4-19-2015

Cross-disciplinary detection and analysis of network motifs

Ngoc Tam L. Tran
University of Connecticut

Luke DeLuccia
Hobart and William Smith Colleges

Aidan F. McDonald
University of Puget Sound

Chun Hsi Huang
University of Connecticut

Cross-Disciplinary Detection and Analysis of Network Motifs

Libertas Academica
FREEDOM TO RESEARCH

Ngoc Tam L. Tran[1], Luke DeLuccia[2], Aidan F. McDonald[3] and Chun-Hsi Huang[1]

[1]Department of Computer Science and Engineering, University of Connecticut, Storrs, CT, USA. [2]Department of Computer Science, Hobart and William Smith Colleges, Geneva, NY, USA. [3]Department of Computer Science, University of Puget Sound, Tacoma, WA, USA.

ABSTRACT: The detection of network motifs has recently become an important part of network analysis across all disciplines. In this work, we detected and analyzed network motifs from undirected and directed networks of several different disciplines, including biological network, social network, ecological network, as well as other networks such as airlines, power grid, and co-purchase of political books networks. Our analysis revealed that undirected networks are similar at the basic three and four nodes, while the analysis of directed networks revealed the distinction between networks of different disciplines. The study showed that larger motifs contained the three-node motif as a subgraph. Topological analysis revealed that similar networks have similar small motifs, but as the motif size increases, differences arise. Pearson correlation coefficient showed strong positive relationship between some undirected networks but inverse relationship between some directed networks. The study suggests that the three-node motif is a building block of larger motifs. It also suggests that undirected networks share similar low-level structures. Moreover, similar networks share similar small motifs, but larger motifs define the unique structure of individuals. Pearson correlation coefficient suggests that protein structure networks, dolphin social network, and co-authorships in network science belong to a superfamily. In addition, yeast protein–protein interaction network, primary school contact network, Zachary's karate club network, and co-purchase of political books network can be classified into a superfamily.

KEYWORDS: network motifs, biological networks, social networks, directed network, undirected network, network motif detection

CITATION: Tran et al. Cross-Disciplinary Detection and Analysis of Network Motifs. *Bioinformatics and Biology Insights* 2015:9 49–60 doi: 10.4137/BBI.S23619.

RECEIVED: January 07, 2015. **RESUBMITTED:** March 05, 2015. **ACCEPTED FOR PUBLICATION:** March 05, 2015.

ACADEMIC EDITOR: Thomas Dandekar, Associate Editor

TYPE: Original Research

FUNDING: This work was supported in part by the National Science Foundation (NSF) (OCI-1156837 to LD, AFM, C-HH), and U.S. Department of Education Graduate Fellowships in Areas of National Need (GAANNs) (P200A130153 to NTLT). The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: ntt10001@engr.uconn.edu

Introduction

Network motif. Network motifs are defined as statistically significant, over-represented subgraphs contained in the larger superstructure of a network.[1] This is based on the idea that randomized networks are not expected to express these motifs beyond fluctuations.[2] Network motifs are sometimes referred to as the building blocks of complex networks.[3] This is because these small building blocks fit together in a specific way to give a network its properties. As networks develop and evolve, the repetition of particular motifs has been thought to be a result of positive selection for these interaction patterns due to their functional or structural properties.[4] One of the main goals of researching network motifs is to gain insight into how the aggregate of small group interactions forms the macroscopic behavior we see in complex networks. Network motifs have several applications. They can be used to categorize networks into superfamilies[5] or to identify application protocols.[6] Network motifs have also been used in the character overlay graph for building evolutionary trees using parsimony methods.[7] Further, network motifs provide the key to better understand the functional roles of some genes in gene regulation.[8]

Networks and disciplines. Complex networks are a convenient method of representing real-life phenomena through nodes and connecting edges. Creating a network from the source often simplifies the original properties. However, relevant and significant results can still be obtained. Broadly, we chose to analyze network motifs from undirected and directed networks of several different disciplines, including biological network, social network, ecological network, as well as other networks such as airlines, power grid, and co-purchase of political books networks. Table 1 contains all networks studied and grouped by different disciplines.

In biological discipline, the *Caenorhabditis elegans* neural network has nodes representing neurons, and two neurons are connected with an edge if there exists at least one synapse or gap junction between them.[9] In the yeast protein–protein interaction network, every node represents a specific protein in the yeast, and an edge joins two nodes if the proteins interact in some way.[10] Nodes in *Escherichia coli* transcription network represent operons and edges, which are directed from an operon that encodes a transcription factor to an operon that it regulates.[11] In diseasome network, nodes represent specific

Table 1. Network datasets from online data sources. Networks are listed by discipline.

NETWORK DATASET	FORMAT	DISCIPLINE	NETWORK TYPE	DATA SOURCE	REFERENCE
C. elegans neural network	GEXF	Biological	Directed	Gephi Wiki Datasets	9
Yeast	GEXF	Biological	Undirected	Gephi Wiki Datasets	10
E. coli transcription network	Text	Biological	Directed	Uri Alon's Complex Networks	11
Diseasome	GEXF	Biological	Undirected	Gephi Wiki Datasets	12
Protein structure 1	Text	Biological	Undirected	Uri Alon's Complex Networks	5
Protein structure 2	Text	Biological	Undirected	Uri Alon's Complex Networks	5
Protein structure 3	Text	Biological	Undirected	Uri Alon's Complex Networks	5
Cypress dry season	Text	Ecological	Directed	Pajek datasets	13
Everglades graminoids wet season	Text	Ecological	Directed	Pajek datasets	13
Dolphin social network	GML	Social	Undirected	University of Michigan Network Data	14
Primary school contact network	GEXF	Social	Undirected	http://www.plosone.org/article/fetchSingleRepresentation.action?uri=info:doi/10.1371/journal.pone.0023176.s003	15
Co-authorships in network science	GML	Social	Undirected	Gephi Wiki Datasets	16
Zachary's karate club	GML	Social	Undirected	Gephi Wiki Datasets	17
Unknown airlines	GRAPHML	Other	Directed	Gephi Wiki Datasets	18
US air 97	NET	Other	Directed	Gephi Wiki Datasets	18
Power grid	GML	Other	Undirected	University of Michigan Network Data	9
Co-purchase of political books	GML	Other	Undirected	University of Michigan Network Data	19

diseases and edge connecting nodes if they share at least one gene.[12] Finally, all three protein structure networks have nodes representing α or β helices, and they are connected if the helices are within 10 Å of each other.[5] Diseasome, protein structures, and yeast are undirected networks, while *C. elegans* and *E. coli* are directed networks.

The ecological discipline has two food web datasets: Cypress Dry Season and Everglades Graminoids Wet Season.[13] They are network analyses of the trophic dynamics in South Florida ecosystems. In these networks, nodes represent the major components of the ecosystem, and edge represents the transfer of material or energy among the major components.[13]

The social discipline consists of four undirected networks. The dolphin social network has nodes representing individual dolphins in the community, and edge connecting two nodes indicates that two individual dolphins have direct contact with each other.[14] In the primary school contact network, nodes represent teachers, parents, or students, and edge represents face-to-face interaction between two individuals.[15] Nodes in the co-authorships network are researchers, and edge connecting two nodes implies that two researchers have co-authored an article in the field of network science.[16] The final social network depicts a friendship network in a karate club, with nodes representing individuals and edges specifying friendships.[17]

The last four networks in Table 1 are neither social nor biological. The directed networks in this category are airline traffic data from two different airlines: unknown airlines and US Air 97, which contains North American transportation atlas data.[18] In these networks, nodes represent airports, and edge represents a flight that connects two airports. The undirected networks in this category are power grid[9] and co-purchase of political books[19] that were published around the 2004 election. The power grid network represents the topology of the western states' power grid of the United States, with nodes representing generators, transformers, or substations, and edge representing the high-voltage transmission line between them.[9] The network of co-purchase of political books has nodes representing books and edge connecting books that are frequently co-purchased by the same buyers.[19] We believe this diverse set of networks is a reasonable collection for drawing significant results.

