Supplementary Information for "Topology-Function Conservation in Protein-Protein Interaction Networks"

Darren Davis [1], Ömer Nebil Yaveroğlu [1,2], Noël Malod-Dognin [2], Aleksandar Stojmirovic [3,4], and Nataša Pržulj [2]

[1] *Calit2, University of California, Irvine, CA, USA*

[2] *Department of Computing, Imperial College London, UK*

[3] *National Center for Biotechnology Information (NCBI), USA*

[4] *Janssen Research and Development, LLC, Spring House, PA, USA*

S.1 Effects of Different Thresholds for Protein and GO Term Annotation Filtering

As described in the main manuscript, we apply two filters on the input datasets of our analysis: (1) we exclude proteins with degree less than 4, and (2) we exclude GO terms that annotate less than 5 proteins or more than 5% of the proteins in the protein-protein interaction network (PIN). It is a common practice to exclude low degree proteins from systematic analyses of PINs, because the low degree proteins are likely to correspond to incomplete parts of the PINs (Wang and Wu, 2013). We set the degree threshold to 4, so that all 2- to 5-node graphlet orbits can appear on the considered nodes. The reason for excluding the GO terms that annotate less than 5 nodes is that, such GO terms do not provide enough variance for CCA to identify significant topology-function relationships. The topological patterns identified for such GO terms are not generic and their topological patterns are not stable, as they may change a lot if a single additional protein is annotated with the GO term. The effects of choosing different thresholds for these filters are discussed in detail below.

To understand the effects of different degree thresholds, we perform our experiments with degree thresholds of 2, 3, 4 and 5. Table S.3 presents the number of proteins that are considered for each of these degree thresholds and Table S.5 presents the number of identified topology-function relationships together with the comparison of these results with those obtained with degree threshold 4. As shown in the tables, different degree thresholds cause a trade-off between the count and the level of confidence on the identified topology-function relationships: with lower degree thresholds, we obtain higher number of topology-function relationships but the confidence in these relationships is lower, since they are obtained from more

noisy interaction datasets. When higher degree thresholds are used, the number of identified topology-function relationships are lower, but the confidence in these relationships is higher since they are obtained from high confidence interactions in the PINs. Interestingly, the topology-function relationships that are identified in the high confidence part of the PIN are also identified when the incomplete part of the PIN is considered (i.e., when low degree nodes are included into the analysis) – see the "Intersection/Total" columns of Table S.5. In the manuscript, we decided to report the results on the degree threshold of 4, since it captures all the high confidence topology-function relationships without including too many lower confidence topology-function relationships. Degree threshold 4 also guarantees that all 2- to 5-node graphlet patterns can be observed.

Similarly, to understand the effects of different GO term annotation thresholds, we perform our experiments with GO annotation thresholds of 3, 5, 10, 15 (while keeping the degree threshold fixed at 4). The number of considered GO terms with each threshold is summarized in Table S.4 and the number of topology-function relationships identified with each threshold together with the comparison of these results with those obtained with GO term threshold 5 are provided in Table S.6. With increasing thresholds of GO term filtering, more biological process terms are identified to have significant topology-function relationships, while the number of identified relationships are slightly less for cellular component terms. This is due to fact that when GO terms with low variance (i.e., terms annotating a smaller number of proteins) are filtered out, CCA can identify the topological patterns of the higher variance GO terms more accurately without noising the identified weights with the estimates including the low variance GO terms. It should also be noted that the identified topology-function relationships are consistent at different GO term annotation thresholds. Therefore, the value for this parameter should be decided considering the trade-off between the higher number of GO terms analysed and higher number of topology-function relationships obtained. With the aim of covering as many GO terms as possible in our analysis, we decided to use the threshold of 5 rather than higher thresholds.

S.2 Computation of the Association Matrix

The association matrix aims to encode all the topology-function relationships that are identified by the canonical correlation analysis (CCA) and transforms the graphlet degree vectors of proteins to functional annotations based on a least-squares fit model. The first variable set of the CCA, $\mathbb{R}^t$, describes the

topological characteristics of the proteins based on the graphlet degree vectors. For each protein, the graphlet degree vector is a 73-dimensional vector ($t = 73$), where each element corresponds to the graphlet degree of one orbit from 2- to 5-node graphlets. The second variable set of CCA, $\mathbb{R}^f$, describes the functions of proteins based on their GO term annotations. CCA is performed separately for the three GO annotation categories, i.e., biological process (BP), molecular function (MF) and cellular component (CC). The sizes of the GO annotation vectors (i.e., the numbers of GO terms) that are considered for each run of the CCA are summarized in Table S.1, as indicated by the total number of GO terms. For example, the number of functional annotation features, f, that are considered for the biological process terms of human is equal to $1,439$.

Applying CCA on these feature sets, the association matrix is computed as $W_1 \times S \times W_2^+$, where W_1 is the matrix of canonical weights for the topological variables, S is a diagonal matrix of canonical correlations and W_2^+ is the Moore-Penrose pseudoinverse matrix of canonical weights for the functional annotation variables. By multiplying the graphlet degree vectors with W_1, the graphlet degree variates are computed. Since the canonical weights that are identified by the CCA guarantees that the graphlet degree variates are maximally correlated with the GO Term variates, multiplying the graphlet degree variates with the diagonal matrix of canonical correlations, S, estimates the GO term variates based on a least-squares fit model. In other words, multiplication of the graphlet degree vectors with $W_1 \times S$ produces GO term variates that are approximated based on a linear model. Finally, multiplying the estimated GO Term variates with W_2^+, the estimated GO term variates are transformed back into GO term annotations of proteins and topology-based GO annotations are obtained.

