Complementarity of network and sequence information in homologous proteins

Vesna Memišević[1], Tijana Milenković[1], and Nataša Pržulj[2,*]

[1]Department of Computer Science, University of California, Irvine, CA 92697-3435, USA
[2]Department of Computing, Imperial College London, London, SW7 2AZ, UK
[*]Corresponding author (e-mail: natasha@imperial.ac.uk)

Summary

Traditional approaches for homology detection rely on finding sufficient similarities between protein sequences. Motivated by studies demonstrating that from non-sequence based sources of biological information, such as the secondary or tertiary molecular structure, we can extract certain types of biological knowledge when sequence-based approaches fail, we hypothesize that protein-protein interaction (PPI) network topology and protein sequence might give insights into different slices of biological information. 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 the protein's role in the inner working of the cell than analyzing sequences of individual genes. Hence, we believe that one could lose much information by focusing on sequence information alone.

We examine whether the information about homologous proteins captured by PPI network topology differs and to what extent from the information captured by their sequences. We measure how similar the topology around homologous proteins in a PPI network is and show that such proteins have statistically significantly higher network similarity than non-homologous proteins. We compare these network similarity trends of homologous proteins with the trends in their sequence identity and find that network similarities uncover almost as much homology as sequence identities. Although none of the two methods, network topology and sequence identity, seems to capture homology information in its entirety, we demonstrate that the two might give insights into somewhat different types of biological information, as the overlap of the homology information that they uncover is relatively low. Therefore, we conclude that similarities of proteins' topological neighborhoods in a PPI network could be used as a complementary method to sequence-based approaches for identifying homologs, as well as for analyzing evolutionary distance and functional divergence of homologous proteins.

1 Introduction

Homology detection is an important problem in computational biology [1], as it can be used to estimate the closeness of genomes of different species. Proteins are "homologs" if they descend from a common ancestor, i.e., if they are related by evolutionary process of divergence. Homologs are a superset of paralogs and orthologs. "Paralogs" are genes in a same species that evolve from a common ancestor through gene duplication events. "Orthologs" are genes in different species that evolve from a common ancestor through speciation events. The notions of paralogy and orthology are closely linked. If, for example, a gene duplication occurred after the speciation event that separated species of interest, then orthology becomes a relationship between sets of paralogs (or "co-orthologs") resulting from the duplication, rather than between

individual genes [2]. Furthermore, one needs to distinguish between "in-paralogs" i.e., paralogs within the same species, as defined above, and "out-paralogs," i.e., paralogs that result from a duplication event prior to the last speciation event [3].

Traditional approaches identify homologs by finding sufficient similarities between their sequences. *Sequence alignment* [4, 5] is a way of arranging protein sequences to identify regions of similarity between them. Sequence alignment generates an alignment by starting at the ends of two protein sequences, attempting to match all possible pairs of amino acids between the sequences. *Global* alignments attempt to align every amino acid in two sequences and are generally useful for similar sequences of roughly equal size. On the other hand, *local* alignments attempt to find regions of local similarity between sequences and are generally useful for less similar sequences. Clearly, there could be a large number of ways to align two sequences. To find the best alignment, one needs to use a scoring scheme to evaluate the goodness of the alignment; the scoring scheme takes into consideration matches, mismatches, and gaps between the aligned sequences. The alignment with the highest sequence similarity score amongst all alignments is chosen, and the resulting *sequence similarity score* describes the extent to which two sequences are related. Since, for example, longer sequences are more likely to produce higher sequence similarity scores than shorter sequences, one needs to evaluate the statistical significance of the observed sequence alignment score, by measuring how likely it is to obtain the score at random.

The extent of similarity between two sequences can also be expressed in terms of their percent *sequence identity*. Given an alignment of two sequences, sequence identity is the percentage of amino acids in the shorter sequence that are matched to exactly the same amino acids in the longer sequence. The degree to which sequences differ is qualitatively related to their evolutionary distance. Roughly, high identity between two sequences suggests that the sequences have a comparatively young most recent common ancestor, while low identity suggests that the divergence is more ancient.

Analogous to sequence-based comparisons, network comparisons across species have also been used to identify proteins with similar functions and detect orthologs [1, 6, 7, 8, 9, 10, 11, 12]. However, almost all of these network comparison methods rely mostly on sequence information and use only limited network topological information. The challenge is to identify functional orthologs solely from network topology.

We hypothesize that network topology and protein sequence might give insights into different slices of biological information and thus, one could lose much information by focusing on sequence alone. It has already been shown that non-sequence based sources of biological information might be more appropriate to extract certain types of biological knowledge than sequence-based ones [13, 14, 15]. For example, it has been argued that the primary structure (i.e., sequence) information may give only limited insights into RNA, and that the use of information on the secondary and tertiary structure of RNA is essential; this is especially true because these higher-order structures play the dominant role in RNA function [13]. Additionally, it has been demonstrated that it is not the primary sequences, but the well conserved secondary structural patterns across species that are the relevant property of ribosomal RNAs [14]. Moreover, a secondary structure-based approach has been found to be more effective at finding new functional RNAs than sequence-based alignments [15]. Similarly, in the context of duplicated proteins, patterns of their interactions in PPI networks may provide new insights into the evolutionary fate and functional role of each protein, complementing the knowledge learned

from their genomic sequences [16]. Furthermore, wiring patterns of duplicated proteins in a PPI network are expected to give insights into their evolutionary distances, since the number of interacting partners shared by yeast paralogs has been shown to decrease rapidly over evolutionary time scales [17, 18], even when their coding sequences are almost perfectly conserved [16]. Thus, PPI networks present opportunity to systematically study the evolutionary distance and functional divergence of retained gene duplicates with respect to their interaction patterns [16]. Furthermore, we hypothesize that, 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 the inner working of the cell than analyzing sequences of individual genes. For these reasons, we examine whether the information about homologous proteins captured by their network topology differs and to what extent from the information captured by their sequences.

