Bioinformatics Advances, 2023, **00**, vbad140
https://doi.org/10.1093/bioadv/vbad140
Advance Access Publication Date: 4 October 2023
Original Article

Systems biology

Identification of disease modules using higher-order network structure

Pramesh Singh [1,2], **Hannah Kuder**[3], **Anna Ritz** [1,*]

[1]Biology Department, Reed College, Portland, OR 97202, United States
[2]Data Intensive Studies Center, Tufts University, Medford, MA 02155, United States
[3]Physics Department, Reed College, Portland, OR 97202, United States

*Corresponding author. Biology Department, Reed College, 3203 SE Woodstock Blvd, Portland, OR 97202, United States. E-mail: aritz@reed.edu

Associate Editor: Sofia Forslund

Abstract

Motivation: Higher-order interaction patterns among proteins have the potential to reveal mechanisms behind molecular processes and diseases. While clustering methods are used to identify functional groups within molecular interaction networks, these methods largely focus on edge density and do not explicitly take into consideration higher-order interactions. Disease genes in these networks have been shown to exhibit rich higher-order structure in their vicinity, and considering these higher-order interaction patterns in network clustering have the potential to reveal new disease-associated modules.

Results: We propose a higher-order community detection method which identifies community structure in networks with respect to specific higher-order connectivity patterns beyond edges. Higher-order community detection on four different protein–protein interaction networks identifies biologically significant modules and disease modules that conventional edge-based clustering methods fail to discover. Higher-order clusters also identify disease modules from genome-wide association study data, including new modules that were not discovered by top-performing approaches in a Disease Module DREAM Challenge. Our approach provides a more comprehensive view of community structure that enables us to predict new disease–gene associations.

Availability and implementation: https://github.com/Reed-CompBio/graphlet-clustering.

1 Introduction

Understanding how genes and proteins interact with each other is a fundamental problem in molecular biology. Recent advancements in high-throughput experiments and computational techniques have enabled accurate inference of the underlying molecular interaction networks. Many complex diseases are caused by a number of genes or proteins interacting with one another (Oti *et al.* 2006), yet identifying such a group (also called a disease module) in a molecular interaction network such as a protein–protein interaction (PPI) network (an *interactome*) is computationally challenging. A common way to find these groups is to use *community detection* (or *clustering*) methods that aim to find densely connected subsets of nodes in a given network. There is another class of methods to discover these groups which takes as input the PPI network and a known set of disease genes and builds these groups (Ghiassian *et al.* 2015, Levi *et al.* 2021). However, in this article we focus on the former. A number of different community detection algorithms have been developed and used extensively over the years for this task (Fortunato 2010, Choobdar *et al.* 2019). Recently, the DREAM Disease Module Identification Challenge systematically assessed 75 community detection algorithms to detect modules across six different PPI networks that are enriched in genome-wide association study (GWAS) data from 180 diseases (Choobdar *et al.* 2019). While these methods have been useful in detecting disease groups in biological networks, they differ significantly from each other and show varying performance (e.g. number of significant modules and their sizes), suggesting that optimal detection of these disease modules remains a challenging task.

Despite their differences, nearly all community detection algorithms focus on identifying groups of nodes that are densely connected by edges. Thus, these methods rely on pairwise relationships between nodes while neglecting higher-order interaction patterns among more than two nodes. The importance of higher-order structure within biological networks has been emphasized by many recent studies (Benson *et al.* 2016, Agrawal *et al.* 2018, Rubel *et al.* 2022). Therefore, it is worthwhile to extend the idea of communities to higher-order communities that are groups of nodes connected by pattern involving more than two nodes. However, besides a few methods (Arenas *et al.* 2008, Benson *et al.* 2016) which define communities of *motifs* (Milo *et al.* 2002), not much attention has been paid to investigating higher-order communities within networks.

Graphlets (Fig. 1) are small induced subgraphs and have been used to characterize the higher-order topology of biological networks (Pržulj *et al.* 2004, Hočevar and Demšar 2017, Agrawal *et al.* 2018, Rubel *et al.* 2022). A recent paper defines graphlet-induced networks (Windels *et al.* 2019), in

Received: 23 May 2023; **Revised:** 18 September 2023; **Editorial Decision:** 25 September 2023; **Accepted:** 3 October 2023

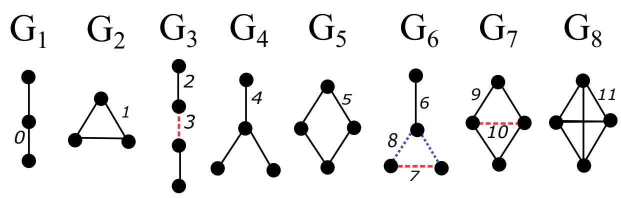

Figure 1. All graphlets of size 3 and 4 (G_1–G_8). The distinct edge positions (0–11) (edge orbits) are shown with a different line style and color.

which, all nodes that comprise a graphlet in an PPI network are connected into a clique, resulting in a network with dense subregions. However, the density of the resulting network can negatively affect the quality of network-based clustering (or *partition*).

In this article, we develop a new graphlet-based community detection method that generalizes the conventional edge-based communities and identifies groups of nodes that are connected through specific graphlets. In contrast to Windels *et al.* (2019), our approach involves transforming a given network to retain edges that participate in a particular graphlet and subsequently applying a random-walk based clustering algorithm to this transformed network which is computationally more advantageous. This way, by restricting the random-walk to edges of interest we can distinguish the parts of the network with a high concentration of a particular graphlet. We show that different graphlets admit quantifiably different clusterings, and comparing these clusters from four different interactomes with expert curated pathway databases, we find that higher-order graphlets detect biologically relevant functional groups that are missed by the edge-based, classic clustering algorithm. Further, using GWAS trait datasets and disease association datasets, we show that specific graphlets admit clusters that are enriched for specific trait and disease-associated genes that edge-based clustering algorithms do not capture. Thus, leveraging the higher-order connectivity of networks in community detection applications can reveal relationships among disease genes that were previously unknown.

2 Methods

2.1 Graphlets

Graphlets are defined as connected, induced, non-isomorphic subgraphs of a specific size (Pržulj *et al.* 2004). Graphlets describe the structure of a network without requiring the specification of a null model and thus differ from motifs (Milo *et al.* 2002). The edges of every graphlet are partitioned into a set of automorphism groups called orbits such that two edges belong to the same orbit if they map to each other in some isomorphic projection of the graphlet onto itself (Hočevar and Demšar 2017) (Fig. 1). There are 30 graphlets up to five nodes that have 67 edge orbits (see Supplementary Fig. S1). Existing software such as ORCA (Hočevar and Demšar 2017) can count, for every edge, the number of edge orbits of each type.