S.3 Defining the Orbit Contribution Strength Measure

We identify the orbits that are statistically significantly linked with the wiring patterns of a GO term by computing the orbit contribution strengths, i.e., Pearson's correlations between the graphlet degrees and the topology-based GO Annotations (Fig. 2.B). For this task, we choose to use the topology-based annotations rather than observed GO annotations due to the two following reasons:

1. The protein-protein interaction and GO term annotation datasets are highly noisy and incomplete. Assuming that a GO term is linked with certain wiring patterns (i.e., orbits) and these patterns are consistent for all proteins that are annotated with the GO term, the noise in the protein interaction

and GO term annotation datasets would make their wiring patterns heterogeneous. For this reason, due to the incompleteness and noisiness of the datasets, it is not always possible to observe similar topological characteristics for the proteins annotated with a given GO term. This makes it harder to characterize the topology-function relationships of a given GO term. Instead of directly extracting these topological characteristics from the data, we first estimate the most likely topological profile of a GO term using canonical correlation analysis. Then, without referring to the data, we aim to understand the estimated topological profile by defining the orbit contribution strength profiles. This approach filters out the heterogeneous wiring characteristics caused by the noise in the datasets and links the most likely topological characteristics of the GO terms.

2. The observed GO annotations are binary (i.e., 1 when the protein is annotated with the GO term, and 0 otherwise), while the topology-based GO annotations are real-valued (each value corresponding to the strength of the association between an orbit and a GO term). The variance in the topology-based GO annotations makes them a better choice for quantifying the linear dependencies using Pearson's correlation.

S.4 Significance Testing

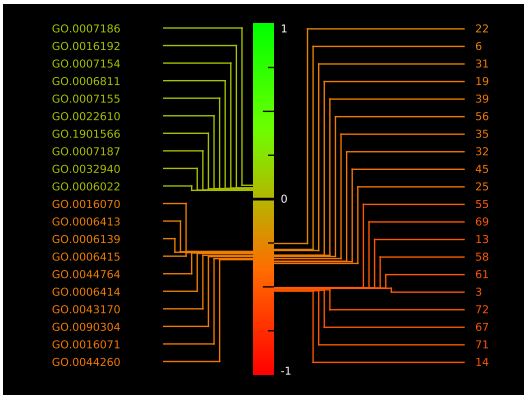
We use permutation tests to examine whether the two strength measures and cross-species similarities of topology-function associations are higher than expected by chance. We apply our methodology for $10,000$ permutations, randomly re-assigning graphlet degree vectors to proteins in each permutation. The permutations make any topology-function association a matter of random chance, while preserving the distribution of graphlet degrees and GO annotations. We estimate the p-values for each of the measures (i.e., structure association strengths, orbit contribution strengths, multi-species structure association strengths, and orbit contribution similarities) as the proportion of permutations where the obtained values exceed the observed values. The permutation tests avoid imposing the assumption that the data are normally distributed. We adjust the estimated p-values using Benjamini-Hochberg correction (Benjamini and Hochberg, 1995) for the statistical errors caused by multiple testing.

S.5 Predicting GO Term Annotations from Graphlet Degrees

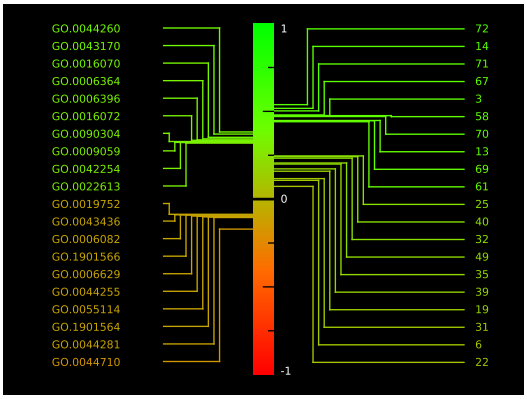
We analyse the linear dependencies between graphlet degree vectors and GO term annotations using CCA. Fig. S.1 summarizes the results of our CCA by providing the canonical correlations for the first set of canonical variates and illustrating the variables with highest and lowest ranked cross-loadings (i.e., Pearson's correlations of variables and the canonical variates of the other set of attributes) for each experiment. As shown in Fig. S.1, the highest canonical correlations range between 0.239 and 0.433 depending on the experiment type. These canonical correlations are not identified to be statistically significant as the p-values approximated by both Wilks' theorem and Hotelling's T-squared distribution tests are greater than 0.05. There could be two reasons for CCA to fail maximizing the canonical correlations further: (1) There exist many GO terms that are annotated with small number of proteins, and therefore, there are many GO term features for which the variance is extremely low, and (2) As the GO term features are binary-valued (being 1 when the protein is annotated with the GO term and being 0 otherwise), the GO term features are bound to have low variance.

Even though no statistically significant canonical correlations are identified by CCA, this only shows that CCA is not able to find strong linear dependencies between the weighted sum of all variables; i.e., canonical covariates. If the CCA results are carefully mined by looking into the pairwise relations between each biological function and graphlet orbit pair, we can still obtain the most prominent topological characteristics of the proteins associated with a given GO term from the CCA. The second step of our methodology is designed exactly for this purpose, and it identifies the statistically significant graphlet orbit and GO term relationships by mining the association matrix produced by CCA.

