OPEN ACCESS Freely available online

PLoS one

Dominating Biological Networks

Tijana Milenković[1], Vesna Memišević[2], Anthony Bonato[3], Nataša Pržulj[4]*

1 Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, Indiana, United States of America, **2** Department of Computer Science, University of California Irvine, Irvine, California, United States of America, **3** Department of Mathematics, Ryerson University, Toronto, Ontario, Canada, **4** Department of Computing, Imperial College London, London, United Kingdom

Abstract

Proteins are essential macromolecules of life that carry out most cellular processes. Since proteins aggregate to perform function, and since protein-protein interaction (PPI) networks model these aggregations, one would expect to uncover new biology from PPI network topology. Hence, using PPI networks to predict protein function and role of protein pathways in disease has received attention. A debate remains open about whether network properties of "biologically central (BC)" genes (i.e., their protein products), such as those involved in aging, cancer, infectious diseases, or signaling and drug-targeted pathways, exhibit some topological centrality compared to the rest of the proteins in the human PPI network. To help resolve this debate, we design new network-based approaches and apply them to get new insight into biological function and disease. We hypothesize that BC genes have a topologically central (TC) role in the human PPI network. We propose two different concepts of topological centrality. We design a new *centrality measure* to capture complex wirings of proteins in the network that identifies as TC those proteins that reside in dense *extended* network neighborhoods. Also, we use the notion of *domination* and find dominating sets (DSs) in the PPI network, i.e., sets of proteins such that every protein is either in the DS or is a neighbor of the DS. Clearly, a DS has a TC role, as it enables efficient communication between different network parts. We find statistically significant enrichment in BC genes of TC nodes and outperform the existing methods indicating that genes involved in key biological processes occupy topologically complex and dense regions of the network and correspond to its "spine" that connects all other network parts and can thus pass cellular signals efficiently throughout the network. To our knowledge, this is the first study that explores domination in the context of PPI networks.

Citation: Milenković T, Memišević V, Bonato A, Pržulj N (2011) Dominating Biological Networks. PLoS ONE 6(8): e23016. doi:10.1371/journal.pone.0023016

Editor: Franca Fraternali, King's College London, United Kingdom

Received February 23, 2011; **Accepted** July 11, 2011; **Published** August 26, 2011

Funding: This work was supported by the National Science Foundation Faculty Early Career Development (CAREER) Program IIS-0644424, the National Science Foundation Cyber-Enabled Discovery and Innovation OIA-1028394, and the Serbian Ministry of Education and Science Project III44006 grants, as well as grants from Natural Sciences and Engineering Research Council and Mathematics of Information Technology and Complex Systems. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: natasha@imperial.ac.uk

Introduction

A *network* (or a *graph*) is a set of *nodes* (or *vertices*), and *edges* (or *links*) between the nodes. Networks enable studying the properties of complex systems that emerge from interactions among individual parts. Hence, networks have been used to model and analyze many real-world phenomena in numerous domains. Examples include social, technological, transportation, information, financial, ecological, chemical, and biological systems. We focus on molecular interaction networks, with the goal of understanding complex cellular functioning by studying cells as inter-connected systems rather than as a collection of individual constituents [1]. Nodes in these networks represent biomolecules, such as genes, proteins, or metabolites, and edges connecting the nodes indicate functional, physical, or chemical interactions between the corresponding biomolecules. Since proteins execute the genetic code and carry out most biological processes, we focus on protein-protein interaction (PPI) networks. In these networks, nodes correspond to proteins and undirected edges represent physical interactions between them.

We have been witnessing the exponential growth of the amounts of available PPI network data, along with the development of computational approaches for studying and modeling of these data. High-throughput screens for interaction detection, such as yeast two-hybrid (Y2H) assays [2–8], affinity purification coupled to mass spectrometry (AP/MS) [9–12], genome-wide chromatin immunoprecipitation, correlated m-RNA expression, and genetic (synthetic-lethal) and suppressor networks [13,14], have yielded partial networks for many model organisms [2–5,11–13] and humans [6,7], as well as for bacterial [15–17] and viral [18–20] pathogens. Numerous biological network datasets are now publicly available in several databases, including Saccharomyces Genome Database (SGD) [21], the Database of Interacting Proteins (DIP) [22], Human Protein Reference Database (HPRD) [23], and the Biological General Repository for Interaction Datasets (BioGRID) [24].

Proteins are essential macromolecules of life, and hence, understanding their function and their role in disease is of importance. Since proteins aggregate to perform a function instead of acting in isolation, and since PPI networks model interactions between proteins, analyzing PPI network topology is expected to uncover new biology. Therefore, it is not surprising that prediction of protein function [25–27] and the role of protein networks in disease [1,28–32] from the topology of PPI networks have received attention in the post-genomic era.

Nonetheless, there is still a debate about whether network properties of "biologically central" genes or proteins, such as those involved in aging, cancer and infectious diseases caused by

bacterial or viral pathogens (*e.g.*, HIV, herpesvirus, hepatitis, and influenza), exhibit some "topological centrality" compared to the rest of the proteins in the PPI network [1,28–31,33–35]. Many approaches have focused on examining only simple topological properties of these proteins, such as their direct neighborhoods in a PPI network. For example, the key assumption of many studies is that proteins that are direct neighbors are more likely to perform the same function than those that are not [25,26], or that a neighbor of a disease-causing gene is likely to cause either the same or a similar disease [1,34]. Another example is the observed correlation between a protein's essentiality and its *degree centrality* (the larger the degree of a node, the more "degree-central" the node) in a PPI network of baker's yeast [36]. However, the controversy arose in the light of newer and more complete PPI network data for which this correlation was not observed [37,38] and it appears to hold only for literature-curated [39] and smaller in scope Y2H PPI networks [3], possibly because these data sets are biased towards essential proteins [38]. Also, degree alone might be a weak measure of network topology, as it captures limited network topology, i.e., only direct neighborhood of a node [27,31,40]. A similar controversy arose when cancer genes were initially shown to have greater connectivities and centralities compared to non-cancer genes, indicating central roles of cancer genes within the interactome [33], but it was later demonstrated that most of disease genes do not show a tendency to code for proteins that are hubs [29], although a recent study again reached the conclusion that cancer proteins have different network topologies, e.g., higher degrees, than "control" genes [35]. Apart from this, general conclusions are that disease genes have high connectivity and are centrally positioned within the PPI network [1]. In addition, it has been suggested that aging genes tend to have higher degrees than non-aging ones [41,42], as well as that the majority of viral and bacterial pathogens show tendency to interact with high-degree proteins, or with "bottleneck" proteins that are central to many paths in the PPI network [43].

Measures of network topology that are more constraining than degrees might help resolve these controversies. Hence, various topological centrality concepts have been formulated. Examples include the *betweenness centrality* [35], according to which nodes that occur in many of the shortest paths in a network have high centrality, and the *subgraph centrality*, which counts the number of closed walks of different lengths in the network starting and ending at the node in question and according to which nodes that participate in a large number of such walks have high centrality [44,45].

In addition, we have recently designed a graphlet-based measure of network topology; graphlets are small *induced* subgraphs of a large network [46,47]. As opposed to *partial* subgraphs (e.g., network *motifs* [48]), graphlets are *induced*, meaning that they contain *all* edges between the nodes of the subgraph that are present in the large network. This measure generalizes the degree of a node that counts the number of edges that the node touches, where an edge is the only 2-node subgraph, into the *graphlet degree vector* (GDV) that counts the number of different graphlets that the node touches, for all 2–5-node graphlets. Hence, GDV of a node describes the topology of its up to 4-deep neighborhood. This is an effective measure: going to distance of 4 around a node captures a large portion of a network due to the small-world nature of many real networks [49]. For this reason, and since the number of graphlets on n nodes increases exponentially with n, we believe that using larger graphlets would unnecessarily increase the computational complexity of the method. We designed the similarity measure between GDVs of different nodes, *GDV-similarity*, to quantify the topological similarity of the extended neighborhoods of two nodes. We used this constraining measure of network topological similarity to demonstrate that: in PPI networks, biological function of a protein and its local network structure are closely related [27,50]; from topology of PPI networks we can extract biological information that cannot always be extracted from sequence and hence, topology could be used as a complementary method to sequence-based methods for homology detection [51]; topology around cancer and non-cancer genes is different and can be used to successfully predict new cancer genes in melanogenesis-related pathways [31,40]; purely topological network alignments can be used to extract protein function and species phylogeny [52,53].