2.2 Graphlet-induced community structure

To identify communities that are enriched for specific graphlets (graphlet-induced clusters or modules), we make use of the Markov Clustering algorithm (MCL) (Van Dongen 2000) with a modified initial transition matrix.

2.2.1 Markov clustering algorithm

Given an adjacency matrix A, a random walk on a network can be defined by a transition matrix P where the probability of transitioning from node i to node j is $p_{ij} = a_{ij} / \sum_j a_{ij}$. To find groups of densely connected nodes (i.e. clusters) in a network, the standard MCL simulates a random walk and successively applies *expansion* and *inflation* operators on the transition matrix P (Van Dongen 2000). The expansion operation spreads random flow while the inflation operation makes strong links stronger and weak links weaker which reduces the flow between clusters. As the algorithm progresses, the network gets divided into disconnected subnetworks. The procedure is repeated until the transition matrix converges, i.e. it does not change with further expansion or inflation operations (Van Dongen 2000). Finally, connected subgraphs that remain after convergence are extracted as clusters. The granularity of clusters can be tuned by varying the inflation parameter. In general, a larger inflation parameter results in more fine-grained clusters.

We used an existing implementation of MCL (Van Dongen 2008). For large networks, this algorithm does not follow the MCL procedure exactly and uses approximations for speed thus we sometimes find isolated nodes in clusters. As a final step, we identify and eliminate any such nodes and retain only the connected component within the cluster.

2.2.2 Modified transition matrix

We modify the transition matrix P to feed graphlet-specific transition matrices into MCL. For a given graphlet G_k, the probability to transition from node i to node j is nonzero only when the edge from i to j participates in graphlet G_k. Let O_k be the set of edge orbits that define graphlet G_k (e.g. $O_k = \{2, 3\}$ for graphlet G_3 in Fig. 1). Given an undirected network G with adjacency matrix A, we first count, for every edge, the number of edge orbits of each type using ORCA (Hočevar and Demšar 2017). We then make a modified adjacency matrix $A^{(k)}$, where

$$a_{ij}^{(k)} = \begin{cases} a_{ij} & \text{if edge } (i,j) \text{ has at least one orbit in } O_k \\ 0 & \text{otherwise.} \end{cases}$$

As a consequence of this edge selection, a subgraph of the original network is retained. Further, the graphlet-specific matrix for G_0 is simply the original adjacency matrix A. The graphlet-specific transition matrix $P^{(k)}$ for graphlet G_k is then

$$p_{ij}^{(k)} = a_{ij}^{(k)} / \sum_j a_{ij}^{(k)}.$$

For nodes i that have no incident edges we set $p_{ii}^{(k)} = 1$, implying that a walker starting at node i stays at node i. The idea is illustrated in Fig. 2 for graphlet G_2 (triangle).

MCL is then used to find communities using the graphlet-specific transition matrices, which finds the communities that are connected through G_k. We only keep clusters of size 3 or larger for further analysis. For the remainder of this article,

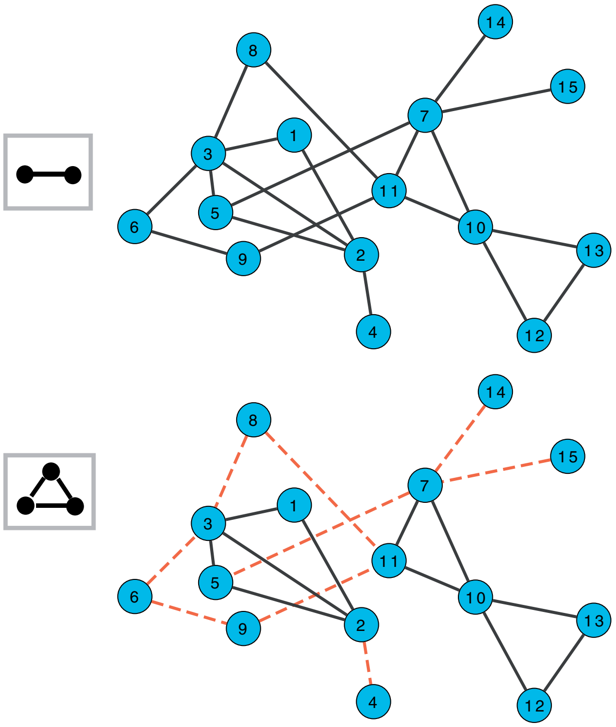

Figure 2. An illustration of graphlet-induced network for G_2. Under the standard MCL, transition all edges are allowed and are shown by black lines (top). For graphlet G_2 (triangle), red dashed edges represent transitions that are no longer allowed for G_2 (bottom).

we will refer to the transformed network obtained by a modified transition network as a *graphlet-induced network*.

2.3 Retaining nonredundant graphlet-induced networks

Our approach returns 30 different clusterings of the same PPI network, one for each of the graphlets up to five nodes (G_0–G_{29}). However, some of graphlet-specific transition matrices $P^{(k)}$ are in fact not very different from the original transition matrix P. As a result, some of the MCL clusterings are similar not because they clustered different networks in the same way, but because they are essentially clustering the same network. We identify and ignore these redundant graphlets that do not alter the network. For each network, we retain graphlets G_k where the graphlet-induced adjacency matrix is $<95\%$ similar, indicating that more than 5% of the edges are dropped because they do not participate in the same graphlet (Supplementary Fig. S4). The numbers of retained graphlets for each network are shown in Table 1.

2.4 Methods for comparison

We note that there are many available clustering algorithms, and many of them can be adapted to use graphlet-induced networks as described above. Running MCL with G_0 is equivalent to the original MCL algorithm, since the G_0-induced subnetwork of G is simply G. Thus, we use the G_0 MCL as a comparison to traditional MCL.

We also compare our results to six additional methods from two bodies of work. In the first method we consider a different way of constructing the graphlet-induced network described in Windels *et al.* (2019), which has similar goals to our work. In contrast to our approach, the approach of Windels *et al.* (2019) defines two nodes to be adjacent in a graphlet-induced network if they share a graphlet regardless of whether or not they are connected in the original network. Thus, the main difference between the two approaches is that Windels *et al.* (2019) transforms the network by turning a graphlet into a fully connected clique which makes the induced networks denser than the original whereas we remove edges in a targeted manner which results in a sparser induced network than the original. The approach of Windels *et al.* (2019) introduces a large number of new edges in sparse networks [that contain large number of sparse graphlets (e.g. G_3, G_9, G_{10}, etc.)] since it adds the missing links in the graphlets under consideration. To evaluate the influence of the graphlet-transformed networks, we run MCL on the transformed networks for both approaches. Due to the density of the transformed network using the graphlet-induced network of Windels *et al.* (2019), we limited our comparison to graphlets up to four nodes.

