On the Planarity of Validated Complexes of Model Organisms in Protein-Protein Interaction Networks

Kathryn Cooper[1], Nathan Cornelius[1], William Gasper[1], Sanjukta Bhowmick[2], Hesham Ali[1]
[1] *College of Information Science & Technology, University of Nebraska at Omaha, NE USA*
[2] *College of Engineering, University of North Texas, Denton, TX, USA*
kmcooper@unomaha.edu

Abstract. Leveraging protein-protein interaction networks to identify groups of proteins and their common functionality is an important problem in bioinformatics. Systems-level analysis of protein-protein interactions is made possible through network science and modeling of high-throughput data. From these analyses, small protein complexes are traditionally represented graphically as complete graphs or dense clusters of nodes. However, there are certain graph theoretic properties that have not been extensively studied in PPI networks, especially as they pertain to cluster discovery, such as planarity. Planarity of graphs have been used to reflect the physical constraints of real-world systems outside of bioinformatics, in areas such as mapping and imaging.

Here, we investigate the planarity property in network models of protein complexes. We hypothesize that complexes represented as PPI subgraphs will tend to be planar, reflecting the actual physical interface and limits of components in the complex. When testing the planarity of known complex subgraphs in *S. cerevisiae* and selected mammalian PPIs, we find that a majority of validated complexes possess this planar property. We discuss the biological motivation of planar versus nonplanar subgraphs, observing that planar subgraphs tend to have longer protein components. Functional classification of planar versus nonplanar complex subgraphs reveals differences in annotation of these groups relating to cellular component organization, structural molecule activity, catalytic activity, and nucleic acid binding. These results provide a new quantitative and biologically motivated measure of real protein complexes in the network model, important for the development of future complex-finding algorithms in PPIs. Accounting for this property paves the way to new means for discovering new protein complexes and uncovering the functionality of unknown or novel proteins.

Keywords: *planar graphs, PPI networks, protein complexes, DDI networks*

1 Introduction

1.1 A Brief History and Motivation

In the early stages of bioinformatics research, many studies focused on data generation approaches along with standard analysis of this data, to take advantage of the rapid advancement of biomedical technologies. The lack of data availability in the early days of bioinformatics meant that every attempt was made to take full advantage of all

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50371-0_48

available data. These large, aggregated databases make datasets from multiple research groups and experiments available but has also led to certain practices that impedes the quality of the data if attention is not paid to details of the dataset provenance. Such practices include aggregation of data collected under different experimental conditions or incorporating relationships obtained via prediction rather than observed experiments.

Recently, with the massive explosion of available data in the bioscience and medical domains, the attention has shifted towards a focus on validation, data quality, and in-depth data analysis. To achieve these objectives, there is a need to develop advanced validation mechanisms to assess the quality of the large currently available biological data. We posit that an important step in this direction is to study underlying properties or features associated with current datasets and use these futures to validate the various databases and assess the quality of their data items. In this work, we explore how studying the underlying structural properties of biological networks can lead to a better understanding of the nature of the network data. In particular, we look into the impact of the physical aspects that are associated with protein interaction networks and how the physical restrictions of the interactions enforce certain properties in such networks. Our primary hypothesis is that protein complexes are likely to form planar underlying structures when represented as a subgraph of a protein-protein interaction network, particularly if their domains or subcomponents are large. Proving such hypothesis will open the door to a new direction in utilizing the large amount of data associated with biological networks and objectively assess their quality.

1.2 Overview of Network Modeling of PPIs

Modeling of protein-protein interaction (PPI) networks has grown in popularity since 1999 with the advancement of open source community databases for sharing PPI data, a rapidly growing body research on the link between network models and biological functionality (Barabasi & Albert, 1999; Barabasi & Oltvai, 2004; Jeong, H., Mason, Barabasi, & Oltvai, 2001), and the development of algorithms and tools for clustering proteins to identify common functionality (Barabasi, A. L. & Albert, 1999; Barabasi, A. L. & Oltvai, 2004; Jeong, H., Mason, Barabasi, & Oltvai, 2001, Brohee & van Helden, 2006). A number of popular algorithms designed specifically for clustering proteins from PPI networks are now available, including (but certainly not limited to) ClusterONE for finding overlapping protein complexes (Nepusz, Yu, & Paccanaro, 2012), HC-Pin for functional complex discovery (Wang, Li, Chen, & Pan, 2011), Altaf-Ul-Amin's 2006 algorithm for detecting complexes in large PPI networks (Altaf-Ul-Amin, Shinbo, Mihara, Kurokawa, & Kanaya, 2006), PRODISTIN for prediction of cellular function in PPI complexes (Brun, Herrmann, & Guénoche, 2004), as well as MCODE (Bader & Hogue, 2003), MINE (Rhrissorrakrai & Gunsalus, 2011), and SPICi (Jiang & Singh, 2010). All of these aforementioned approaches are a part of a large majority of clustering algorithms built for protein-protein interaction networks that use a density measure or function to some extent to identify clusters or complexes within a protein-protein interaction network. While nearly all of the aforementioned literature notes explicitly in their work that density is not the only factor with weight in clustering edges in a protein-protein interaction network, a majority of algorithms can simplify

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50371-0_48

protein complex identification with the justification that complexes are represented as densely connected clusters in a PPI network. This is typically done using a hard-clustering approach (Pu, Vlasblom, Emili, Greenblatt, & Wodak, 2007), but performance is mixed.