2 Methods

2.1 Our approach

We analyze the high-confidence physical PPI network of yeast *S. cerevisiae* [19], containing 1,621 proteins and 9,074 interactions. To assess topological properties of proteins in the PPI network, we rely on their "graphlet degree signatures" (or just "signatures", for brevity), topological descriptors of proteins' extended PPI network neighborhoods that capture their interconnectivities out to a distance of 4 (see Section 2.2 for details). We also rely on our "signature similarity" measure that compares signatures of two proteins and thus measures the topological resemblance of their network neighborhoods, where a higher signature similarity between two proteins corresponds to a higher topological similarity between their extended network neighborhoods (see Section 2.2 for details). When protein signatures are computed, all proteins and interactions from the network are taken into consideration. However, in all of our subsequent analyzes, we focus only on proteins with more than three interacting partners, since poorly connected proteins are more likely to be in incomplete parts of a PPI network [20, 21]. In the yeast PPI network, 920 out of 1,621 proteins have degrees higher than 3.

We analyze the data sets available in Clusters of Orthologous Groups (COGs) of Proteins System [22] and KEGG Orthology (KO) System [23]. These databases contain groups of proteins from different organisms that consist of orthologs and in-paralogs, as explained below. For simplicity, henceforth we refer to these groups as "groups of orthologous proteins."

The groups of orthologous proteins in COGs System were formed as follows [22]. All pairwise sequence comparisons among proteins encoded in different genomes were performed, and for each protein, the best hit ("BeT") in each of the other genomes was detected. The identification of an orthology group was based on consistent patterns, such as a triangle, in the graph of BeTs. If a gene from one of the compared genomes had BeTs in two other genomes, it was highly unlikely that the respective genes were also BeTs for one another unless they were indeed orthologs. Thus, the consistency between BeTs resulting in triangles did *not* depend on the absolute level of similarity between the compared proteins. The groups were then produced by merging adjacent triangles that had a common side, and thus, the resulting groups typically contained both orthologs from different species and paralogs from the same species.

The groups of orthologous proteins in KO System were formed as follows [23]. The bit scores were defined for each gene against all other genes in KEGG, which were 1 if the p-value of the sequence alignment score was lower than 1.0^{-8}, and 0, otherwise. Then, the similarity profile of each gene was defined with a vector of bit scores with respect to all other genes. Next, a graph was constructed in which nodes corresponded to protein sequences and edges were labeled with correlation coefficients of above profiles for the corresponding sequences. Similar to COGs, groups in KEGG were then constructed automatically by searching in the graph for cliques with an appropriate definition for the profiles of similarity scores. Thus, again, the memberships in groups did *not* depend on the absolute level of similarity between the compared proteins.

In our study, we focus on yeast genes only. For each of the COGs and KO groups, we extract all possible pairs of yeast proteins in the group. For simplicity, henceforth we call all of the extracted protein pairs "orthologous pairs." COGs System contains 4,004 yeast genes divided into 4,852 COGs. These 4,004 genes result in 8,014 pairs of yeast orthologs, where proteins in each protein pair belong to a same COG. KO System contains 2,123 yeast genes divided into 1,540 KO groups. These 2,123 genes result in 2,354 pairs of yeast orthologs, where proteins in each protein pair belong to a same KO group. (Note that a protein can have multiple orthologs, and thus, it can belong to more than one pair; this is why there are more pairs than genes in each of the COGs and KO Systems.) There are 9,643 unique orthologous pairs in COGs and KO Systems together, of which 175 pairs are found in the yeast PPI network and have degrees higher than 3. These 175 pairs are composed of 181 unique proteins.

We analyze topological signatures of these 175 orthologous protein pairs in the yeast PPI network with the hypothesis that their connectivity patterns in the PPI network are more similar than those of non-orthologous proteins. In addition, we compute their sequence identities; we do so by using Smith-Waterman local alignment algorithm with BLOSUM50 substitution matrix as the scoring scheme. Then, we compare signature similarity trends of these orthologous protein pairs with their sequence identity trends to find out if more information about their orthology can be captured by sequence identities or by signature similarities.

2.2 Graphlet degree signatures and signature similarities

To determine topological similarity between two proteins in the PPI network, we use the similarity measure of nodes' local neighborhoods, as described by Milenković and Pržulj [21]. This measure generalizes the degree of a node, which counts the number of edges that the node touches, into the vector of *graphlet degrees*, that counts the number of graphlets that the node touches, for all 2-5-node graphlets (see Supplementary Figure S1). Since it is topologically relevant to distinguish between, for example, nodes touching graphlet G_1 (Supplementary Figure S1) 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 2-5-node graphlets [24]. The full vector of 73 coordinates is the *signature* of a node that describes the topology of its neighborhood and captures its interconnectivities out to distance 4 (see [21] for details).

The signature of a node is a highly constraining measure of local topology in the node's vicinity in the network and comparing the signatures of two nodes is a demanding measure of their network similarity. The node signature similarities are computed as follows. For a node u,

u_i denotes the i^{th} coordinate of its signature vector, i.e., u_i is the number of times node u touches an orbit i. 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 signifying its "importance" (see [21] for details). 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: $S(u,v) = 1 - D(u,v)$ (see [21] for details). Clearly, a higher signature similarity between two nodes corresponds to a higher topological similarity between their extended neighborhoods (out to distance 4).

