactors they will have in common, and vice-versa. In addition to PPI networks, GEO is a well-fitting theoretical null model for other biological networks, e.g., brain function networks[71] and protein structure networks.[72]

Accepting GEO as the optimal null model for PPI networks, we compute the probability of obtaining the EC of 11.72% in our alignment of yeast2 and human1 to be 8.4×10^{-3}. We do so by aligning with GRAAL pairs of GEO networks of the same size as yeast2 and human1 and by applying the following form of the Vysochanskij–Petunin inequality: $P(|X - \mu| \geq \lambda\sigma) \leq \frac{4}{9\lambda^2}$. Since GEO networks that are aligned have *the same* number of nodes and edges as the data, it is reasonable to assume that the distribution of their alignment scores is unimodal. Thus, we use the Vysochanskij–Petunin inequality, since it is more precise than Chebyshev's inequality for unimodal distributions. More details are supplied in the Supplementary Information.

Acknowledgments

We thank M. Rašajski for computational assistance. This project was supported by the NSF CAREER IIS-0644424 grant.

References

1. Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., Sakaki, Y.: A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci U S A **98**(8) (2001) 4569–4574
2. Ito, T., Tashiro, K., Muta, S., Ozawa, R., Chiba, T., Nishizawa, M., Yamamoto, K., Kuhara, S., Sakaki, Y.: Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. Proc Natl Acad Sci U S A **97**(3) (2000) 1143–7
3. Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, E., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleish, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., Rothberg, J.M.: A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. Nature **403** (2000) 623–627
4. Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., Timm, J., Mintzlaff, S., Abraham, C., Bock, N., Kietzmann, S., Goedde, A., Toksoz, E., Droege, A., Krobitsch, S., Korn, B., Birchmeier, W., Lehrach, H., Wanker, E.: A human protein-protein interaction network: A resource for annotating the proteome. Cell **122** (2005) 957–968
5. Rual, J., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G.F., Gibbons, F.D., Dreze, M., Ayivi-Guedehoussou, N., Klitgord, N., Simon, C., Boxem, M., Milstein, S., Rosenberg, J., Goldberg, D.S., Zhang, L.V., Wong, S.L., Franklin, G., Li, S., Albala, J.S., Lim, J., Fraughton, C., Llamosas, E., Cevik, S., Bex, C., Lamesch, P., Sikorski, R.S., Vandenhaute, J., Zoghbi, H.Y., Smolyar, A., Bosak, S., Sequerra, R., Doucette-Stamm, L., Cusick, M.E., Hill, D.E., Roth, F.P., Vidal, M.: Towards a proteome-scale map of the human protein-protein interaction network. Nature **437** (2005) 1173–78
6. Simonis, N., Rual, J.F., Carvunis, A.R., Tasan, M., Lemmens, I., Hirozane-Kishikawa, T., Hao, T., Sahalie, J.M., Venkatesan, K., Gebreab, F., Cevik, S., Klitgord, N., Fan, C., Braun, P., Li, N., Ayivi-Guedehoussou, N., Dann, E., Bertin, N., Szeto, D., Dricot, A., Yildirim, M.A., Lin, C., de Smet, A.S., Kao, H.L., Simon, C., Smolyar, A., Ahn, J.S., Tewari, M., Boxem, M., Milstein, S., Yu, H., Dreze, M., Vandenhaute, J., Gunsalus, K.C., Cusick, M.E., Hill, D.E., Tavernier, J., Roth, F.P., Vidal, M.: Empirically controlled mapping of the Caenorhabditis elegans protein-protein interactome network. Nat Meth **6**(1) (December 2008) 47–54
7. Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M., Seraphin, B.: A generic protein purification method for protein complex characterization and proteome exploration. Nature Biotechnol. **17** (1999) 1030–1032
8. Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreault, M., Muskat, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A.R., Sassi, H., Nielsen, P.A., Rasmussen, K.J., Andersen, J.R., Johansen, L.E., Hansen, L.H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., Sorensen, B.D., Matthiesen, J., Hendrickson, R.C., Gleeson, F., Pawson, T., Moran, M.F., Durocher, D., Mann, M., Hogue, C.W., Figeys, D., Tyers, M.: Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry. Nature **415**(6868) (2002) 180–3
9. Gavin, A.C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.M., Cruciat, C.M., Remor, M., Hofert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M.A., Copley, R.R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G., Superti-Furga, G.: Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature **415**(6868) (2002) 141–7
10. Krogan, N.J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A.P., Punna, T., PeregrÃn Alvarez, J.A.M., Shales, M., Zhang, X., Davey, M., Robinson, M.D., Paccanaro, A., Bray, J.E., Sheung, A., Beattie, B., Richards, D.P., Canadien, V., Lalev, A., Mena, F., Wong, P., Starostine, A., Canete, M.M., Vlasblom, J., Wu, S., Orsi, C., Collins, S.R., Chandran, S., Haw, R., Rilstone, J.J., Gandi, K., Thompson, N.J., Musso, G., St Onge, P., Ghanny, S., Lam, M.H.Y., Butland, G., Altaf-Ul, A.M., Kanaya, S., Shilatifard, A., O'Shea, E., Weissman, J.S., Ingles, C.J., Hughes, T.R., Parkinson, J., Gerstein, M., Wodak, S.J., Emili, A., Greenblatt, J.F.: Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature **440** (2006) 637–643