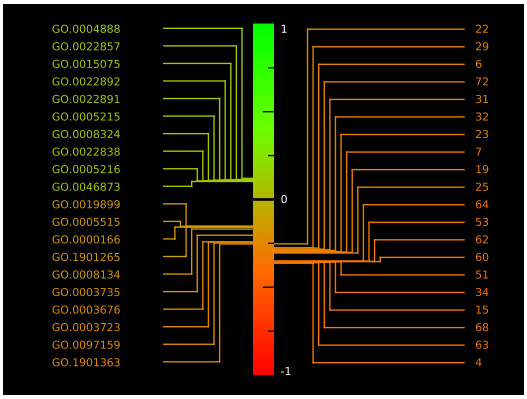
The transformation defined by the association matrix could be further used for predicting the GO term annotations of the proteins. However, the canonical correlations and the cross-loadings indicate that these predictions would not be promising at the current state of the methodology. Enriching the graphlet degree statistics with other types of biological properties such as protein sequence or protein structure, we could define a better GO term annotation prediction algorithm based on CCA. We leave this task for future work since we focus on characterizing wiring patterns of biological functions and identifying orthologous topological patterns in the scope of this study.



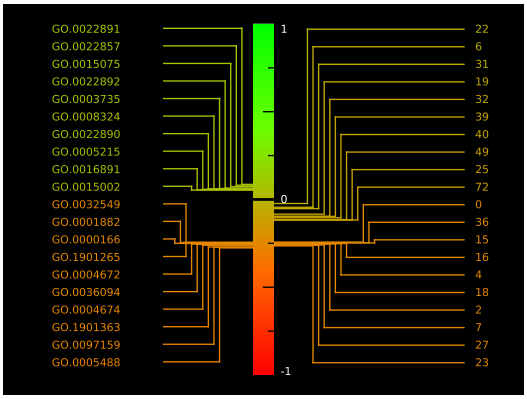
(A) Human - Biological Process
Canonical Correlation = 0.433



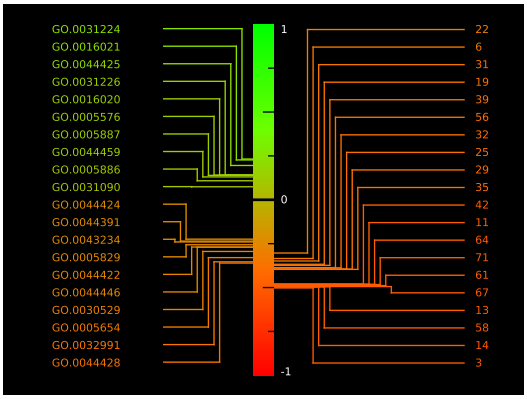
(B) Yeast - Biological Process
Canonical Correlation = 0.416



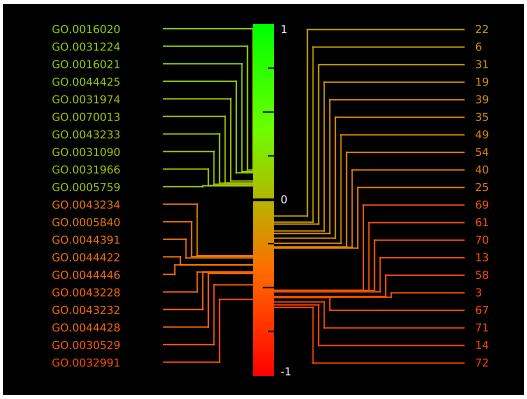
(C) Human - Molecular Function
Canonical Correlation = 0.345



(D) Yeast - Molecular Function
Canonical Correlation = 0.432



(E) Human - Cellular Component
Canonical Correlation = 0.239



(F) Yeast - Cellular Component
Canonical Correlation = 0.247

Figure S.1. The 10 highest and lowest ranked topology-function relationships as identified by the first set of canonical weights. The variables on the left side correspond to different GO terms and the variables on the right side correspond to the graphlet degrees of different orbits. The connection points of the variables to the color bar corresponds to their cross-loadings (i.e., the Pearson's correlation between the variable and the weighted sum of the variables on the other side).

S.6 On the Significance of Link between Topology and Function

We identify 15 biological process and 9 cellular component terms to have significantly preserved topology-function relationships for yeast and human (Section 3.1 of the main manuscript). We tested the statistical significance of these relationships with the permutation experiments, as explained in Supp. Section S.4. In those experiments, the link between topology and function are broken by randomly perturbing the graphlet degree vector and GO term annotation vector matches. The permutation experiments successfully test the statistical significance of the identified topology-function association of a GO term.

Here, we provide a further evaluation of the significance of the number of GO terms that are identified to have significant topology-function relationships. In order to further show that the $15+9 = 24$ GO terms are not identified to have significant topology-function relationships by chance, we perform randomization experiments with a different perturbation scheme than in Supplementary Section S.4.

In these randomization experiments, we perturb the features of yeast and human datasets (i.e., the annotations of each GO term and the graphlet degrees of each orbit) considering each feature independently, and we analyse the resulting randomized datasets in the same way as we analyse the unperturbed yeast and human PIN data. We perform 500 such randomization experiments for the PINs of yeast and human, and count the number of statistically significant topology-function relationships identified from each of these experiments. Figure S.2 summarizes the number of statistically significant topology-function relationships identified from these 500 randomization experiments. The results show that it is very unlikely to identify significant topology-function relationships on randomized datasets.

In particular, on a single species, we observe the following patterns:

- For randomized dataset of yeast (Figure S.2–A),
  - For biological process terms, 76.35% of the randomizations lead to no significant topology-function relationship, 21.64% of the randomizations identify 1 significant topology-function relationship, 1.80% of the randomizations identify 2 significant topology-function relationships and 0.20% of the randomizations identify 3 significant topology-function relationships.
  - For molecular function terms, 93.99% of the randomizations lead to no significant topology-function relationship, 5.81% of the randomizations identify 1 significant topology-function relationships and 0.20% of the randomizations identify 2 significant topology-function relationships.

  - For cellular component terms, 97.40% of the randomizations lead to no significant topology-function relationship, 2.41% of the randomizations identify 1 significant topology-function relationships and 0.20% of the randomizations identify 2 significant topology-function relationships.