1.3 3D Structure of Protein Complexes *in vivo*

Inherently, any clustering algorithm that uses density as a major component of its algorithm makes an assumption that a denser subgraph is the desired outcome, which may not always be the case. As a protein complex grows in size (in length of protein complex components and/or number of interaction partners), it becomes more and more unlikely that all components of a protein complex will have space to physically interface with one another. Inherently, a protein chain in its tertiary or quaternary form can typically only be bound to one partner per interface at a time (Keskin, Gursoy, Ma, & Nussinov, 2008). It is known that the stability of protein-protein interactions can be measured by affinity as transient or permanent if they are part of a non-obligate PPI complex (Acuner Ozbabacan, Engin, Gursoy, & Keskin, 2011). Further information is known about the stability and permanence of protein-protein interactions; for example, interactions between homodimeric proteins tend to be more stable in their PPI interfaces than heterodimers (Jones & Thornton, 1996) and also tend to be easier to predict (Keskin et al., 2008). One reasoning behind this is that the interfaces of heterodimers tend to be flatter than homodimers (Jones & Thornton, 1996).

Note that a PPI network is only a model. For example, due to the nature of the techniques used to infer PPIs at the systems level (such as tandem affinity purification, mass spectrometry, or older techniques such as the Y2H experimental system,), a protein complex as it is found within its quaternary form in the cellular machine may not necessarily be accurately represented by the PPI network. Many of these techniques present a protein of interest (bait) and determine through affinity which other proteins (prey) interact with it outside of their normal functioning in the cell, meaning that the PPIs measured represent physical interactions but not their spatial arrangement or temporal stability (Uetz et al., 2000). Therefore, a number of factors, such as protein interactor length, binding affinity, experimental system used to determine the interaction, and stability of the interaction may or may not be represented in a PPI network.

1.4 Planarity in graph theory

The term “planar graph” denotes a well-known graph theoretic property indicating that a graph is planar if it can be embedded on a plane without having its edges cross. This notion differs from “planarity” that has been used to describe shape and size of a protein’s interface with another within its 3D structure (Janin, Bahadur, & Chakrabarti, 2008; Jones & Thornton, 1996). Henceforth ,when referring to planarity or planar graphs, we refer to the graph theoretic definition, as in Definition 1 below.

Definition 1. *A graph G = (V,E) has a planar embedding if it can be drawn on a plane without crossing any of its edges. A graph is planar if it has at least one planar embedding (West, 1996).*

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50371-0_48

In this paper, we assume complexes are represented within PPI networks as an induced subgraph $G = (V,E)$ where G is a simple graph, meaning it contains neither self-loops nor multiple edges, and edges representing interaction relationships are binary (0 = does not exist, 1 = exists). Subgraphs are not required to be connected graphs. (See the example given in Figure 1).

Interestingly, there appears to be no prior research into the planarity of subgraphs representing protein complexes mined from protein-protein interaction networks. A 2010 model submitted to arXiv notes that while some interactions are too complex to be reliably represented in the "protein – edge – protein" format of the PPI, it is possible to model the relationship between PPI network topology and relative protein abundance on the assumptions that there exists a subset of protein interactions tend to be flat, stable, and ordered (Heo, Maslov, & Shakhnovich, 2010). However, a search for applications of planar graphs reveals no prior research in biological networks.

1.5 Characterizing a graph as planar

There exist several algorithms for testing whether a given graph is planar. The most well-known ones use a direct application of Kuratowski's basic planarity theorem which states that a graph is planar if and only if it does not contain K_5, $K_{3,3}$, or any of their subdivisions as an induced subgraph. Note that K_5 denotes a complete graph (clique) of five vertices and $K_{3,3}$ denotes a complete bipartite graph of six vertices with three vertices in each set. However, Kuratowski's method is expensive to test in practice, particularly for large graphs. Linear time planarity testing algorithms include expanding a smaller planar graph by adding paths (path addition, Hopcroft & Tarjan, 1974), vertices (vertex addition, Even & Tarjan, 1976) or edge (edge addition, Boyer & Myrvold, 2004). Parallel algorithms for planarity testing have also been developed

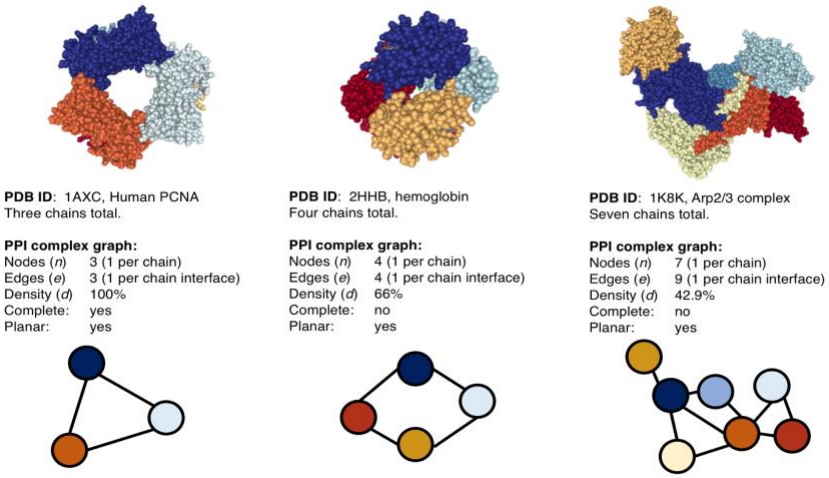

Figure 1. Three examples of 3D protein structures and a dummy graph model of their PPI representation. The top row shows a given protein complex with its different protein components highlighted with a different color; the bottom row provides an example of how that complex might be represented graphically. Note that interactions/edges in the graphical model are drawn where there is a physical interface within the 3D protein structure. In the middle, we provide dummy examples of measures of number of nodes, edges, and density, as well as planarity and completeness of each complex.