Methods

We used the network motif detection tool FANMOD (FAst Network MOtif Detection)[20] for detecting motifs in all networks in Table 1.

Datasets. The network data analyzed in this research were collected from a variety of online sources: Pajek datasets,[13] Gephi Wiki Datasets,[18] Uri Alon's Complex Networks,[21] and University of Michigan Network Data.[22] Our collection contains 6 directed networks and 11 undirected networks. The detailed dimension for each network can be found in Table 2.

The network data collected in various formats including GML, GRAPHML, GEXF, NET, and adjacency list in Text format. A sample of each format can be found in Supplementary Table 1. These formats can be useful while

Table 2. Network size. Networks are listed by discipline.

NETWORK	DISCIPLINE	NUMBER OF NODES	NUMBER OF EDGES
C. elegans neural network	Biological	297	2359
Yeast	Biological	2361	7182
E. coli transcription network	Biological	418	519
Diseasome	Biological	2821	2673
Protein structure 1	Biological	95	213
Protein structure 2	Biological	53	123
Protein structure 3	Biological	97	212
Cypress dry season	Ecological	71	640
Everglades Graminoids Wet Season	Ecological	69	916
Dolphin social network	Social	62	159
Primary school contact network	Social	236	5899
Co-authorships in network science	Social	1461	2742
Zachary's karate club	Social	34	78
Unknown airlines	Other	234	2101
US air 97	Other	332	2126
Power grid	Other	4941	6594
Co-purchase of political books	Other	104	441

using network visualization programs such as Gephi[23] or Cytoscape,[24] but they cannot be read by FANMOD. FANMOD only analyzes data from a simple text file in which each line represents an edge of the network (adjacency list). Therefore, we wrote simple programs in both Java and Python in order to convert the data in different formats to the format that FANMOD accepts.

FANMOD. The fast network motif detection tool, FANMOD, created by Rasche and Wernicke, uses an algorithm called RAND-ESU in order to enumerate and sample subgraphs in given networks.[20] This algorithm is faster than its competitors such as mfinder[25] and MAVisto[26] in attempting to accomplish the same task.

Network motif detection with FANMOD involves three main steps[20]:

1. Search the input network for subgraphs and determine how often each subgraph occurs.
2. Analyze the subgraphs by establishing which are isomorphic, and then group the subgraphs together appropriately.
3. Determine which of these groups occurs more commonly than in randomly generated networks.

In FANMOD, step 1 is customizable by the user in two ways. The first is where the size of the subgraph can be chosen from three up to eight nodes. We chose to run experiments on each network starting at motifs of size three and continuing until we reached a size motif that could not finish in our allotted time within 1 week. The second customization option is where the program can fully enumerate the subgraphs in a given network or only sample a specific number of subgraphs. Although the latter scheme decreases the runtime of a network analysis, it does not provide a very accurate conclusion about the network because some subgraphs are not included in the search described in step (1). As a result, we chose to fully enumerate the subgraphs in each experiment we conducted.

FANMOD allows for other customizable features as well, such as specifying how many randomly generated networks should be compared to the input network. We decided to keep the number of networks at the recommended value of 1,000. In addition, FANMOD allows the user to alter the random network generation process. However, we chose to run the experiments with this process unchanged. Finally, FANMOD supports the analysis of both undirected and directed networks, which the user can specify when entering the input file.

Once FANMOD completely runs through an experiment on a specified network, it allows the user to generate HTML files to show the statistical data that was collected. FANMOD gives the user the option of customizing this feature as well, but we chose to leave it unchanged. This means that our HTML files generated for each size motif for each network were arranged by descending z-score, with a z-score greater than 2 as the minimum. Formula (1) represents the process of computing the z-score of a network motif.[5]

$$Z_i = \frac{Nreal_i - \{Nrand_i\}}{std\left(Nrand_i\right)} \quad (1)$$

where $Nreal_i$ is the number of times subgraph type i appears in the network, $\{Nrand_i\}$ is the mean of its appearances in the set of random networks, and $std(Nrand_i)$ is the standard deviation of its appearances in the set of random networks.

Experiments. We analyzed all the networks in Table 1 using FANMOD with the settings described above. Motifs up to size five were found for each network except for the primary school contact network, while sizes above five were able to be completed only on smaller networks such as karate, dolphin, and protein structure networks. HTML files were generated for each possible motif size of every network in order to visualize the motifs and their corresponding z-scores.

After extensive experimentation and data collection, we compiled the top three motifs with the highest z-scores for each motif size and network in Supplementary Tables 2 and 3, respectively. These tables contain significant motifs for undirected and directed networks, respectively. We picked the z-score as our delimiter because motifs with greater z-scores are more statistically significant than those with low scores. These tables allowed us to analyze our data in multiple dimensions. The first step before analysis was separating the undirected and directed networks. This is because no significant conclusions can be drawn between these opposite graphs. Next, we refined the tables further by subdividing the networks into the disciplines of biological, ecological, social, and others. In this way, we were able to compare networks of the same discipline across motif size, and also compare the motif structure across different disciplines. It also allowed us to analyze how motif topology changes as the size of a motif grows. In addition, these tables allowed us to view the similarities between smaller motifs and larger motifs in the same network, as well as across different disciplines. Finally, the tables allowed us to look at the most significant motifs of different sizes and manually count the number of smaller motifs found in larger motifs of the same network.

The output files and the export HTML files generated from FANMOD were used in much of the analysis as well. Because each HTML file is systematically created in the same pattern, we were able to parse the files using Java programs in order to perform further analysis. Using these programs, we observed the motifs found most frequently among undirected networks, and then repeated the process for directed networks. Subsequently, we collected the z-scores from the output files and used these z-scores to compute significance profiles [Formula (2)] for each motif found in each network.[5]

$$SP_i = \frac{Z_i}{\sqrt{\sum_i Z_i^2}} \quad (2)$$

where Z_i is the z-score of a subgraph i, and

SP (significance profile) is the vector of z-scores normalized to length 1.

Supplementary Tables 4 and 5 contain the z-scores collected from FANMOD and the significance profiles calculated for the top three significant motifs in undirected and directed networks.

Results and Discussion

Motif size and structure. One of the questions we set out to address when we started this work was how motif topology changes as the node number increases, specifically if larger motifs contain smaller motifs within them. We analyzed this by determining the number of times the most significant three-node motif occurred in the most significant motif of larger size (four to eight nodes) in the same network. For 15 out of 17 networks, the most significant four-node motif contained at least one, and up to four, of the most significant three-node motifs. When the motif size increased to five nodes, 15 out of 17 networks contained at least one instance of the most significant three-node motif. Figure 1 illustrates this observation.