- For randomized dataset of human (Figure S.2–B),
  - For biological process terms, 85.57% of the randomizations lead to no significant topology-function relationship, 13.63% of the randomizations identify 1 significant topology-function relationship and 0.80% of the randomizations identify 2 significant topology-function relationships.
  - For molecular function terms, 93.99% of the randomizations lead to no significant topology-function relationship, 5.61% of the randomizations identify 1 significant topology-function relationships and 0.40% of the randomizations identify 2 significant topology-function relationships.
  - For cellular component terms, 96.59% of the randomizations lead to no significant topology-function relationship, 3.21% of the randomizations identify 1 significant topology-function relationships and 0.20% of the randomizations identify 2 significant topology-function relationships..

Then, between the two species, no statistically significant conserved topology-function relationship are identified on any of the randomized datasets (Figure S.2–C).

Comparing these with what we observed on the experiments performed with real yeast and human datasets, we find support that neither the topologically orthologous functions nor the single-species topology functions are identified by chance. These results once more prove that the wiring patterns of proteins in protein-protein interaction networks are linked with their biological functions.

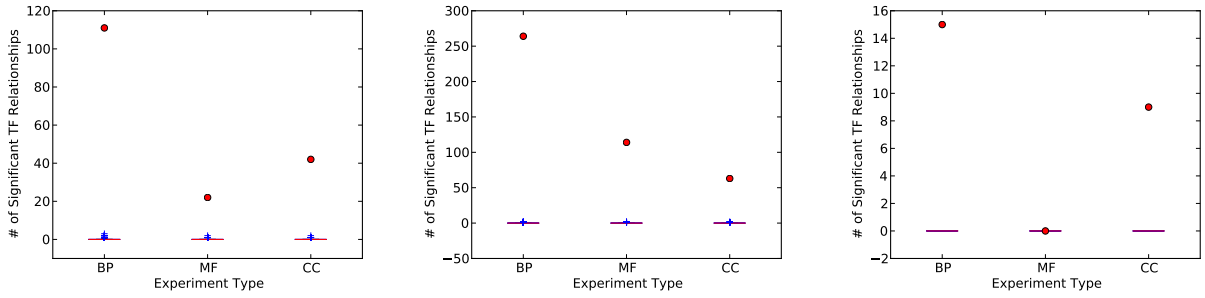


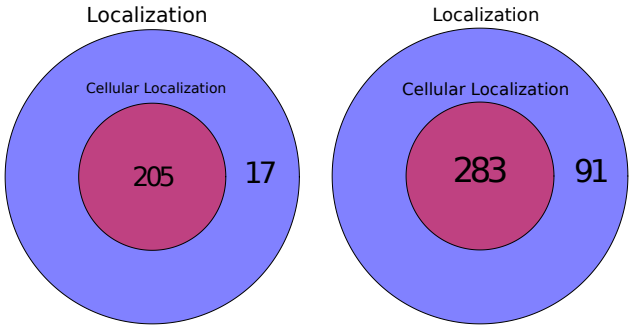
Figure S.2. The number of statistically significant topology-function relationships identified from randomized datasets. The number of statistically significant topology-function relationships identified from the real datasets are presented with the red circles. The lines at the value of 0 on y-axis are boxplots representing the first quartile, median and third quartile of the randomization experiments; these statistics are all 0 for the randomized experiments.**Panel A** presents the significant topology-function relationships for yeast. **Panel B** presents the significant topology-function relationships for human. **Panel C** presents the number of conserved topology-function relationships for yeast and human.

S.7 Redundancies among GO Terms with Conserved Topology-Function Association

We say that two GO terms are "redundant" if they annotate similar sets of proteins and have similar meanings. We do not apply any filtering on the analysed GO terms based on their redundancies before performing the canonical correlation analysis to identify all topology-function relationships. We identify 15 biological process terms and 9 cellular component terms that are linked with statistically significant topology-function relationships with our analysis. For simplicity, we group the patterns of these GO terms into non-redundant functional groups based on the redundancies defined above and we summarize our results in Fig. 3. In this section, we list the GO terms that form these non-redundant functional groups. We also provide the number of proteins that are commonly annotated with the GO terms in the non-redundant functional groups (Fig. S.3 – S.9). Finally, we present an extended version of Fig. 3 in Fig. S.10, where we present the topology-function relationships separately for each GO term.

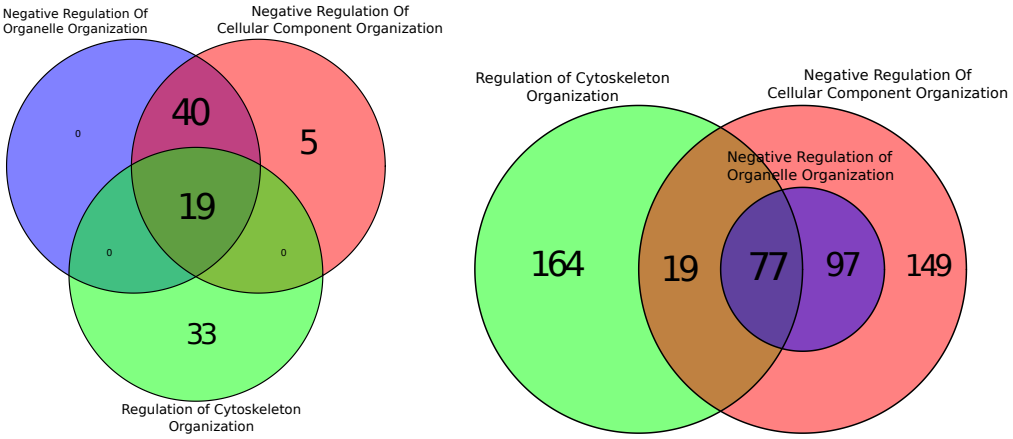
Among the 15 biological process terms that we observe to have significantly conserved topology-function relationships, we identify 7 non-redundant functional groups of GO terms. The first non-redundant functional group contains two GO terms that are linked with "Localization" (Fig. S.3). "Cellular Localization" term is a child of "Localization" term in GO hierarchy, and so, all of its proteins are