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50371-0_48

(Klein & Reif, 1988). Several graph softwares such as the Boost Graph Library (Siek, Lumsdaine, & Lee, 2002) and Library of Efficient Data Types (LEDA) (Mehlhorn & Näher, 1989) include algorithms for testing the planarity of graphs.

2 Results

In this work we investigate the planarity of known protein complexes as represented by induced subgraph in the well-characterized model organism *S. cerevisiae* and other mammalian model organisms. We provide evidence that a large portion of these complexes are planar in our datasets. To highlight our work, we provide the following results and their supporting evidence for two manually curated datasets with a combined total of 808 known complexes in *S. cerevisiae*, CYC2008 (n = 408) and YHTP2008 (n = 400), and other mammalian complexes from the Comprehensive Resource of Mammalian Protein Compleses (CORUM) dataset (See Methods for more detailed information). Briefly, we extracted the induced subgraph for each complex from the PPI network by pulling all intra-protein interactions available from the Biological General Repository for Interaction Datasets (BioGRID) database for all proteins in the complex lists provided by the datasets. Only interactions that are classified as "physical" were analyzed to reflect the spatial nature of the interaction, so only "physical" experimental system edges were kept.

Table 1. The average lengths in amino acid (AA) residues for all proteins in planar and nonplanar subgraphs in *S. cerevisiae* datasets CYC2008 and YHTP2008 is given below. This table also includes the absolute value of difference (Δ) in averages between planar and nonplanar protein lengths. An unpaired Wilcoxon Rank Test was performed on the lengths of the proteins in each dataset (planar vs. nonplanar) and the averages are significantly different (p-value<<<0.001).

	Avg. Protein Length (AA)			
Dataset	**Planar**	**Non- planar**	**Δ (AA)**	**P-value**
CYC2008	546.62	463.39	83.24	1.17 E-07
YHTP2008	598.90	520.72	78.18	2.13 E-06

2.1 Protein complexes as a graph tend to be planar

We applied a planarity checking algorithm (see Methods) to the 3,129 validated complex subgraphs from yeast and other model organisms to characterize each one as either "planar" or "nonplanar". We find that 2,619 (83.6%) were planar graphs, and the remaining 510 subgraphs (16.3%) were nonplanar. Further, for each subgraph in our dataset, 100 random graphs with the same number of nodes (n) and edges (m) were also evaluated for planarity. Interestingly, we observe that 99.38% of planar subgraphs maintained their planar quality even when edges were randomly shuffled within their structure. This consistency would imply that the planar nature of the subgraphs is primarily a result of size and density. We hypothesize that this relationship between planarity of an induced subgraph and complex size may be a result of the inherent properties of the interactors in the complex.

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50371-0_48

The length of a protein involved in a planar subgraph is longer on average than the length of a protein involved in a nonplanar complex. Each subgraph used is made up of a list of ORF ids and interactions. There were 506 planar ORFs and 1,415 nonplanar ORFs in the CYC2008 dataset, and 854 planar ORFs and 1,223 nonplanar ORFs in the YHTP2008 dataset. Lengths in AA residues for proteins involved in both planar and nonplanar subgraphs were retrieved from the Saccharomyces Genome Database using their ORF IDs. Lengths of these proteins were compared, and on average, proteins in planar subgraphs tended to be ~78 to 83 AA longer than proteins involved in nonplanar subgraphs (in the YHTP2008 and CYC2008, respectively) as shown in Table 1. The differences in means were found to be significant (p-value <0.001) in both datasets using a Wilcoxon Rank unpaired test. However, it can be argued that any subgraph with $n = 4$ proteins or less will automatically be planar as there is a planar embedding for all iterations and subgraphs of a K_4 graph. When we examine the planarity of only subgraphs with 5 or more nodes in all datasets we find that only 31.22% of subgraphs total are planar (combined dataset, $n = 236$), and the remaining subgraphs ($n = 520$, 68.78%) are not planar, as shown in Table 2. Unfortunately, this result is not significant by a paired t-test (p-value > 0.01) and so does not provide sufficient evidence to speculate on the biological motivation, if any, versus circumstantial or coincidental planar quality of subgraphs. We can speculate, however, that as subgraph size (by node count) grows, it is likely that a subgraph will lose its planar quality, further investigated "Density in validated protein subgraphs in S. cerevisiae" section.

Table 2. Count of valid planar and nonplanar subgraphs in the datasets where the number of nodes is greater than or equal to 5. Planar/nonplanar column refers to those subgraphs labeled as such by our algorithm. Each column has a count for the number of subgraphs characterized as such, and the percent of the total that it represents for that dataset.