Next, we compare our approach to the five top-performing DREAM challenge algorithms (Choobdar *et al.* 2019), which include a kernel clustering approach (method K1), a random-walk based method (method R1), a local agglomerative clustering (method L1), and two methods optimizing modularity (methods M1 and M2). When comparing using the DREAM challenge algorithms, we evaluate the communities based on the DREAM Challenge inputs of 180 GWAS datasets. Besides the known limitations of GWAS (Tam *et al.* 2019), we also note that we have not optimized our graphlet-induced MCL to perform well for the DREAM Challenge inputs, but this provides a nice baseline compared to the state-of-the-art.

2.5 Data sources

2.5.1 Interactomes

We applied our method to four interactomes (Table 1): InWeb (Li *et al.* 2017), an interactome from the Stanford Network Analysis Project (SNAP) (Agrawal *et al.* 2018), HuRI (Luck *et al.* 2020), which include aggregated physical PPIs, and a subset of the STRING (Szklarczyk *et al.* 2021) network of edges weighted 0.8 or larger on a scale from 0 to 1, which includes both direct physical interactions as well as functional associations. These interactomes range in size from 63 000 edges to nearly 400 000 edges and contain widely different number of graphlets within them (Supplementary Fig. S2). Given such variability in network structure, a single choice of inflation parameter may not be appropriate for all the interactomes. Thus, we performed a parameter sweep for each interactome by running MCL with varying inflation parameters (between 1.0 and 8.0) and plotted both the number of clusters returned as well as the number of nodes in the largest cluster for each network and inflation parameter (Supplementary Fig. S3). The inflation parameter was chosen such that the size of the largest cluster is of the order of hundreds of nodes (Table 1).

2.5.2 Biological process and disease gene sets

To assess the performance of different graphlet-induced modules, we compare them to pathways (represented as gene sets) from the Human Molecular Signatures Database (MSigDB) (Subramanian *et al.* 2005). Specifically, we considered a collection of 292 pathway gene sets from BioCarta (Nishimura 2001), 186 pathway gene sets from the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto 2000), and 196 pathway gene sets from the Pathway Interaction Database (PID) (Schaefer *et al.* 2009). Curated by domain

Table 1. PPI networks used in this study.[a]

Interactome	Network statistics				Retained	Graphlet-MCL clustering		
	Nodes	Edges	Avg. Deg.	Weighted	Graphlets	Inflation	Largest cluster	No. of clusters
STRING (Szklarczyk *et al.* 2021) (weights > 0.8)	10 375	213 996	41.2	Yes	1,3,4,5,7,9,10,11,12,15, 16,17,19,20,21,22,24,25,27,28,29	8.0	798	917
InWeb (Li *et al.* 2017)	12 420	397 309	64.0	Yes	5,8,20,22,24,26,27,28,29	3.0	99	1354
SNAP (Agrawal *et al.* 2018)	21 557	338 636	31.4	No	2,7,8,18,22,24,26,27,28,29	3.0	572	1735
HuRI (Luck *et al.* 2020)	9094	63 242	14.0	No	2,5,7,8,18,20,21,22,23,24,25, 26,27,28,29	3.0	131	1013

[a] Retained graphlets are those whose graphlet-induced transition matrices are sufficiently different from the original (G_0) network. For each network, the selected inflation parameter, the size of the largest cluster, and the total number of clusters that contain at least three nodes is shown (see Supplementary Table S1 for runtime of the algorithm).

experts, these gene sets are canonical representations of a biological process.

We also compared the graphlet-induced modules with disease gene sets. We used 519 annotated disease gene sets (Agrawal *et al.* 2018) from DisGeNET (Piñero et al. 2015) which integrates expert-curated databases that cover information on Mendelian and complex diseases and 34 gene-level cancer datasets from The Cancer Genome Atlas (TCGA) mutations, curated by OncoVar (Wang *et al.* 2021).

2.5.3 GWAS trait datasets

We also evaluated disease–gene associations using GWAS data, which offer a complementary perspective to the disease gene sets. We used a collection of 180 GWASs Datasets of disease-related human phenotypes from the DREAM challenge (Choobdar *et al.* 2019), which cover a wide range of molecular processes.

2.6 Module assessment

2.6.1 Evaluating clustering similarity

We first evaluate the similarity of MCL clusterings from the same PPI network using different graphlet-induced networks. To compare clusterings from different graphlet-induced MCL runs, as well as compare graphlet-induced MCL to the other approaches, we use the Adjusted Rand index (ARI) which controls for cluster matching due to random chance. In computing the ARI, we only include the clusters that have at least three nodes.

2.6.2 Hypergeometric P-value-based enrichment

To evaluate the enrichment of the pathway and disease gene sets, we use measures based on the hypergeometric *P*-value. For every module/gene set pair, we calculate the hypergeometric *P*-value adjusted by the Benjamini–Hochberg method (Benjamini and Hochberg 1995) for multiple hypothesis testing. We then calculate the following measures using a *P*-value cutoff of < 0.05:

- The *gene set coverage* is the number of gene sets for which a significant module was found. We also calculate the *gene set percentage* as the fraction of genes (out of all genes in the gene sets) in the significant gene sets.
- The *module coverage* is the number of modules that are significantly enriched for some gene set. We also calculate the *module percentage* as the fraction of genes in the significantly enriched modules (out of genes in all modules within the specified size).

2.6.3 DREAM challenge enrichment

To evaluate the enrichment of GWASs datasets, we use the DREAM Challenge's framework of using Pascal (Lamparter *et al.* 2016). Pascal first obtains gene scores by aggregating single nucleotide polymorphism *P*-values from GWAS, while correcting for linkage disequilibrium structure. It then combines the scores of genes that belong to the same pathways to obtain pathway scores (*P*-values) using the chi-squared method described in Lamparter *et al.* (2016). These pathway level *P*-values are further adjusted for multiple testing using the Benjamini–Hochberg correction (Choobdar *et al.* 2019). Finally, these adjusted *P*-values are used to determine the number of significant modules for a given method. Note that the DREAM Challenge's criteria of number of significant modules is equivalent to our measure of *module coverage* described above. In the DREAM challenge methods are ranked according to the number of significant modules that they discover. Since each significant module can be associated with more than one GWAS, we also report the number of significant GWAS datasets (*trait coverage*, similar to gene set coverage in the previous section).