This study

Here, we present novel network-based approaches applied towards a deeper understanding of biological function and disease. We aim to further study and understand currently poorly described mechanisms by which "biologically central" genes interact with each other and with other genes in the cell. We define as *biologically central (BC)* the genes that belong to one of the following four gene *categories*: aging (A) genes, cancer (C) genes, HIV-interacting (HIV) genes, and pathogen-interacting (PI) genes. Our hypothesis is that BC genes, i.e., their protein products (henceforth, we use terms "gene" and "protein" interchangeably), will have a topologically central role in the human PPI network. We use two different concepts to define "topological centrality": *graphlet degree centrality* and *domination* (defined below).

Previously, we defined GDV-similarity of nodes' neighborhoods that is independent of the densities of these neighborhoods: nodes with identical GDVs have the maximum GDV-similarity, regardless of whether they reside in dense or sparse neighborhoods. Here, we propose a new centrality measure, *graphlet degree centrality (GDC)*, to measure the density and complexity of nodes' neighborhoods by counting the number of different graphlets that the node touches. According to GDC, nodes in dense and complex 4-deep neighborhoods will have higher centralities than nodes in sparse 4-deep neighborhoods. GDC is a different and more constraining measure of network topology than the degree centrality (DC), as illustrated in Figure 1: GDC ranks highly a low-degree gene if its 4-deep neighborhood is dense and gives a low rank to a high-degree gene if its 4-deep neighborhood is sparse (details are below). GDC is conceptually different than the betweenness centrality (BWC), which does not measure topological denseness at all. Subgraph centrality (SC) measures the number of closed walks (which can be thought of as partial subgraphs) that the node touches and it has been shown to be more highly correlated with the lethality of proteins in the PPI network of baker's yeast than DC [44]. Unlike SC, GDC counts induced subgraphs rather than partial ones and in a more rigorous way: while SC counts an edge that a node touches many times, as a 2-edge closed walk (going from node A to node B along edge AB and returning from B to A along the same edge), as a 4-edge closed walk (going from node A to node B and back to A twice), as a 6-edge closed walk (going from A to B and back to A three times) etc., GDC counts the edge only once and only as an edge, rather than as different subgraph structures.

For each of the four centrality measures (DC, BWC, SC, and GDC), we identify the most central genes (explained below) in the human PPI network [28] and measure the *enrichment* of these genes in BC genes (i.e., the percentage of the most central genes that are BC genes), with the goal of finding the centrality measure that is the most discriminative in uncovering BC genes; ideally, the most discriminative measure would have all of the most central genes to be BC genes. We find that: (1) enrichments in BC genes of the

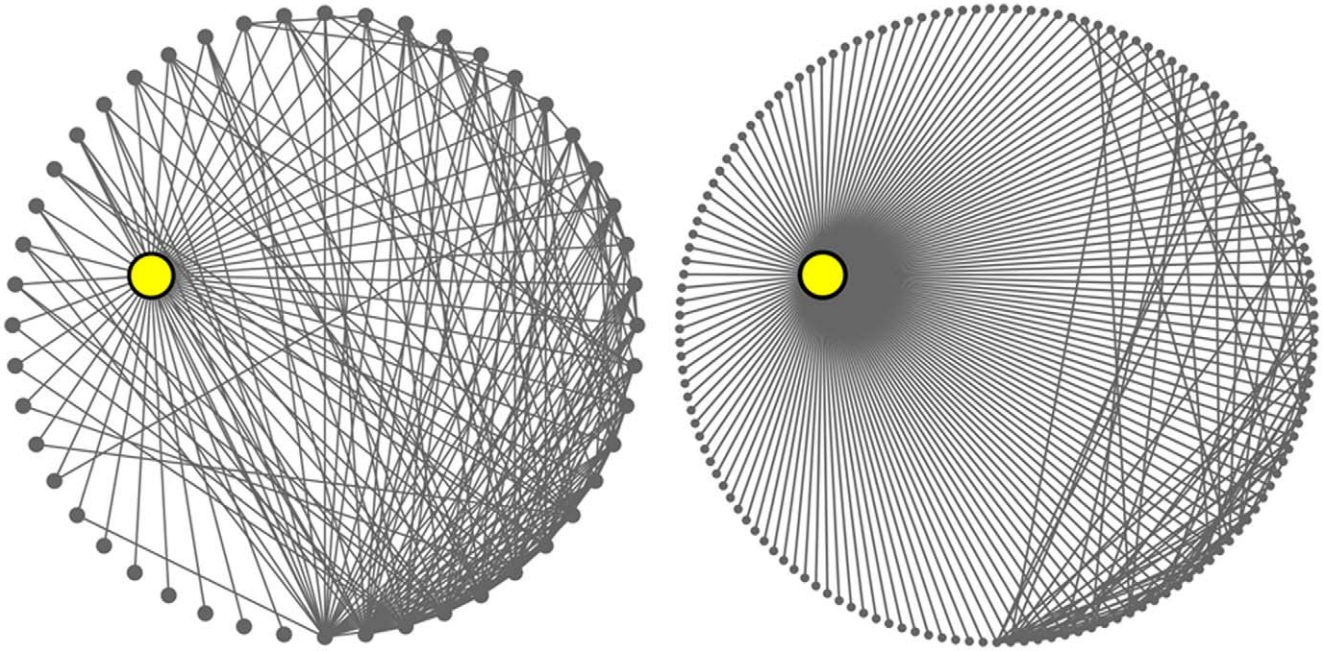


Figure 1. An illustration of the differences between DC and GDC. *Left:* Direct neighborhood of ZAP90, a cancer and HIV gene, in the human PPI network [28]. Its degree is 48 and it is ranked as the top 187^{th} gene with respect to DC. *Right:* Direct neighborhood of PRKACA, an HIV gene, in the network. Its degree is 145 and it is ranked as the top 20^{th} gene with respect to DC. Both proteins have the same GDC and are ranked as top 92^{nd} genes with respect to GDC. Hence, GDC rewards the ranking of a low-degree gene if its 4-deep neighborhood is dense (ZAP90) and penalizes the ranking of a high-degree gene if its 4-deep neighborhood is sparse (PRKACA). (For the esthetics of the figure, we only show 1-deep neighborhoods.)
doi:10.1371/journal.pone.0023016.g001

most GDC-central genes are much higher than those of non-GDC-central genes, (2) the observed enrichments in BC genes of the most GDC-central genes are statistically significant, while those of non-GDC-central genes are not, (3) BC genes that are GDC-central have higher and statistically significant enrichments in known drug targets than BC genes that are non-GDC-central, and (4) GDC is at least as discriminative as the next best centrality measure.

Second, we hypothesize that genes that are vital for normal cellular functioning might correspond to the "spine" of the network that connects all parts of the network. The field of telecommunications and the domain of the efficient design of routing protocols for wireless networks in particular, uses the notion of a *dominating set* (DS) to find the most central set of nodes in wireless networks that would be used for efficient data routing and lead to bandwidth increase and energy savings; in wireless networks, nodes correspond to computers and routers, and edges correspond to links between them [54–57]. A dominating set of a network is a set of nodes such that every node in the network is either in the DS or is a direct neighbor of a node in the DS. Hence, the nodes in the dominating set act as a "gateway" in the network, since all nodes in the network are at most one step away from them and the transfer of the information to all nodes can be quick and cheap. The challenge is to identify a minimum order DS, a DS of the minimum size (i.e., the minimum number of nodes). This problem is NP-hard. Thus, approximate (heuristic) algorithms are sought.

Given the topologically central role of nodes in a DS, we hypothesize that a good DS algorithm might capture a set of proteins in a PPI network that are involved in important biological processes and mechanisms crucial for cell vitality, i.e., that DSs of PPI networks might contain BC proteins and signaling pathways (SPs). We test this by constructing a connected dominating set in the human PPI network with an algorithm that is commonly used in telecommunications [57]. We are interested in connected DSs only since signaling pathways are connected. Other algorithms for finding connected DSs are used in telecommunications as well (e.g., [54,56,58,59]), but are not applicable to biological networks, because they require nodes to be assigned meaningful numerical IDs, e.g., IP addresses in computer networks; clearly, proteins in PPI networks do not have numerically meaningful labels. Also, several algorithms for finding disconnected (i.e., independent; see Methods) DSs exist [60,61], but they are inappropriate for our study for the above mentioned reasons. In addition to applying the existing DS algorithm of Rai *et al.* [57], we design a new and simpler DS algorithm that outperforms the algorithm of Rai *et al.* on our data (explained below). Note that the main focus of this study is not to create a state-of-the-art algorithm for finding DSs, but instead, to demonstrate, as a proof of concept, that a DS of a PPI network found by a very simple algorithm indeed captures biologically vital proteins. Any further algorithmic improvements are likely to yield more optimal DSs and hence improve the biological results.