Additionally, for 15 out of 17 networks, the frequency of the three-node motif occurring in the larger motif either increased or remained constant as the motif size increased from four to five. This suggests that, as motif size increases, larger motifs contain smaller motifs as a subgraph. We do not have the results for directed graphs for motifs with six or more nodes because FANMOD was unable to finish within the allotted time, but this trend is expected for larger motifs in directed networks. In addition, undirected networks do not have a clear pattern as the motif size exceeds five nodes, because some networks continue to contain more of the most significant three-node motif and some contain fewer.

Interdisciplinary motifs. One of the most surprising things about researching motifs among different disciplines is the unexpected similarities and dissimilarities between the motifs of different networks. These features were observed for undirected and directed networks in the following.

Undirected networks. Significant three-node motifs. One major similarity that is apparent from looking at the undirected networks is that all 11 networks have the same significant three-node motif (ID 238), as shown in Figure 2. There are only two possibilities for three-node motifs: an interconnected triangle (three-node in Fig. 2) and a triangle with one edge removed. There is no instance of the latter in any of the undirected networks for a three-node motif. The explanation for this is unique for each network.

The interconnected triangle in the diseasome network suggests that a disease is commonly caused by two genes, and one gene is usually a culprit of at least two diseases. It also suggests that there is a common link between three different diseases.

In protein structures, the interconnected triangle implies that proteins frequently have no outlying α or β helices; if a helix is within 10 Å of two other helices, those two are frequently within 10 Å of each other. This could indicate the presence of communities in the structure of proteins, in which helices of the community are closely packed with other helices of the community.

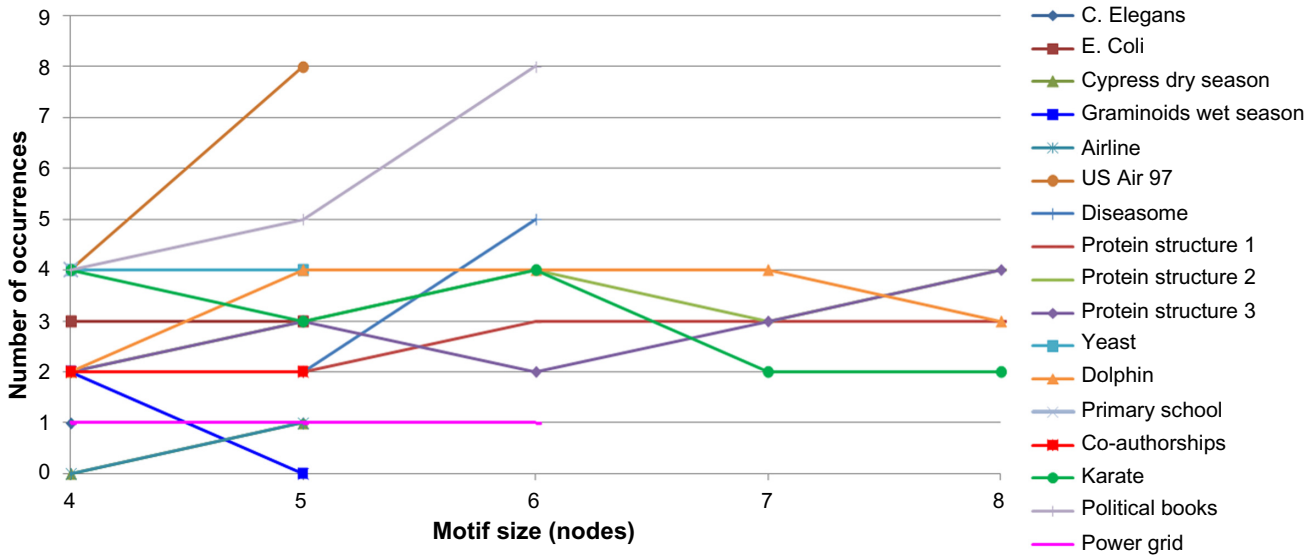

Figure 1. Occurrences of the most significant three-node motif within the most significant larger motifs for the same network.

In a protein–protein interaction network such as yeast, the interconnected triangle motif is known as the protein clique, which is the most abundant motif that makes up the entire network.[27] These proteins interact as a multicomponent machine.[27]

Social networks frequently have this interconnected triangle motif due to an intrinsic property called homophily. Homophily is a tendency in which we tend to be similar to our friends. If friendship exists between A and B and between A and C, then this intrinsic property suggests that B and C are likely similar to A. Thus, they are likely similar to each other.[28]

This principle explains that in social networks a node two connections away from a certain node is also connected to that node.

Commonly, Amazon shoppers who bought any one book in the interconnected triangle also bought the other two. The interconnected triangle in a power grid network suggests that two connected generators, transformers, or substations also connected to a common generator, transformer, or substation. This could be a common structure for avoiding power failure.

Significant four-node to eight-node motifs. Another similarity is that 6 (diseasome, three protein structures, dolphin, and co-authorship) out of 11 undirected networks have the same most and second significant four-node motifs (IDs 13278 and 4958). Additionally, four undirected networks (yeast, primary school contact, karate, and political books) have the same most, second, and third significant four-node motifs. These similarities can be seen in Figures 2 and 3.

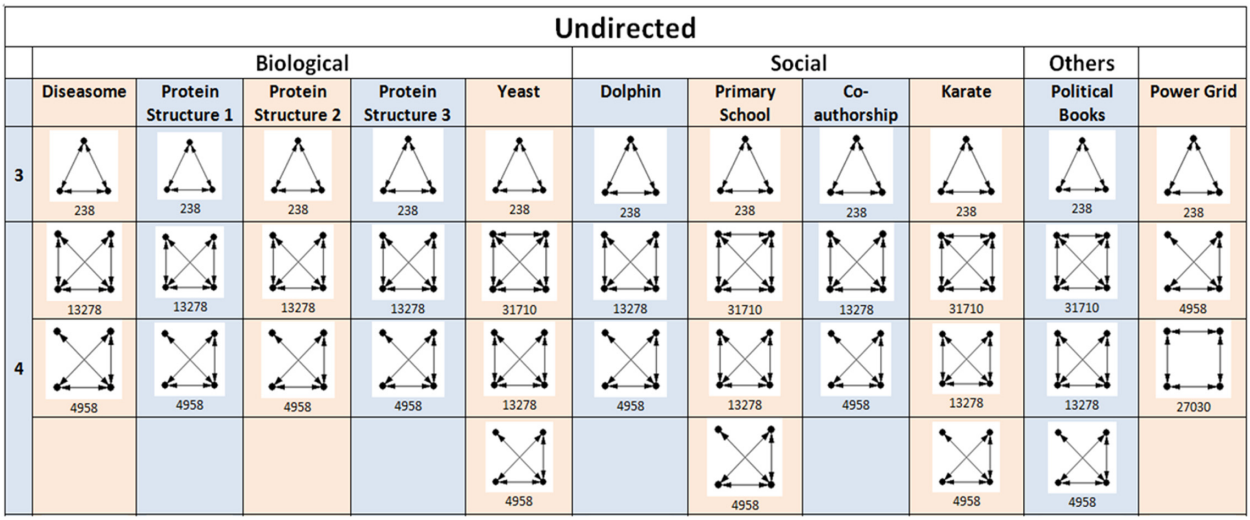

Figure 2. Illustrations of significant three- and four-node motifs for undirected networks. Motif's ID generated by FANMOD is included for each motif. Motifs are listed by significance in descending order.