3 Results

Orthologous proteins are assumed to perform the same or similar biological function. Since we hypothesize that network topology could give insights into orthology, we examine whether proteins with a higher signature similarity are more likely to share a common biological function (Figure 1).

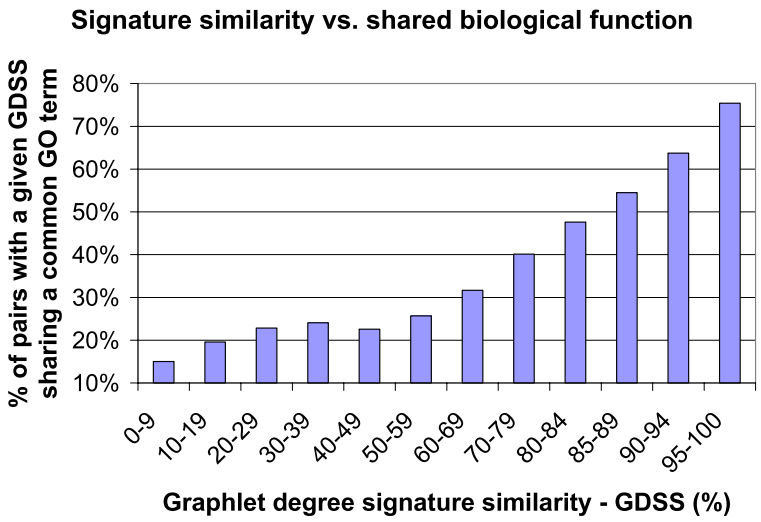


Figure 1: The relationship between topological graphlet degree signature similarities (GDSS) and the level of biological similarities of orthologous protein pairs in the baker's yeast PPI network. For each signature similarity interval illustrated on x-axis (0%-9%, 10%-19%, ..., 90%-94%, and 95%-100%), we show the percentage of protein pairs having signature similarities within the interval that share a common function.

We analyze GO annotation data [25] downloaded in September 2009. We show that with increased signature similarity, a protein pair is more likely to share a common GO term (Figure 1). For example, out of all protein pairs in the yeast PPI network with signature similarities below 10%, only about 15% of the pairs share a common GO term, whereas the same is true for more than 25% of protein pairs with signature similarities of between 50% and 59%, and for more than 75% of protein pairs with signature similarities of above 95%. Thus, we further demonstrate that proteins' topological signatures in the PPI network are closely related to their biological function [21]. Moreover, our result is encouraging, given that the GO data that we consider is the "complete" annotation data, containing all GO annotations, independent of GO

evidence code, in which most of the annotations were obtained computationally by sequence-based analyses [25]. Thus, since proteins' topological similarities within the PPI network can recover the biological information obtained by analyzing their sequence similarities, next we examine whether they can give insights into their homology relationships as well.

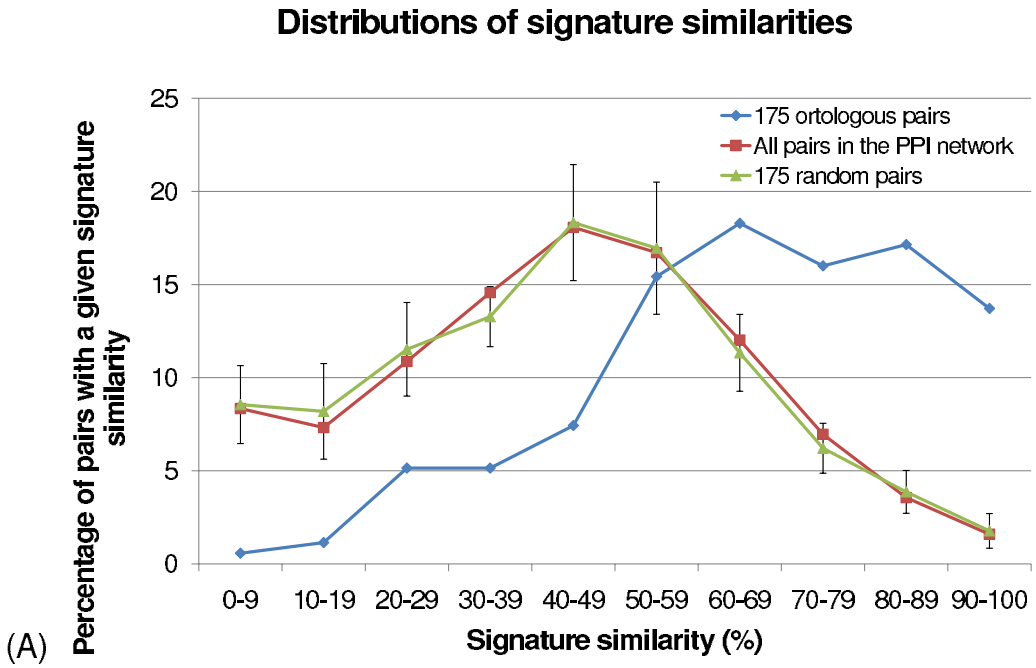
We demonstrate that proteins' graphlet degree signature similarities are capable of capturing homology. 175 pairs of orthologous proteins have very different signature similarity distribution, with higher signature similarities than all protein pairs in the PPI network (Figure 2 A).

By "approximating" the distribution of signature similarities between all protein pairs in the network presented in Figure 2 A with the normal distribution having the same average and standard deviation as the data, and by finding Z-scores and their corresponding p-values for different signature similarity thresholds, we find that the statistically significant signature similarity threshold is 85%, with p-value lower than 0.05. That is, the probability that a protein pair randomly selected among all pairs in the PPI network would have a signature similarity of above 85% is lower than 0.05. We find that a large percentage of orthologous pairs, i.e., more than 20% of them, have statistically significant signature similarities of above 85%.