We apply DS algorithms to the human PPI network [28] and measure the size of the resulting DSs, as well as their enrichments in BC and SP genes. We find that: (1) the enrichments in BC and SP genes of nodes of DSs are much higher than the enrichments of nodes outside of DSs; (2) the enrichments in BC and SP genes of nodes of DSs are statistically significant, while those of nodes outside of DSs are not; and (3) BC and SP genes that are in DSs have much higher and statistically significant enrichments in known drug targets than BC and SP genes that are not in DSs. Hence, we confirm our hypothesis that DSs capture biologically vital proteins and also drug targets.

Furthermore, we demonstrate not only that each of the two measures of topological centrality, GDC and DS, captures a

statistically significant biological signal, i.e., BC and drug target genes (as described above), but also that the combination of the two centralities is even more discriminative in capturing these genes. To our knowledge, this is the first study that uses dominating sets to analyze PPI networks.

Methods

Data sets

We analyze the human PPI network of Radivojac *et al.* that contains 41,456 physical interactions between 9,141 proteins [28], as well as the human PPI networks from BioGRID [24], that contains 30,513 physical interactions between 8,581 proteins, and from HPRD [23], that contains 36,811 physical interactions between 9,449 proteins (we downloaded them in June 2010). Since we obtained qualitatively similar results for all three networks, for simplicity we report only on the PPI network of Radivojac *et al.* [28]; we chose this network, since it has the largest number of interactions.

As mentioned above, *biologically central (BC)* genes that we analyze include: aging, cancer, HIV, and pathogen-interacting genes. We obtained them from the following databases. *Aging genes (A)* are human genes implicated in the process of aging that are available from AnAge Databank - Human Ageing Genomic Resources (http://genomics.senescence.info/) [62]. *Cancer genes (C)* are human genes implicated in cancer that are available from: Cancer Gene Database (http://ncicb.nci.nih.gov/projects/cgdcp), Cancer Genome Project – the Cancer Gene Census (http://www.sanger.ac.uk/genetics/CGP/Census/) [63], GeneCards (http://www.genecards.org/) [64], Kyoto Encyclopedia of Genes and Genomes (KEGG) (http://www.genome.jp/kegg/disease/) [65], and Online Mendelian Inheritance in Man (OMIM) (http://www.ncbi.nlm.nih.gov/sites/entrez?db = omim) [66]. *HIV genes (HIV)* are human genes known to interact with genes of the HIV virus [63] that are available from HIV-1-Human Protein Interaction Database (http://www.ncbi.nlm.nih.gov/RefSeq/HIVInteractions/) [67]. Finally, *pathogen-interacting genes (PI)* are human genes known to interact with genes of pathogens [43]. The data are downloaded in 2009 and 2010.

In the human PPI network, there are 2,101 BC genes in total, of which 237 are aging genes, 887 are cancer genes, 1,132 are HIV genes, and 500 are PI genes. Figure 2 illustrates the overlap of different BC gene categories in the network. The overlap is low and there are only 20 BC genes that are simultaneously aging, cancer, HIV, and PI genes.

Signaling pathways (SPs) that we analyze include the human: MAP kinase interactome [68], cancer and immune pathways from NetPath [69], and all human signaling pathways from KEGG [65]. The data are downloaded in November 2010. In the PPI network, there are 2,253 SP genes, 911 of which are also BC genes. Given that there is a total of 2,101 BC genes in the network, the total number of BC and SP genes together is $2253+2101-911=3443$.

The drug target data was downloaded from DrugBank [70].

Centrality measures

Related work. Several notions of node centrality have been used in the past. *Degree centrality (DC)* of a node is the number of its neighbors, i.e., its degree. Alternatively, DC can be normalized by dividing the degree with $n-1$, where n is the number of nodes in the network. *Betweenness centrality (BWC)* of a node is the sum, over all node pairs i and j in the network, of the percentage of all shortest paths between i and j in the network that go through the node of interest. *Subgraph centrality (SC)* of a node is a weighted sum

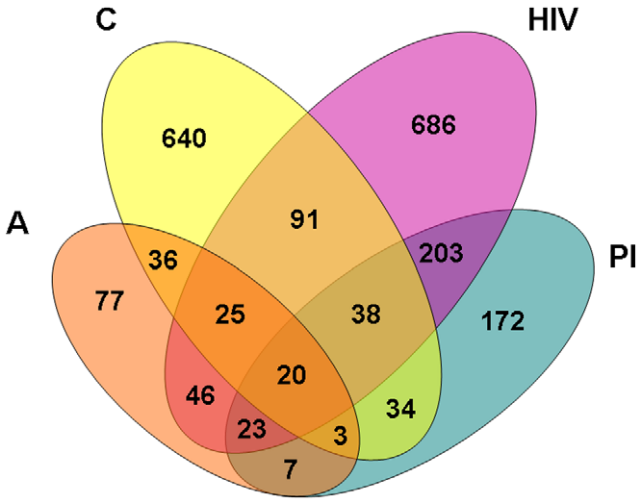
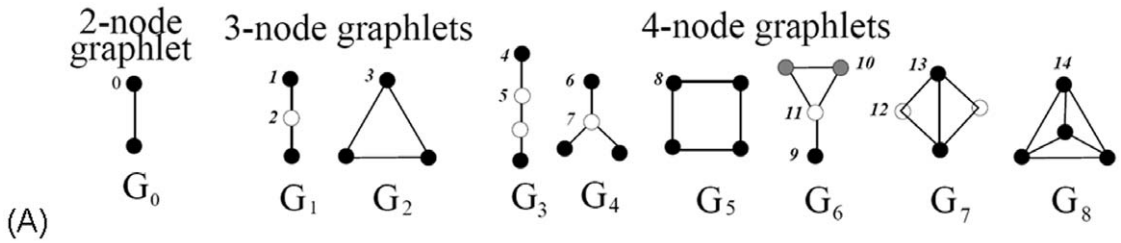


Figure 2. The overlap of BC genes from the four categories in the human PPI network.
doi:10.1371/journal.pone.0023016.g002

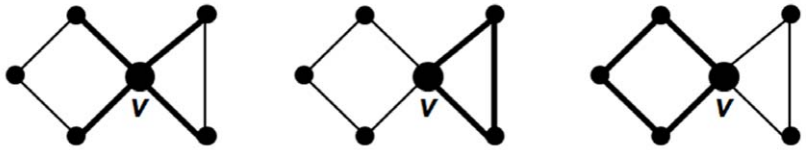
of the numbers of all closed walks of different lengths in the network starting and ending at the node. These closed walks are related to partial subgraphs of a network, e.g., a closed walk with four nodes can "go through" different subgraphs on four nodes, such as along the same edge AB twice (as described above: from node A to node B along edge AB, then back to A along the same edge and then again from A to B and back to A along the same edge), or along a 4-node cycle ABCD that includes edge AB (along the "square" from node A to node B to node C to node D and back to A; this is regardless of whether edges CA and DB that "go along the diagonal of the square" exist) etc. The above mentioned sum is weighted so that the contribution of the closed walks decreases as the length of the walks increases, i.e., shorter walks (smaller subgraphs) have higher weight.

Graphlet degree centrality. We introduce a new node centrality measure as follows. *Graphlets* are small, connected, induced, non-isomorphic subgraphs of a large network (Figure 3 A) [46,47]. Previously, we generalized the degree of a node, that counted how many edges the node touched, into the *graphlet degree vector (GDV)*, that counted how many graphlets of a given type, such as a triangle or a square, the node touched (Figure 3 B) [27]. In Figure 3 B, this is illustrated by a node being touched by an edge (the leftmost illustration), a triangle (the middle illustration), or a square (the rightmost illustration). More precisely, coordinates of a GDV count how many times a node is touched by a particular symmetry group (*automorphism orbit*, see [47] for details) within a graphlet (Figure 3 B). Clearly, the degree of a node is the first coordinate in GDV, since an edge is the only 2-node graphlet. There is a total of 73 orbits in all 2–5-node graphlets. Thus, the GDV of a node, describing its up to 4-deep neighborhood (i.e., 2–5-node graphlets around it), has 73 coordinates [27]. An example of a GDV of a node that contains all 73 orbits can be found in [52].