Table 2	**Planar**		**NonPlanar**		
Dataset	**Count**	**%**	**Count**	**%**	**Total**
CYC2008	11	10.48%	94	89.52%	105
YHTP2008	18	20.69%	69	79.31%	87
Bovine	1	100.00%	0	0.00%	1
Dog	0	0.00%	0	0.00%	0
Human	156	30.83%	350	69.17%	506
Mouse	36	85.71%	6	14.29%	42
Rabbit	0	0.00%	0	0.00%	0
Rat	14	93.33%	1	6.67%	15
Total	**236**	**31.22%**	**520**	**68.78%**	**756**

2.2 Function of proteins involved in planar & nonplanar subgraphs in yeast

A measured difference in planar versus nonplanar subgraphs leads one to question if the planar quality of subgraphs in these *S. cerevisiae* datasets is biologically motivated, circumstantial, or coincidental. To further probe this question, we annotated the planar and nonplanar datasets for CYC2008 and YHTP2008 using the PANTHER Functional Classification Tool for the GO Biological Process tree, the GO Molecular Function tree, and the PANTHER Protein Class ontology. Here, we report those annotations with a

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50371-0_48

strong representation (>5% of hit against input) and/or annotations which differed (not necessarily significantly) between planar and nonplanar subgraphs. The goal of this exercise was to determine if there were biologically motivated differences on a broader level between proteins involved in planar versus nonplanar subgraphs. We observed that there were specific annotations within each classification that showed differences between planar and nonplanar graphs (Figure 2).

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50371-0_48

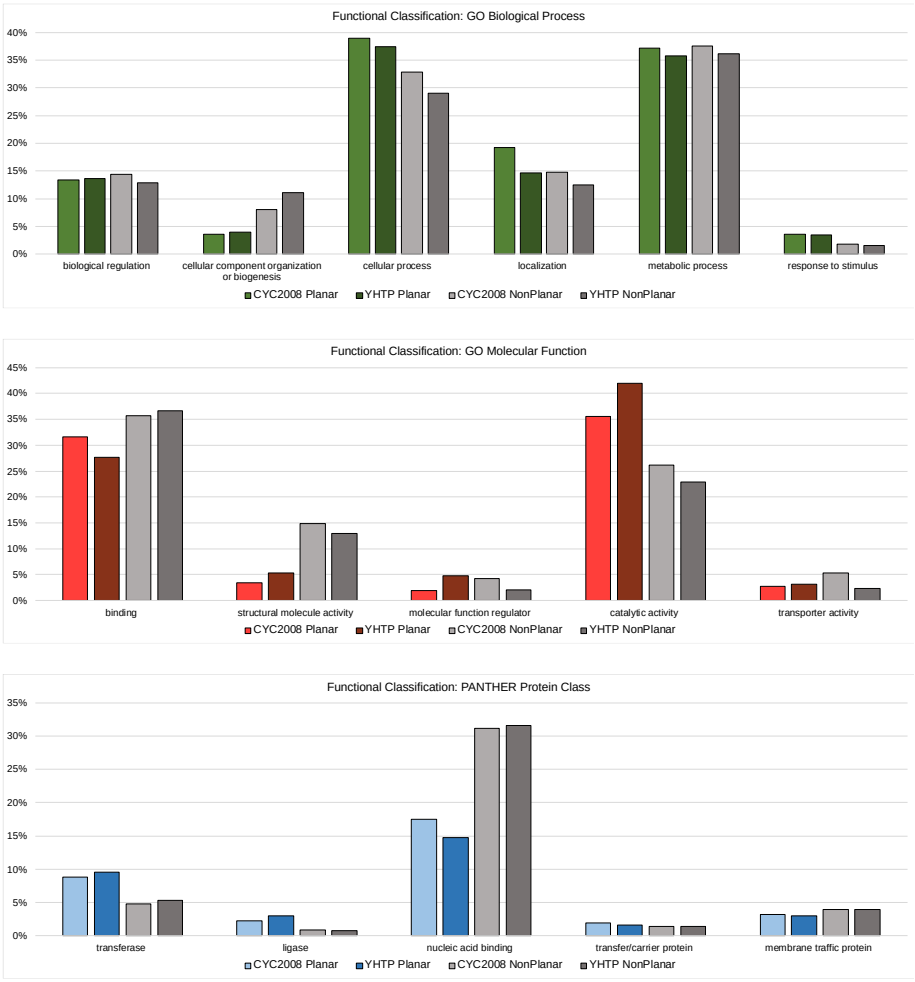

Figure 2. Selected PANTHER functional classification results for GO Biological Process (top), GO Molecular Function (middle), and PANTHER Protein Class (bottom). The x-axis represents the annotation label or name given, and the y-axis represents the % of input against hit, or effectively the number of proteins in the given dataset labeled with that annotation versus the total number of proteins in the dataset.

Specifically, we observe differences in the GO Biological Process result for *cellular component organization or biogenesis*, where planar-involved proteins have a lower representation than nonplanar involved proteins. We also see a minor difference in the GO Biological Process result for *response to stimulus* (GO:0005198), where planar-involved proteins have a higher annotation rate than non-planar-involved proteins. When examining the GO Molecular Function result, there is a larger difference between planar proteins (3.4% and 5.4% for CYC2008 and YHTP2008, respectively) and nonplanar proteins (14.9% and 12.9% for CYC2008 and YHTP2008, respectively) in

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50371-0_48

the *structural molecule activity* annotation. Per the Gene Ontology website, this annotation is defined as "the action of [the] molecule contributes to the structural integrity of a complex or its assembly within or outside of a cell." Interestingly, this would imply that proteins found in nonplanar subgraphs in yeast are more likely to play a role in *structural molecule activity*. We also observe a higher rate of planar proteins annotated with the GO term *catalytic activity* (GO:0003824), (35.5% and 42.0% in planar CYC2008 and YHTP2008 versus 26.2% and 22.9% in nonplanar CYC2008 and YHTP2008, respectively). This annotation is typically given to molecules that catalyze biochemical reactions. Finally, we observe a difference in annotation rates in the PANTHER Protein Class annotation for *nucleic acid binding* (PC00171), with rates of 17.5% and 14.5% for planar CYC2008 and YHTP2008, respectively, compared to 31.1% and 31.6% for nonplanar CYC2008 and YHTP2008. This annotation designates molecules that bind to nucleic acids (i.e. DNA or RNA), which would imply that proteins involved in planar subgraphs are less likely to engage DNA or RNA binding compared to their nonplanar counterparts.