We examine how likely it is to observe at random such high signature similarities that we observe for orthologous proteins. We select randomly 175 protein pairs from the PPI network and compare their signature similarities to those of the orthologous pairs, repeating the procedure of random pair selection 30 times (Figure 2 A). Clearly, signature similarity distribution for the 175 orthologous pairs is different than that for 175 random pairs. Moreover, the distribution for random pairs is very similar to the distribution for all protein pairs in the network. This suggests that the signature similarity measure indeed successfully captures homology.

To test the robustness of the signature similarity measure to noise in PPI networks, we randomly add, delete, and rewire 10%, 20%, and 30% of edges in the PPI network, repeating each randomization procedure 30 times. Then, we recompute in these randomly perturbed PPI networks the signatures for all proteins in the 175 orthologous pairs and we find the signature similarity distributions for the orthologous pairs in such randomized networks (see Supplementary Figure S2). None of the random perturbations introduces a big change to the distribution of signature similarities of orthologous pairs, demonstrating that our approach is robust to noise. It is interesting that random edge additions and rewirings slightly increase the percentage of pairs with high signature similarities (Supplementary Figure S2). Explaining why this happens is a subject of future research.

Next, we calculate sequence identities for 175 orthologous protein pairs, all protein pairs from the PPI network, and 175 protein pairs randomly selected from the PPI network, where the procedure of random pair selection is repeated 30 times. About 70% of our orthologous pairs have sequence identities under 35% (Figure 2 B). Our result is consistent to that of Rost [26] establishing that the vast majority of homologs has such low sequence identities. Somewhat similar distribution is observed for all protein pairs in the network (Figure 2 B), as well for 175 protein pairs randomly selected from the network. For all three distributions, almost all pairs have low sequence identities, below 40%; the only exception are 20% of orthologous pairs that have very high sequence identities of above 90%. These observations could be explained as follows. In the context of sequence alignments, the region of sequence identity between 20% and 35% is referred to as the "twilight zone" [27, 26]. Sequence identities above 35% imply that the sequences of interest are highly similar, evolutionary close, and probably homologous.



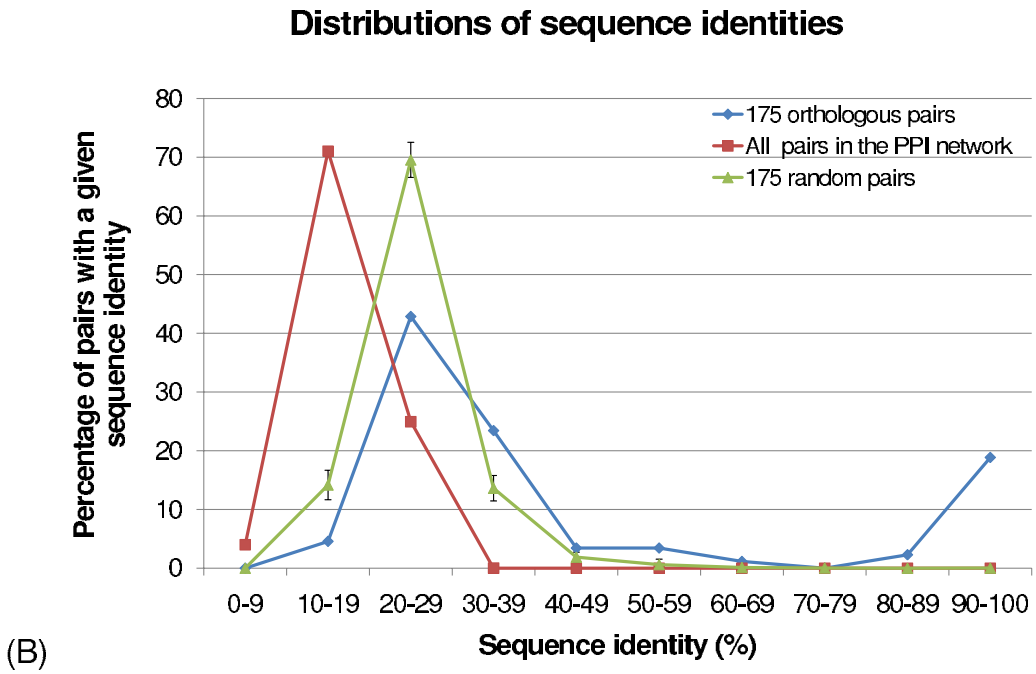


Figure 2: (A) Signature similarity distributions and (B) sequence identity distributions for 175 orthologous protein pairs (blue), all protein pairs in the PPI network with degrees higher than 3 (red), and 175 protein pairs randomly selected from the network (green). On the horizontal axis, signature similarities and sequence identities are binned as follows: 0%-9%, 10%-19%, ..., 80%-89%, and 90%-100%. The procedure of random pair selection (green) was repeated 30 times, and the signature similarities and sequence identities for a given bin were averaged over the 30 runs; green points represent these averages, and bars around the points represent the corresponding standard deviations.

This could explain why orthologous protein pairs have sequence identities above this threshold, whereas all protein pairs and random pairs do not (Figure 2 B). However, high similarity and

evolutionary closeness can not be claimed with certainty for sequences with identities in the twilight zone, since when considering all sequence pairs with identities between 20% and 35%, the population of non-homologous protein pairs explodes [27, 26]; thus, any sequence-based homology predictions on protein pairs with sequence identities below 35% would result in a large number of false positives. This could explain why the sequence identity distributions for orthologous protein pairs are similar to those of all and random pairs for sequence identities below 35% (Figure 2 B). Since we show that sequence identity trends for orthologous protein pairs do not differ much from those for randomly selected pairs and that about 70% of our orthologous pairs have sequence identities below the twilight zone threshold of 35%, we confirm that sequence identities alone cannot be used to detect homology as they would fail to identify the majority of homologous pairs.