We introduce a new node centrality measure, *graphlet degree centrality (GDC)*, which measures the density of the node's extended network neighborhood. Hence, nodes that reside in dense extended network neighborhoods will have higher GDCs than nodes that reside in sparse extended network neighborhoods. In particular, we define GDC as follows. For a node v, we denote by v_i the i^{th} coordinate of its GDV, i.e., v_i is the number of times node v touches an orbit i. Then, GDC of node v is computed as follows:



(A)



Orbit	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
GDV (*v*)	4	2	5	1	0	4	0	2	1	0	0	2	0	0	0

(B)

Figure 3. Graphlets, automorphism orbits, and GDVs. (**A**) All 9 graphlets with 2, 3 and 4 nodes, denoted by G_0, G_1,…,G_8; they contain 15 topologically unique node types, called automorphism orbits, denoted by 0, 1, 2, …, 14. In a particular graphlet, nodes belonging to the same orbit are of the same shade (see [47] for details). (**B**) An illustration of the GDV of node *v*; it is presented in the table for orbits 0 to 14: *v* is touched by 4 edges (orbit 0), end-nodes of 2 graphlets G_1 (orbit 1), etc. The figure is taken from [53].
doi:10.1371/journal.pone.0023016.g003

$$GDC(v) = \sum_{i=0}^{72} w_i \times log(v_i + 1),$$

where w_i is the weight of orbit i that accounts for dependencies between orbits, as in [27]; e.g., counts of orbit 3, a triangle, will affect counts of all orbits that contain a triangle. Hence, for each orbit, we count how many orbits affect it and assign a higher weight w_i ($w_i \in [0,1]$) to the orbits that are not affected by many other orbits (see [27] for details). We use *log* in the formula because the coordinates i and j of the GDV of node v can differ by several orders of magnitude and we do not want the GDC to be entirely dominated by orbits with very large values. We add 1 to v_i in the formula to prevent the logarithm function to go to infinity for an orbit count of 0. Finally, we scale the value of the $GDC(v)$ to (0,1] by dividing it with the maximum $GDC(u)$ over all nodes u in the network.

Algorithms for finding dominating sets

Let $G(V,E)$ be a network, where V is the set of nodes of G and E is the set of edges of G. A *dominating set (DS)* of graph G is a subset $S \subseteq V$ of the nodes such that for all nodes $v \in V$, either $v \in S$ or a neighbor u of v is in S. A dominating set is said to be *minimal* if it contains no proper subset that is dominating and it is said to be *minimum* if it is of the smallest cardinality. The cardinality of a *minimum* dominating set of graph G, $\gamma(G)$, is called the *domination number* of G. It has been shown that for graph G with $|V|$ nodes:

$$\lceil \frac{|V|}{1+\delta_{max}(G)} \rceil \leq \gamma(G) \leq |V| - \delta_{max}(G), \qquad (1)$$

where $\delta_{max}(G)$ is the maximum node degree in G [60]. Identifying a minimum DS is NP-hard, and hence, approximate (heuristic) algorithms are sought.

Heuristic algorithms result in either an independent DS or a connected DS. A subset P of V is said to be an *independent set* if no two vertices in P are adjacent. A connected DS is a DS in which each node is connected to at least one other node that is in the DS. (Note that if a graph consists of several connected components, a DS of such a graph would be connected within each component, but disconnected across components.) In the context of biological networks, we are interested in connected DSs.

First, we implement an existing algorithm by Rai *et al.* for constructing a connected DS of graph $G(V,E)$ that is commonly used in telecommunications [57]. We call this algorithm "DS-RAI". It consists of three phases: (1) constructing an independent DS named S, (2) finding a set of nodes $C \subseteq V \backslash S$ to connect nodes in S by constructing the Steiner tree between the nodes in S, and (3) pruning the DS defined on nodes $S \cup C$ to reduce the number of nodes in the DS. More specifically, the algorithm works as follows. In phase 1, each node is colored white. A white node u that is connected to most other white nodes is taken from V, colored black meaning that it is a "dominator," and added to S. All neighboring nodes of u are colored gray meaning that they are "dominatees" and added to $V \backslash S$. Previous steps are repeated on the remaining white nodes in V until all nodes of V are either colored black and added to S, or colored gray and added to $V \backslash S$. In phase 2, a gray node from $V \backslash S$ that is connected to the largest number of black nodes in S is selected, colored dark gray meaning it is a "connector," and added to C. The algorithm then checks whether node set $S \cup C$ is connected and if so, it stops; otherwise, the algorithm selects the next gray node from $V \backslash S$ that is connected to the largest number of black nodes in S and repeats the entire process until node set $S \cup C$ becomes connected. In phase 3, "redundant" nodes are deleted from the connected DS defined on $S \cup C$ to reduce its size as follows. Let $G[V']$ denote a subgraph of G induced on a subset of nodes $V' \subseteq V$. The algorithm selects a node u with the minimum degree in $G[S \cup C]$ and checks whether the DS defined on $S \cup C \backslash \{u\}$ remains a

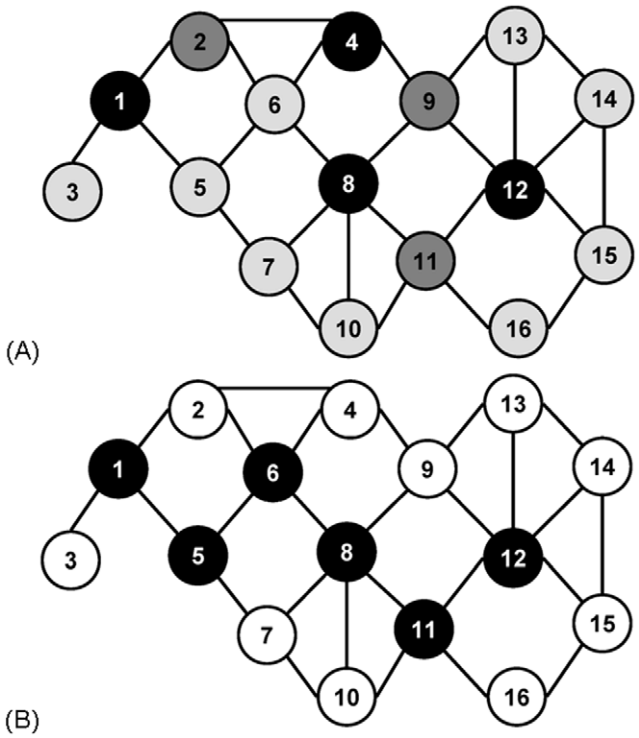


Figure 4. An illustration of DSs in a toy network. The DSs were obtained by (**A**) DS-RAI and (**B**) DS-DC algorithms. The example in panel A is taken from [57], and the authors describe the algorithm as follows. In phase 1, nodes 1, 4, 8, 12, and 16 are colored black as members of an independent DS. In phase 2, nodes 2, 9, and 11 are colored dark grey as connectors that connect nodes in the independent DS resulting from phase 1. In phase 3, the connected DS resulting from phase 2 is pruned to reduce it size by removing node 16 from the DS (no other nodes can be removed without violating the requirement of producing a connected DS of the graph). In panel B, all nodes are initially in the DS and then nodes are visited in order of their increasing degrees and removed from the DS if the resulting DS is a valid connected DS of the graph. That is, nodes are removed in the following order: 3, 16, 2, 4, 7, 10, 13, 14, 15, and 9. The resulting DS therefore contains the remaining nodes: 1, 5, 6, 8, 11, and 12. Clearly, the DS produced by DS-DC (black nodes in panel B) is smaller than the DS produced by DS-RAI (black and dark grey nodes in panel A).
doi:10.1371/journal.pone.0023016.g004

connected DS of G. If so, the node u is removed from $S \cup C$. Otherwise, it remains in $S \cup C$. This is repeated for all nodes in $S \cup C$, in the order of their increasing degrees. The node set resulting from node removals from $S \cup C$ in step 3 is the final DS produced by DS-RAI algorithm. An illustration is presented in Figure 4 A.

The algorithm breaks all ties uniformly at random. Interestingly, the algorithm is robust to this randomness: we run the algorithm on the human PPI network 30 times using different random seeds, which results in 94.2% overlap between the resulting 30 DSs. The average DS size over the 30 runs is $1{,}817 \pm 1$ nodes, out of which 1,711 (i.e., 94.2%) appear in all of the 30 DSs. Hence, given that such a large proportion of any DS is in all DSs, any DS is representative of all of them. Therefore, we continue further analyses of one of the DSs.