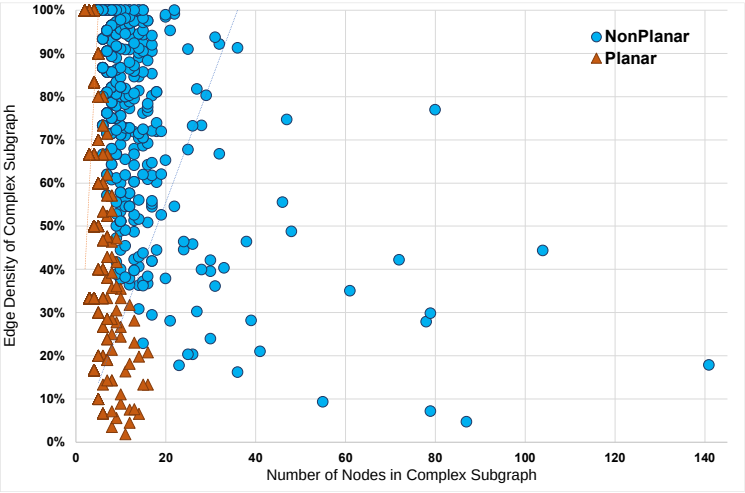

Figure 3. Scatterplots of node count (x-axis) versus edge density (y-axis) of planar and nonplanar complexes for all *S. cerevisiae* and CORUM complexes combined. The plots do not reflect the volume of planar complexes (there are far more planar complexes than nonplanar, but they all have similar node size and density and as such overlap in the graph). Both plots show that the majority of complexes are small - there are no planar complexes beyond n = 16

2.3 Density in validated protein subgraphs in *S. cerevisiae*

When comparing the relative size of each complex, we find that there appears to be a natural boundary for planar subgraphs in terms of node size. In Figure 3, we plot the number of nodes (x-axis) against the edge density (y-axis) for each cluster and include planar and nonplanar labels. Although the number of planar subgraphs far outweighs the number of nonplanar subgraphs, it is apparent in both datasets that the more nodes a subgraph has, the less likely it is to be planar. In both datasets, there are no subgraphs

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50371-0_48

with more than n = 16 nodes that are planar. The average edge densities for all subgraphs with enough entries to measure statistical significance and $n > 4$ nodes is reported in Table 3. We also observe that for both planar and nonplanar subgraphs, there are a not-insignificant number of known, validated subgraphs that have lower than average edge density, which furthers the argument while density should certainly play a role determining complex membership when performing clustering on PPI networks, using it alone will exclude some portion of real complexes in the data.

Table 3. The average edge density of complexes in all evaluated datasets where there are enough complexes to measure statistical significance. Average edge density in complex subgraphs with n = 5 nodes or more only is reported, with associated p-value for a Student's T-test of unequal variance between means between planar and nonplanar complex subgraphs.

	Average Edge Density of Complex Subgraphs		
Dataset	**Non-Planar**	**Planar**	**P-value**
CYC2008	89.72%	77.10%	0.0175
Human (CORUM)	79.33%	49.18%	8.3326E-33
Mouse (CORUM)	72.88%	26.26%	0.0021
YHTP2008	77.86%	63.57%	0.0168

Table 4. Count of valid planar and nonplanar subgraphs in the 3did DDI datasets, where a subgraph consists of individual protein complexes, nodes represent domains with a protein, and edges represent interactions between domains. Each column has a count for the number of subgraphs characterized as such, and the percent of the total that it represents for that dataset

	Planar		**NonPlanar**		
	Count	**%**	**Count**	**%**	**Totals**
All complexes	733	84.64%	133	15.36%	866
Complexes with >4 DDIs	231	63.46%	133	36.54%	364

2.4 Planarity of domain-domain interactions

It could be argued that the planarity or lack thereof in protein subgraphs can be attributed to the domain-domain interactions of the proteins themselves, not the entirety of the protein. Domain-domain interactions (DDIs) are the physical contact points for protein-protein interactions, where one protein component of a complex may have many interactions, and the domains of a protein are where proteins physically interface with themselves and other subcomponents. Therefore, we captured all known and validated DDIs for *S. cerevisiae* from the 3did dataset (https://3did.irbbarcelona.org/) and examined the planarity of known DDI's within a validated RSCB PDB protein complex. We find that regardless of inclusion of 'small' complexes (<=4 DDIs or less), the majority of complexes have DDIs that form planar subgraphs, the opposite of what is found with examining PPI complexes (
Table 4).