also annotated with "Localization". The second non-redundant functional group consists of three GO terms that are linked with "Regulation of Cellular Organization" (Fig. S.4). The third non-redundant functional group consists of four GO terms that are linked with "Acetylation" (Fig. S.5). The number of proteins that are commonly annotated with these four Acetylation-related terms is very high, making these four terms almost identical. Two other non-redundant functional groups consist of four GO terms (two terms for each group) that are linked with "Transcription" (Fig. S.6). We separately consider the "Transcription Initiation" and "Transcription Elongation" terms in Fig. 3, because the intersection between these two sets is small and the topological patterns that we identify for these two groups are slightly different. The last two non-redundant functional groups consist of single GO terms, one containing only the "Proteasome Assembly" term and the other containing only "Protein Modification By Small Protein Removal" term.



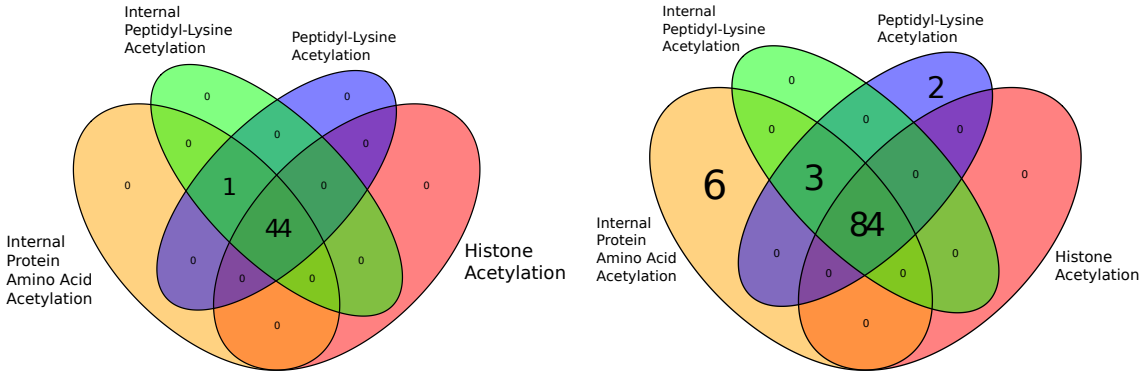
(A) Yeast **(B)** Human

Figure S.3. GO Terms related with *Localization*: Venn Diagrams illustrating the number of proteins that are commonly annotated by "Localization" (GO:0051179) and "Cellular Localization" (GO:0051641).



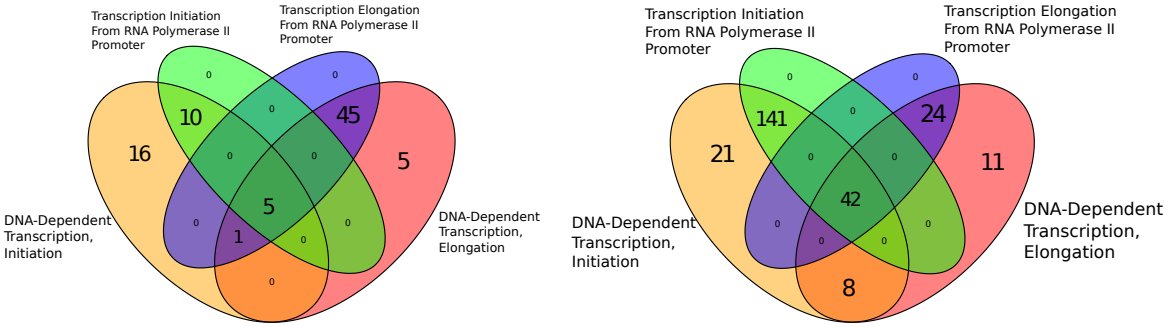
(A) Yeast **(B)** Human

Figure S.4. GO Terms related with *Regulation of Cellular Organization*: Venn Diagrams illustrating the number of proteins that are commonly annotated by "Negative Regulation Of Organelle Organization" (GO:0010639), "Negative Regulation Of Cellular Component Organization" (GO:0051129) and "Regulation Of Cytoskeleton Organization" (GO:0051493).



(A) Yeast **(B)** Human

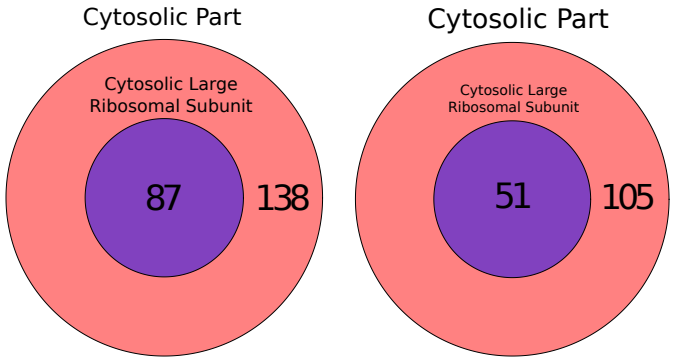
Figure S.5. GO Terms related with *Acetylation*: Venn Diagrams illustrating the number of proteins that are commonly annotated by "Internal Protein Amino Acid Acetylation" (GO:0006475), "Histone Acetylation" (GO:0016573), "Internal Peptidyl-Lysine Acetylation" (GO:0018393) and "Peptidyl-Lysine Acetylation" (GO:0018394).