3 Results

3.1 Graphlets admit different clusters

We first quantify the similarity in the community structure found by clustering different nonredundant graphlet-induced networks with MCL. The variability in the ARI scores between different MCL clusterings for each of the four PPI networks shows that a number of these clusterings are substantially different from the original G_0-based clustering and thus contain distinct information within them [Fig. 3 (see Supplementary Fig. S5 for all 30 graphlets)]. Specifically, about 57% pairs of community structures in STRING, 44% in InWeb, 26% in SNAP, and 58% in HuRI have an ARI score of <0.8. For each nonredundant graphlet, we also characterize the topological structure within clusters in terms of *network transitivity*. Generally, clusters obtained by denser graphlets exhibit higher transitivity (Supplementary Fig. S6).

3.2 Pathway enrichment of clusters

We next assess the biological relevance of clusters obtained by higher-order graphlets. Since we expect genes within pathways to be near each other in PPI networks, we focus on the gene set percentage of the three pathway databases (BioCarta, KEGG, and PID).

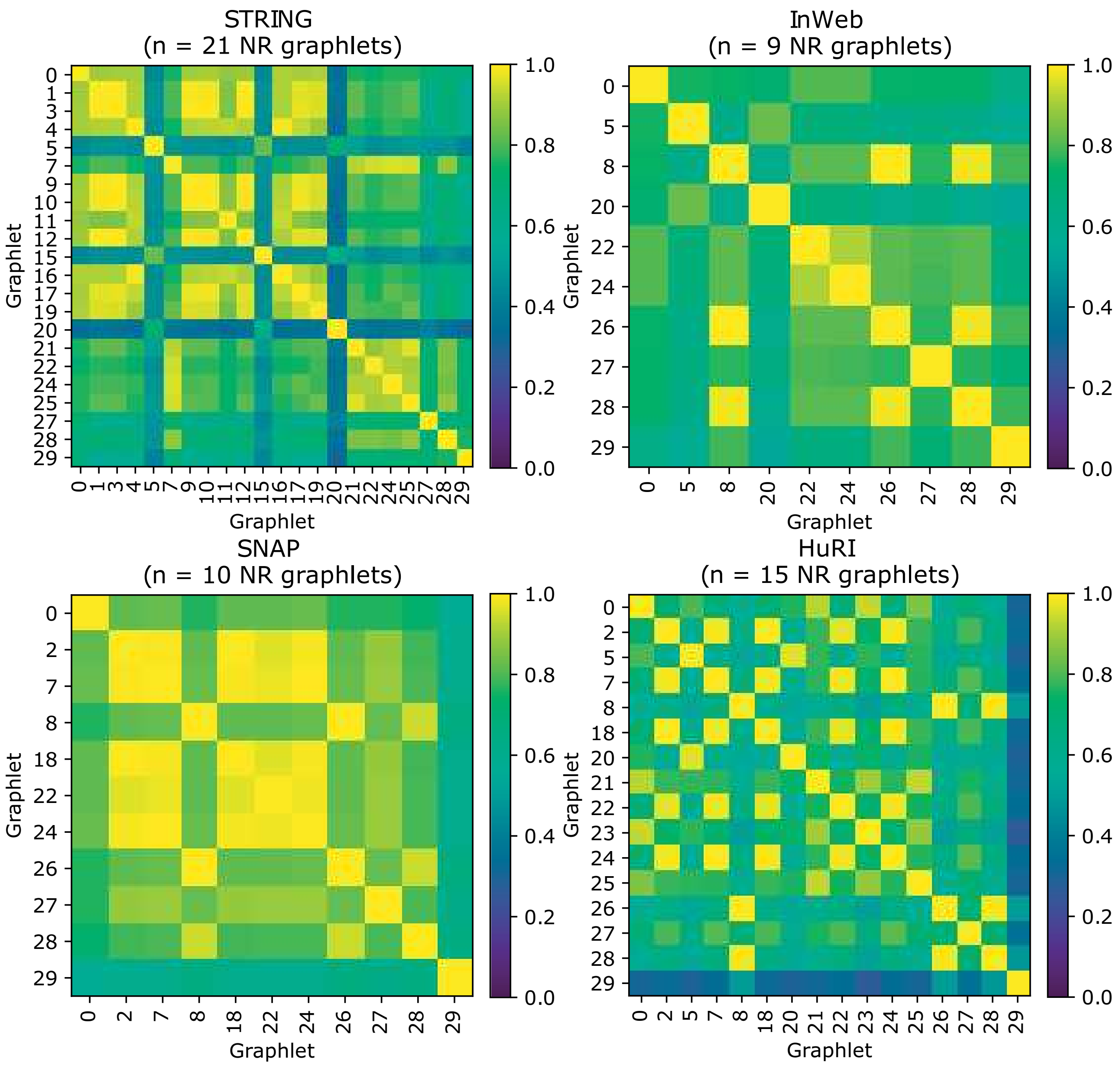

Figure 3. ARI scores for nonredundant (NR) graphlet-induced clustering of the interactomes. A higher ARI indicates a higher level of similarity between the two clusterings.

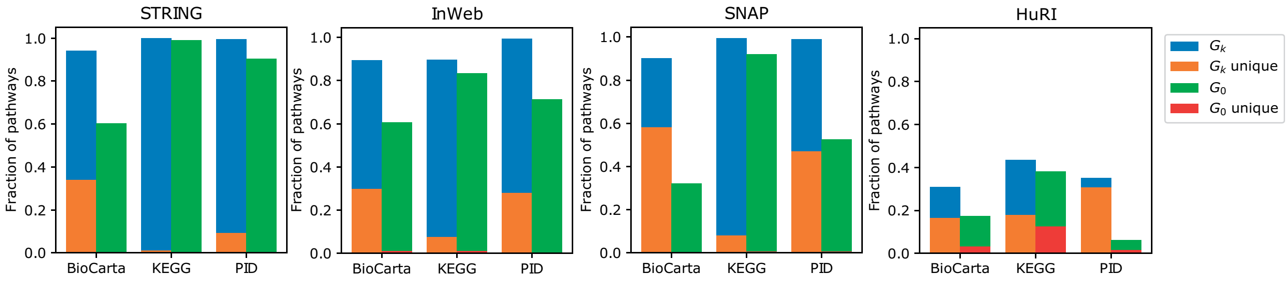

Figure 4. Fraction of pathways from different databases that are significantly associated with at least one module discovered by higher-order clustering. It shows the fractional coverage obtained by G_0 (green) and other graphlets G_k (blue) for each interactome. The height of red and orange bars shows what fraction of these pathways are unique to the two sets G_0 and G_k, respectively. We note that a pathway can be associated with multiple modules.

In all combinations of PPI network and pathway database, the gene set percentage is larger for higher-order graphlet-induced clusterings ($G_k = G_1 - G_{29}$) than clustering based on G_0 (Fig. 4). This is even true for the HuRI PPI network, which also identifies a good number of pathway-enriched modules using G_0 that are not found in the higher-order clusterings (red bars). Importantly, this analysis shows that higher-order graphlets can find unique pathway associations (orange bars) that are not detected by G_0. In general, a module can be associated with more than one pathway and not all modules find a significant association (see Supplementary Fig. S7 for fraction of enriched clusters).