We re-examined complex subgraphs from *S. cerevisiae* at the DDI level, identifying 352 PSCB PDB complexes with known DDIs and their corresponding planarity. In this

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50371-0_48

work, the length of the DDI in amino acid residues is measured, and the results show that planar complexes have a longer physically interacting regions (35.4 AA residues on average, $n = 173$) than nonplanar ones (29.9 AA residues on average, $n = 176$). The difference between the means is statistically significant ($p<0.0005$ using an unpaired Student's t-test with unequal variance). These results suggest that on average, planar interactions at the complex level correspond with longer DDI interactions.

3 Methods

3.1 Data download and pre-processing

We chose to begin our study of planarity in PPI networks in *Saccharomyces cerevisiae* due to the extensive body of research on PPIs in the organism itself (Fields & Song, 1989; Ho et al., 2002; Krogan et al., 2006; Schwikowski, Uetz, & Fields, 2000; Uetz et al., 2000; Von Mering et al., 2002a; Von Mering et al., 2002b), including the sentinel paper by Jeong et al. in 2001 examining centrality and essentiality in yeast PPIs (Jeong, Hawoong, Mason, Barabási, & Oltvai, 2001). We used two datasets of protein complexes described by Pu et al 2007, curated through a multi-step procedure of clustering densely connected subunits of the yeast PPI network, and mapping to a high-quality consolidated PPI network (Pu et al., 2007). The result of this work is two catalogs of protein complexes in yeast, the first focusing on literature-validated, heteromeric protein complexes derived from small-scale experimentation (CYC2008, $n = 408$) and the second focusing on complexes derived from high-throughput assays (YHTP2008, $n = 400$) with interactions supported by literature(Pu, Wong, Turner, Cho, & Wodak, 2008). These complexes and their components given as ORF id numbers are available as node lists from http://wodaklab.org/cyc2008 in multiple file formats and were downloaded in September 2018. We also included the Comprehensive Resource of Mammalian Protein Complexes (CORUM) non-redundant complex dataset downloaded on June 27, 2019 from their website https://mips.helmholtz-muenchen.de/corum/#download. This website contains over 4,000 validated protein complexes from *H. sapiens*, *B. Taurus* (bovine), *C. familiaris* (dog), *M. musculus* (mouse), *R. norvegicus* (rat), and *O. cuniculus* (rabbit). The complexes from the Wodak and CORUM datasets were then mapped to their respective protein-protein interaction networks downloaded from BioGRID's August 2018 release (3.4.164,file BIOGRID-ORGANISM-3.4.164.tab.zip) to elucidate their network structure. One subgraph of this PPI was generated for each *S. cerevisiae* protein complex in the YHTP2008 and CYC2008 datasets, and the same was performed for all *H. sapiens* datasets in the *H. sapiens* BioGRID PPI, for the *C. familiaris* (dog) dataset and the *C. familiaris* PPI, and so on. Complex subgraphs were generated in the following manner: First, the set of the proteins involved in each individual complex were extracted from the two complex datasets. Then, for each set of proteins, the interaction network was searched for edges such that both nodes coincident with the edge were in the given complex. Duplicate edges and self-loops were removed from this network before evaluation of planar structure. Edges are undirected. Edges and nodes not explicitly named in the PPI catalogs were removed. The resulting simple subgraph induced by this process was then

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50371-0_48

extracted and stored for further analysis. The result was a total of 3,129 protein subgraphs (as separate connected network components) with nodes and edges as they exist in PPI format.

We also collected domain-domain interactions for proteins for which high-resolution three-dimensional structures are known in *S. cerevisiae* using the 3did database. We downloaded all 4,451 PDB IDs for complexes in yeast on January 16, 2020. The DDI's contained in the 3did dataset were mapped to their PDB ID using Pfam domains. Each PDB ID represented one complex with domains represented as nodes and domain-domain interactions mined from 3did represented edges. These complex subgraphs were then analyzed for planarity by our algorithm implementation.

3.2 Planarity testing

The planarity of the subgraphs was tested using the Boyer and Myrvold planarity test (Boyer *et al.* 2004), an **O**(*n*) planarity test based on embeddings via edge addition and Kuratowski subdivisions. This algorithm returns a result of "True" if the graph *G* given as input is planar and "False" if it is not. In addition to the testing of the planarity of the subgraphs themselves, for each individual subgraph a series of 100 random graphs with the same number of nodes (*n*) and edges (*m*) were also evaluated for planarity. These random graphs were created by generating *m* random edges where the endpoints of each edge were randomly chosen from the set of all *n* nodes. The generated edges were filtered to prevent duplicate edges and self-loops, resulting in *m* unique, unordered pairs of distinct nodes. Our code for checking the planarity of subgraphs has been made available at https://github.com/ndcornelius/complex-graphs.

3.3 Validation for nonplanar subgraphs

The complete interaction datasets for the model organism datasets were downloaded from BioGrid on May 20, 2019 from the version 3.5.172 release archive. ORF ids were used to identify nodes and all other data included was stored as node or edge attributes. We only wanted to investigate physical interactions so we removed any "genetic" Experimental System types. As an example, the *S. cerevisiae* network as downloaded, after removal of genetic interactions, self-loops, multiple edges, and direction, included 6,313 nodes and 110,596 edges (0.56% edge density). There was a total of 17 different types of physical experimental systems included in the BIOGRID filtered network that resulted, and all 17 measure a physical interaction of protein to protein or RNA with varying levels of quality based on experimental system (types available upon request).