This conclusion is consistent with the explanation given in Section 2.1 about how the orthology relationships in COGs and KO groups do not depend on the absolute level of similarity between sequences of the compared proteins. Instead, they depend on identification of consistent patterns, such as a triangle, in the graph of the best matches. It would be interesting to examine whether signature similarity measure could be used to identify orthologs in an equivalent manner, by finding all-to-all pairwise signature similarities between proteins in PPI networks of different species, identifying for each protein in a species the best hits in each of the other species, and searching for triangles or larger cliques in the graph of best hits. This issue is a subject of future research.

We compare signature similarity distribution (Figure 2 A) and sequence identity distribution (Figure 2 B) for orthologous protein pairs. The statistically significant threshold for signature similarities is at 85%. About 20% of orthologs have signature similarities above this threshold. Sequence pairs are considered to have high levels of similarity if their identities are above the twilight zone threshold of 35% [26]. About 30% of orthologs have sequence identities above this threshold. Thus, sequence identities seem to uncover slightly higher level of homology information than signature similarities. However, both measures fail to capture this information in its entirety. Moreover, the overlap between the 20% of orthologous pairs with high signature similarities and the 30% of orthologous pairs with high sequence identities is about 60% of the smaller set, additionally verifying that signature similarities and sequence identities aim to detect somewhat complementary slices of homology information.

4 Discussion

Non-sequence based sources of biological information have already been used to extract important biological knowledge [13, 14, 15]. For example, secondary and tertiary structure-based approaches have been found to be more effective at functionally describing RNAs than sequence-based approaches, as they were able to extract biological information that could not have been discovered from pure sequences [13, 14, 15]. Similarly, it is possible that protein sequence and network topological information give different biological insights. Protein's 3-dimensional structure is expected to closely relate with the number and type of its potential interacting partners in the PPI network. Although high proteins' sequence similarity correlates with their functional and structural similarity, sequence-similar proteins can have functions and structures that differ significantly from one another [28, 29, 30, 31]. Thus, restricting analysis to sequences may mask important structural and functional information. On the other hand, sequence-similar

but structurally-dissimilar proteins are expected to have different PPI network topological characteristics. Moreover, while the vast majority of protein pairs with sequence identity higher than 30%-35% are found to be structurally similar, the most similar protein structure pairs appear to have less than 12% pairwise sequence identity; the average sequence identity between all pairs of similar structures is between 8% and 10% [26]. Thus, entirely different protein sequences can produce very similar structures [32, 30]. When such proteins are expected to share a common function, a sequence-based function prediction would fail, where a network topology-based one would not.

We further support our argument by demonstrating that sequence identity of 100% does not necessarily imply signature similarity of 100%. That is, some of our orthologous protein pairs with 100% identical sequences do not have the same set of interacting partners in the PPI network. In Figure 3 A and B, we show two such orthologous protein pairs, RPL12A and RPL12B, and RPS18A and RPS18B, respectively, and their direct network neighbors.

Both pairs of proteins have sequence identities of 100%. However, the topologies around proteins in each orthologous pair are different, with signature similarities of 65% and 50% respectively. The two orthologous protein pairs (RPL12A and RPL12B, and RPS18A and RPS18B) are ribosomal proteins. 59 of the yeast ribosomal proteins, including these two orthologous protein pairs, retained two genomic copies. Prior studies have suggested that duplicated proteins are functionally redundant [28]. However, it has been shown that paralogous ribosomal proteins have different genetic requirements for their assembly and localization, and that they are functionally distinct [28]. Our analysis suggests that such functional distinction of these proteins is possible (perhaps even likely), since we show that their topological similarities in the PPI network are low (Figure 3), as well as that network topology is closely related to biological function (Figure 1).

Furthermore, it has been shown that after gene duplication, the number of interacting partners in a PPI network that are shared by the resulting yeast homologs appears to decrease rapidly as a function of their evolutionary distance [17, 18]. Thus, signature similarities between duplicated proteins may provide new insights into their evolutionary and functional divergence, complementing the knowledge that can be extracted from their sequences [16]. Therefore, the PPI network topology might present a novel and independent source of biological information, as well as an opportunity to systematically study the evolutionary distance and functional divergence of duplicated proteins; this will especially be true once PPI data becomes entirely accurate and complete [16].

5 Conclusion

We examine whether homology information captured by PPI network topology differs and to what extent from the information captured by protein sequences. We analyze topological signatures of homologous protein pairs in the yeast PPI network and show that they have statistically significantly higher signature similarities than non-homologous protein pairs. We show that their signature similarities are robust to noise in the PPI network. We compare signature similarity trends of homologous protein pairs with their sequence identity trends and find that sequence identities and signature similarities uncover similar levels of homology information. Although none of the two methods, network topology or sequence identity, seems to capture homology information in its entirety, we demonstrate that they might give insights into somewhat

