GRPhIN: Graphlet Characterization of Regulatory and Physical Interaction Networks

Altaf Barelvi[1,†], Oliver Anderson[1,†], and Anna Ritz[1,*]

[1]Biology Department, Reed College, Portland, Oregon, USA
[†]These authors contributed equally
[*]Corresponding Author: aritz@reed.edu

Abstract

Graphs are powerful tools for modeling and analyzing molecular interaction networks. Graphs typically represent either undirected physical interactions or directed regulatory relationships, which can obscure a particular protein's functional context. Graphlets can describe local topologies and patterns within graphs, and combining physical and regulatory interactions offer new graphlet configurations that can provide biological insights. We present GRPhIN, a tool for characterizing graphlets and protein roles within graphlets in mixed physical and regulatory interaction networks. We describe the graphlets of mixed networks in *B. subtilis*, *C. elegans*, *D. melanogaster*, *D. rerio*, and *S. cerevisiae* and examine local topologies of proteins and subnetworks related to the oxidative stress response pathway. We found a number of graphlets that were abundant in all species, specific node positions (orbits) within graphlets that were over-represented in stress-associated proteins, and rarely-occurring graphlets that were enriched in stress subnetworks. These results showcase the potential for using graphlets in mixed physical and regulatory interaction networks to identify new patterns beyond a single interaction type.

Keywords: graphlets, protein interaction networks, gene regulatory networks, oxidative stress response

1 Introduction

Graphs are powerful tools for computationally representing complex biological networks [1–4], and are effective in identifying differences between healthy and disease states in organisms [5]. Biological networks can take many forms, and researchers often study networks with isolated interaction types. Some of the most commonly studied network types are protein-protein interaction (PPI) networks and gene regulatory networks (GRNs). PPI networks represent all the physical and molecular interactions of proteins within a system or organism [6] and are represented graphically with proteins as nodes and undirected edges describing their interactions. GRNs differ from PPIs in both the biological context they provide and their graphical structure. GRNs describe directed interactions between transcription factors (TF) and their target genes and are typically used to study TF-regulated development, diseases, and processes [7].

Many tools represent PPIs and GRNs in isolation, but this can obscure the complete functional context of a particular protein [8]. PPI networks and GRNs do not exist separately: proteins are TFs, genes encode proteins, and physical and regulatory interactions coexist and mix, forming

distinct patterns that networks with only one interaction type cannot capture. Wang et al. [9] found that physical interactions can participate in regulatory interactions and can even predict long-range enhancer-promoter interactions. Thus, identifying and quantifying patterns and proteins within a heterogeneous network would be valuable in understanding the full biological context of a specific protein or the regulation of a pathway or process.

Characterizing biological networks beyond simple summary statistics has been an ongoing research area. Graphlets are small, connected subnetworks that can describe the local network organization of a graph [10, 11]. Graphlets have been used in many applications of molecular interaction networks [12], including protein function prediction [13], characterizing disease genes [14], and describing the local topologies of signaling pathways [15], among many others.

Since their introduction in the early 2000s, efficient algorithms have been developed to count graphlets and the specific node positions within each graphlet (called *orbits*) [16, 17]. Graphlets have been defined on network extensions such as directed graphs [18–20], signed graphs [21], and probablistic graphs [22]. While some research describes graphlets for heterogeneous networks, which contain different types of relational information, this area is less well-developed. In 2004, Yeger-Lotem and colleagues counted graphlets that contained a mixture of physical and regulatory interactions in yeast, and identified five graphlets that were over-represented compared to a random rewiring model [8]. In 2017, Somnez and Can extended graphlets to directed multi-label graphs [23], where edges may have different edge types. In 2020, Dimitrova and colleagues characterized graphlets in undirected multiplex networks with different edge types [24]. To our knowledge, no previous research except for the 2004 work [8] has explicitly considered the case where there are a mix of directed and undirected edges, which is required when combining undirected physical and directed regulatory interactions.

In this paper, we adapt the multi-label graphlet identification work presented in [23] and apply it to mixed networks that contain physical undirected interactions and directed regulatory interactions. Through a new tool called **G**raphlet Characterization of **R**egulatory and **Ph**ysical **I**nteraction **N**etworks (GRPhIN, pronounced "griffin"), we characterize the graphlet distributions of networks of five different species (*B. subtilis*, *C. elegans*, *D. rerio*, *D. melanogaster*, and *S. cerevisiae*), and identify graphlets with a mix of physical and regulatory edges that are conserved across all species. Additionally, we characterize oxidative stress response, a pathway that is conserved across the five species, both by using the node positions within graphlets as well as by identifying subnetworks that represent the stress response pathways. GRPhIN gives us new insights into local physical and regulatory patterns within these networks and within the stress response pathway.

2 Methods

We first formally define the graph that we generate from species-specific data. For a particular species, we identify undirected physical relationships (undirected edges) and directed regulatory relationships (directed edges) for a set of proteins (nodes). We build a regulatory and physical interaction (RPI) graph by taking the union of the undirected and directed edges, creating a multigraph (where two nodes may have more than one edge). Here, nodes represent either a protein or the gene that encodes the protein (in the case of a regulatory target). We then re-label the edges in the RPI graph into one of five edge types (Figure 1A), making this graph a multi-label graph with only one edge between any two nodes. Note that this graph still contains a mixture of undirected (e.g., physical-only), bi-directed (e.g., two co-regulating edges), and directed edges.