3.4 Functional analysis of *S. cerevisiae* subgraphs with >4 nodes

The four sets of ORF ids from both datasets (CYC2008, planar and nonplanar as well as YHTP2008, planar and nonplanar) was analyzed to characterize functionality with the online PANTHER Classification system (version 14.1) using their functional classification tool using the *S. cerevisiae* reference genome. This tool reports a number of measures, including an annotation label or name according to the ontology being used, the accession number of that annotation, and the "% hit against input", or the number of IDs in the input against the total number of IDs in the input. We performed

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50371-0_48

functional classification of all four subsets across five ontologies: GO Biological Process, GO Molecular Function, GO Cellular Component, PANTHER Protein Class, and PANTHER Pathway. In this work, we report those annotations with a strong representation (>5% of hit against input) and/or annotations which differed (not necessarily significantly) from planar to nonplanar subgraphs. The goal of this exercise was to identify any broad functional differences between planar/non-planar complexes quantitatively .

3.5 Comparison of protein length in planar vs. nonplanar subgraphs

Subgraphs were sorted into two types (planar and nonplanar) and gene lists (using ORF as an id) for each complex were generated using in-house Python scripts. Thus, we were able to compile a list of all ORF ids for proteins involved in planar and nonplanar subgraphs for both the CYC2008 and YHTP2008 datasets. We used the *Saccharomyces* Genome Database (www.yeastgenome.org) to pull protein lengths for all ORF ids in planar and nonplanar subgraphs. Average protein lengths (in AA residues) for each group were calculated, and within datasets, length of proteins involved in planar and nonplanar subgraphs were compared using an unpaired Wilcoxon rank-sum test.

4 Discussion

Bioinformatics as a scientific discipline has gone through various stages of maturity in the last few decades. In its next stage, it is anticipated that rigorous validation and verification studies will play significant roles in solidifying major Bioinformatics findings and will increase their impact in advancing biomedical research. The reported work of this paper represents a step in this direction by employing biologically motivated concepts to analyze and measure of the quality of the widely-used biological networks.

Subgraph density has long been a measure of importance when determining the functional potential of a network structure in protein-protein interaction networks. While there is no doubt that density plays a role in finding complexes in protein-protein interaction networks, there are other underlying physical properties of proteins in complex that can be revealed with application of more advanced graph theoretic concepts. In this work, we have applied a planarity checking algorithm to 2 datasets of known PPI complexes in *S. cerevisiae* and found that a majority of protein subgraphs possess this planar property. We have identified a relationship between this planar property that may be linked to physical and spatial constraints of protein interactions at the cellular level and should be investigated with further studies. In the reported results, we find that in the broad majority of planar subgraphs, the planar embedding is not random. We also find that proteins in planar subgraphs tend to be 78-83 amino acid residues longer than proteins in nonplanar subgraphs. We do identify some functional properties of these subgraphs that differ between planar and nonplanar proteins. However, this is a preliminary study and we do realize that further work is needed to determine if this difference is significant. In the future, we plan to expand our research to more high confidence PPI datasets and more model organisms to further confirm our

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50371-0_48

original hypothesis, that *due to the physical nature of protein interactions, protein subgraphs are likely to form planar underlying structures, particularly if their domains or subcomponents are large.*

This research is important for the study of protein-protein interaction networks for several reasons. First, it offers a new structural measure that is readily identifiable from the network structure, without any biological annotation or input. This could allow for the improvement or development of protein complex finding algorithms by uncovering subgraphs that were previously undiscoverable because they were not necessarily dense (for example, having 40% edge density versus 75%), but have this planar component. Secondly, it opens the door to further analysis of structure of domain-domain interactions, a subfield of protein-protein interaction research; we preliminarily find that DDI networks in yeast also maintain this planar component, perhaps even more stringently. Thirdly, it allows for the re-use and re-analysis of existing PPI datasets with the justification that this planar property may reveal previously unknown or partially known protein complexes, opening the door for discovery from our existing community databases. We look forward to expanding our proposed work and investigating further this interesting planar property in PPI networks.

Acknowledgements: Bhowmick's research was supported by NSF CFF #1916084

References

1. Acuner Ozbabacan, S. E., Engin, H. B., Gursoy, A., & Keskin, O. (2011). Transient protein–protein interactions. *Protein Engineering, Design and Selection, 24*(9), 635-648.
2. Altaf-Ul-Amin, M., Shinbo, Y., Mihara, K., Kurokawa, K., & Kanaya, S. (2006). Development and implementation of an algorithm for detection of protein complexes in large interaction networks. *BMC Bioinformatics, 7*(1), 207.
3. Bader, G. D., & Hogue, C. W. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics, 4*, 2.
4. Barabasi, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science (New York, N.Y.), 286*(5439), 509-512. doi:7898 [pii]
5. Barabasi, A. L., & Oltvai, Z. N. (2004). Network biology: Understanding the cell's functional organization. *Nature Reviews.Genetics, 5*(2), 101-113. doi:10.1038/nrg1272
6. Boyer, J. M., & Myrvold, W. J. (2004). On the cutting edge: Simplified O (n) planarity by edge addition. *J.Graph Algorithms Appl., 8*(2), 241-273.
7. Brohee, S., & van Helden, J. (2006). Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics, 7*, 488. doi:10.1186/1471-2105-7-488
8. Brun, C., Herrmann, C., & Guénoche, A. (2004). Clustering proteins from interaction networks for the prediction of cellular functions. *BMC Bioinformatics, 5*(1), 95.
9. Even, S, & Tarjan, RE (1976). Computing an st-numbering. *Theoret Comp Sci, 2*(3), 339-344.
10. Fields, S., & Song, O. (1989). A novel genetic system to detect protein–protein interactions. *Nature, 340*(6230), 245.
11. Heo, M., Maslov, S., & Shakhnovich, E. I. (2010). Protein abundances and interactions coevolve to promote functional complexes while suppressing non-specific binding. *arXiv Preprint arXiv:1007.2668,*
12. Ho, Y., Gruhler, A., Heilbut, A., … Boutilier, K. (2002). Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry. *Nature, 415*(6868), 180.