3.3 Disease gene enrichment

We then moved on to assessing the graphlet-enriched clusters with respect to identifying disease modules. We do not expect genes from each disease to be near each other in the PPI networks, especially for complex diseases—thus, we considered the total number of diseases that are enriched (gene set coverage) rather than the percentage of diseases found for each dataset. We find the number of significant disease modules for G_0 and higher-order G_k-based clusterings (Fig. 5). Each of these sets of clusterings finds unique disease associations and with the exception of HuRI, the number of unique associations in higher-order clusterings is consistently higher across all interactomes and disease databases. These results indicate that modules found by other graphlet-based methods can provide a large number of new disease associations that the traditional approach does not find.

To show that these modules can provide functional predictions for un-annotated genes, we examined four modules that are enriched in distinct diseases from DisGeNET in the SNAP PPI network (Fig. 6). Notably, none of these diseases are revealed in the G_0-based clustering. Even the best corresponding modules in the G_0-based clustering have adjusted P-values that do not cross the significance threshold (Supplementary Table S2). These networks show the edges that participate in the corresponding graphlet in the SNAP PPI network.

Figure 6a is associated with thrombosis—a condition characterized by formation of clots inside blood vessels. The genes colored blue are already associated with thrombosis according to DisGeNet. Five of the gray genes (*GGCX, F2RL3, F11, F7*, and *GP5*) are known to play a role in blood coagulation (Megy *et al.* 2019). However, the gene *SERPINB6* does not have a known link to thrombosis and it may be a promising candidate for further investigation. Similarly, the association of the blue nodes with Chronic Myeloid Leukemia (Fig. 6b) is already known (Sheng-Fung *et al.* 2004, Hanoun *et al.* 2012), while the gray genes are potentially new. In Fig. 6c, genes *C5* and *CPN1* are our predicted associations for age related macular degeneration besides the known involvement of the blue genes (Lu *et al.* 2018). Figure 6d shows the disease module associated with Glioblastoma. Deregulation of NOTCH receptors and their ligands (nodes *NOTCH1, NOTCH2, NOTCH3, JAG1, JAG 2, DLL1, DLL3*, and *DLL4*) are known to play a role in Glioblastoma (Fiaschetti *et al.* 2014); but the role of *MFNG* is not established.

We also compared the graphlet-induced clusters from the SNAP PPI network to those using the approach of Windels *et al.* (2019) for graphlets up to four nodes. For the disease gene sets from DisGeNET (Piñero et al. 2015), we find that our approach results in a larger number of enriched disease modules for each graphlet $G_1 - G_7$ (Supplementary Fig. S8). Running the clustering algorithm was prohibitively slow for the other graphlet-induced networks constructed according to Windels *et al.* (2019), limiting our ability to compare across all four networks.

3.4 GWAS enrichment

We evaluate the performance of our method to find disease/trait associated modules in the InWeb interactome using 180 GWAS datasets from the DREAM Challenge (Choobdar *et al.* 2019). We chose to use InWeb for this analysis since it was also used in the DREAM Challenge. InWeb is also sparse enough and can be efficiently analyzed without the need to discard the low weight edges for speed as we did in STRING (Table 1).

Our method is able to capture a number of trait-associated modules in InWeb (Fig. 7). While none of the graphlet-based methods outperform K1, M1, and M2, five of these methods are at least as good as L2 and R1 (Fig. 7, top). We observe a similar trend when comparing the number of GWAS associations discovered by each of these methods. DREAM challenge methods K1 and M2 still find more GWAS associations, which is not necessarily surprising since the DREAM challenge methods are customized to perform well on the challenge datasets. Nevertheless, all graphlet-based methods are at least as good as the L2 (showing lowest performance of the top five methods) (Fig. 7, bottom). Interestingly, the modules found by each higher-order graphlet reveal 17 significant associations that are not found by G_0 (shown in green in Fig. 7). In addition, we also find that there are five unique associations that higher-order graphlets identify but are not found by G_0 or any of the five top-performing DREAM

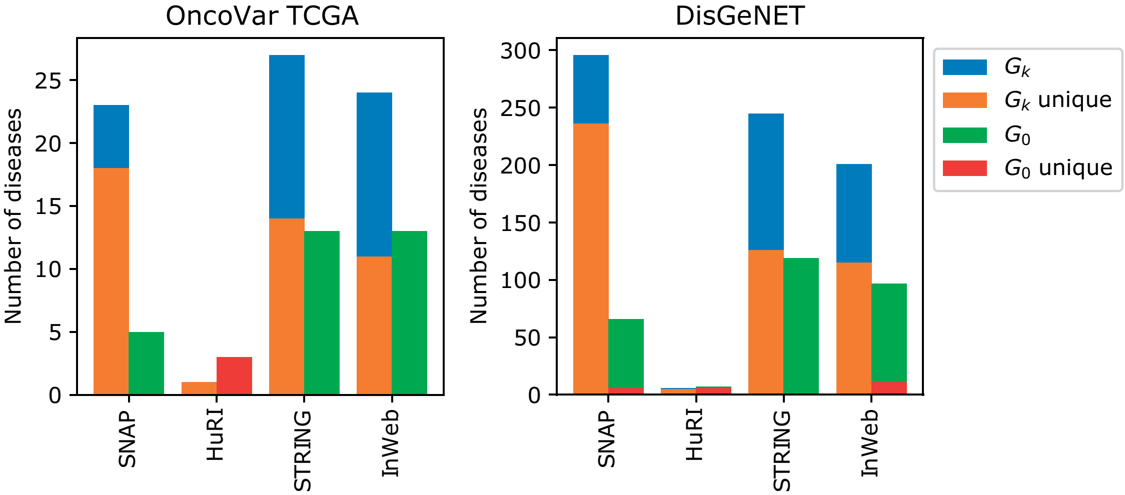

Figure 5. Number of significantly associated unique diseases from different disease association databases discovered by higher-order clustering. Like in Fig. 4, although each disease can be associated with more than one module, it is counted only once.