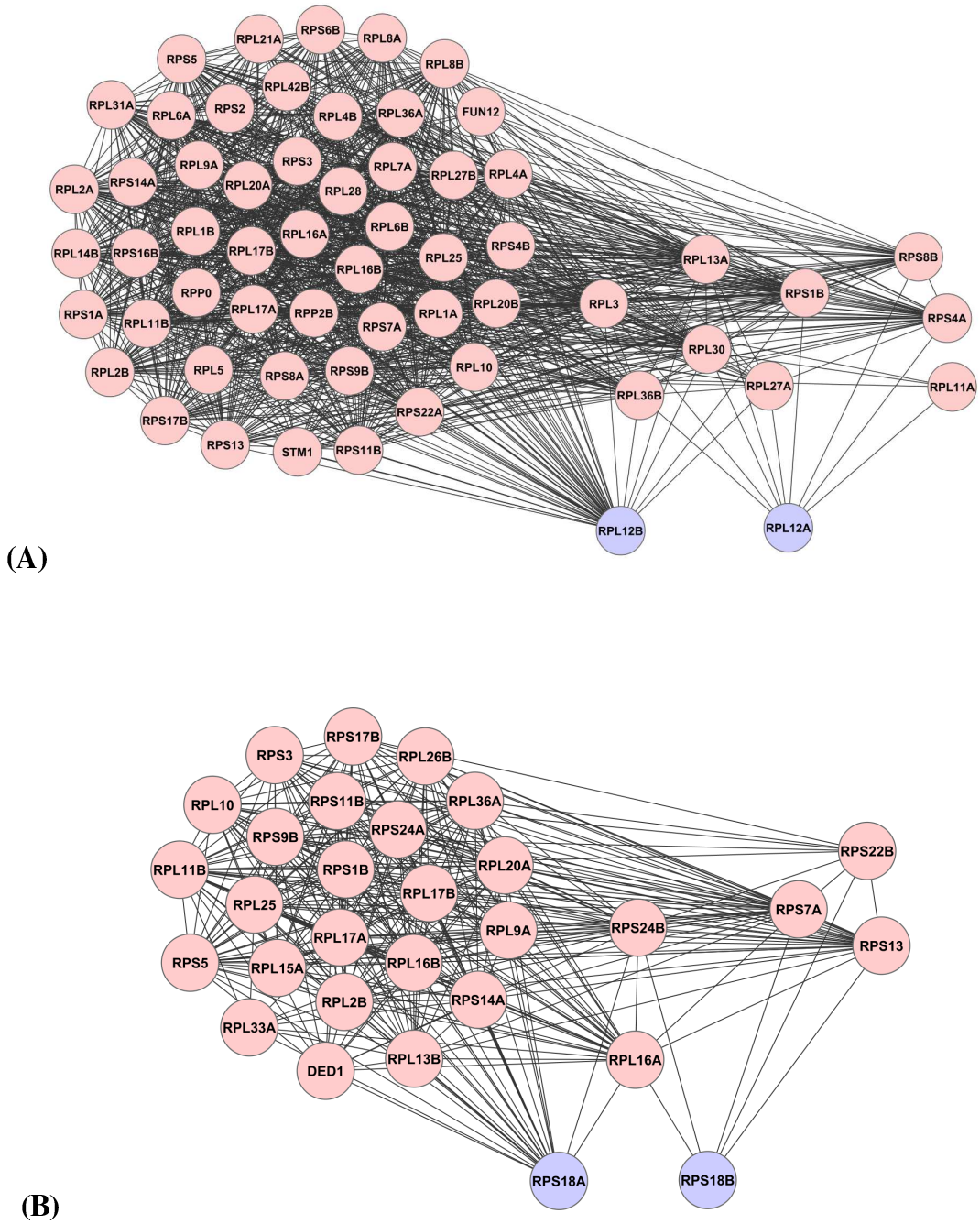


Figure 3: (A) Subnetwork containing orthologous proteins RPL12B and RPL12A (blue), their direct neighbors (pink), and all interactions in the yeast PPI network that exist between these nodes. The subnetwork contains 59 nodes and 1,262 interactions. RPL12B interacts with 54 proteins, while RPL12A interacts with 9 proteins. Proteins RPL12B and RPL12A have identical sequences, but graphlet degree signature similarity in the PPI network of 65%. (B) Subnetwork containing orthologous proteins RPS18A and RPS18B (blue), their direct neighbors (pink), and all interactions in the yeast PPI network that exist between these nodes. The subnetwork contains 30 nodes and 346 interactions. RPS18A interacts with 25 proteins, while RPS18B interacts with 5 proteins. Proteins RPS18A and RPS18B have identical sequences, but graphlet degree signature similarity in the PPI network of 50%.

different types of biological information. Therefore, we conclude that similarities of proteins' topological signatures in the PPI network could potentially be used as a complementary method

to sequence-based approaches for identifying homologs.

Acknowledgements

This project was supported by the NSF CAREER IIS-0644424 grant.

References

[1] N. Yosef, R. Sharan, and W.S. Noble. Improved network-based identification of protein orthologs. *Bioinformatics*, 24(16):i200–i206, 2008.

[2] W.M. Fitch. Distinguishing homologous from analogous proteins. *Syst Zool*, 19:99–111, 1970.

[3] M. Remm, C. Storm, and E. Sonnhammer. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.*, 314:1041–1052, 2001.

[4] P.G. Higgs and T.K. Attwood. *Bioinformatics and Molecular Evolution*. Blackwell, 2005.

[5] D.W. Mount. *Bioinformatics - Sequence and Genome Analysis*. Cold Spring Harbor Laboratory Press, 2004.

[6] B. P. Kelley, Y. Bingbing, F. Lewitter, R. Sharan, B. R. Stockwell, and T. Ideker. PathBLAST: a tool for alignment of protein interaction networks. *Nucl. Acids Res.*, 32:83–88, 2004.

[7] J. Berg and M. Lassig. Local graph alignment and motif search in biological networks. *PNAS*, 101:14689–14694, 2004.

[8] J. Flannick, A. Novak, S. S. Balaji, H. M. Harley, and S. Batzglou. Graemlin: General and robust alignment of multiple large interaction networks. *Genome Res*, 16(9):1169–1181, 2006.