11. Gavin, A.C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L.J., Bastuck, S., Dumpelfeld, B., Edelmann, A., Heurtier, M.A., Hoffman, V., Hoefert, C., Klein, K., Hudak, M., Michon, A.M., Schelder, M., Schirle, M., Remor, M., Rudi, T., Hooper, S., Bauer, A., Bouwmeester, T., Casari, G., Drewes, G., Neubauer, G., Rick, J.M., Kuster, B., Bork, P., Russell, R.B., Superti-Furga, G.: Proteome survey reveals modularity of the yeast cell machinery. Nature (440) (2006) 631636
12. Collins, S., Kemmeren, P., Zhao, X., Greenblatt, J., Spencer, F., Holstege, F., Weissman, J., Krogan, N.: Toward a comprehensive atlas of the physical interactome of saccharomyces cerevisiae. Molecular and Cellular Proteomics **6**(3) (2008) 439–450
13. Sharan, R., Ideker, T.: Modeling cellular machinery through biological network comparison. Nature Biotechnology **24**(4) (Apr 2006) 427–433
14. Colizza, V., Flammini, A., Serrano, M.A., Vespignani, A.: Detecting rich-club ordering in complex networks. Nature Physics **2** (2006) 110–115
15. Guimera, R., Sales-Pardo, M., Amaral, L.A.N.: Classes of complex networks defined by role-to-role connectivity profiles. Nature Physics **3** (2007) 63–69
16. Cook, S.: The complexity of theorem-proving procedures. In: Proc. 3rd Ann. ACM Symp. on Theory of Computing: 1971; New York, Assosiation for Computing Machinery (1971) 151–158
17. Venkatesan, K.e.a.: An empirical framework for binary interactome mapping. Nature Methods **6**(1) (2009) 83–90
18. Singh, R., Xu, J., Berger, B.: Global alignment of multiple protein interaction networks. Proceedings of Pacific Symposium on Biocomputing 13 (2008) 303–314
19. Zaslavskiy, M., Bach, F., Vert, J.P.: Global alignment of protein-protein interaction networks by graph matching methods. Bioinformatics **25**(12) (2009) i259–i267
20. Kelley, B.P., Bingbing, Y., Lewitter, F., Sharan, R., Stockwell, B.R., Ideker, T.: PathBLAST: a tool for alignment of protein interaction networks. Nucl. Acids Res. **32**(Web Server issue) (2004) W83–W88
21. Berg, J., Lassig, M.: Local graph alignment and motif search in biological networks. PNAS **101** (2004) 14689–14694
22. Flannick, J., Novak, A., Balaji, S., Harley, H., Batzglou, S.: Graemlin general and robust alignment of multiple large interaction networks. Genome Res **16**(9) (2006) 1169–1181
23. Liang, Z., Xu, M., Teng, M., Niu, L.: NetAlign: a web-based tool for comparison of protein interaction networks. Bioinformatics **22**(17) (2006) 2175–2177
24. Berg, J., Lassig, M.: Cross-species analysis of biological networks by Bayesian alignment. Proceedings of the National Academy of Sciences **103**(29) (2006) 10967–10972
25. Sharan, R.: Conserved patterns of protein interaction in multiple species. Proc. Natl. Acad. Sci. USA **102** (2005) 1974–1979
26. Koyuturk, M., Kim, Y., Topkara, U., Subramaniam, S., Szpankowski, W., Grama, A.: Pairwise alignment of protein interaction networks. Journal of Computational Biology **13**(2) (March 2006) 182–199
27. Singh, R., Xu, J., Berger, B.: Pairwise global alignment of protein interaction networks by matching neighborhood topology. In: Research in Computational Molecular Biology. Springer (2007) 16–31
28. Flannick, J., Novak, A.F., Do, C.B., Srinivasan, B.S., Batzoglou, S.: Automatic parameter learning for multiple network alignment. In: RECOMB. (2008) 214–231
29. Liao, C.S., Lu, K., Baym, M., Singh, R., Berger, B.: Isorankn: spectral methods for global alignment of multiple protein networks. Bioinformatics **25**(12) (2009) i253–258
30. Komili, S., Farny, N.G., Roth, F.P., Silver, P.A.: Functional specificity among ribosomal proteins regulates gene expression. Cell **131**(3) (2007) 557–571
31. Watson, J.D., Laskowski, R.A., Thornton, J.M.: Predicting protein function from sequence and structural data. Current opinion in structural biology **15**(3) (2005) 275–284
32. Whisstock, J.C., Lesk, A.M.: Prediction of protein function from protein sequence and structure. Q Rev Biophys **36**(3) (2003) 307–340
33. Kosloff, M., Kolodny, R.: Sequence-similar, structure-dissimilar protein pairs in the pdb. Proteins **71**(2) (2008) 891–902
34. Memišević, V., Milenković, T., Pržulj, N.: Complementarity of network and sequence structure in homologous proteins. (2009) under review.
35. Laurents, D.V., Subbiah, S., Levitt, M.: Different protein sequences can give rise to highly similar folds through different stabilizing interactions. Protein Sci **3**(11) (1994) 19381944
36. The Gene Ontology Consortium: Gene ontology: tool for the unification of biology. Nature Genetics **25** (2000) 25–29
37. Pržulj, N., Corneil, D.G., Jurisica, I.: Modeling interactome: Scale-free or geometric? Bioinformatics **20**(18) (2004) 3508–3515
38. Pržulj, N., Corneil, D.G., Jurisica, I.: Efficient estimation of graphlet frequency distributions in protein-protein interaction networks. Bioinformatics **22**(8) (2006) 974–980
39. Pržulj, N.: Biological network comparison using graphlet degree distribution. Bioinformatics **23** (2007) e177–e183
40. Milenković, T., Pržulj, N.: Uncovering biological network function via graphlet degree signatures. Cancer Informatics **6** (2008) 257–273

41. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature **393** (1998) 440–442
42. Altschul, S.F., Gish, W., Miller, W., Lipman, D.J.: Basic local alignment search tool. Journal of Molecular Biology **215** (1990) 403–410
43. Radivojac, P., Peng, K., Clark, W.T., Peters, B.J., Mohan, A., Boyle, S.M., D., M.S.: An integrated approach to inferring gene-disease associations in humans. Proteins **72**(3) (2008) 1030–1037
44. Stark, C., Breitkreutz, B., Reguly, T., Boucher, L., Breitkreutz, A., Tyers, M.: BioGRID: A general repository for interaction datasets. Nucleic Acids Research **34** (2006) D535–D539
45. Peri, S., Navarro, J.D., Kristiansen, T.Z., Amanchy, R., Surendranath, V., Muthusamy, B., Gandhi, T.K., Chandrika, K.N., Deshpande, N., Suresh, S., Rashmi, B.P., Shanker, K., Padma, N., N iranjan, V., Harsha, H.C., Talreja, N., Vrushabendra, B.M., Ramya, M.A., Yatish, A.J., Joy, M., S hivashankar, H.N., Kavitha, M.P., Menezes, M., Choudhury, D.R., Ghosh, N., Saravana, R., Chandran, S., Mohan, S., Jonnalagadda, C.K., Prasad, C.K., Kumar-Sinha, C., Deshpande, K.S., Pandey, A.: Human protein reference database as a discovery resource for proteomics. Nucleic Acids Res **32** (2004) D497–501
46. Milenković, T., Lai, J., Pržulj, N.: Graphcrunch: a tool for large network analyses. BMC Bioinformatics **9**(70) (2008)
47. Barabási, A., Albert, R.: Emergence of scaling in random networks. Science **286**(5439) (1999) 509–512
48. Pržulj, N., Higham, D.: Modelling protein-protein interaction networks via a stickiness index. Journal of the Royal Society Interface **3**(10) (2006) 711–716
49. Labarga, A., Valentin, F., Andersson, M., Lopez, R.: Web services at the european bioinformatics institute. Nucleic Acids Research **35**(Web Server issue) (2007) W6–W11
50. Forst, C., Schulten, K.: Phylogenetic analysis of metabolic pathways. J Mol Evol **52** (2001) 471–489
51. Heymans, M., Singh, A.: Deriving phylogenetic trees from the similarity analysis of metabolic pathways. Bioinformatics **19** (2003) i138–i146
52. Zhang, Y., Li, S., Skogerb, G., Zhang, Z., Zhu, X., Zhang, Z., Sun, S., Lu, H., Shi, B., Chen, R.: Phylophenetic properties of metabolic pathway topologies as revealed by global analysis. BMC Bioinformatics **7:252** (2006)
53. Suthram, S., Sittler, T., Ideker, T.: The plasmodium protein network diverges from those of other eukaryotes. Nature **438** (2005) 108–112
54. Webb, E.C.: Enzyme nomenclature 1992: recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes, the University of Michigan (1992)