Next, we introduce a new, simple, one-step algorithm for constructing a connected DS, that we call "DS-DC": it starts with $S = V$, selects a node u with the minimum degree in $G[S]$, removes u from S only if the DS defined on $S \backslash \{u\}$ remains a connected DS of G, and repeats the above steps for all nodes in S in order of their increasing degrees. An illustration is presented in Figure 4 B. Clearly, DS-DC is much simpler than DS-RAI. Also, as illustrated in Figure 4, DS-DC results in a smaller DS than DS-RAI (the same holds for real-world PPI networks, as demonstrated in Section 0). Finally, we introduce a modification of DS-DC in which nodes from S are visited in order of their increasing GDCs instead of degrees, which we call "DS-GDC" algorithm.

Statistical significance of enrichments

For a given protein set X of size $|X|$, we measure its enrichment in BC (and SP) genes. We compute the statistical significance (p-value) of observing a given enrichment by measuring the probability that the same enrichment would be observed in a randomly chosen set of $|X|$ proteins in the PPI network. This probability is computed as follows by using the following notation: the total number of proteins in the network is $|V|$; the number of proteins in set X is $|X|$; the number of proteins in set X that are BC (SP) genes is $|f|$; there are $|F|$ proteins in the entire PPI network that are BC (SP) genes. Then, the enrichment is $|f|/|X|$, and the ***p*-value**, i.e., the probability of observing the same or higher enrichment purely by chance, is obtained by using the hypergeometric distribution formula for sampling without replacement:

$$p-value = 1 - \sum_{i=0}^{|f|-1} \frac{\binom{|F|}{i}\binom{|V|-|F|}{|X|-i}}{\binom{|V|}{|X|}}. \quad (2)$$

Results and Discussion

GDC captures BC genes

For each of the four centralities (DC, BWC, SC, and GDC) and each of the four categories of BC genes (A, C, HIV, and PI), we find in the human PPI network the top k% of the most central genes ($k = 1,2,3,...,100$%) and measure how many BC genes they contain. For example, we measure how many cancer genes (C) are in the top 1%, the top 2%, the top 3% etc. most central genes with respect to each of the four centrality measures. We do the same for aging (A), HIV, and PI genes. For a given centrality measure, BC gene category, and k, we quantify the accuracy of the centrality measure in capturing BC genes by computing precision and recall. Precision can be seen as a measure of exactness: it is the percentage of the top k% of the most central genes that are BC genes. Recall can be seen as a measure of completeness: it is the percentage of BC genes of the network that are in the top k% of the most central genes. We need to determine a threshold for k that results in the best combination of precision and recall. Since when varying the values of k, every decrease in precision corresponds to increase in recall, we choose as the threshold for k the point where precision and recall cross (Figure 5). We do this for each of the four centrality measures and each of the four BC gene categories. If the threshold is found to be K, we denote as "central" those genes that are amongst the top K% of the most central genes and as "non-central" all the remaining genes in the network. We find that the thresholds are 3, 10, 12, and 6, for A, C, HIV, and PI genes, respectively, for each of the four centrality measures.

We compute the BC gene enrichments of central and non-central genes. We find that with respect to GDC, enrichments in each of the four BC gene categories are much higher for central genes, ranging between 23.5% and 36.4%, than enrichments for

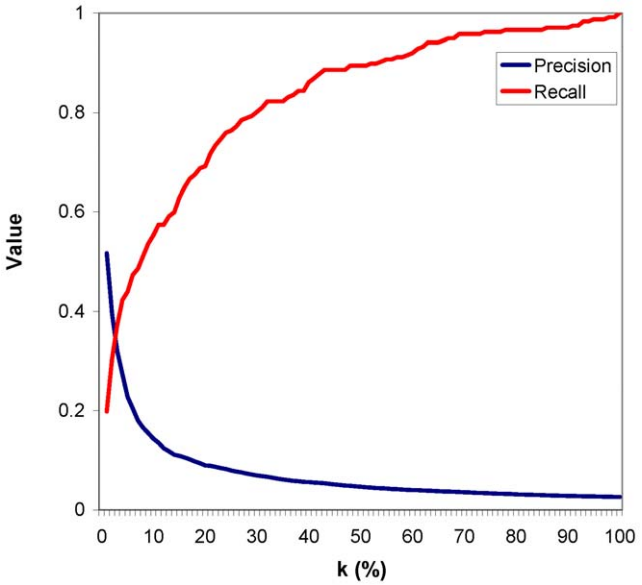


Figure 5. Precision and recall for aging genes in the human PPI network. They were computed for the top k% of the most GDC-central genes ($k = 1,2,...,100$). Here, precision and recall cross at $k = 3$.
doi:10.1371/journal.pone.0023016.g005

non-central genes, ranging between 1.6% and 9.5% (Figure 6 A). These enrichments are statistically significant for central genes, with p-values $\leq 10^{-11}$, while for non-central genes they are not, with p-values $= 1$ (see Methods). As expected, if we choose lower k, e.g., 1%, precision is even higher (although recall is lower): out of the top $1\% = 91$ of the most GDC-central proteins in the network, 55% (i.e., 47 of them) are aging genes, 45% (i.e., 41 of them) are cancer genes, 71.5% (i.e., 65 of them) are HIV genes, and 42.9% (i.e., 39 of them) are PI genes (Figure 7).

Also, we measure the enrichment in drug targets of BC genes (i.e., of each of the four BC gene categories: "A", "C", "HIV", and "PI" defined above) that are GDC-central and of BC genes that are non-GDC-central. We hypothesize that higher GDC of nodes in the PPI network reflects their functional importance. Proteins that are targeted by drugs are clearly functionally important. Hence, we examine whether the sets of BC genes that are GDC-central contain more drug targets than the sets of BC genes that are non-GDC-central. Indeed, we find that enrichments in drug targets are higher for BC genes that are GDC-central than for BC genes that are non-GDC-central (Figure 6 B). Furthermore, these enrichments in drug targets are statistically significant for GDC-central BC genes (with the exception of GDC-central HIV genes), with p-values ≤ 0.047, while for non-GDC-central BC genes they are not, with p-values ≥ 0.9 (see Methods).

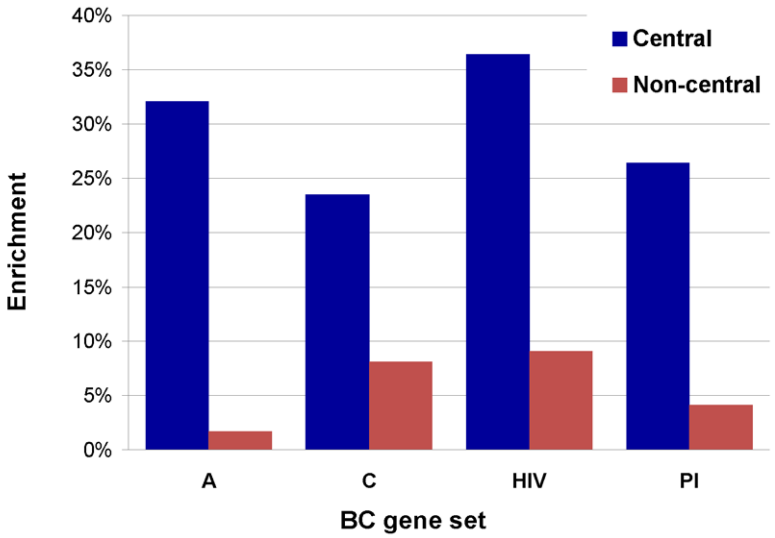
In addition to the above demonstration that GDC captures statistically significant biological signal, we compare its performance against the performance of the three other centrality measures (DC, BWC, and SC). We do so by determining which measure is the most discriminative in the sense that it uncovers the largest number of BC genes amongst the top K% of the most central genes (K is computed as above) and hence results in the highest enrichments. As shown in Figure 6 C, GDC is at least as good as other centrality measures for all categories of BC genes, except for cancer genes, for which SC has a slightly higher enrichment, but GDC still outperforms DC and BC. GDC always outperforms DC, confirming our hypothesis that GDC, as a more constraining measure of network topology, could capture the biological signal better. SC also outperforms DC for aging genes, but interestingly not for HIV and PI genes. Hence, although GDC and SC both capture deeper network topology than DC and are conceptually similar in the sense that they both count a number of subgraphs that a node participates in, unlike GDC, SC is not always more discriminative than DC.