(A) Yeast **(B)** Human

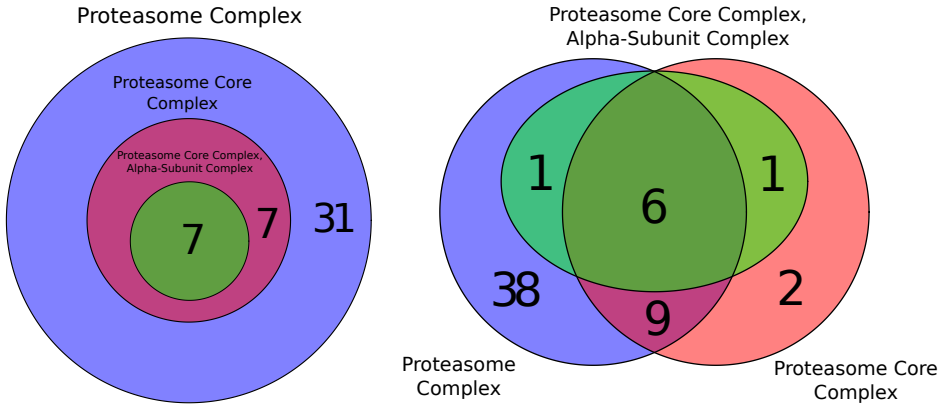
Figure S.6. GO Terms related with *Transcription*: Venn Diagrams illustrating the number of proteins that are commonly annotated by "DNA-Dependent Transcription, Initiation" (GO:0006352), "DNA-Dependent Transcription, Elongation" (GO:0006354), "Transcription Initiation from RNA Polymerase II Promoter" (GO:0006367) and "Transcription Elongation from RNA Polymerase II Promoter" (GO:0006368).

Among the 9 cellular component terms that we identify to have significantly conserved topology-function relationships, we observe 2 non-redundant functional groups. The first functional group consists of two GO terms that are linked with "Cytosolic Part" (Fig. S.7). "Cytosolic Part" is the ancestor of "Cytosolic Large Ribosomal Subunit" in GO hierarchy, and hence, it annotates all of the "Cytosolic Large Ribosomal Subunit" annotated proteins. The second functional group contains seven protein complexes, namely, "Proteasome Complex," "Proteasome Core Complex," "Proteasome Core Complex, Alpha-Subunit Complex," "Swi/Snf Complex," "Swi/Snf Superfamily-type Complex," "Mediator Complex" and "Baf-Type Complex". Three of these protein complexes are linked with proteasome (Fig. S.8). The "Proteasome Complex" term annotates almost all of the proteins that are annotated with "Proteasome Core Complex" and "Proteasome Core Complex, Alpha-Subunit Complex". Two of the remaining protein complexes are linked with "Swi/Snf Complex" (Fig. S.7). This redundancy is caused by an "is-a" relation, "Swi/Snf Complex" being a descendant of "Swi/Snf Superfamily-Type Complex" in GO hierarchy. We combine all these protein complexes into a single functional group, since the topological patterns of these terms are all similar and the reason for this pattern is the simple fact that these proteins appear as protein complexes which are mostly identified by mass spectrometry experiments that produce dense subgraph patterns.



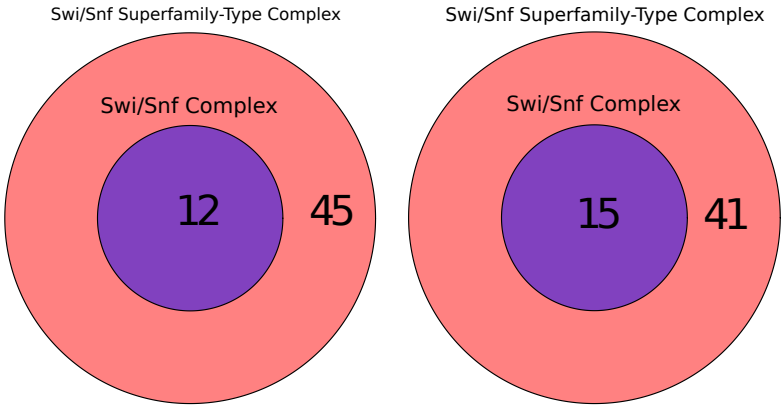
(A) Yeast **(B)** Human

Figure S.7. GO Terms related with *Cytosolic Part*: Venn Diagrams illustrating the number of proteins that are commonly annotated by "Cytosolic Part" (GO:0044445) and "Cytosolic Large Ribosomal Subunit" (GO:0022625).



(A) Yeast **(B)** Human

Figure S.8. GO Terms related with *Proteasome*: Venn Diagrams illustrating the number of proteins that are commonly annotated by "Proteasome Complex" (GO:0000502), "Proteasome Core Complex" (GO:0005839) and "Proteasome Core Complex, Alpha-Subunit Complex" (GO:0019773).



(A) Yeast **(B)** Human

Figure S.9. GO Terms related with *Swi/Snf Complex*: Venn Diagrams illustrating the number of proteins that are commonly annotated by "Swi/Snf Complex" (GO:0016514) and "Swi/Snf Superfamily-Type Complex" (GO:0070603).

For completeness, we present an expanded version of Fig. 3 in Fig. S.10, in which we present the orbit contribution strength profile of each GO term separately. In Fig. S.10, the names of the GO terms that represent the functional groups in Fig. 3 are written in bold and the functional groups are separated by horizontal lines. Note that we manually chose the most general, or the most representative GO term from each functional group when deciding on the group representatives.