55. Agrafioti, I., Swire, J., Abbott, J., Huntley, D., Butcher, S., Stunpf, M.P.: Comparative analysis of the saccharomyces cerevisiae and caenorhabdits elegans protein interaction networks. BMC Evolutionary Biology **5**(23) (2005)
56. Kanehisa, M., Goto, S.: Kegg: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. **28** (2000) 27–30
57. Pennisi, E.: Modernizing the tree of life. Science **300**(5626) (2003) 1692 – 1697
58. Keeling, P., Luker, M., Palmer, J.: Evidence from beta-tubulin phylogeny that microsporidia evolved from within the fungi. Mol. Biol. Evol. **17**(1) (2000) 23–31
59. Tanriverdi, S., Widmer, G.: Differential evolution of repetitive sequences in cryptosporidium parvum and cryptosporidium hominis. Infect. Genet. Evol. **6**(2) (2006) 113–22
60. Xu, P., Widmer, G., Wang, Y., Ozaki, L., Alves, J., Serrano, M., Puiu, D., Manque, P., Akiyoshi, D., Mackey, A., Pearson, W., Dear, P., Bankier, A., Peterson, D., Abrahamsen, M., Kapur, V., Tzipori, S., Buck, G.: The genome of cryptosporidium hominis. Nature **431**(7012) (2004) 1107–12
61. Out, H., Sayood, K.: A new sequence distance measure for phylogenetic tree construction. Bioinformatics **19**(16) (2003) 2122–2130
62. Thorne, T., Stumpf, M.: Generating confidence intervals on biological networks. BMC Bioinformatics **8**(1) (2007) 467
63. Snijders, T.A.: Markov chain monte carlo estimation of exponential random graph models. Journal of Social Structure **3**(2) (2002) 2–40
64. Kuchaiev, O., Pržulj, N.: Learning the structure of protein-protein interaction networks. Pacific Symposium on Biocomputing (2009) 39–50
65. Higham, D., Rašajski, M., Pržulj, N.: Fitting a geometric graph to a protein-protein interaction network. Bioinformatics **24(8)** (2008) 1093–1099
66. Jeong, H., Mason, S.P., Barabási, A.L., Oltvai, Z.N.: Lethality and centrality in protein networks. Nature **411**(6833) (2001) 41–2
67. Stumpf, M.P.H., Wiuf, C., May, R.M.: Subnets of scale-free networks are not scale-free: Sampling properties of networks. Proceedings of the National Academy of Sciences **102** (2005) 4221–4224
68. Han, J.D.H., Dupuy, D., Bertin, N., Cusick, M.E., Vidal, M.: Effect of sampling on topology predictions of protein-protein interaction networks. Nature Biotechnology **23** (2005) 839–844
69. de Silva, E., Thorne, T., Ingram, P., Agrafioti, I., Swire, J., Wiuf, C., Stumpf, M.P.: The effects of incomplete protein interaction data on structural and evolutionary inferences. BMC Biol. **4**(39) (2006)
70. Pržulj, N., Kuchaiev, O., Stevanović, A., Hayes, W.: Geometric evolutionary dynamics of protein interaction networks. Pacific Symposium on Biocomputing (2010) to appear