To evaluate whether GDC captures statistically significant biological signal and outperforms other centrality measures irrespective of the chosen thresholds k, for each centrality measure, we compute the area under precision-recall curve (AUPR) as the threshold is varied between 0% and 100% in increments of 1%. The results obtained from AUPRs corresponding to different centrality measures are mostly consistent with the results obtained at selected thresholds where precision and recall cross (described above): for HIV and PI genes, AUPRs for GDC are the highest, followed by AUPRs for DC, SC, and BWC, respectively; for A and C genes, AUPRs for SC are the highest, followed by AUPRs for GDC, DC, and BWC, respectively. Hence, as was the case for individual thresholds (see above), GDC always outperforms DC, while SC outperforms DC only for A and C genes. Hence, GDC is always more discriminative than DC, while SC is not always more discriminative than DC, even though SC captures a deeper network topology compared to DC. The values of AUPRs for GDC are: 0.27 for A, 0.2 for C, 0.34 for HIV, and 0.2 for PI genes. These somewhat law values are not surprising, since in biological applications, the number of positive examples (here, the known BC genes) is much smaller than the number of negative examples (here, all proteins in the network that are currently not known to be BC genes). Furthermore, we do not know true negatives (genes that are true non-BC genes). Since we expect that many currently unreported BC genes will turn out to be BC genes in the future, AUPRs are likely to increase as this happens. Moreover, the observed AUPRs are statistically significant: we compute, at each value of recall, the probability of observing a given precision and we find that the probabilities of observing a given number of BC genes among k% of randomly chosen genes are in the range $0.03 - 10^{-13}$ for k up to 90% (clearly, for k close to 100%, results become statistically insignificant, which is expected, since we choose as GDC-central all genes in the network).

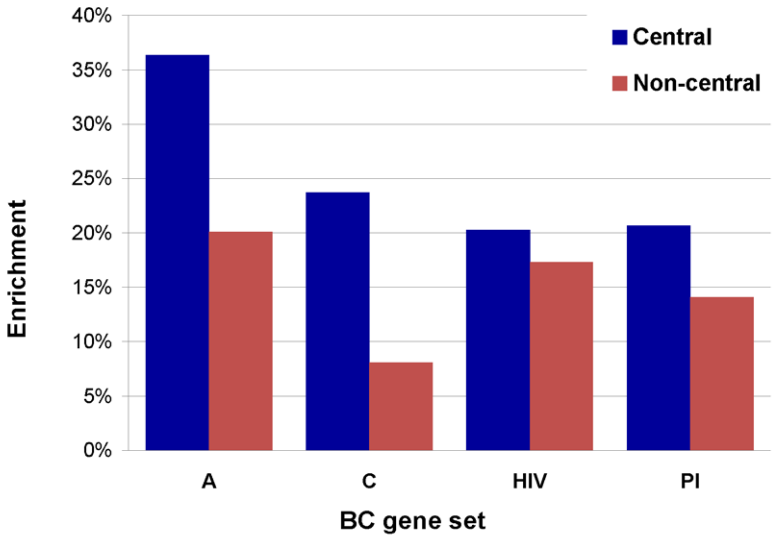
Dominating sets capture BC genes, signaling pathways, and drug targets

We find DSs in the human PPI network by using the three DS algorithms described above, DS-RAI, DS-DC, and DS-GC (see Methods). We find that the overlap between the three resulting DSs is large, containing 1,720 nodes, out of the total of 1,834 nodes in DS-RAI, 1,815 nodes in DS-DC, and 1,828 nodes in DS-GC DSs (Figure 8). Both of our algorithms, DS-DC and DS-GDC, produce smaller DSs than DS-RAI. Also, each of them produces a DS that captures a huge portion of the DS produced by DS-RAI. Using GDC to guide our algorithm does not seem to result in a smaller DS then when we use DC and thus, we continue our analysis on the DS created by DS-DC.

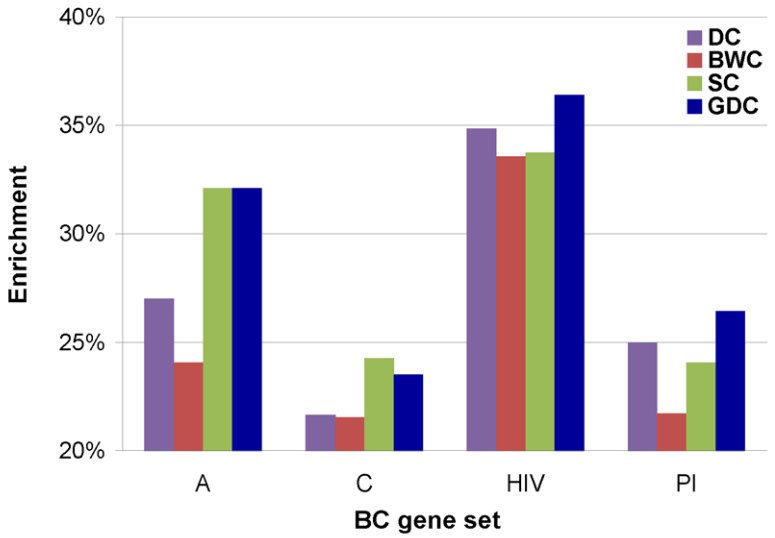
For the DS created by DS-DC algorithm and for its complement (the set of proteins in the network that are not in the DS, "non-DS"), we calculate their enrichments in BC genes, genes that are members of signaling pathways (SP), genes that are in the union of BC and SP genes ("BC or SP"), and genes that are both BC and SP genes ("BC and SP"). We find that the enrichments are much higher for the DS than for non-DS (Figure 9 A). Furthermore, the enrichments for the DS are statistically significant, with p-values $\leq 10^{-11}$, while for non-DS they are not, with p-values $= 1$ (see Methods).


40%
35%
30%
25%
20%
15%
10%
5%
0%
Enrichment
Central
Non-central
A
C
HIV
PI
BC gene set

(A)


40%
35%
30%
25%
20%
15%
10%
5%
0%
Enrichment
Central
Non-central
A
C
HIV
PI
BC gene set

(B)


40%
35%
30%
25%
20%
Enrichment
DC
BWC
SC
GDC
A
C
HIV
PI
BC gene set

(C)

Figure 6. The performance of GDC and its comparison with other centrality measures. (**A**) Enrichments in BC genes of the top k% of the most GDC-central genes (denoted by "Central", blue bars) and all remaining genes (denoted by "Non-central", red bars) in the human PPI network. (**B**) Enrichment in drug targets of BC genes that are GDC-central ("Central") and BC genes that are non-GDC-central ("Non-central"). (**C**) Enrichments in BC genes of the top k% of the most central genes in the human PPI network, with respect to the four centrality measures (DC, BWC, SC, and GDC), broken into the four BC gene categories (aging (A), cancer (C), HIV (HIV), and pathogen-interacting (PI) genes). In all panels, the values of k where precision and recall cross (as illustrated in Figure 5) are used; k equals 3, 10, 12, and 6, for A, C, HIV, and PI genes, respectively, for each of the four centrality measures.
doi:10.1371/journal.pone.0023016.g006

Furthermore, we measure the enrichment in drug targets of BC and SP genes (i.e., of gene categories "BC", "SP", "BC or SP", and "BC and SP" defined above) that are in the DS, and of BC and SP genes that are not in the DS. If our hypothesis that the topological positioning of nodes in the DS indeed reflects their functional importance is correct, then BC and SP genes that are in the DS should contain more drug targets than BC and SP genes that are not in the DS, since proteins that are targeted by drugs are clearly important for normal cellular functioning. Indeed, we find that enrichments in drug targets are much higher for BC and SP genes that are in the DS than for BC and SP genes that are not in the DS (Figure 9 B). Furthermore, these enrichments for the BC and SP genes that are in the DS are statistically significant, with p-values $\leq 10^{-4}$, while for SP and BC genes that are not in the DS they are not, with p-values $= 0.9998$ (see Methods).

Functional analysis of topologically central genes

For each category of BC genes (A, C, HIV, and PI genes), we compute enrichment of GDC-central and non-GDC-central genes in each of the Gene Ontology (GO) terms [71]. We consider all GO terms belonging to each of the three GO categories: molecular function (MF), biological process (BP), and cellular component (CC). Of the total of 1,359 MF, 3,925 BP, and 736 CC GO terms present in the human PPI network, 117 MF, 379 BP, and 27 CC GO terms are statistically significantly enriched (see Methods) in all 4 BC gene categories of GDC-central genes, while 4 MF, 10 BP, and 4 CC GO terms are statistically significantly enriched in non-GDC-central genes. Interestingly, there is no overlap between GO terms that are enriched in central genes and GO terms that are enriched in non-central genes.

Similar results are obtained for central and non-central genes with respect to membership in the dominating set (DS). DS-central genes are statistically significantly enriched in 153 MF, 574 BP, and 44 CC GO terms, while non-DS-central genes are statistically significantly enriched in 7 MF, 7 BP, and 0 CC GO terms, with no overlap between GO terms of central and non-central genes.