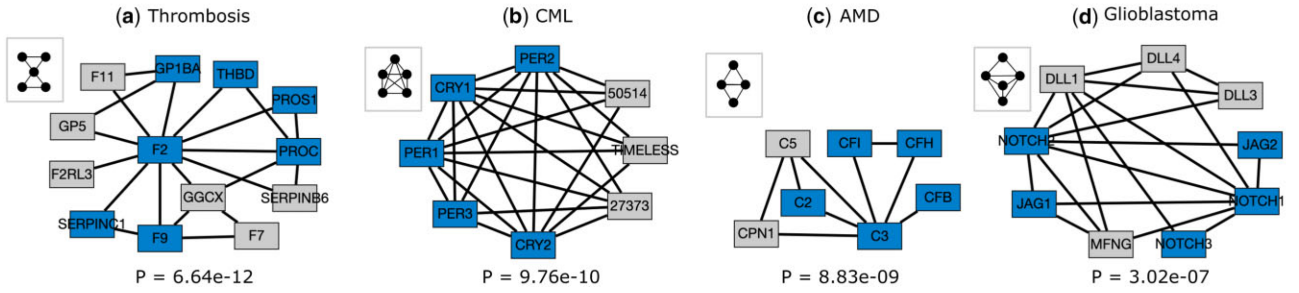

Figure 6. Disease modules discovered by graphlet-aware community detection using specific graphlets for Thrombosis (a, graphlet G_{18}), Chronic Myeloid Leukemia (b, graphlet G_{29}), Age related macular degeneration (c, graphlet G_7), and Glioblastoma (d, graphlet G_{26}). Blue nodes indicate genes present in the DisGeNet disease set and gray nodes are not annotated to the disease. The hypergeometric P-value is indicated below each module.

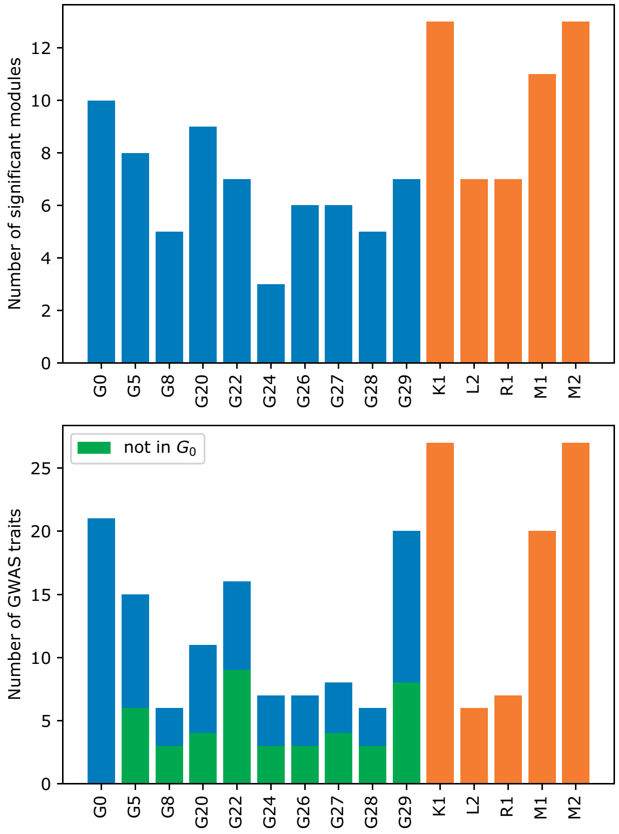

Figure 7. Number of modules significantly associated a GWAS trait (top) and the number of significantly associated GWAS traits found (bottom) in the InWeb interactome using different (nonredundant) graphlet-based clustering. Results obtained by different (nonredundant) graphlet-based clustering are shown in blue and top five methods from DREAM challenge submission are shown in orange whereas green indicates the set of unique associations identified by higher-order graphlets that are not found by any of the G_0-based clusters.

Table 2. GWAS traits identified by different higher-order graphlet-based clusterings that are not identified either by our G_0-based method or any of the top five DREAM challenge methods.

GWAS trait	Graphlet MCL	Module Size	Adj. P-value
Coronary artery disease (Nikpay *et al.* 2015)	22	3	9.78e−6
Body mass index (Horikoshi *et al.* 2015)	29	13	1.01e−4
Type 2 diabetes (Morris *et al.* 2012)	27	6	7.86e−5
Overweight (Berndt *et al.* 2013)	24	5	4.02e−5
Alzheimer's disease (Lambert *et al.* 2013)	5	3	9.41e−6

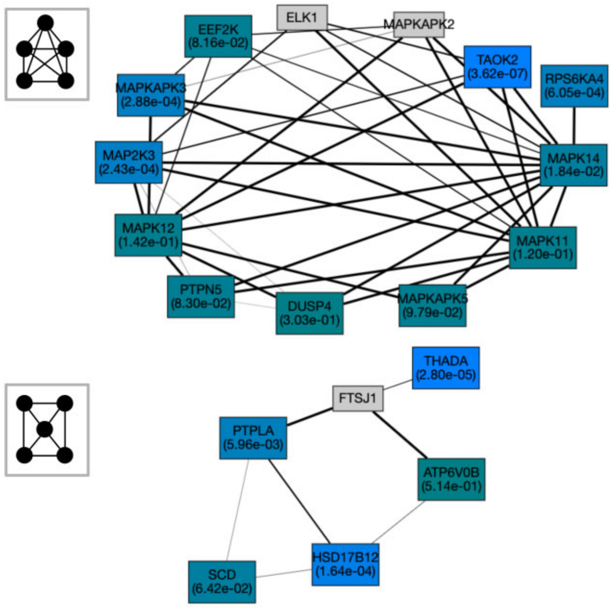

Figure 8. Modules detected by G_{29}-based and G_{27}-based community detection which are associated with the traits BMI (top) and type 2 diabetes (bottom) with module P-values 1.01e−4 and 7.86e−5, respectively, computed by Pascal (Lamparter *et al.* 2016) (see Supplementary Table S3). The gene P-values in each module are indicated by different colors. The lighter shades represent smaller gene P-values.

methods (Table 2). These five datasets represent four distinct disease classes, including anthropometric, cardiovascular, glycemic, and neurodegenerative diseases (Supplementary Table S3).

Two notable examples of a module uniquely identified by graphlet-based method are shown in Fig. 8 (the other three modules are visualized in Supplementary Fig. S9). First, the module of size 13 associated with body mass index (BMI) contains genes with statistically significant Pascal gene scores (adjusted $P < .05$). The involvement of genes in obesity/BMI is supported by other studies. A recent study links variants in *TAOK2* to human obesity (Agrawal *et al.* 2021). Genes *MAP2K3* and *MAPKAPK3* are found to play a role in BMI (Bian *et al.* 2013, Shao *et al.* 2022). Other genes in MAPK signaling that are not significant according to the gene scores (*RPS6KA4, DUSP4, MAPKAPK5*) have also been associated with obesity (Ow and Kuznetsov 2015). *EEF2K* is also predicted to be a novel target for obesity (Joshi *et al.* 2021). However, this module also contains genes (e.g. *ELK1, MAPKAPK2*) with no gene score that could potentially be associated with BMI.