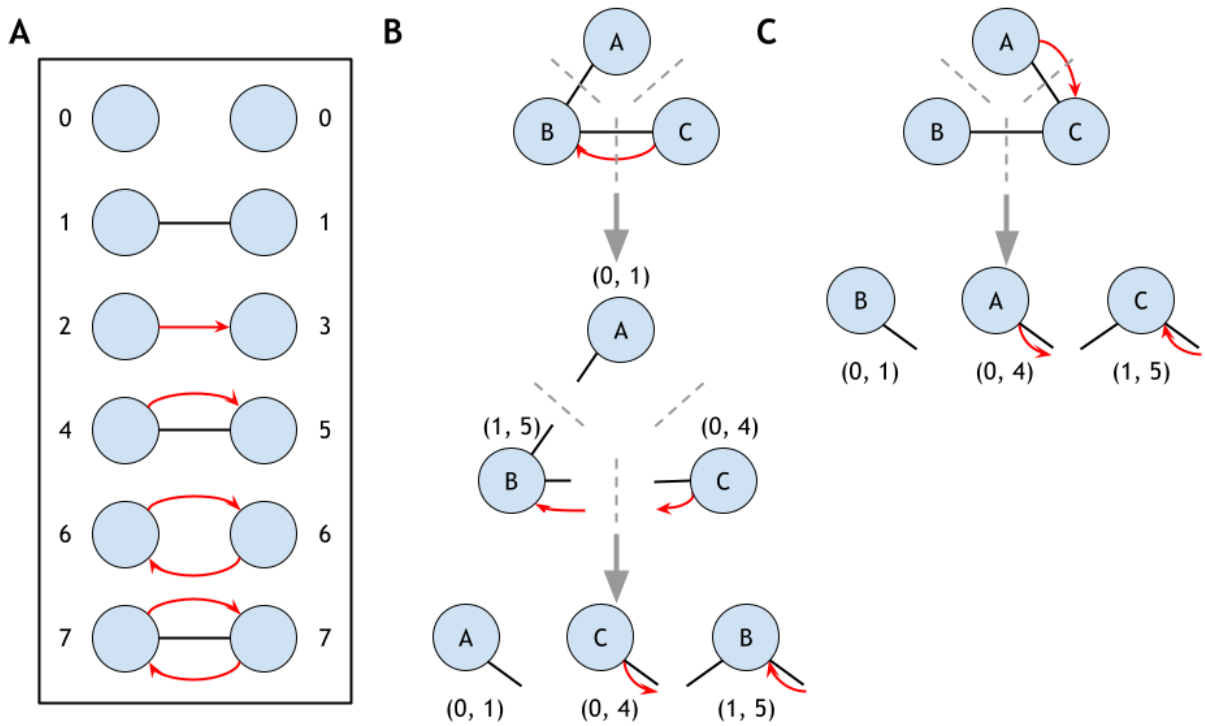


Figure 1: Orbit labeling and graphlet sorting. A) Unique edges and orbit in an RPI graph (orbit 0 indicates a non-edge; orbits 1–7 are mapped to the five distinct edge types). B) Sorting node orbits in ascending order maintains a unique graphlet structure. C) Sorting isomorphisms result in the same unique set of integer pairs.

2.1 Regulatory and Physical Interaction (RPI) Graphlets

Given a graph G, graphlets are connected, induced, non-isomorphic subgraphs of G described in previous work [10, 11, 17]. Graphlets are typically restricted to subgraphs of a specific size (e.g., 2-4 nodes). We define regulatory and physical interaction (RPI) graphlets, which are connected, induced, non-isomorphic subgraphs of the RPI graph. There are five two-node RPI graphlets (corresponding to the five edge types, Figure 1A), but the number of three-node graphlets is much larger than graphlets on directed or undirected graphs: there are 98 three-node graphlets, including 28 that contain two edges ("line graphlets") and 70 that contain three edges ("triangle graphlets"). Supplementary Figure 6 shows all graphlets; a larger version is available at https://github.com/Reed-CompBio/grphin. We restrict our attention to these 2- and 3-node graphlets due to the combinatorial complexity of the enumeration.

For each RPI graphlet, we define the non-automorphic node positions (orbits) which make up that graphlet. Two nodes are equivalent if some automorphism maps one node onto the other; we partition graphlets into equivalence classes defined as orbits [25]. For example, in Figure 1A: orbit 4 reflects a node with a physical edge and outgoing regulatory edge, while orbit 5 reflects a node with a physical edge and incoming regulatory edge. There are seven orbits for the 2-node graphlets and 259 orbits for the 3-node graphlets. While previous work enumerated the orbits, we index the orbits by the specific graphlet for clarity (described in more detail below).

2.1.1 RPI Graphlet Encoding

Since graphlets are connected, each node has exactly one or two incident edges. These incident edges can be described by a node's orbits in the 2-node RPI graphlets, labeled as index 1–7 (we will call orbits in the 2-node RPI graphlets *edge orbits*). If a node has only one incident edge, we assign a 0 to indicate it is not connected to the other node. Thus, each 3-node RPI graphlet is described as a triple of integer pairs that describe the edge adjacencies among the nodes (Figure 1B/C). For a graphlet G with labeled nodes a, b, and c:

$$\big(\underbrace{(o_{a\to b}, o_{a\to c})}_{\text{node } a}, \underbrace{(o_{b\to a}, o_{b\to c})}_{\text{node } b}, \underbrace{(o_{c\to a}, o_{c\to b})}_{\text{node } c}\big), \tag{1}$$

where

$$o_{x\to y} \begin{cases} \text{edge orbit of } x \text{ in } (x,y) & \text{if edge } (x,y) \text{ exists} \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Note that the first two pairs completely determine the graphlet (e.g., once you know the incident edges of a and b, then c is automatically determined), but we keep this notation for ease of exposition. We order the orbits within these pairs: that is, let $o_a = \text{sort}(o_{a\to b}, o_{a\to c})$ for graphlet with nodes a, b, and c (the top portion of Figure 1B illustrates this idea). Without loss of generality, we assume that a, b, c are ordered such that $o_a < o_b < o_c$. Thus, the final encoding for graphlet G with nodes labeled a, b, c is

$$G(a, b, c) = (o_a, o_b, o_c)\,. \tag{3}$$

This ordering is unique to each 3-node RPI graphlet and will be used to break symmetries when counting. Figure 2B-C show how two isomorphic graphlets, when sorted, results in the same graphlet encoding.