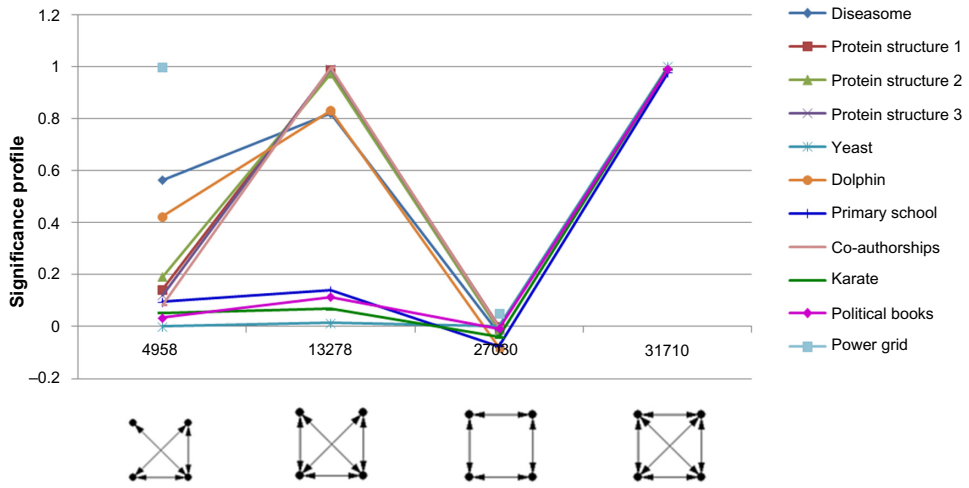

Figure 3. Significant four-node motifs for undirected network. Vertical axis shows significant profile. Horizontal axis shows motifs and their associated IDs.

Figure 3 shows a clear pattern for all significant four-node motifs found in each undirected network.

Figures 4–7 show the patterns of all significant motifs for each motif size from five to eight nodes in undirected networks. As the motif size increases, the patterns become more unclear. It is apparent from these graphs that, once the motif size exceeds four nodes, the graphs lose the pattern that exists with smaller motifs that have three or four nodes. This could mean that many of the undirected networks are similar at the basic three- and four-node motifs level structure but not at the higher level structure, which includes motifs with five or more nodes. This suggests that the dissimilarity increases between these networks as the motif size increases.

Correlation between undirected networks. We further observed the correlation between undirected networks using the Pearson correlation coefficient (PCC)[29] and compared their significant motifs. The PCC scores were obtained based on significance profiles, which were calculated using Equation (2) for each network. Table 3 shows the PCC scores for all undirected networks.

All three protein structures in Table 1 came from the following molecules: Diels–Alder catalytic antibodies, suppressors of tumorigenicity, and aldehyde ferredoxin oxidoreductase molecules.[5] The observations on their significant motifs revealed the following characteristics: Although the sizes and superstructures of these networks are all unique, the low-level community structure of each network is the same. All three networks have identical significant three-node motifs (ID 238). They also share the most and second significant four-node motifs (IDs 13278 and 4958). However, larger motifs across three protein structures have few similarities between them. Protein structures 2 and 3 share the third significant five-node motif (ID 9997502), which is the second significant motif of protein structure 1. Besides, two protein structures

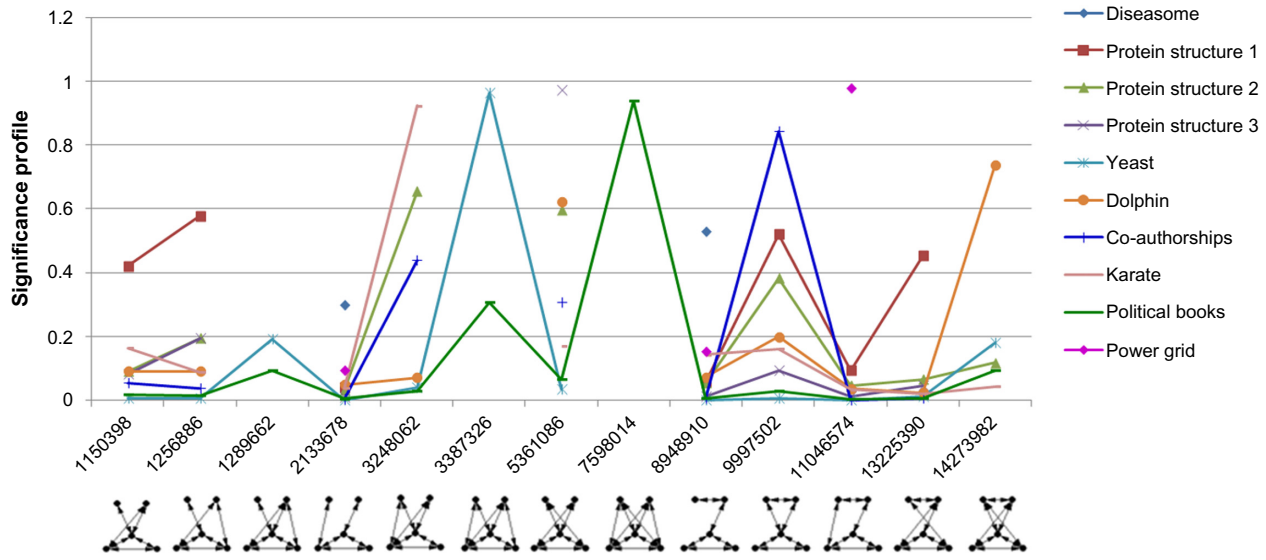

Figure 4. Significant five-node motifs for undirected networks.

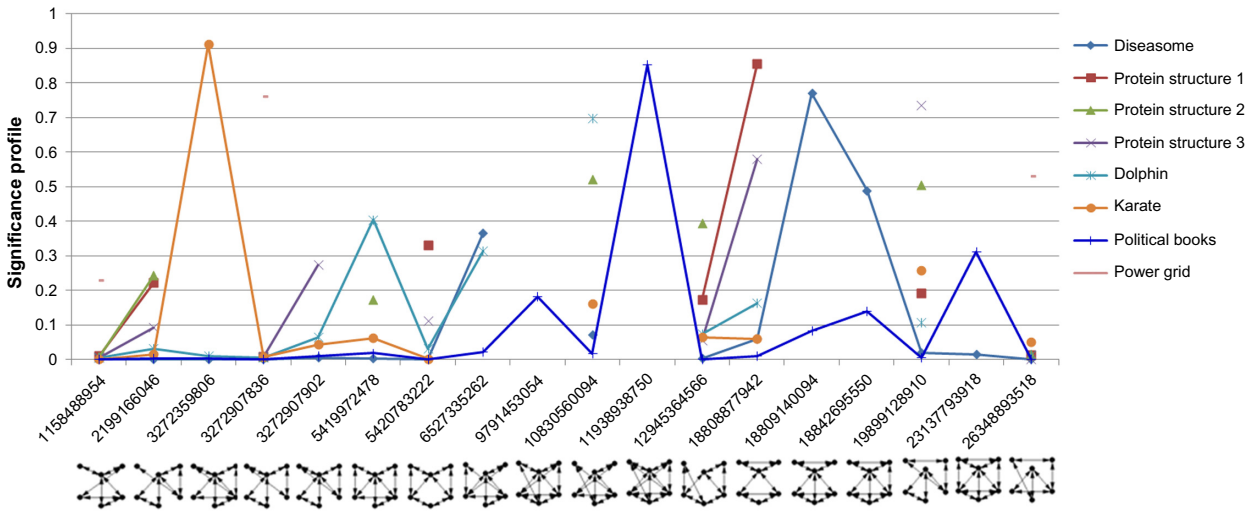

Figure 5. Significant six-node motifs for undirected network.