71. Kuchaiev, O., Wang, P.T., Nenadić, Z., Pržulj, N.: Structure of brain functional networks. 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (2009)
72. Milenković, T., Filippis, I., Lappe, M., Pržulj, N.: Optimized null model for protein structure networks. PLoS ONE **4**(6) (2009) e5967

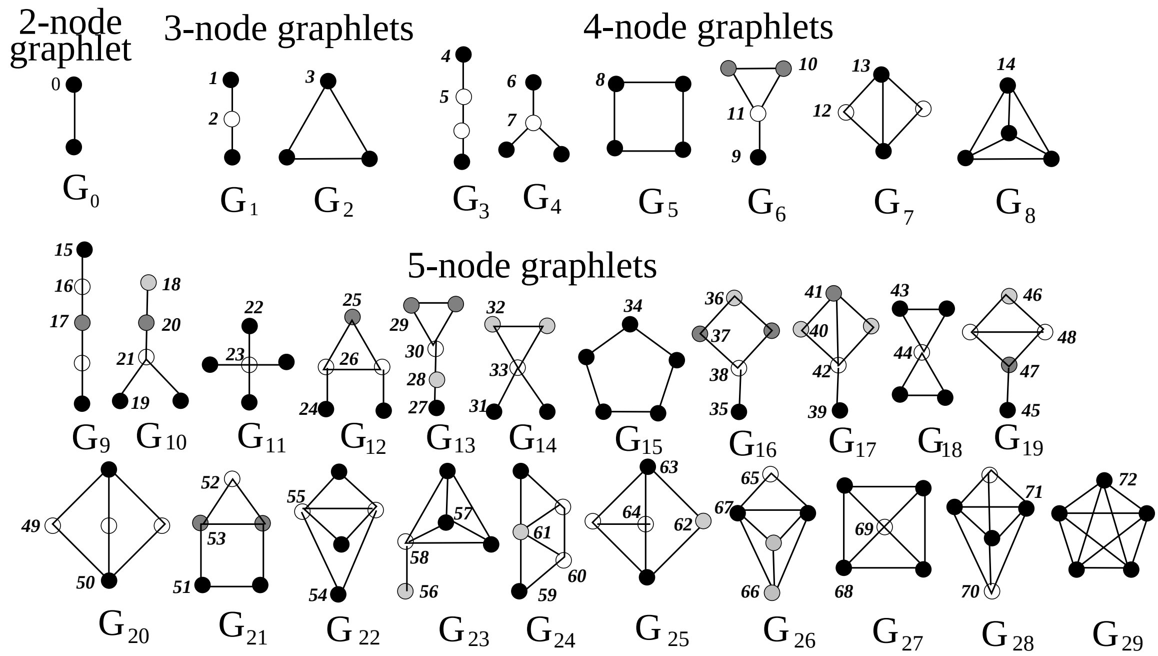
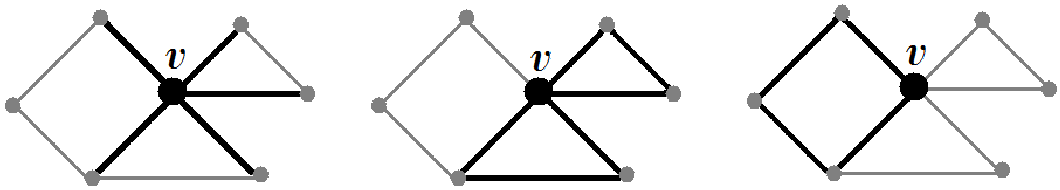
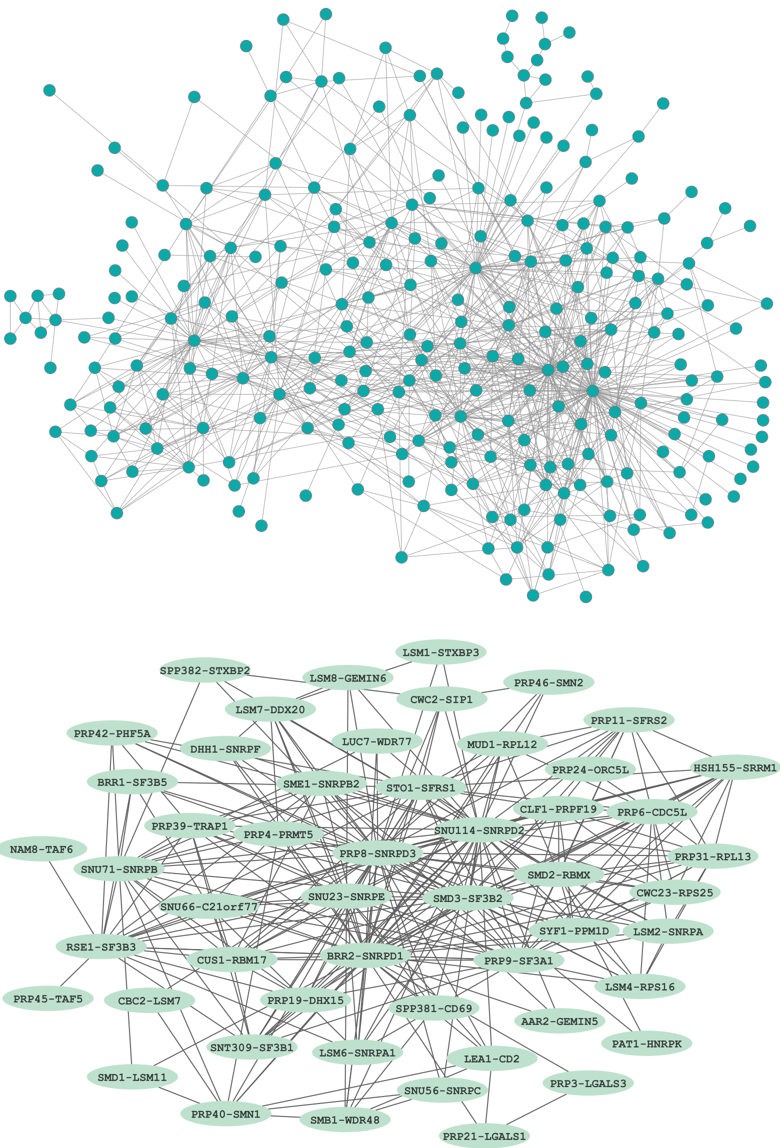