2.1.2 Counting RPI Graphlets and Orbits

Many analytic and enumeration-based algorithms that are extensions of exact subgraph counting exist to count graphlets (see [25] for a review). However, due to the mixed-graph nature of RPI graphs, we developed GRPhIN to count graphlets. Since we are concerned with 2- and 3-node RPI graphlets, we take an enumeration approach similar to counting directed multi-label graphlets from Sonmez and Can [23].

To count 3-node RPI graphlets, we first arbitrarily order the nodes in the RPI graph G and enumerate all connected triples of nodes i, j, k. By definition, the subgraph of each connected node triple i, j, k is isomorphic to some 3-node RPI graphlet. For each node i, j, and k, we encode the pairs of node adjacencies and sort each pair to produce o_i, o_j, o_k. Then, we find a mapping $f : \{i, j, k\} \mapsto \{a, b, c\}$ such that:

$$G(f(i), f(j), f(k)) = G(a, b, c) \tag{4}$$

for some graphlet G (Figure 1B,C). We increment the count of graphlet G for the triple i, j, k, and we increment the orbits of $f(i)$,$f(j)$, and $f(k)$ for the nodes i, j, and k. The output of GRPhIN is (1) a count of every 2-node and 3-node RPI graphlet in the entire RPI graph (called a graphlet distribution vector in previous work [13, 25]) and (2) a count of every orbit in every RPI graphlet for each node in the RPI graph. GRPhIN is available at the https://github.com/Reed-CompBio/grphin.

2.2 Algorithm Running Time

We used five different species networks that had varying network density and size (see below). In the most connected and largest network, *S. cerevisiae*, the run time was less than 50 minutes, and on the smallest network, *B. subtilis*, was 10 seconds. The rest of the networks ran in less than 10 minutes. The runtime of the algorithm increased as the average number of node neighbors increased, requiring more computation to calculate all three node combinations. All computation was done on a Linux machine with an AMD Ryzzen Threadripper 3990x 64-Core Processor CPU, 257 GB memory, and two Nvidia GeForm RTX 3060 GPUs. Since this tool is Python3 based and did not utilize multi-threading, the added performance increase was due to the additional CPU cores and memory rather than the GPUs.

2.3 Species-Specific Networks

	Proteins		**Interactions**	
Organism	**Total**	**Stress**	**PPI**	**Reg.**
B. subtilis	3,163	18	6,441	5,634
C. elegans	4,106	103	13,915	78,223
D. melanogaster	12,823	145	233,054	17,530
D. rerio	16,606	154	45,003	25,960
S. cerevisiae	7,644	124	164,432	237,315

Table 1: Regulatory and physical interaction (RPI) networks. PPI: protein-protein interaction. Reg.: regulatory. Stress proteins were collected from the Gene Ontology (see Methods).

We collected experimentally, text-mined, and database-validated physical and regulatory interaction data for five non-human model organisms: *B. subtilis*, *C. elegans*, *D. melanogaster*, *D. rerio*, and *S. cerevisiae* (Table 1). These interactions come from species-specific databases (e.g., DRoID, FlyBase, WormBase, ZFin, and SubtiWiki) and broader databases that contain species-specific components (e.g., STRING, BIOGRID, and TFLink). These publicly available networks are described in more detail in the ProteinWeaver paper [26].

2.4 Stress Response Proteins

To identify proteins involved in stress responses present in all species, we examined gene ontology (GO) annotations [27, 28]. Using AMIGO [29], we exported a list of proteins annotated with the GO term "response to oxidative stress" (GO:0006979), a process common to the five species studied (Table 1).

Evaluating Over-Representation of Orbits To evaluate whether stress proteins tend to occupy certain roles (orbits) within RPI graphlets, we performed a random sampling method for each species network separately, following a process to Agrawal et al. [14]. We first calculated the observed median orbits of the collection of stress proteins. We then calculated the median orbits of 1,000 randomly sampled subsets of nodes of the same size as the stress proteins. An orbit is significantly over-represented in the stress proteins if less then 5% of the random medians are equal to or larger than the observed median, corresponding to a p-value of 0.05.

2.5 Stress Response Subnetworks

For each species, we performed a random walk on the RPI network restarting at the oxidative stress response proteins ($\alpha = 0.85$) to generate a list of proteins near stress response proteins in the graph. We recovered the highest-scoring nodes from the RWR until 80% of the stress response proteins were included and generated the induced subnetwork (which we will call the *oxidative stress subnetwork* for each species). We then counted RPI graphlets for each stress response subnetwork.

Evaluating Over-Representation of Oxidative Stress Subnetworks Enrichment ratio scores, $\mathbb{E}$, indicate whether the number of graphlets observed in the subnetwork is surprising compared to the full interactome of a particular organism [30]. The score is the ratio of the graphlets in the subnetwork to the expected number of graphlets based on the structure of the RPI network. Enrichment ratio scores greater than one ($\mathbb{E} > 1$) suggest graphlet enrichment, while scores less than one indicate the graphlet occurs less frequently than expected. Given a graph $G = (V, E)$, a subgraph $G' = (V', E') \subseteq G$, and a graphlet g, let $C(G, g)$ be the count of graphlet g in graph G. Let $|E|$ and $|E'|$ be the number of edges in the global network and subnetwork, respectively.

$$\mathbb{E} = \frac{C(G', g)}{\mu}, \text{ where} \tag{5}$$

$$\mu = \frac{|E'|C(G, g)}{|E|}. \tag{6}$$

Here, μ represents the expected number of graphlets in the subnetwork, calculated by normalizing the number of edges in the subnetwork to the global network. Degree-based scaling accounts for uneven connectivity among nodes, as some nodes act as hubs with many connections while others have very few. This scaling ensures that μ reflects the network structure without bias [31]. For all enrichment scores, we calculate Z-scores and use Python's `scipy` module to calculate an associated p-value [32–34].