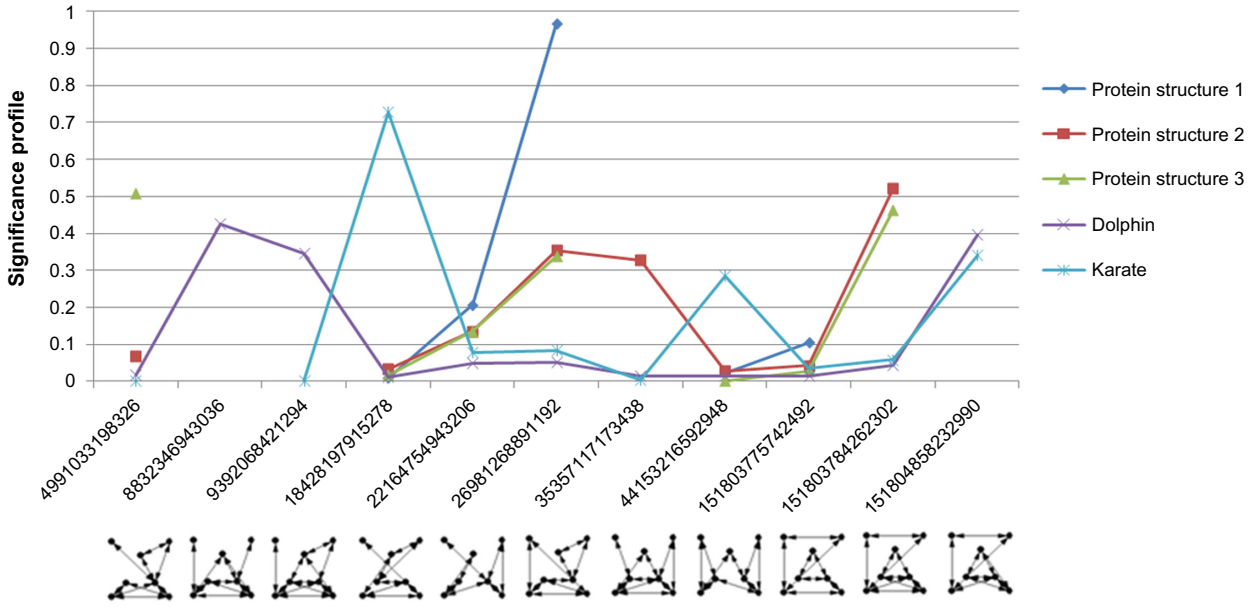

Figure 6. Significant seven-node motifs for undirected network.

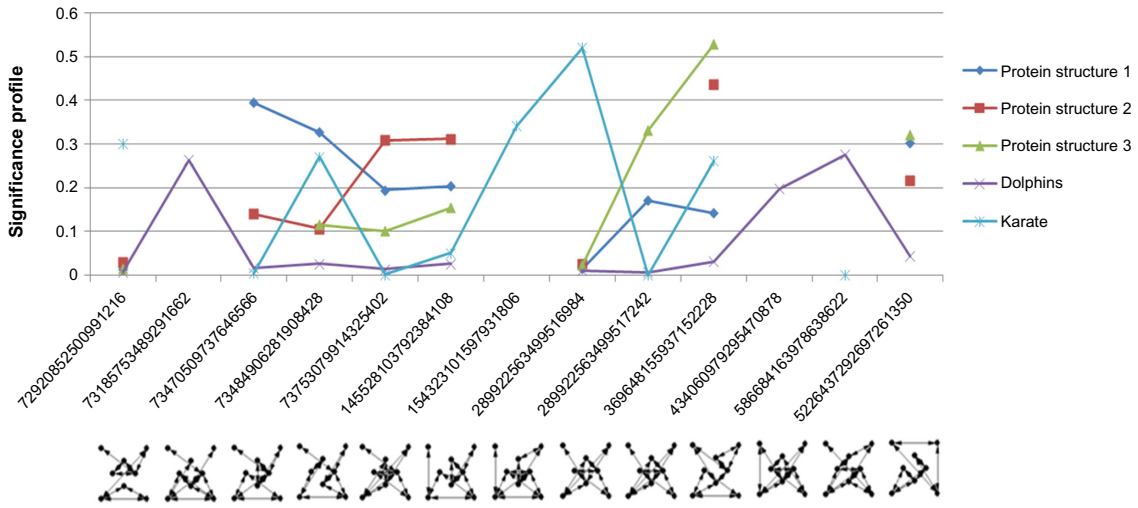

Figure 7. Significant eight-node motifs for undirected network.

Table 3. Pearson correlation coefficient scores for undirected networks. Bold face shows strong relationship between networks.

	DISEASOME	PROTEIN STRUCTURE 1	PROTEIN STRUCTURE 2	PROTEIN STRUCTURE 3	YEAST	DOLPHIN	PRIMARY SCHOOL	CO-AUTHORSHIPS	KARATE	POLITICAL BOOKS	POWER GRID
Diseasome	1										
Protein structure 1	0.5528	1									
Protein structure 2	0.5652	**0.8230**	1								
Protein structure 3	0.5290	**0.7913**	**0.8523**	1							
Yeast	0.3283	0.4108	0.4254	0.4311	1						
Dolphin	0.6375	0.6690	0.6834	**0.7333**	0.5356	1					
Primary school	0.5804	0.5331	0.5450	0.5321	**0.9852**	0.6186	1				
Co-authorships	0.5950	**0.8025**	**0.8917**	**0.7172**	0.4204	**0.7529**	0.5279	1			
Karate	0.3166	0.3824	0.4993	0.3760	**0.7596**	0.3045	**0.9958**	0.4733	1		
Political books	0.4740	0.4737	0.5119	0.4965	**0.8901**	0.5537	**0.9943**	0.5196	0.6619	1	
Power grid	0.4978	0.5175	0.5138	0.4436	0.3615	0.6046	0.4154	0.4196	0.4021	0.3742	1

share two significant five-node motifs (IDs 1256886 and 5361086). Two protein structures also share two significant six-node motifs (IDs 18808877942 and 19899128910). All three networks have a common significant seven-node motif (ID 26981268891192). In addition, two protein structures share a significant seven-node motif (ID 151803784262302). Further, two protein structures share the most and third significant eight-node motif (IDs 369648155937152228 and 5226437292697261350). These observations can be seen in Figures 4–7 and in Supplementary Table 2. The analysis suggests that all three protein structures share a common blueprint for arranging α and β helices at the small community level. Once they exceed this level, differences arise, leading to unique properties of different protein structures. In addition, PCC scores showed strong positive relationships among these networks. Protein structure 1 has strong positive relationships with protein structures 2 and 3 (PCC scores 0.8230 and 0.7913, respectively). Protein structure 2 also has a strong positive relationship with protein structure 3 (PCC score 0.8523). Thus, it suggests that these protein structures belong to the same family.