Hence, central genes appear to group by functions that are different than functions of non-central genes. Biological functions with the most significant enrichments that are present among all groups of central genes (but none of which is present among any of the groups of non-central genes) include many processes critical for normal cellular functioning, such as: regulation of cell cycle,

Gene	A	C	HIV	PI
TP53	X	X	X	
SRC		X	X	X
CREBBP	X	X	X	
ESR1	X	X	X	
EP300	X	X	X	X
GRB2	X		X	X
SMAD3			X	X
MAPK1			X	
RELA	X		X	
PIK3R1	X	X	X	
FYN			X	X
EGFR	X	X	X	X
AR	X	X		
YWHAG			X	
BRCA1	X	X	X	X
SHC1	X		X	X
HDAC1	X	X	X	X
SMAD2		X	X	
JUN	X	X	X	X
HSP90AA1	X			
RB1	X	X	X	
SMAD4		X	X	X
PTPN11	X	X	X	
STAT3	X	X	X	X
YWHAZ	X		X	
LCK		X	X	X
PLCG1			X	
CSNK2A1				
CHUK			X	
PXN			X	
PRKCA	X	X	X	

Gene	A	C	HIV	PI
SP1	X		X	X
CBL		X	X	X
LYN			X	X
CTNNB1	X	X	X	
PTPN6				
RAF1			X	X
NFKB1	X		X	X
MAPK3	X		X	
JAK2	X	X		
NR3C1	X		X	
HSPA1A	X			
PTK2	X	X	X	
YWHAB		X	X	
ABL1	X	X		
SYK		X		X
PRKCD	X		X	
STAT1		X	X	X
CDC2	X		X	X
INSR	X	X		
TBP	X		X	X
CASP3		X	X	X
AKT1	X	X	X	
IKBKG		X		
IKBKB			X	
RASA1				X
HDAC2	X	X		X
GNB2L1			X	
YWHAH		X	X	
NFKBIA	X	X	X	
IRS1	X			

Gene	A	C	HIV	PI
ACTB			X	X
PTK2B	X	X	X	
VAV1			X	
CRK				
TGFBR1				
NCOA1		X	X	X
SMAD1				
UBE2I	X		X	
STAT5A	X	X	X	
SMARCA4			X	
RXRA				X
JAK1			X	
SUMO1	X			X
YWHAE			X	
SIN3A	X			X
TRAF2				X
NCOR2		X		X
PCAF				X
PDGFRB	X	X		
PML	X	X	X	X
PPARBP				
CEBPB	X		X	X
CDK2			X	X
CDC42	X			X
KIT		X	X	
HTATIP				
MYC	X	X	X	X
CAV1			X	
PRKCZ			X	
PRKDC	X		X	X

Figure 7. The top 1% (i.e., 91) GDC-central genes. If a gene is an aging ("A"), cancer ("C"), HIV ("HIV"), or pathogen-interacting ("PI") gene, there is an "X" in the corresponding entry.
doi:10.1371/journal.pone.0023016.g007

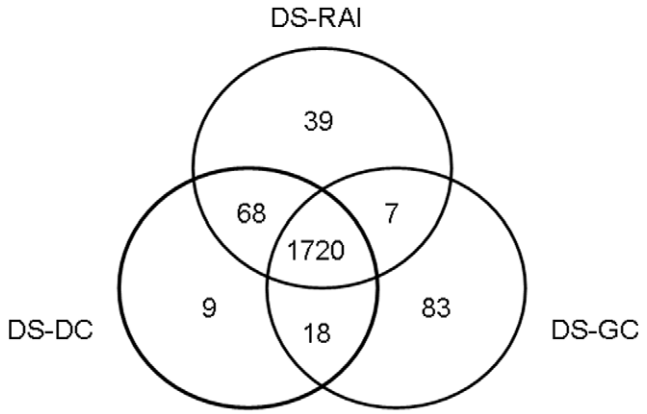


Figure 8. Overlap of the three DSs created by DS-RAI, DS-DC, and DS-GDC algorithms applied to the human PPI network.
doi:10.1371/journal.pone.0023016.g008

apoptosis, multicellular organism growth, telomere maintenance, innate immune response, regulation of cell differentiation, signal transduction, activity of many signaling pathway cascades (e.g., MAPK, I-kappaB kinase/NF-kappaB, EGFR, FGFR, IGFR, androgen receptor, nerve growth factor receptor, T cell receptor, toll-like receptor, etc.), phosphorylation, response to DNA damage, blood coagulation, regulation of cell proliferation, T cell activation and co-stimulation, response to tumor necrosis factor, response to drug, interspecies interaction between organisms etc.

Implications

GDC captures the density and topological complexity of up to 4-deep network neighborhood around a node. Since we have demonstrated significant enrichment of GDC-central proteins in BC genes, this means that genes that are involved in key biological processes occupy topologically complex and dense parts of the human PPI network. Similarly, since we have demonstrated significant enrichment of DSs in BC and SP genes, this indicates that proteins that are vital for normal cellular functioning reside on the "spine" of the network that dominates, i.e., connects, all other parts of the network. Hence, the notion of network domination seems to capture the topology required for passing cellular signals efficiently throughout the network.

We hypothesize that GDC-central proteins and proteins in DSs of PPI networks could represent potential candidates for therapeutic intervention, since targeting GDC-central proteins with drugs would have more significant impacts on the network than targeting proteins that reside in sparse and non-complex network regions and since the topology of a DS can enable quick propagation of drug effects through the entire network. Indeed, we find that the enrichment in drug targets of genes that are GDC-central *or* are in the DS (this is the union of the set of genes that are GDC-central and the set of genes that are in the DS) is 11.4% and it is statistically significant, with p-value of 1.3×10^{-4}. Furthermore, the enrichment in drug targets of genes that are simultaneously GDC-central *and* are in the DS (this is the intersection of the set of genes that are GDC-central and the set of genes that are in the DS) is even higher, it is 31.7%; this enrichment is also statistically significant, with p-value of 0. Hence, not only that each of the two concepts of topological centrality, GDC and DS, captures a statistically significant percentage of drug targets, but also when the two centralities are combined, the percentage of drug targets that they capture is significant and even higher.

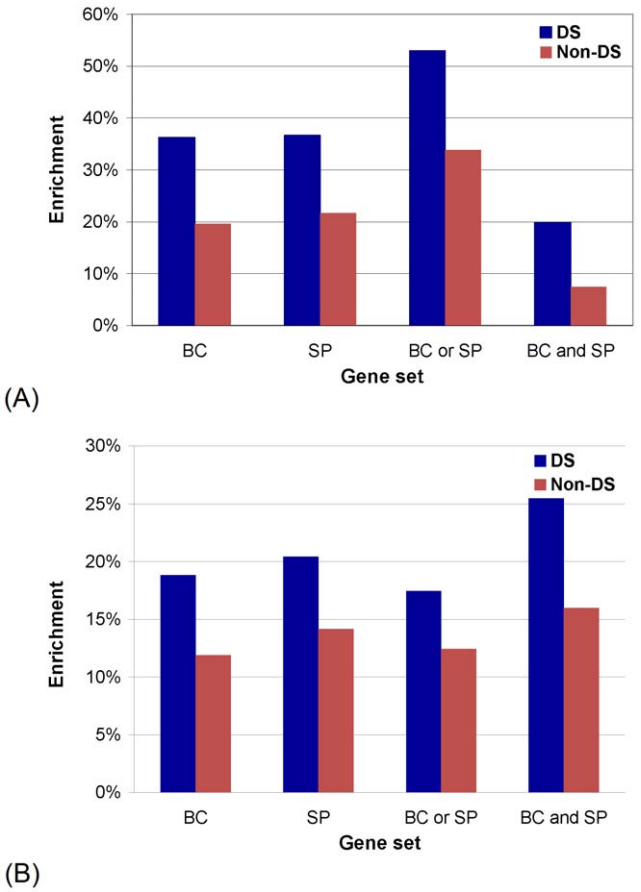


Figure 9. "Biological centrality" of the DS. (**A**) Enrichment in BC and SP genes of the dominating set ("DS") and its complement ("non-DS") in the human PPI network. (**B**) Enrichment in drug targets of BC and SP genes that are in the dominating set ("DS") and BC and SP genes that are not in the dominating set ("Non-DS").
doi:10.1371/journal.pone.0023016.g009

Concluding remarks

We propose a new centrality measure, graphlet degree centrality (GDC), to simultaneously measure the density and complexity of a node's extended neighborhood by counting the number of different graphlets that the node touches. We find that: (1) the enrichments in BC genes are much higher for GDC-central genes than for non-GDC-central genes; (2) the observed enrichments are statistically significant for GDC-central genes, while for non-GDC-central genes they are not; (3) BC genes that are GDC-central have higher and statistically significant enrichments in known drug targets than BC genes that are non-GDC-central; and (4) GDC outperforms other centrality measures in the sense that it uncovers the largest number of BC genes among the most central genes and is thus the most discriminative centrality measure.