Fig. 1. All the connected graphs on up to 5 nodes. When appearing as an induced subgraph of a larger graph, we call them *graphlets*. They contain 73 topologically unique node types, called "automorphism orbits." In a particular graphlet, nodes belonging to the same orbit are of the same shade. Graphlet G_0 is just an edge, and the degree of a node historically defines how many edges it touches. We generalize the degree to a 73-component "graphlet degree" vector that counts how many times a node is touched by each particular automorphism orbit.[39]



Orbit	0	1	2	3	4	5	6	7	8	9	10	11	12…20	21	22…25	26	27…29	30	31	32	33	34…37	38	39…43	44	45…52	53	54…72
GDV(v)	5	2	8	2	0	5	0	4	1	0	1	6	0…0	2	0…0	2	0…0	2	0	0	4	0…0	2	0…0	1	0…0	1	0…0

Fig. 2. An illustration of how the degree of node v in the leftmost panel is generalized into its "graphlet degree vector," or "signature," that counts the number of different graphlets that the node touches, such as triangles (middle panel) or squares (rightmost panel). Values of the 73 coordinates of the graphlet degree vector of node v, $GDV(v)$, are presented in the table.



(A)

(B)

Fig. 3. The alignment of yeast2 and human1 PPI networks. An edge between two nodes means that an interaction exists in both species between the corresponding protein pairs. Thus, the displayed networks appear, in their entirety, in the PPI networks of both species. (A) The largest *common connected subgraph* (CCS) consisting of 900 interactions amongst 267 proteins. (B) The second largest CCS consisting of 286 interactions amongst 52 proteins; each node contains a label denoting a pair of yeast and human proteins that are aligned.

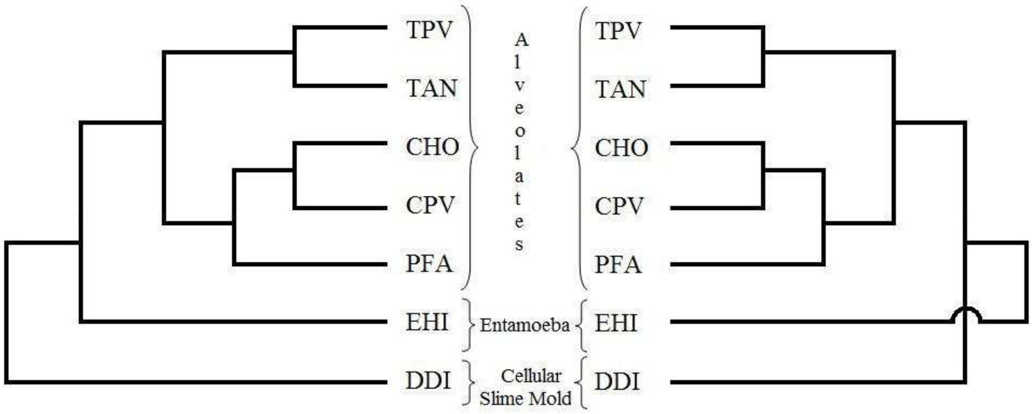


Fig. 4. Comparison of the phylogenetic trees for protists obtained by genetic sequence alignments and by GRAAL's metabolic network alignments. Left: The tree obtained from genetic sequence comparison.[57] Right: The tree obtained from GRAAL. The following abbreviations are used for species: CHO - Cryptosporidium hominis, DDI - Dictyostelium discoideum, CPV - Cryptosporidium parvum, PFA - Plasmodium falciparum, EHI - Entamoeba histolytica, TAN - Theileria annulata, TPV - Theileria parva. The species are grouped into the following classes: "Alveolates," "Entamoeba," and "Cellular Slime mold."