3 Results

3.1 RPI Graphlet Characterization of Species Networks

We first examined the distribution of graphlet counts across the five species networks. Of the two-node graphlets (the edge types illustrated in Figure 1A), the edge type that denotes a single physical interaction and the edge type that denotes a single regulatory interaction dominate the RPI graphlet counts: only 0.03-1.1% of the species networks contain the two mixed edge types. These rare edge types appear in some significant RPI graphlets, however, and many mixed RPI graphlets consist of single-physical or single-regulatory edges, as we will see.

Despite the differences in the RPI graph sizes and total abundance of RPI graphlets in each graph, there is a distinct pattern of three-node RPI graphlets across species (Figure 2A). The most abundant RPI graphlets are either a line of two physical edges (index 0) a line with two outgoing regulatory edges from a middle node (index 23) (Figure 2B). Interestingly, Graphlet 15, which includes a physical and regulatory mixed edge, is also abundant across species (except for *D. melanogaster*, where it is present but not one of the most abundant graphlets). While less abundant, there are also a number of triangle graphlets that are commonly found in all networks (Figure 2B). Note that there is some dependence between the line and triangle graphlets (e.g., many of the triangle graphlets contain subgraphs that are isomorphic to the most common line graphlets).

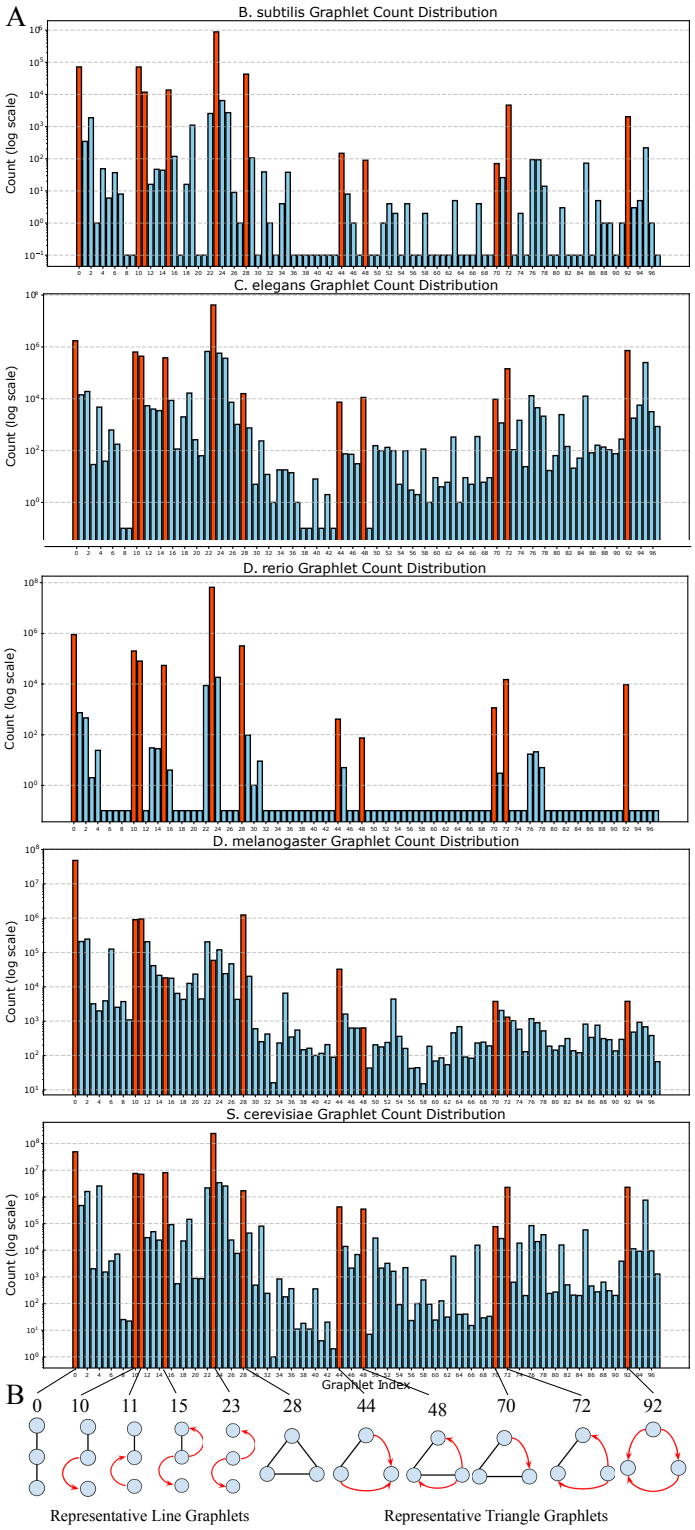
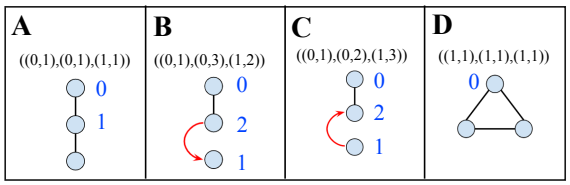


Figure 2: RPI Graphlet characterizations of the five species networks. (A) Graphlet count distributions of the 98 graphlets for each species network. Graphlet index indicates the sorted ordering of the graphlet encoding. (B) Representative 3-node graphlets that occur the most across all five species (red bars in panel A).