The dolphin social network also has the same significant three- and four-node motifs with three protein structures (ID 238 for three-node; IDs 13278 and 4958 for four-node). In addition, it shares the second and third significant five-node motifs with protein structure 2 (IDs 5361086 and 9997502, respectively). Besides, it shares the most significant six-node motif with protein structure 2 (ID 10830560094). These observations can be seen in Supplementary Table 2. Furthermore, the PCC score revealed a strong positive relationship between the dolphin social network and protein structure 3 (PCC score 0.7333). However, there are less strong positive relationships between the dolphin social network with protein structure 1 and protein structure 2 (PCC scores 0.6690 and 0.6834, respectively). These observations suggest that the dolphin social network shares low-level community structure (three and four nodes) with three protein structures. It also suggests that the dolphin social network and three protein structures belong to a superfamily.[5]

The co-authorships network also shares significant three- and four-node motifs with three protein structures (ID 238 for three-node, IDs 13278 and 4958 for four-node). It also has a common significant five-node motif with three protein structures (ID 9997502). Besides, it shares the significant five-node motif (ID 5361086) with two protein structures. In addition, it shares two significant five-node motifs with the dolphin social network (IDs 9997502 and 5361086). The PCC scores also revealed strong positive relationships between the co-authorships network and three protein structures (PCC scores 0.8025, 0.8917, and 0.7172 with protein structure 1, 2, and 3, respectively). This observation suggests that the co-authorships network also shares low-level community structure (three and four nodes) with three protein structures. It also suggests that the co-authorships network and three

protein structures belong to a superfamily. Additionally, the co-authorships network has a strong positive relationship with the dolphin social network (PCC score 0.7529). Thus, the analysis suggests that the co-authorships network, the dolphin social network, and three protein structure networks belong to the same superfamily.

The karate, yeast, primary school contact, and co-purchase of political books networks have the same significant three- and four-node motifs (ID 238 for three-node, IDs 31710, 13278 and 4958 for four-node). Thus, it suggests that these networks share a low-level community structure. In addition, the PCC score showed a strong positive relationship between karate and yeast networks (PCC score 0.7596). There is also a very strong positive relationship between karate and primary school contact networks (PCC score 0.9958). Furthermore, the co-purchase of political books network has a strong positive relationship with yeast, and it has another very strong positive relationship with the primary school contact network (PCC scores 0.8901 and 0.9943, respectively). However, the co-purchase of political books network has a less strong positive relationship with karate (PCC score 0.6619). Hence, the observations suggest that karate, yeast, primary school contact, and co-purchase of political books networks belong to the same superfamily.

The superfamily identified above contains different networks across different disciplines, but these networks are similar because they share similar low-level structures based on the observations of significant motifs and they have strong positive relationship based on PCC scores. The reason why these networks have similar motifs could be that they are naturally formed to perform similar tasks.[5] Thus, it suggests that research and results can be used to learn and share among these networks.

Although several undirected networks share a common significant three-node motif, the function of this motif may be specific to each network. The detailed function of this motif for each network is beyond the scope of this work.

Besides the common significant motifs, each network has its own set of motifs that are unique to that network. The reason could be that these motifs play a role in characterizing the unique structure of individual networks. Supplementary Table 6 shows some insignificant motifs specific to each undirected network. For example, motif ID 213597653354134 was found only in protein structure 1, and motif ID 72649290795 is exclusive to the dolphin social network.

Directed networks. Significant three- to five-node motifs. The similarities observed in undirected networks for three- and four-node motifs do not exist in directed networks. All directed networks have few common significant three- to five-node motifs. This can be seen in Figures 8–10 and in Supplementary Table 3. These figures show little or no clear pattern for these networks. Thus, it suggests little or no similarity between them.

C. elegans neural network. This network has both one- and two-directional edges as interaction between neurons can be a one-way or two-way interaction. The top three significant three-node motifs in this network are motif IDs 238, 166, and 46. The feed-forward loop (motif ID 38) was reported as an over-represented three-node motif for *C. elegans* in previous work.[30] This motif was also detected and reported as the fifth significant three-node motif by FANMOD. It is interesting that the most common structural three-node motifs in *C. elegans* do not include the well-known feed-forward loop, which plays an essential role in information processing. The structure of the most significant three-node motif (ID 238) suggests that there are interactions between neurons in information processing. The top three significant four-node motifs detected for this network are motif IDs 25566, 27340, and 990. The well-known bi-fan motif (ID 204) was previously reported as an over-represented motif for *C. elegans*.[30] However, this motif is not over-represented in this network. All top three significant five-node motifs contain one or more significant three-node motifs.

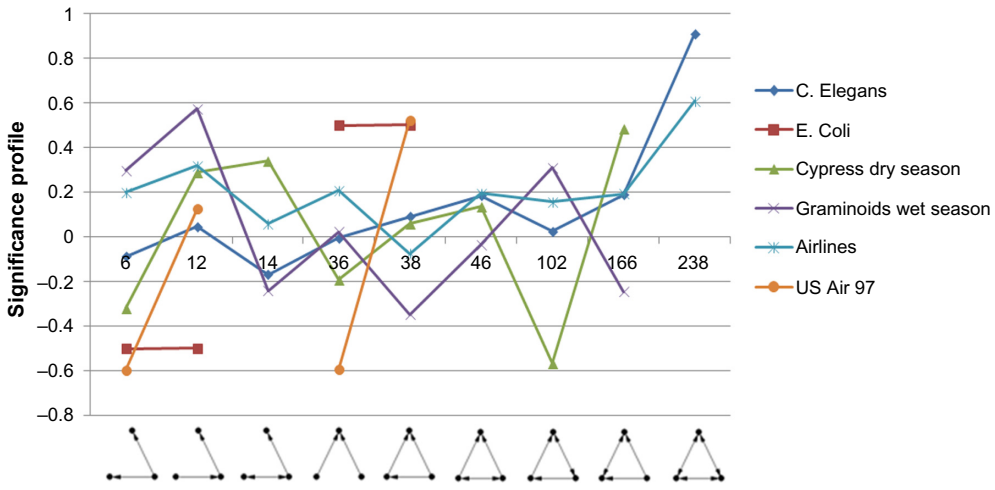

Figure 8. Significant three-node motifs for directed network.

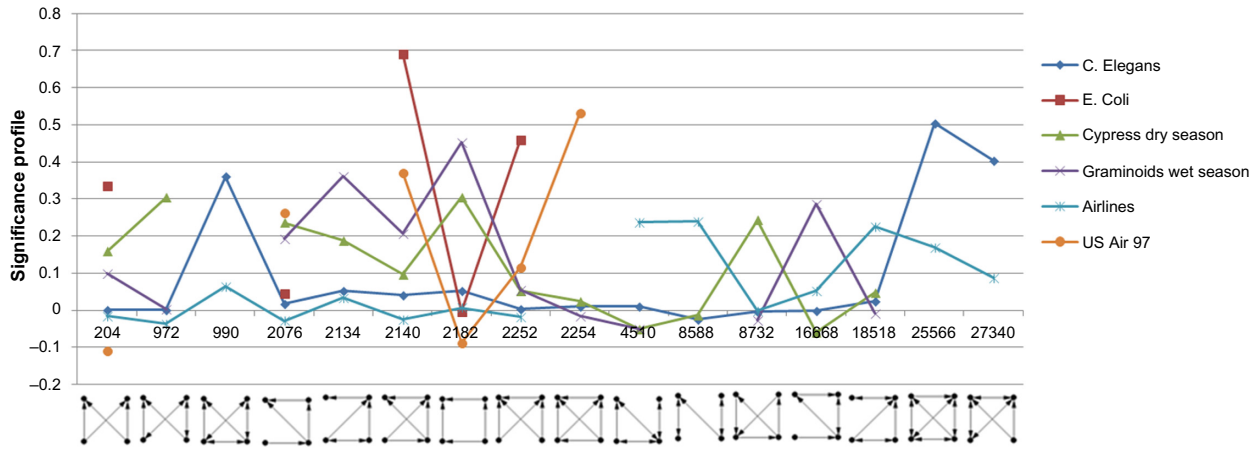

Figure 9. Significant four-node motifs for directed network.