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50371-0_48

13. Hopcroft, J., & Tarjan, R. (1974). Efficient planarity testing. *Journal of the ACM (JACM), 21*(4), 549-568.
14. Janin, J., Bahadur, R. P., & Chakrabarti, P. (2008). Protein–protein interaction and quaternary structure. *Quarterly Reviews of Biophysics, 41*(2), 133-180.
15. Jeong, H., Mason, S. P., Barabasi, A. L., & Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature, 411*(6833), 41-42. doi:10.1038/35075138 [doi]
16. Jiang, P., & Singh, M. (2010). SPICi: A fast clustering algorithm for large biological networks. *Bioinformatics (Oxford, England), 26*(8), 1105-1111. doi:10.1093/bioinformatics/btq078 [doi]
17. Jones, S., & Thornton, J. M. (1996). Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences, 93*(1), 13-20.
18. Keskin, O., Gursoy, A., Ma, B., & Nussinov, R. (2008). Principles of protein–protein interactions: What are the preferred ways for proteins to interact? *Chemical Revs, 108*(4), 1225-1244.
19. Klein, P. N., & Reif, J. H. (1988). An efficient parallel algorithm for planarity. *Journal of Computer and System Sciences, 37*(2), 190-246.
20. Krogan, N. J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Tikuisis, A. P. (2006). Global landscape of protein complexes in the yeast *S. cerevisiae*. *Nature, 440*(7084), 637.
21. Mehlhorn, K., & Näher, S. (1989). LEDA a library of efficient data types and algorithms. Paper presented at the *Int'l Symposium on Mathematical Foundations of Computer Science*, 88-106.
22. Nepusz, T., Yu, H., & Paccanaro, A. (2012). Detecting overlapping protein complexes in protein-protein interaction networks. *Nature Methods, 9*(5), 471.
23. Pu, S., Vlasblom, J., Emili, A., Greenblatt, J., & Wodak, S. J. (2007). Identifying functional modules in the physical interactome of saccharomyces cerevisiae. *Proteomics, 7*(6), 944-960.
24. Pu, S., Wong, J., Turner, B., Cho, E., & Wodak, S. J. (2008). Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Research, 37*(3), 825-831.
25. Rhrissorrakrai, K., & Gunsalus, K. C. (2011). MINE: Module identification in networks. *BMC Bioinformatics, 12*, 192. doi:10.1186/1471-2105-12-192; 10.1186/1471-2105-12-192
26. Schwikowski, B., Uetz, P., & Fields, S. (2000). A network of protein–protein interactions in yeast. *Nature Biotechnology, 18*(12), 1257.
27. Siek, J., Lumsdaine, A., & Lee, L. (2002). *The boost graph library* Addison-Wesley.
28. Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., … Pochart, P. (2000). A comprehensive analysis of protein–protein interactions in *S. cerevisiae*. *Nature, 403*(6770), 623.
29. Voevodski, K., Teng, S., & Xia, Y. (2009). Finding local communities in protein networks. *BMC Bioinformatics, 10*(1), 297.
30. Von Mering, C., Krause, R., Snel, B., Cornell, M., … Bork, P. (2002a). Comparative assessment of large-scale data sets of protein–protein interactions. *Nature, 417*(6887), 399.
31. Wang, J., Li, M., Chen, J., & Pan, Y. (2011). A fast hierarchical clustering algorithm for functional modules discovery in protein interaction networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 8*(3), 607-620.
32. West, D. B. (1996). *Introduction to graph theory* Prentice hall Upper Saddle River, NJ.
33. Fig 1. Image of 1AXC (Gulbis, J. M., Kelman, Z., … Kuriyan, J. (1996). Structure of the C-terminal region of p21WAF1/CIP1 complexed with human PCNA. *Cell*, *87*(2), 297-306) created with Protein Workshop (J.L. Moreland, A. Gramada, … P.E. Bourne (2005) The Molecular Biology Toolkit (MBT). *BMC Bioinformatics* 6:21).
34. Fig 1. Image of 2HHB (Fermi, G., Perutz, M. F., Shaanan, B., & Fourme, R. (1984). The crystal structure of human deoxyhaemoglobin at 1.74 Å resolution. *Journal of molecular biology*, *175*(2), 159-174.) created with Protein Workshop (J.L. Moreland, A. Gramada, … P.E. Bourne (2005) The Molecular Biology Toolkit (MBT). *BMC Bioinformatics* 6:21).

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50371-0_48

35. Fig. 1 Image of 2HHB (Robinson, R. C., Turbedsky, K., Kaiser, D. A., Marchand, J. B., Higgs, H. N., Choe, S., & Pollard, T. D. (2001). Crystal structure of Arp2/3 complex. *science*, *294*(5547), 1679-1684.) created with Protein Workshop (J.L. Moreland, A. Gramada, … P.E. Bourne (2005) The Molecular Biology Toolkit (MBT). *BMC Bioinformatics* 6:21).

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50371-0_48