3.2 Orbit Distributions of Stress vs. Non-Stress proteins

We then identified specific orbits that tend to be occupied by proteins related to oxidative stress for each species. Significantly over-enriched orbits may indicate that there is a particular role the stress proteins play within the network. We found 0 significant orbits in *B. subtilis*, 9 significant orbits in *C. elegans*, 2 significant orbits in *D. rerio*, 22 significant orbits in *D. melanogaster*, and 48 significant orbits in *S. cerevisiae* ($p < 0.05$ with 1,000 random samples; see Methods). Since *B. subtilis* has no significant orbits (presumably because there were only 18 stress proteins), we focus on the remaining four species for the rest of this section.



Orbit	*B. sub.*	*C. ele.*	*D. mela.*	*D. rer.*	*S. cer.*
A0	*1.5*	**883**	**10,599**	**5**	**15,510**
D0	*0*	**1**	**74**	**1**	**92**
A1	*0*	**4**	**412.5**	*0*	**181**
B2	*0*	**6**	**14**	*0*	**145**
C0	*0*	**27**	**52**	*0*	**390**
C1	*0*	**43**	**569**	*0*	**305**

Table 2: Orbits found to be significant in at least three of the species networks. (Top) The four graphlets that contain the six significant graphlets in the table below (orbits in blue). (Bottom) Median orbit counts for the stress proteins in each species. Orbits are indicated by their graphlets A-D in the top image. Bold: significant ($p < 0.05$).

Two of the significant orbits were common to all four species, and four of the significant orbits are common to *C. elegans*, *D. melanogaster*, and *S. cerevisiae* (Table 2). Most of these orbits are part of graphlets through protein-protein interactions, indicating that the stress proteins are enriched in physical interactions (rather than regulatory interactions). However, the center orbit in the graphlet B from Table 2 is an example where the node role involves a physical connection to one node and regulation of another node.

3.3 Graphlet Enrichment in Oxidative Stress Subnetworks

		Edges		Graphlets	
Org.	**Nodes**	**PPI**	**Reg.**	**Unique**	**Enriched**
B. sub.	27	34	39	17	3
C. ele.	123	452	2686	88	56
D. mela.	141	1436	217	84	18
D. rer.	49	24	87	12	4
S. cer.	177	2492	5709	96	72

Table 3: Oxidative stress response subnetwork sizes and number of of unique and enriched 3-node graphlets ($\mathbb{E} > 1$ and $p < 0.05$).

We generated subnetworks that connected 80% of the oxidative stress response proteins within each species network (see Methods). All 98 unique 3-node graphlets were present in at least one of the oxidative stress subnetworks, and 80 unique graphlets were found to be significantly enriched in at least one subnetwork (Table 3). *S. cerevisiae* had the largest number of enriched graphlets out of all species, and it also had the most diverse array of enriched graphlets, with 72 of the 98 unique graphlets were enriched in the stress response subnetworks compared to the full RPI network. Conversely, *B. subtilis* had the lowest count of enriched graphlets in the stress response subnetwork, presumably because it was the smallest subnetwork in terms of nodes and edges. While all graphlets were present in at least one subnetwork, only nine graphlets were present in the the subnetworks of all five species: two graphlets with only physical interactions, four graphlets with only regulatory interactions, and three graphlets with mixed regulatory and physical interactions.

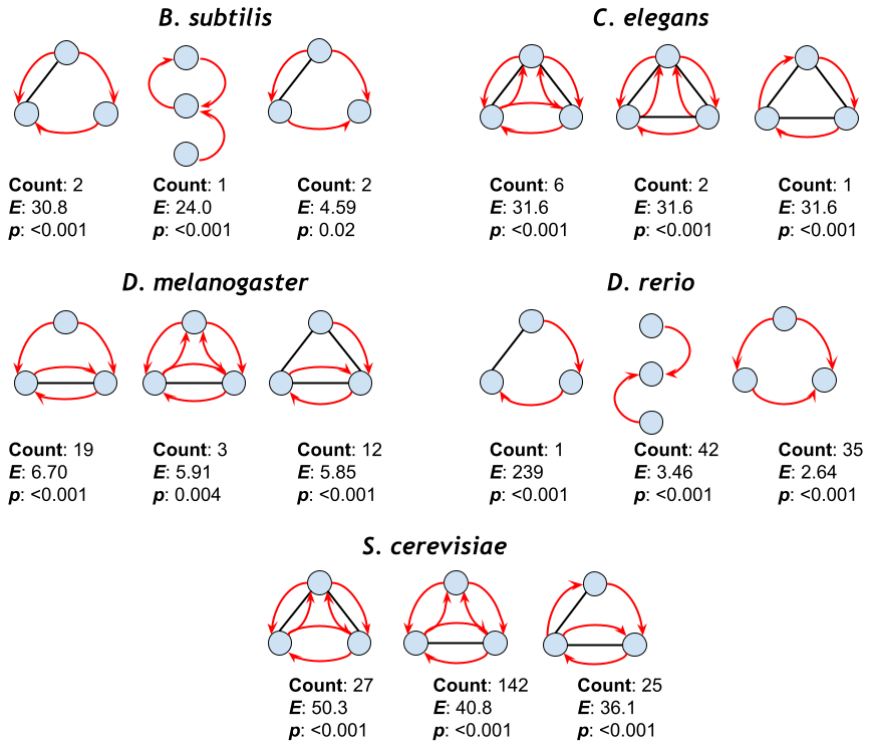


Figure 3: The three most enriched graphlets in each species' oxidative stress response networks (all have p-values < 0.05).

The most enriched graphlets (with a p-value of < 0.05) differed across all species' oxidative stress response networks (Figure 3). Thirteen of the 15 enriched 3-node graphlets in Figure 3, are triangle graphlets, and 11 of the graphlets contain mixed edge types (which comprise a very small percentage of the total network, as we reported earlier). Thus, even though there are relatively few numbers of these mixed-edge-type graphlets, they are surprisingly enriched within the stress subnetworks.We again note a similarity among the enriched graphlets, since some graphlets differ by a single edge type.