E. coli transcription network. This network does not have a bidirectional edge because transcription factor regulates gene or other transcription factors in one-way direction. There are two significant three-node motifs in this network: motif ID 38, which is a feed forward loop, and motif ID 36, which has two transcription factors that co-regulate a gene. In transcription network, the feed-forward loop is known as two transcription factors co-regulating a gene, with one transcription factor regulating the other. This motif was found previously as the most significant motif in the *E. coli* transcription network.[31] FANMOD also reported this motif as the most significant three-node motif for this network. Two of the top three significant four-node motifs contain feed-forward loops (IDs 2140 and 2252). The third significant four-node motif (ID 204) is a bi-fan, which is known as two transcription factors that co-regulate two genes. All top three significant five-node motifs contain one or more significant three-node motifs.

Food web networks. In food web networks, the direction of one directional edge points from a predator to its prey. If it is a bidirectional edge, then the species can be both a predator and a prey. Both one- and bidirectional edges exist in food web networks, as one species can hunt other species and vice versa.

The Cypress Dry Season food web has motif ID 166 as the most significant three-node motif. This motif indicates that two preys of a common predator also prey each other. The second significant three-node motif (ID 14) indicates that one of the two predators preying each other also preys another species. The third significant three-node motif is a cascade motif (ID 12), which was discovered previously in food webs.[3] This motif shows that a prey of a predator is also a predator of another species. The most significant four-node motif in this network is a bi-parallel (ID 2182), which was also discovered previously.[3] This motif indicates that two preys of a common predator also are predators of a common prey. The second significant four-node motif (ID 972) contains the second significant three-node motif. The third significant four-node motif (ID 8732) also contains the third significant cascade three-node motif. All top three significant five-node motifs contain one or more significant three-node motifs.

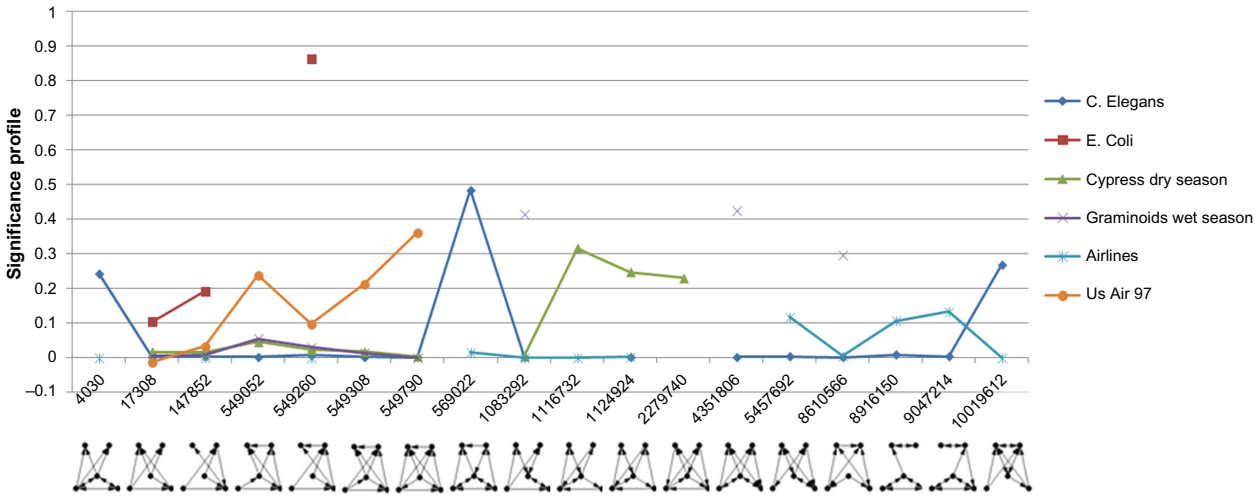

Figure 10. Significant five-node motifs for directed network.

The Everglades Graminoids Wet Season food web has as the most significant three-node motif a cascade motif (ID 12). The second significant three-node motif (ID 102) indicates two species that prey each other: one is a predator and the other is a prey of a common species. The third significant three-node motif (ID 6) shows that a predator preys two species. This network also shares the most significant four-node motif, which is a bi-parallel with the Cypress Dry Season food web. The second and third significant motifs for four and five nodes contain instances of significant three-node motifs.

Airline networks. In unknown airline network, the most significant three-node motif is a fully connected bidirectional-edge triangle (ID 238). This motif implies that a round trip commonly has the length of two or three flights. The second significant three-node motif is a cascade (ID 12), which indicates that two airports connect through a common airport via one-way trip. The third significant three-node motif (ID 36) shows that two flights have a common destination. The most and second significant four-node motifs (IDs 8588 and 4510) contain the three-node motif ID 14. The third significant four-node motif (ID 18518) has an instance of three-node motif ID 46. Both motif IDs 14 and 46 indicate that two airports are connected via a one-way or a two-way flight. The top three significant five-node motifs (IDs 9047214, 5457692, and 8916150) contain instance of either the most or second significant three-node motif. This airline network has both one- and bidirectional edges, meaning that a one-way trip or a round trip between two consecutive airports is possible. In general, this airline network reveals that a round trip commonly has two or three flights and two airports are commonly connected via one or two flights.

In US Air 97 network, the most significant three-node motif is a feed-forward loop (ID 38), which shows that two airports are connected via a one-way flight or two one-way flights. This airline network shares the second significant three-node motif (cascade motif ID 12) with the unknown airlines network. The most significant four- and five-node motifs (IDs 2254 and 549790) contain feed-forward loops. The second and third significant motifs for four and five nodes (IDs 2140 and 2076 for four-node, IDs 549052 and 549308 for five-node) contain instances of feed-forward loops and cascades. This airline network does not have a bidirectional edge, meaning that round trip between two consecutive airports is not possible. In general, this airline network shows that two airports are connected via a one-way flight or two one-way flights. The structure of this airline network also reveals that common round trip is not possible with two or three flights.

Correlation between directed networks. We also observed the correlation between directed networks using the PCC method. The correlation scores in Table 4 show no strong relationship between these networks. However, some inverse relationships exist between some networks. For example, the unknown airlines and US Air 97 have an inverse relationship (PCC score −0.4921). The cause of this inverse relationship could be that the unknown airlines network offers service that is not offered by US Air 97. The unknown airlines network also has a weak inverse relationship with the *E. coli* transcription network (PCC score −0.3822). A weak inverse relationship also exists between the food web Everglades Graminoids Wet Season and *E. coli* transcription network (PCC score −0.3183). In addition, another weak inverse relationship was also found between the food web Everglades Graminoids and US Air 97 (PCC score −0.2392).

The analysis suggests that directed networks are distinct compared to undirected networks. However, these networks have a common characteristic: that is, larger motifs contain three-node motifs as their subgraphs.

Conclusions and Future Work

We detected and analyzed network motifs in undirected and directed networks from several different disciplines. The comparisons between significant motifs in undirected and directed networks showed that larger motifs contain three-node motifs as their subgraphs. Therefore, it suggests that the three-node motif is a building block of larger motifs. The analysis based on PPC scores and significant motifs revealed that directed networks are distinct, while the analysis based on significant motifs showed similar low-level structure in multiple undirected networks. In addition, three protein structure networks share similar low-level community structure at three and four nodes, but as the motif size increases, differences arise. Hence, it suggests that similar networks share similar small motifs, but larger motifs define the unique structure of individuals.