[9] Z. Liang, M. Xu, M. Teng, and L. Niu. NetAlign: a web-based tool for comparison of protein interaction networks. *Bioinformatics*, 22(17):2175–2177, 2006.

[10] J. Berg and M. Lassig. Cross-species analysis of biological networks by Bayesian alignment. *Proceedings of the National Academy of Sciences*, 103(29):10967–10972, 2006.

[11] J. Flannick, A. F. Novak, C. B. Do, B. S. Srinivasan, and S. Batzoglou. Automatic parameter learning for multiple network alignment. In *RECOMB*, pages 214–231, 2008.

[12] M. Zaslavskiy, F. Bach, and J. P. Vert. Global alignment of protein-protein interaction networks by graph matching methods. *Bioinformatics*, 25(12):i259–i267, 2009.

[13] D. Gautheret, F. Major, and R. Cedergren. Pattern searching/alignment with rna primary and secondary structures: an effective descriptor for trna. *Comput. Appl. Biosci.*, 6(4):325–331, 1990.

[14] C. R. Woese, R. Gutell, R. Gupta, and H. F. Noller. Detailed analysis of the higher-order structure of 16s-like ribosomal ribonucleic acids. *Microbiological reviews*, 47(4):621–669, 1983.

[15] C.H. Webb, N.J. Riccitelli, D.J. Ruminski, and A. Lupták. Widespread occurrence of self-cleaving ribozymes. *Science (New York, N.Y.)*, 326(5955):953+, 2009.

[16] O. Ratmann, C. Wiuf, and J.W. Pinney. From evidence to inference: probing the evolution of protein interaction networks. *HFSP Journal*, 3(5):290–306, 2009.

[17] A. Wagner. The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol Biol Evol*, 18(7):1283–1292, 2001.

[18] X. He and J. Zhang. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics*, 169(2):1157–1164, 2005.

[19] S.R. Collins, P. Kemmeren, X.C. Zhao, J.F. Greenblatt, F. Spencer, F.C. Holstege, J.S. Weissman, and N.J. Krogan. Toward a comprehensive atlas of the physical interactome of saccharomyces cerevisiae. *Molecular and Cellular Proteomics*, 6(3):439–450, 2008.

[20] C. Brun, C. Herrmann, and A. Guénoche. Clustering proteins from interaction networks for the prediction of cellular functions. *BMC Bioinformatics*, 5, 2004.

[21] T. Milenković and N. Pržulj. Uncovering biological network function via graphlet degree signatures. *Cancer Informatics*, 6:257–273, 2008.

[22] R.L. Tatusov, N.D. Fedorova, J.D. Jackson, A.R. Jacobs, B. Kiryutin, E.V. Koonin, D.M. Krylov, R. Mazumder, S.L. Mekhedov, A.N. Nikolskaya, B.S. Rao, S Smirnov, A.V. Sverdlov, S. Vasudevan, Y.I. Wolf, J.J. Yin, and D.A. Natale. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, 4(41), 2003.

[23] M. Kanehisa and S. Goto. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, 28:27–30, 2000.

[24] N. Pržulj. Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23:e177–e183, 2007.

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

[26] B. Rost. Twilight zone of protein sequence alignments. *Protein Engineering*, 12(2):85–94, 1999.

[27] R. F. Doolittle. Similar amino-acid sequences - chance or common ancestry? *Science*, 214(4517):149–159, 1981.

[28] S. Komili, N. G. Farny, F. P. Roth, and P. A. Silver. Functional specificity among ribosomal proteins regulates gene expression. *Cell*, 131(3):557–571, 2007.

[29] J. D. Watson, R. A. Laskowski, and J. M. Thornton. Predicting protein function from sequence and structural data. *Current opinion in structural biology*, 15(3):275–284, 2005.

[30] J. C. Whisstock and A. M. Lesk. Prediction of protein function from protein sequence and structure. *Q Rev Biophys*, 36(3):307–340, 2003.

[31] M. Kosloff and R. Kolodny. Sequence-similar, structure-dissimilar protein pairs in the pdb. *Proteins*, 71(2):891–902, 2008.

[32] D. V. Laurents, S. Subbiah, and M. Levitt. Different protein sequences can give rise to highly similar folds through different stabilizing interactions. *Protein Sci*, 3(11):1938–1944, 1994.

Supplementary figures for:

Complementarity of network and sequence information in homologous proteins

Vesna Memišević[1], **Tijana Milenković**[1], **and Nataša Pržulj**[2,*]

[1]Department of Computer Science, University of California, Irvine, CA 92697-3435, USA
[2]Department of Computing, Imperial College London, London, SW7 2AZ, UK
[*]Corresponding author (e-mail: natasha@imperial.ac.uk)