None of the nine graphlets found in all subnetworks were significantly enriched in all five species. Howevever, 13 graphlets were significantly enriched in at least three species (including 11 graphlets with mixed interactions). Four of these mixed graphlets also appeared at least once in the top three most enriched graphlets by species (Figure 4). As a case study, we visualized the graphlet in

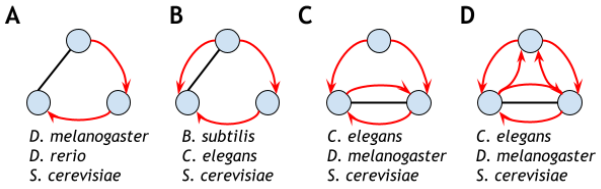


Figure 4: RPI graphlets common to at least three species' oxidative stress response networks that appear at least once in Figure 3

Figure 4B within the *B. subtilis* oxidative stress subnetwork (Figure 5). The graphlet in Figure 4B is a feed-forward loop with an additional physical interaction. The *B. subtilis* oxidative stress subnetwork contains RNA polymerase sigma factor *sigA*, which initiates transcription in all growth phases, stress, and sporulation. In two mixed graphlets in *B. subtilis*, *sigA* regulates and forms a physical co-complex with *spxA*, an enzyme known to regulate growth during periods of stress [35]. In one, *sigA* regulates *perR* a hydrogen-and peroxide-activated transcription factor that coregulates *spxA* with *sigA*. In the other, *sigA* coregulates *spxA* with *sigB*, a sigma factor that *sigA* also regulates. In these two cases, the graphlet describes a transcriptional activation complex within a feed-forward loop, a pattern that neither physical interactions nor regulatory interactions would capture alone.

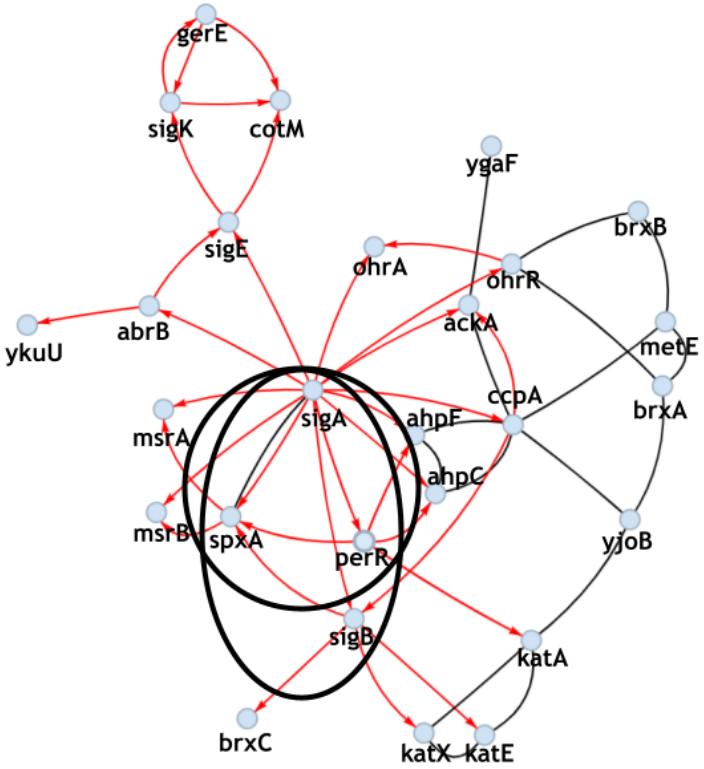


Figure 5: *B. subtilis* oxidative stress response network with the mixed graphlet enriched in three species highlighted (Figure 4B).

4 Discussion

GRPhIN is a new graphlet-counting approach for enumerating mixed patterns of undirected physical and directed regulatory interactions. GRPhIN adopts a graphlet and orbit encoding scheme that breaks isomorphic (graphlet) and automorphic (orbit) symmetries in a two-step sorting process (Figure 1). We use this sorting in an enumerative algorithm, which can count graphlets and orbits quickly in practice (the largest network took less than 50 minutes). However, there are many opportunities to optimize this algorithm, especially by adopting an analytic approach to graphlet counting similar to ORCA [17, 25].

We applied GRPhIN to the networks of five species, identifying commonly abundant graphlets despite their varying size and composition of physical and regulatory interactions (Figure 2). Focusing on the oxidative stress response pathway, we identified node positions (orbits) that tended to be occupied by stress proteins more often than by chance. While none of these orbits were significant across all species, six were significant for three or more networks (Table 2. We then identified graphlets that were enriched in stress subnetworks across all species (Table 3 & Figure 3), including identifying four graphlets with mixed physical and regulatory interactions that were significant in at least one species' subnetwork and present in two others (Figure 4). Finally, we illustrated a graphlet that describes a transcriptional activation complex within a feed-forward loop in *B. subtilis* (Figure 5).

Oxidative stress response is conserved across species [36], and our results highlight roles that stress proteins may play within the larger network. Ongoing work includes biologically characterizing enriched graphlets and over-represented orbits found in the stress subnetworks, such as the graphlet case study we describe in Figure 5. Identifying key proteins' roles within mixed graphlets could provide insights into stress response coordination beyond single network types. In addition, further comparison across species could identify whether specific stress-response graphlets are universally conserved or organism-specific. Oxidative stress is only one type of stress and other stress response pathways may be universally conserved, further analysis into different stress response subnetworks could provide insights into whether graphlets are specific to oxidative stress responses or are enriched in other stress networks.

While GRPhIN has uncovered similarities among the five species networks, additional work is needed to identify conserved graphlets across taxa. The species networks are all in different states of completeness, largely due to the nature of research with those organisms. Further, graphlets have been criticized as being sensitive to network perturbations, though other work has reaffirmed their robustness and stability for describing protein interaction networks [37]. Accounting for the density and completeness of the species networks when identifying conserved graphlets remains future work.