The module related to type 2 diabetes contains three genes with significant *P*-values (*THADA, PTPLA, HSD17B12*) and three with $P < .05$ (Fig. 8). The genes *THADA* and *HSD17B12* have previously been implicated in type 2 diabetes (Zeggini *et al.* 2008, Hachim *et al.* 2020). The other genes in the module are potentially new and could be good candidates for further investigations. This includes the genes *PTPLA* with a statistically significant gene score and *FTSJ* with no assigned gene score.

4 Discussion

Well-established community detection methods for disease module prediction often neglect the higher-order connectivity patterns among genes or proteins of interest and focus only on their pairwise relationships, even though, studies have demonstrated that higher-order structure is biologically relevant. In this article, we have presented a generalized community detection method that incorporates higher-order structures in the form of graphlets. While one can focus on partition with respect to a particular graphlet that may be relevant for a given problem, by providing an ensemble of partitions, each corresponding to a different graphlet (including G_0), our approach provides a comprehensive view of network communities. Each of the nonredundant graphlet-based clusterings offers a unique perspective and thus, compliments the traditional (G_0) clustering method. Using a collection of diverse expert-curated association datasets (pathways, diseases, and GWAS), we further demonstrated usefulness of our approach in identifying disease modules in four different interactomes. Our analysis shows strong evidence that the higher-order graphlet-based clusters can reveal unique biological associations that traditional methods cannot. We note that for both pathway and disease gene enrichment, the clustering methods do not perform well on HuRI. This can potentially be attributed to the fact that the HuRI network is much sparser and has a very different local structure compared to the other three interactomes as it uses yeast two-hybrid (Y2H) screens to detect pairwise protein interactions and is likely free from ascertainment bias. Another reason is that HuRI contains fewer nodes and fewer annotated genes for enrichment which affects the statistical significance when performing the hypergeometric test. For pathway enrichment, we could only map about 49% of BioCarta, 44% of KEGG, and 50% of PID genes onto HuRI. These fractions are roughly (99%, 91%, 99%) for SNAP, (91%, 85%, 91%) for STRING, and (89%, 74%, 88%) for InWeb, respectively.

A limitation of our study is that although it can find potentially novel associations using an ensemble of graphlet-based clusters, it does not determine the role or provide an interpretation of a graphlet structure in a specific biological context. While some network motifs (e.g. feed forward loop) and their functions are well-studied, and some others such as paths of length 3, are found to be important in the context of PPI networks (Kovács *et al.* 2019), a clear biological interpretation of a general graphlet structure is still lacking and presents a promising direction for future research. Sometimes dense communities are functionally relevant. For this purpose, community detection with respect to a dense graphlet like G_{29} can provide useful insights. In many of our benchmarks, we observe that G_{29}-based clustering finds more significant associations (Figs 6–8). One way to systematically assess the importance of graphlets in a given interactome can be to look at their over and under-representation in it (Rubel *et al.* 2022).

Our graphlet-based clustering framework uses undirected graphlets and thus works with undirected networks. However, many gene/protein interactions are inherently directed. If we ignore the edge directions, we can still apply our framework to find relevant modules in these networks. However, by neglecting the directionality of edges, we may be missing important context about these interactions. An interesting future research direction will be extend this framework to directed networks by incorporating directed graphlets (Sarajlić *et al.* 2016, Trpevski *et al.* 2016) into our module detection approach.

Complex diseases likely have multiple factors in play, and while the same disease might appear in multiple modules, in some cases, it is possible that one graphlet is insufficient to capture the heterogeneity of disease genes. Prior work has suggested that multiple graphlets are over-represented in the same disease module or pathway (Agrawal *et al.* 2018, Rubel *et al.* 2022). Thus, extending our framework to identify modules with respect to a combination of different graphlets, or identifying and taking a weighted consensus across partitions found by most relevant graphlets may reveal more accurate disease associations.

Finally, it will be interesting to investigate higher-order organization in problems beyond network clustering, for example, pathway reconstruction where the goal is to find a subnetwork that connects genes of interest (Ritz *et al.* 2016) or the problem of detecting active modules given a set of active/seed genes (Levi *et al.* 2021). Even though these networks contain higher-order structure within them (Rubel *et al.* 2022), to our knowledge no method explicitly focuses on optimizing the higher-order structure in these networks. Thus, it may be worthwhile to develop new methods that aim to identify subnetworks that exhibit a desired graphlet profile.

Acknowledgements

We would like to thank Anthony Gitter, Lixing Yi, Max Bennett, Henry Jacques, Larry Zeng, Nina Young, and Alex Richter for fruitful discussions.

Supplementary data

Supplementary data are available at *Bioinformatics Advances* online.

Conflict of interest

None declared.

Funding

This work was supported by the National Science Foundation [NSF-DBI-1750981 to A.R.].

Data availability

The code and sources of datasets used in this study are available at: https://github.com/Reed-CompBio/graphlet-clustering.

References

Agrawal M, Zitnik M, Leskovec J. Large-scale analysis of disease pathways in the human interactome. In: *Pacific Symposium on Biocomputing 2018: Proceedings of the Pacific Symposium*, January 3–7, 2018, Big Island of Hawaii, USA. World Scientific, 2018, 111–22.

Agrawal N, Lawler K, Davidson CM *et al.*; INTERVAL. Predicting novel candidate human obesity genes and their site of action by systematic functional screening in drosophila. *PLoS Biol* 2021;**19**: e3001255.

Arenas A, Fernández A, Fortunato S *et al.* Motif-based communities in complex networks. *J Phys A Math Theor* 2008;**41**:224001.

Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B (Methodological)* 1995;**57**:289–300.

Benson AR, Gleich DF, Leskovec J *et al.* Higher-order organization of complex networks. *Science* 2016;**353**:163–6.

Berndt SI, Gustafsson S, Mägi R *et al.* Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. *Nat Genet* 2013;**45**:501–12.

Bian L, Traurig M, Hanson RL *et al.* MAP2K3 is associated with body mass index in American Indians and Caucasians and may mediate hypothalamic inflammation. *Hum Mol Genet* 2013;**22**:4438–49.

Choobdar S, Ahsen ME, Crawford J *et al.*; DREAM Module Identification Challenge Consortium. Assessment of network module identification across complex diseases. *Nat Methods* 2019;**16**:843–52.

Fiaschetti G, Schroeder C, Castelletti D *et al.* Notch ligands JAG1 and JAG2 as critical pro-survival factors in childhood medulloblastoma. *Acta Neuropathol Commun* 2014;**2**:39.

Fortunato S. Community detection in graphs. *Phys Rep* 2010;**486**: 75–174.