1 Supplementary figures

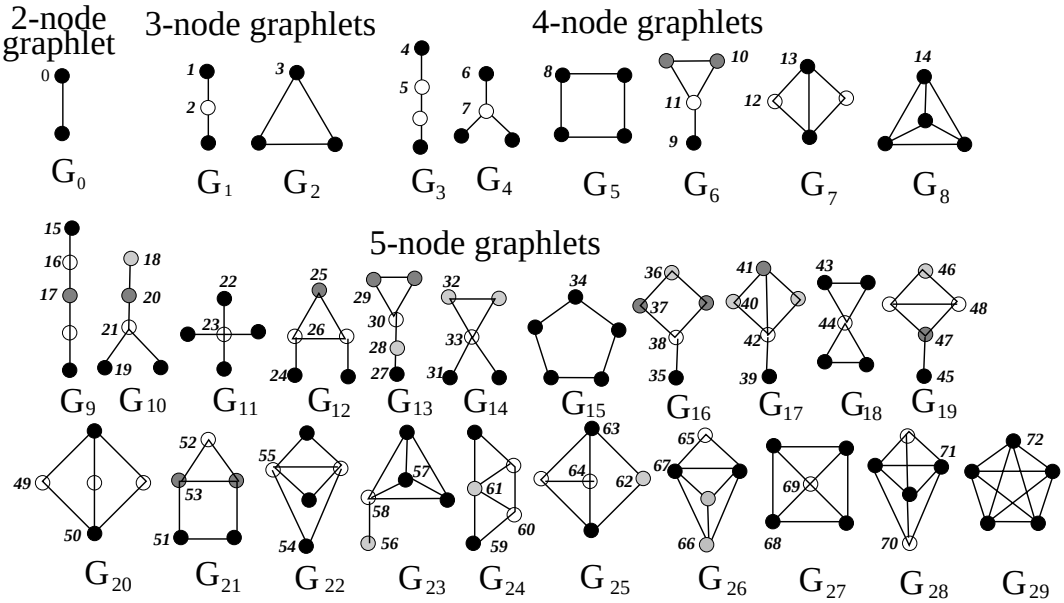


Figure S1: All 30 connected graphs on 2 to 5 nodes. When appearing as an induced subgraph of a larger graph, we call them *graphlets*. They contain 73 topologically unique node types, called "automorphism orbits," numerated from 0 to 72. In a particular graphlet, nodes belonging to the same orbit are 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. The figure is taken from Pržulj *et al.* (2007).

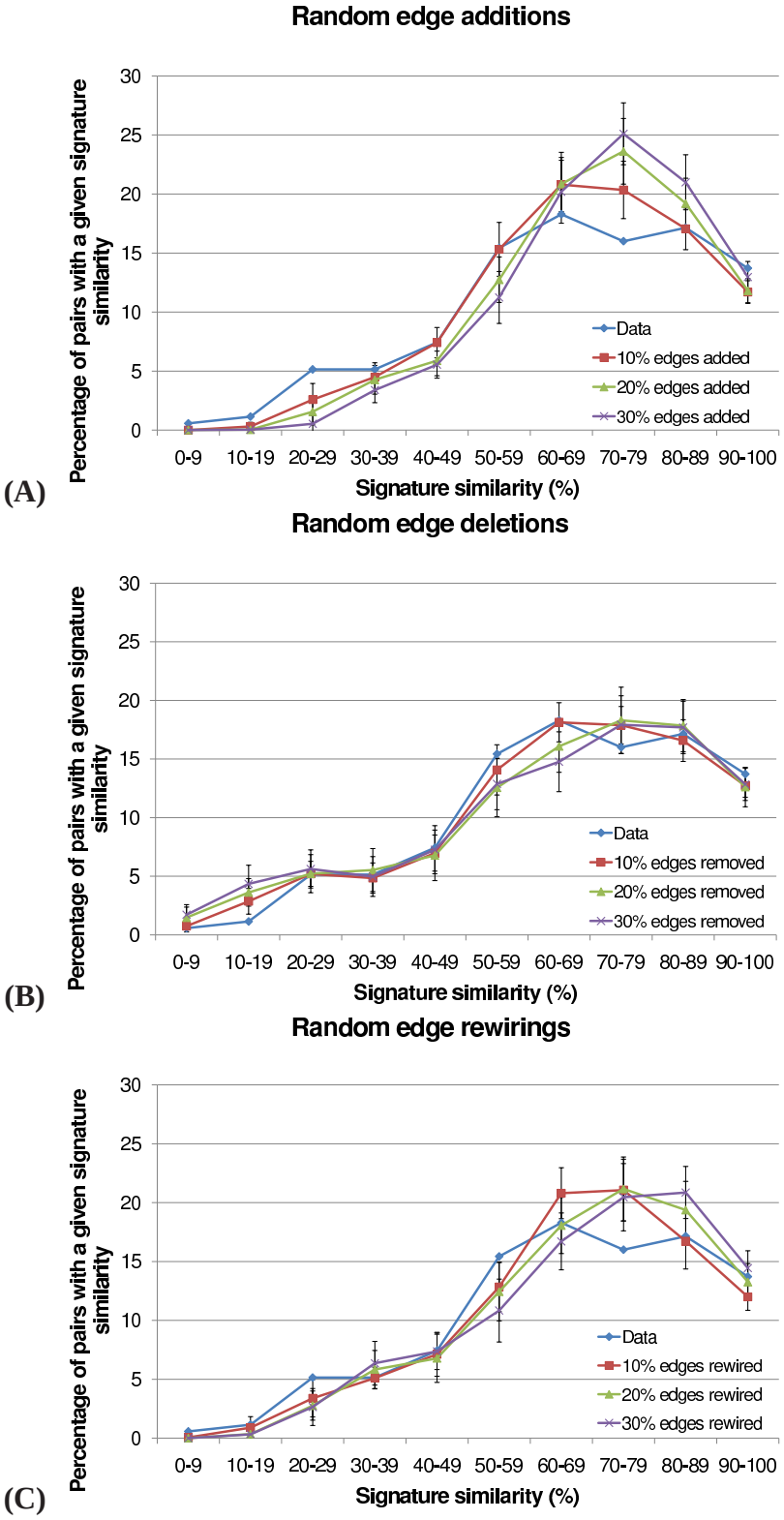


Figure S2: Comparison of signature similarity distribution of 175 orthologous protein pairs in the network (blue) and their signature similarity distributions computed from randomized networks, obtained by randomly (A) adding, (B) deleting, and (C) rewiring 10% (red), 20% (green), and 30% (purple) of edges in the PPI network.