5 Competing interests

No competing interest is declared.

6 Author contributions statement

A.R. conceived the project and supervised all work; A.B. and O.A. implemented all code and conducted all experiments. All authors wrote and reviewed the manuscript.

7 Acknowledgements

This work is supported by the National Science Foundation (NSF: #1750981).

References

[1] Federica Conte et al. A paradigm shift in medicine: A comprehensive review of network-based approaches. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1863(6):194416, June 2020.

[2] Jessica Xin Hu, Cecilia Engel Thomas, and Søren Brunak. Network biology concepts in complex disease comorbidities. *Nature Reviews Genetics*, 17(10):615–629, October 2016.

[3] Bai Zhang, Ye Tian, and Zhen Zhang. Network Biology in Medicine and Beyond. *Circulation: Cardiovascular Genetics*, 7(4):536–547, August 2014.

[4] Giulia Muzio, Leslie O'Bray, and Karsten Borgwardt. Biological network analysis with deep learning. *Briefings in Bioinformatics*, 22(2):1515–1530, March 2021.

[5] Nahid Safari-Alighiarloo et al. Protein-protein interaction networks (PPI) and complex diseases. *Gastroenterology and Hepatology From Bed to Bench*, 7(1):17–31, 2014.

[6] Gavin C. K. W. Koh et al. Analyzing Protein–Protein Interaction Networks. *Journal of Proteome Research*, 11 (4):2014–2031, April 2012.

[7] Daniele Mercatelli et al. Gene regulatory network inference resources: A practical overview. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1863(6):194430, June 2020.

[8] Esti Yeger-Lotem et al. Network motifs in integrated cellular networks of transcription–regulation and protein–protein interaction. *Proceedings of the National Academy of Sciences*, 101(16):5934–5939, April 2004.

[9] Hao Wang, Binbin Huang, and Jianrong Wang. Predict long-range enhancer regulation based on protein–protein interactions between transcription factors. *Nucleic Acids Research*, 49(18):10347–10368, October 2021.

[10] Natasa Pržulj, Derek G Corneil, and Igor Jurisica. Modeling interactome: scale-free or geometric? *Bioinformatics*, 20(18):3508–3515, 2004.

[11] Nataša Pržulj. Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23(2):e177–e183, 2007.

[12] Khalique Newaz and T Milenković. Graphlets in network science and computational biology. *Analyzing network data in biology and medicine: an interdisciplinary textbook for biological, medical and computational scientists*, 193, 2019.

[13] Tijana Milenković and Nataša Pržulj. Uncovering biological network function via graphlet degree signatures. *Cancer informatics*, 6:CIN–S680, 2008.

[14] Monica Agrawal, Marinka Zitnik, and Jure Leskovec. Large-scale analysis of disease pathways in the human interactome. In *PACIFIC SYMPOSIUM on BIOCOMPUTING 2018: Proceedings of the Pacific Symposium*, pages 111–122. World Scientific, 2018.

[15] Tobias Rubel, Pramesh Singh, and Anna Ritz. Reconciling signaling pathway databases with network topologies. In *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2022*, pages 211–222. World Scientific, 2021.

[16] Tomaž Hočevar and Janez Demšar. A combinatorial approach to graphlet counting. *Bioinformatics*, 30(4): 559–565, 2014.

[17] Tomaž Hočevar and Janez Demšar. Combinatorial algorithm for counting small induced graphs and orbits. *PLOS ONE*, 12(2):e0171428, February 2017. ISSN 1932-6203. doi: 10.1371/journal.pone.0171428.

[18] Igor Trpevski, Tamara Dimitrova, Tommy Boshkovski, Nikola Stikov, and Ljupcho Kocarev. Graphlet characteristics in directed networks. *Scientific Reports*, 6(1):1–8, 2016.

[19] Anida Sarajlić, Noël Malod-Dognin, Ömer Nebil Yaveroğlu, and Nataša Pržulj. Graphlet-based characterization of directed networks. *Scientific Reports*, 6(1):1–14, 2016.

[20] David Aparicio, Pedro Ribeiro, and Fernando Silva. Extending the applicability of graphlets to directed networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 14(6):1302–1315, 2016.

[21] Apratim Das, Alex Aravind, and Mark Dale. Algorithm and application for signed graphlets. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 613–620, 2019.

[22] Sergio Doria-Belenguer, Markus K Youssef, René Böttcher, Noël Malod-Dognin, and Nataša Pržulj. Probabilistic graphlets capture biological function in probabilistic molecular networks. *Bioinformatics*, 36(Supplement_2): i804–i812, 2020.

[23] Arzu Burcak Sonmez and Tolga Can. Comparison of tissue/disease specific integrated networks using directed graphlet signatures. In *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 533–534, 2016.

[24] Tamar Dimitrova, Kristijan Petrovski, and Ljupcho Kocarev. Graphlets in multiplex networks. *Scientific reports*, 10(1):1928, 2020.

[25] Pedro Ribeiro, Pedro Paredes, Miguel EP Silva, David Aparicio, and Fernando Silva. A survey on subgraph counting: concepts, algorithms, and applications to network motifs and graphlets. *ACM Computing Surveys (CSUR)*, 54(2):1–36, 2021.

[26] Oliver Anderson, Altaf Barelvi, Aden O'Brien, Ainsley Norman, Iris Jan, and Anna Ritz. Proteinweaver: A webtool to visualize ontology-annotated protein networks. *bioRxiv*, pages 2024–10, 2024.

[27] Suzi A Aleksander et al. The Gene Ontology knowledgebase in 2023. *Genetics*, 224(1):iyad031, March 2023. ISSN 0016-6731. doi: 10.1093/genetics/iyad031.

[28] Michael Ashburner et al. Gene Ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, May 2000.