Ghiassian SD, Menche J, Barabási A-L *et al.* A disease module detection (diamond) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. *PLoS Comput Biol* 2015;**11**:e1004120.

Hachim MY, Aljaibeji H, Hamoudi RA *et al.* An integrative phenotype–genotype approach using phenotypic characteristics from the UAE national diabetes study identifies HSD17B12 as a candidate gene for obesity and type 2 diabetes. *Genes (Basel)* 2020;**11**:461.

Hanoun M, Eisele L, Suzuki M *et al.* Epigenetic silencing of the circadian clock gene CRY1 is associated with an indolent clinical course in chronic lymphocytic leukemia. *PLoS One* 2012;**7**:e34347.

Hočevar T, Demšar J. Combinatorial algorithm for counting small induced graphs and orbits. *PLoS One* 2017;**12**:e0171428.

Horikoshi M, Mägi R, van de Bunt M *et al.*; ENGAGE Consortium. Discovery and fine-mapping of glycaemic and obesity-related trait loci using high-density imputation. *PLoS Genet* 2015;**11**:e1005230.

Joshi H, Vastrad B, Joshi N *et al.* Identification of key pathways and genes in obesity using bioinformatics analysis and molecular docking studies. *Front Endocrinol (Lausanne)* 2021;**12**:628907.

Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 2000;**28**:27–30.

Kovács IA, Luck K, Spirohn K *et al.* Network-based prediction of protein interactions. *Nat Commun* 2019;**10**:1240.

Lambert JC, Ibrahim-Verbaas CA, Harold D *et al.*; Cohorts for Heart and Aging Research in Genomic Epidemiology. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat Genet* 2013;**45**:1452–8.

Lamparter D, Marbach D, Rueedi R *et al.* Fast and rigorous computation of gene and pathway scores from SNP-based summary statistics. *PLoS Comput Biol* 2016;**12**:e1004714.

Levi H, Elkon R, Shamir R *et al.* Domino: a network-based active module identification algorithm with reduced rate of false calls. *Mol Syst Biol* 2021;**17**:e9593.

Li T, Wernersson R, Hansen RB *et al.* A scored human protein–protein interaction network to catalyze genomic interpretation. *Nat Methods* 2017;**14**:61–4.

Lu F, Liu S, Hao Q *et al.* Association between complement factor C2/C3/CFB/CFH polymorphisms and age-related macular degeneration: a meta-analysis. *Genet Test Mol Biomarkers* 2018;**22**:526–40.

Luck K, Kim D-K, Lambourne L *et al.* A reference map of the human binary protein interactome. *Nature* 2020;**580**:402–8.

Megy K, Downes K, Simeoni I *et al.*; Subcommittee on Genomics in Thrombosis and Hemostasis. Curated disease-causing genes for bleeding, thrombotic, and platelet disorders: communication from the ssc of the isth. *J Thrombosis Haemostasis* 2019;**17**:1253–60.

Milo R, Shen-Orr S, Itzkovitz S *et al.* Network motifs: simple building blocks of complex networks. *Science* 2002;**298**:824–7.

Morris AP, Voight BF, Teslovich TM *et al.*; DIAbetes Genetics Replication and Meta-analysis (DIAGRAM) Consortium. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat Genet* 2012;**44**:981–90.

Nikpay M, Goel A, Won H-H *et al.* A comprehensive 1000 genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat Genet* 2015;**47**:1121–30.

Nishimura D. Biocarta. *Biotech Softw Internet Rep Comput Softw J Sci* 2001;**2**:117–20.

Oti M, Snel B, Huynen MA *et al.* Predicting disease genes using protein–protein interactions. *J Med Genet* 2006;**43**:691–8.

Ow GS, Kuznetsov VA. Multiple signatures of a disease in potential biomarker space: getting the signatures consensus and identification of novel biomarkers. *BMC Genomics* 2015;**16**:S2–14.

Piñero J, Queralt-Rosinach N, Bravo A *et al.* Disgenet: a discovery platform for the dynamical exploration of human diseases and their genes. *Database* 2015;**2015**:bav028.

Pržulj N, Corneil DG, Jurisica I *et al.* Modeling interactome: scale-free or geometric? *Bioinformatics* 2004;**20**:3508–15.

Ritz A, Poirel CL, Tegge AN *et al.* Pathways on demand: automated reconstruction of human signaling networks. *NPJ Syst Biol Appl* 2016;**2**:16002–9.

Rubel T, Singh P, Ritz A. Reconciling signaling pathway databases with network topologies. In: *Pacific Symposium on Biocomputing 2022*, January 3–7, 2022, Big Island of Hawaii, USA. World Scientific, 2022, 211–22.

Sarajlić A, Malod-Dognin N, Yaveroğlu ÖN *et al.* Graphlet-based characterization of directed networks. *Sci Rep* 2016;**6**:35098–14.

Schaefer CF, Anthony K, Krupa S *et al.* Pid: the pathway interaction database. *Nucleic Acids Res* 2009;**37**:D674–9.

Shao Y, Tian J, Yang Y *et al.* Identification of key genes and pathways revealing the Central regulatory mechanism of brain-derived glucagon-like peptide-1 on obesity using bioinformatics analysis. *Front Neurosci* 2022;**16**:931161.

Sheng-Fung L, Yang M-Y, Chang J-G *et al.* Downregulation of circadian genes, PER1, PER2, and PER3, in chronic myeloid leukemia. *Blood* 2004;**104**:4317.

Subramanian A, Tamayo P, Mootha VK *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 2005;**102**:15545–50.

Szklarczyk D, Gable AL, Nastou KC *et al.* The string database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res* 2021;**49**:D605–12.

Tam V, Patel N, Turcotte M *et al.* Benefits and limitations of genome-wide association studies. *Nat Rev Genet* 2019;**20**:467–84.

Trpevski I, Dimitrova T, Boshkovski T *et al.* Graphlet characteristics in directed networks. *Sci Rep* 2016;**6**:37057–8.

Van Dongen S. Graph clustering via a discrete uncoupling process. *SIAM J Matrix Anal Appl* 2008;**30**:121–41.

Van Dongen SM. Graph clustering by flow simulation. Ph.D. Thesis, University Utrecht, 2000.

Wang T, Ruan S, Zhao X *et al.* OncoVar: an integrated database and analysis platform for oncogenic driver variants in cancers. *Nucleic Acids Res* 2021;**49**:D1289–301.

Windels SFL, Malod-Dognin N, Pržulj N *et al.* Graphlet laplacians for topology-function and topology-disease relationships. *Bioinformatics* 2019;**35**:5226–34.

Zeggini E, Scott LJ, Saxena R *et al.*; Wellcome Trust Case Control Consortium. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet* 2008;**40**:638–45.