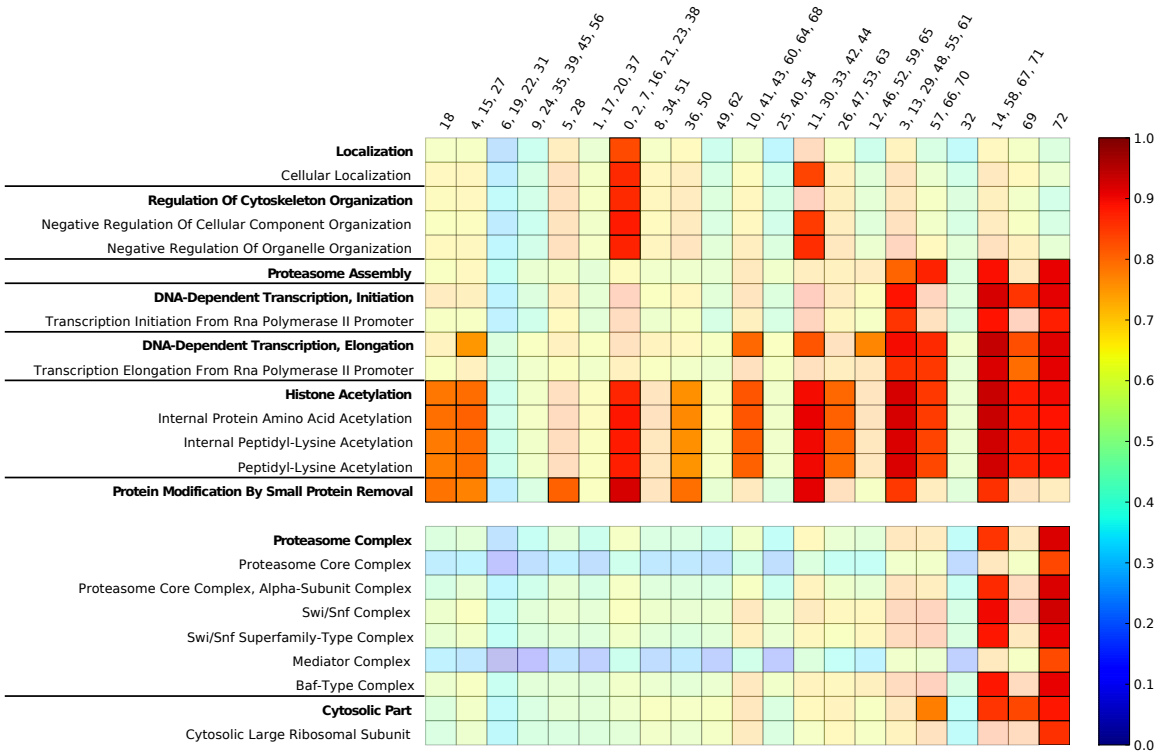


Figure S.10. The orbit contribution strength profiles of the GO terms that have significantly conserved topology-function relationships. Each row of the heatmap corresponds to the average orbit contribution strength profile of a GO term. The heatmap extends Fig. 3 by presenting the orbit contribution strength profiles of each GO term within the functional groups separately. The functional groups are separated by the horizontal lines on the left. The bold GO terms are the representative GO terms of each functional group, which are used for illustrating the orbit contribution strength profiles of the functional groups in Fig. 3. As in Fig. 3, graphlet orbits are grouped based on the similarity of their graphlet degrees (see Fig. 1). Each cell of the heatmap represents the maximum orbit correlation strength in the relevant orbit group. When an orbit group does not have any significant orbit contribution strengths, they are plotted semi-transparently. This is done to highlight the orbit groups that have significant relationships with the GO terms. Note that, cells plotted with solid colors do not mean that all orbits in the relevant group have significant relationships with the GO term, but it means that at least one of the orbits has a significant relationship (for the list of all significant orbits, see Supp. Data 1.)

S.8 Grouping Similar Orbits

Yaveroğlu *et al.* (2014) quantify the similarity between two orbits using Spearman's correlation coefficient between their graphlet degrees and show that the structure of a network can be described using the interplay between a small number of orbits. The orbit pairs that are highly correlated can be grouped together since they describe similar topological characteristics over the nodes. Although we do not

apply this simplification at any experimental step of our method, we utilize it for grouping similar orbits to simplify the illustration in Fig. 3. To identify similar orbit groups, we compute the graphlet degree vectors of all nodes in yeast and human protein-protein interaction networks and compute the pairwise Spearman's correlations between each pair of orbits. We apply single-linkage clustering to uncover the most similar orbit groups. We choose the orbit groups by manually cutting the hierarchical tree considering the similarities of the roles described by orbits. This simplification enables us to easily capture the relationships between GO terms and orbits in Fig. 3. Note that the orbits that are identified to be linked with GO terms are reported individually in Supp. Data 1 to prevent misinterpretation that may be caused by this illustration.

S.9 Additional Case Studies

Case Study 3: Proteasome Complex Proteasome Complex (GO:0000502) term annotates the set of proteins that form a large multisubunit complex that catalyses protein degradation. This complex can be found in the nucleus and cytoplasm of eukaryotes, archaea and some bacteria (Peters *et al.*, 1994). Our analysis on this cellular component shows that the proteins involved in this complex appear in dense regions of the PINs, mostly as cliques (fully connected networks) and mediators connecting other nodes to cliques.

Since this GO term describes a protein complex, it is expected that we observe clique-like patterns associated with this term. This pattern is also verified by (Zhang *et al.*, 2006): their clique searching based functional module identification algorithm finds the proteasome complex as the third highest ranked functional module when applied to the yeast PIN. It is known that the eukaryotic proteasome complex has evolved from a simpler archaebacterial form, being similar in structure and containing only three different peptides (Wollenberg and Swaffield, 2001). Therefore, our results show that the proteasome complex is not only conserved in terms of its sequence, but also it forms similar patterns of interactions that is conserved through evolution from yeast to human.