[29] Seth Carbon et al. AmiGO: online access to ontology and annotation data. *Bioinformatics*, 25(2):288–289, January 2009.

[30] Jesús Lascorz, Kari Hemminki, and Asta Försti. Systematic enrichment analysis of gene expression profiling studies identifies consensus pathways implicated in colorectal cancer development. *Journal of Carcinogenesis*, 10:7, March 2011. ISSN 1477-3163. doi: 10.4103/1477-3163.78268.

[31] S. Itzkovitz, R. Milo, N. Kashtan, G. Ziv, and U. Alon. Subgraphs in random networks. *Physical Review E*, 68 (2):026127, August 2003. doi: 10.1103/PhysRevE.68.026127. Publisher: American Physical Society.

[32] N. Kashtan, S. Itzkovitz, R. Milo, and U. Alon. Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. *Bioinformatics*, 20(11):1746–1758, July 2004. ISSN 1367-4803. doi: 10.1093/bioinformatics/bth163.

[33] Shuo Yu, Jin Xu, Chen Zhang, Feng Xia, Zafer Almakhadmeh, and Amr Tolba. Motifs in Big Networks: Methods and Applications. *IEEE Access*, 7:183322–183338, 2019. ISSN 2169-3536. doi: 10.1109/ACCESS.2019.2960044. Conference Name: IEEE Access.

[34] Pauli Virtanen et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

[35] Daniel F. Rojas-Tapias and John D. Helmann. Roles and regulation of Spx family transcription factors in Bacillus subtilis and related species. *Advances in microbial physiology*, 75:279–323, 2019. ISSN 0065-2911. doi: 10.1016/bs.ampbs.2019.05.003. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7055150/.

[36] Enni Markkanen. Not breathing is not an option: How to deal with oxidative DNA damage. *DNA Repair*, 59:82–105, November 2017. ISSN 1568-7864. doi: 10.1016/j.dnarep.2017.09.007. URL https://www.sciencedirect.com/science/article/pii/S1568786417303439.

[37] Wayne Hayes, Kai Sun, and Nataša Pržulj. Graphlet-based measures are suitable for biological network comparison. *Bioinformatics*, 29(4):483–491, 2013.

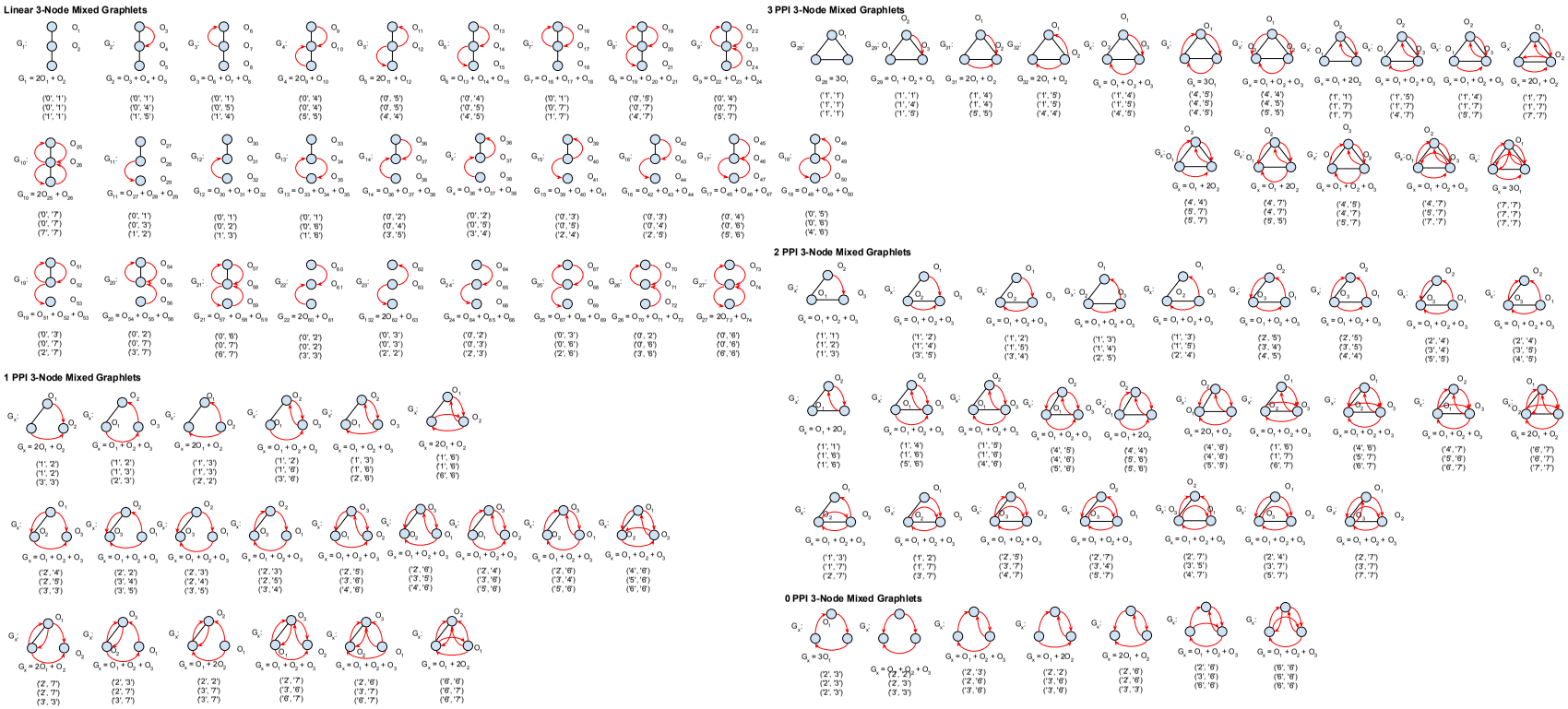


Figure 6: Supplementary Figure of all graphlets and graphlet encodings.