Table 4. Pearson correlation coefficient scores for directed networks. Bold face shows inverse relationship between networks.

	C. ELEGANS	*E. coli*	CYPRESS DRY SEASON	GRAMINOIDS WET SEASON	AIRLINES	US AIR 97
C. elegans	1					
E. coli	0.3470	1				
Cypress Dry Season	0.1330	0.0386	1			
Graminoids Wet Season	0.0040	**−0.3183**	0.0825	1		
Airlines	0.4049	**−0.3822**	0.0115	0.1488	1	
US Air 97	0.4929	0.3478	0.2870	**−0.2392**	**−0.4921**	1

The PPC scores suggest that protein structure networks, the dolphin social network, and the co-authorships network belong to a superfamily. Furthermore, yeast protein–protein interaction network, primary school contact network, karate network, and co-purchase of political books network can be classified into the same superfamily. The PCC scores also revealed an inverse relationship between an unknown airlines and US Air 97 networks. In addition, weak inverse relationships were found between the *E. coli* network and other networks such as unknown airlines network and food web Everglades Graminoids Wet Season network. Further, a weak inverse relationship was also found between US Air 97 and food web Everglades Graminoids Wet Season networks.

Cross-disciplinary research is a vital aspect of motif analysis and comprehension. Further research on this topic can go in many directions. One such direction could be discovering new datasets from these or other disciplines and performing experiments and analyses on such datasets. With the advent of faster and more powerful computation, new networks are becoming available and could be used for future research. Finally, directed networks could be investigated even further by analyzing motifs with six or more nodes.

Acknowledgment

The authors would like to thank the anonymous reviewers for their comments and suggestions, which helped to improve this paper.

Author Contributions

Designed, performed the experiments, and drafted the initial manuscript: LD, AFM. Re-performed the experiments and revised the manuscript critically: NTLT. Directed and helped to draft the manuscript: C-HH. All authors reviewed and approved the final manuscript.

Supplementary Data

Supplementary Table 1. Network file types. Format examples for each network type.

Supplementary Table 2. Significant motifs for undirected networks. Illustrations of top three significant motifs for each motif size and network in undirected networks. Motif ID is included for each motif.

Supplementary Table 3. Significant motifs for directed networks. Illustrations of top three significant motifs for each motif size and network in directed networks. Motif ID is included for each motif.

Supplementary Table 4. *z*-Scores and significance profiles for undirected networks. *z*-Scores from FANMOD and significance profiles calculated using Equation (2) for top three significant motifs in undirected networks.

Supplementary Table 5. *z*-Scores and significance profiles for directed networks. *z*-Scores from FANMOD and significance profiles calculated using Equation (2) for top three significant motifs in directed networks.

Supplementary Table 6. Examples of insignificant motifs for undirected networks. The number in the column for each network represents the *z*-score.

REFERENCES

1. Wong E, Baur B, Quader S, Huang C. Biological network motif detection: principles and practice. *Brief Bioinform*. 2012:13(2):202–15.
2. Arenas A, Fernandez A, Fortunato S, Gomez S. Motif-based communities in complex networks. *J Phys A Math Theor*. 2008;41:1–9.
3. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. Network motifs: simple building blocks of complex networks. *Science*. 2002;298:824–7.
4. Schwobbermeyer H. Network motifs. In: Junker BH, Schreiber F, eds. *Analysis of Biological Networks*. Edited by ed. John Wiley & Sons, Inc; Hoboken, NJ, USA. 2008:85–111.
5. Milo R, Itzkovitz S, Kashtan N, et al. Superfamilies of evolved and designed networks. *Science*. 2004;303(5663):1538–42.
6. Allan EG, Turkett WH, Fulp EW. Using network motifs to identify application protocols. *Global Telecommunications Conference 2009. GLOBECOM 2009. IEEE*; Honolulu, HI, USA. 2009:1–7.
7. Przytycka TM. An important connection between network motifs and parsimony models. *Res Comput Mol Biol*. 2006;3909:321–35.
8. Tsang J, Zhu J, van Oudenaarden A. MicroRNA-mediated feedback and feedforward loops are recurrent network motifs in mammals. *Mol Cell*. 2007;26(5):753–67.
9. Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks. *Nature*. 1998;393:440–2.
10. Bu D, Zhao Y, Cai L, et al. Topological structure analysis of the protein-protein interaction network in budding yeast. *Nucleic Acids Res*. 2003;31(9):2443–50.
11. Shenn-Orr S, Milo R, Mangan S, Alon U. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat Genet*. 2002;31:64–8.
12. Goh K, Cusick ME, Valle D, Childs B, Vidal M, Barabsi A. The human disease network. *Proc Natl Acad Sci USA*. 2007;104(21):8685–90.
13. South Florida Food Webs: (1998). Available at: https://networkdata.ics.uci.edu/. Date last accessed April 1, 2015.
14. Lusseau D, Schneider K, Boisseau OJ, Haase P, Slooten E, Dawson SM. The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. *Behav Ecol Sociobiol*. 2003;54(4):396–405.
15. Stehle J, Voirin N, Barrat A, et al. High-resolution measurements of face-to-face contact patterns in a primary school. *PLoS One*. 2011;6(8):e23176.
16. Newman MEJ. Finding community structure in networks using the eigenvectors of matrices. *Phys Rev E Stat Nonlin Soft Matter Phys*. 2006;74:036104.
17. Zachary WW. An information flow model for conflict and fission in small groups. *J Anthropol Res*. 1977;33(4):452–73.
18. Gephi Wiki Datasets. Available at: https://github.com/gephi/gephi/wiki/Datasets. Date last accessed April 1, 2015.
19. V. Krebs: Books About US Politics. Available at: http://www-personal.umich.edu/~mejn/netdata/. Date last accessed April 1, 2015.
20. Wernicke S, Rasche F. FANMOD: a tool for fast network motif detection. *Bioinformatics*. 2006;22(9):1152–3.
21. Uri Alon's Complex Networks. Available at: http://wws.weizmann.ac.il/mcb/UriAlon/download/collection-complex-networks. Date last accessed April 1, 2015.
22. University of Michigan Network Data. Available at: http://www-personal.umich.edu/~mejn/netdata/. Date last accessed April 1, 2015.
23. Bastian M, Heymann S, Jacomy M. Gephi: an open source software for exploring and manipulating networks. *International AAAI Conference on Weblogs and Social Media*. San Jose, CA, USA. 2009:1–2.
24. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 2003;13(11):2498–504.
25. Kashtan N, Itzkovitz S, Milo R, Alon U. Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. *Bioinformatics*. 2004;20(11):1746–58.
26. Schreiber F, Schwbbermeyer H. MAVisto: a tool for the exploration of network motifs. *Bioinformatics*. 2005;21(17):3572–4.
27. Yeger-Lotem E, Sattath S, Kashtan N, et al. Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction. *Proc Natl Acad Sci U S A*. 2004;101(16):5934–9.
28. Easley D, Kleinberg J. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press; New York, NY, U S A. 2010.
29. Pearson K. Notes on regression and inheritance in the case of two parents. *Proc R Soc Lond*. 1895;58:240–2.
30. Reigl M, Alon U, Chklovskii DB. Search for computational modules in the C. elegans brain. *BMC Biol*. 2004;2:25.
31. Mangan S, Alon U. Structure and function of the feed-forward loop network motif. *Proc Natl Acad Sci U S A*. 2003;100(21):11980–5.