S.10 Supplementary Tables

Table S.1. The number of GO terms with significant topology-function relationships identified by the structure association strengths and orbit contribution strenghts of yeast and human datasets.

Organism	Bio. Process (Sign. / Total)	Mol. Func. (Sign. / Total)	Cell. Comp. (Sign. / Total)
Human	264 / 1,439	114 / 483	63 / 312
Yeast	111 / 1,439	22 / 483	42 / 312

Table S.2. Densities and average clustering coefficients of subnetworks that are induced on the proteins annotated with different cellular component GO terms. We support our claims on the differences in the wiring patterns of different cellular components by computing the densities and the average clustering coefficients of the subnetworks that are induced on the proteins annotated with different cellular component GO terms. The density of a network (Dens.) is the proportion of the node pairs in a network that are connected with edges. The clustering coefficient of a node is the proportion of the node pairs within its neighbourhood that are connected with each other. The average clustering coefficient (C.C.) of a network is then computed by averaging the clustering coefficients of all nodes in the network.

GO Term ID	GO Term Name	Human Dens. (%)	Human C.C.	Yeast Dens. (%)	Yeast C.C.
GO:0005634	Nucleus	0.379	0.149	1.175	0.250
GO:0005737	Cytoplasm	0.345	0.135	0.823	0.230
GO:0016020	Membrane	0.192	0.110	0.555	0.222

Table S.3. The numbers of proteins that are analysed for different protein degree thresholds for human (column 2) and yeast (column 3). Percentages are computed with respect to the 13,410 human and the 5,831 yeast proteins in the complete PINs.

Degree Threshold	Human	Yeast
2	10,786 (80.4%)	5,457 (93.6%)
3	9,251 (69.0%)	5,081 (87.1%)
4	8,192 (61.1%)	4,740 (81.3%)
5	7,369 (55.0%)	4,435 (76.1%)

Table S.4. The numbers of considered GO terms (columns 2-4) for different annotation count thresholds (column 1). Percentages are computed with respect to the 3,211 biological process, 1,797 molecular function and 585 cellular component GO terms that annotate both yeast and human proteins.

Annotation Threshold	Biological Process	Molecular Function	Cellular Component
3	1,927 (60.0%)	703 (39.1%)	417 (71.3%)
5	1,439 (44.8%)	483 (26.9%)	312 (53.3%)
10	1,068 (33.3%)	356 (19.8%)	212 (36.2%)
15	862 (26.9%)	280 (15.6%)	173 (29.6%)

Table S.5. Topologically orthologous GO terms identified with protein degree threshold = 4 are also consistently identified with other protein degree threshold values. For each of the experiments performed on biological process (column 2), molecular function (column 3) and cellular component (column 4) terms, we report the number of topologically orthologous GO terms that our method identifies ("# of Identified" columns) when using different degree threshold values (rows). We also report the ratio of the orthologous GO terms that are identified for degree threshold 4 that are consistently identified at these different thresholds ("Consistent / Total" columns).

	Biological Process		**Molecular Function**		**Cellular Component**	
Degree Threshold	**# of Identified**	**Consistent / Total**	**# of Identified**	**Consistent / Total**	**# of Identified**	**Consistent / Total**
2	37	14 / 15	0	0 / 0	29	9 / 9
3	29	14 / 15	0	0 / 0	18	9 / 9
5	10	9 / 15	0	0 / 0	2	2 / 9

Table S.6. Topologically orthologous GO terms identified with GO term annotation threshold = 5 are also consistently identified with other GO term annotation threshold values. For each of the experiments performed on biological process (column 2), molecular function (column 3) and cellular component (column 4) terms, we report the number of topologically orthologous GO terms that our method identifies ("# of Identified" columns) when using different GO term annotation thresholds (rows). We also report the ratio of the orthologous GO terms that are identified for GO term annotation threshold 5 that are consistently identified with these different thresholds ("Consistent / Total" columns).

	Biological Process		**Molecular Function**		**Cellular Component**	
Degree Threshold	**# of Identified**	**Consistent / Total**	**# of Identified**	**Consistent / Total**	**# of Identified**	**Consistent / Total**
3	16	15 / 15	0	0 / 0	8	8 / 9
10	23	14 / 15	0	0 / 0	8	8 / 9
15	25	14 / 15	0	0 / 0	6	5 / 9

S.11 Supplementary Data

Supplementary Data 1 can be downloaded from:

`http://bio-nets.doc.ic.ac.uk/conservedPPI/supplementary_Data_1.xlsx`

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300.

Peters, J. M., Franke, W. W., and Kleinschmidt, J. A. (1994). Distinct 19 s and 20 s subcomplexes of the 26 s proteasome and their distribution in the nucleus and the cytoplasm. *Journal of Biological Chemistry*, **269**(10), 7709–7718.

Wang, S. and Wu, F. (2013). Detecting overlapping protein complexes in ppi networks based on robustness. *Proteome Science*, **11**(Suppl 1), S18.

Wollenberg, K. and Swaffield, J. C. (2001). Evolution of proteasomal atpases. *Molecular Biology and Evolution*, **18**(6), 962–974.

Yaveroğlu, O. N., Malod-Dognin, N., Davis, D., Levnajic, Z., Janjic, V., Karapandza, R., Stojmirovic, A., and Pržulj, N. (2014). Revealing the hidden language of complex networks. *Scientific Reports*, **4**(4547).

Zhang, C., Liu, S., and Zhou, Y. (2006). Fast and accurate method for identifying high-quality protein-interaction modules by clique merging and its application to yeast. *Journal of Proteome Research*, **5**(4), 801–807.