Given the topologically central role of nodes in a DS, we apply to the human PPI network an existing DS algorithm that is commonly used in telecommunications, with the hypothesis that a DS might capture a set of proteins in a PPI network that are involved in important biological processes and mechanisms crucial for cell vitality. Also, we design a new and simpler DS algorithm that outperforms the existing algorithm on our data. We emphasize that our main focus is not to create a state-of-the-art algorithm for finding DSs, but instead, to demonstrate, as a proof of concept, that a DS of a PPI network found by a very simple

algorithm captures biologically vital proteins. Indeed, we find that: (1) the enrichments in BC and SP genes are much higher for nodes of DSs than for nodes outside of DSs; (2) the observed enrichments are statistically significant for nodes of DSs, while for nodes outside of DSs they are not; (3) BC and SP genes that are in DSs have much higher and statistically significant enrichments in known drug targets than BC and SP genes that are not in DSs; and (4) GDC-central genes that are also in the DS contain the highest, statistically significant percentage of drug targets.

These results imply that nodes in dense and complex neighborhoods that dominate the network are vital for normal cellular functioning and signaling. Hence, they might be targets for new therapeutic exploitation. Further algorithmic improvements would aid in more precise identification of these new targets.

Author Contributions

Conceived and designed the experiments: TM AB NP. Performed the experiments: TM VM. Analyzed the data: TM VM NP. Wrote the paper: TM NP.

References

1. Ideker T, Sharan R (2008) Protein networks in disease. Genome Research 18: 644–652.
2. Ito T, Tashiro K, Muta S, Ozawa R, Chiba T, et al. (2000) Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. Proc Natl Acad Sci U S A 97: 1143–7.
3. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, et al. (2000) A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. Nature 403: 623–627.
4. Li S, Armstrong C, Bertin N, Ge H, Milstein S, et al. (2004) A map of the interactome network of the metazoan c. elegans. Science 303: 540–543.
5. Giot L, Bader J, Brouwer C, Chaudhuri A, Kuang B, et al. (2003) A protein interaction map of drosophila melanogaster. Science 302: 1727–1736.
6. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck F, et al. (2005) A human protein-protein interaction network: A resource for annotating the proteome. Cell 122: 957–968.
7. Rual J, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, et al. (2005) Towards a proteomescale map of the human protein-protein interaction network. Nature 437: 1173–78.
8. Simonis N, Rual JF, Carvunis AR, Tasan M, Lemmens I, et al. (2009) Empirically controlled mapping of the Caenorhabditis elegans protein-protein interactome network. Nature Methods 6: 47–54.
9. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, et al. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415: 141–7.
10. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, et al. (2002) Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry. Nature 415: 180–3.
11. Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, et al. (2006) Proteome survey reveals modularity of the yeast cell machinery. Nature 440: 631–636.
12. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, et al. (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature 440: 637–643.
13. Tong AHY, Lesage G, Bader GD, Ding H, Xu H, et al. (2004) Global mapping of the yeast genetic interaction network. Science 303: 808–813.
14. Collins S, Schuldiner M, Krogan N, Weissman J (2006) A strategy for extracting and analyzing large-scale quantitative epistatic interaction data. Genome Biology 7: R63.
15. Rain JD, Selig L, De Reuse H, Battaglia V, Reverdy C, et al. (2001) The protein-protein interaction map of helicobacter pylori. Nature 409: 211–215.
16. Parrish JR, Yu J, Liu G, Hines JA, Chan JE, et al. (2007) A proteome-wide protein interaction map for campylobacter jejuni. Genome Biology 8: R130.
17. LaCount DJ, Vignali M, Chettier R, Phansalkar A, Bell R, et al. (2005) A protein interaction network of the malaria parasite plasmodium falciparum. Nature 438: 103–107.
18. Uetz P, Dong YA, Zeretzke C, Atzler C, Baiker A, et al. (2006) Herpesviral protein networks and their interaction with the human proteome. Science 311: 239–242.
19. von Brunn A, Teepe C, Simpson JC, Pepperkok R, Friedel CC, et al. (2007) Analysis of intraviral protein-protein interactions of the sars coronavirus orfeome. PLoS ONE 2: e459.
20. Chatr-aryamontri A, Ceol A, Peluso D, Nardozza A, Panni S, et al. (2009) Virusmint: a viral protein interaction database. Nucleic Acids Res 37: D669–D673.
21. Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, et al. (1998) SGD: Saccharomyces Genome Database. Nucleic acids research 26: 73–79.
22. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, et al. (2004) The Database of Interacting Proteins: 2004 update. Nucleic Acids Research 32: D449–51.
23. Peri S, Navarro JD, Kristiansen TZ, Amanchy R, Surendranath V, et al. (2004) Human protein reference database as a discovery resource for proteomics. Nucleic Acids Res 32 Database issue: D497–501.
24. Breitkreutz BJ, Stark C, Reguly T, Boucher L, et al. (2008) The BioGRID Interaction Database: 2008 update. Nucleic Acids Research 36: D637–D640.
25. Sharan R, Ulitsky I, Ideker T (2007) Network-based prediction of protein function. Molecular Systems Biology 3.
26. Schwikowski B, Fields S (2000) A network of protein-protein interactions in yeast. Nature Biotechnology 18: 1257–1261.
27. Milenković T, Pržulj N (2008) Uncovering biological network function via graphlet degree signatures. Cancer Informatics 4: 257–273.
28. Radivojac P, Peng K, Clark WT, Peters BJ, Mohan A, et al. (2008) An integrated approach to inferring gene-disease associations in humans. Proteins 72: 1030–7.
29. Goh K, Cusick ME, Valle D, Childs B, Vidal M, et al. (2007) The human disease network. PNAS 104: 8685–8690.
30. Yidirim MA, Goh KI, Cusick ME, Barabási AL, Vidal M (2007) Drugtarget network. Nature Biotechnology 25.
31. Milenković T, Memišević V, Ganesan A, Pržulj N (2010) Systems-level cancer gene identification from protein interaction network topology applied to melanogenesis-related functional genomics data. Journal of the Royal Society Interface 44: 353–350.
32. Vanunu O, Magger O, Ruppin E, Shlomi T, Sharan R (2010) Associating genes and protein complexes with disease via network propagation. PLoS Computational Biology 6: e1000641.
33. Jonsson P, Bates P (2006) Global topological features of cancer proteins in the human interactome. Bioinformatics 22: 2291–2297.
34. Aragues R, Sander C, Oliva B (2008) Predicting cancer involvement of genes from heterogeneous data. BMC Bioinformatics 9.
35. Sun J, Zhao Z (2010) A comparative study of cancer proteins in the human protein-protein interaction network. BMC Genomics 11: S5.
36. Jeong H, Mason SP, Barabási AL, Oltvai ZN (2001) Lethality and centrality in protein networks. Nature 411: 41–2.
37. Yu H, Braun P, Yıldırım MA, Lemmens I, Venkatesan K, et al. (2008) High-Quality Binary Protein Interaction Map of the Yeast Interactome Network. Science 322: 104–110.
38. Ratmann O, Wiuf C, Pinney J (2009) From evidence to inference: probing the evolution of protein interaction networks. HFSP Journal 3: 290–306.
39. Reguly T, Breitkreutz A, Boucher L, Breitkreutz B, Hon G, et al. (2006) Comprehensive curation and analysis of global interaction networks in saccharomyces cerevisiae. J Biol 5: 10.
40. Ho H, Milenković T, Memišević V, Aruri J, Pržulj N, et al. (2010) Protein interaction network topology uncovers melanogenesis regulatory network components within functional genomics datasets. BMC Systems Biology 4.
41. Promislow DE (2004) Protein networks, pleiotropy and the evolution of senescence. Proc Biol Sci 1545: 1225–1234.
42. Ferrarini L, Bertelli L, Feala J, McCulloch AD, Paternostro G (2005) A more efficient search strategy for aging genes based on connectivity. Bioinformatics 21: 338–348.
43. Dyer MD, Murali TM, Sobral BW (2008) The landscape of human proteins interacting with viruses and other pathogens. PLoS Pathog 4: e32.
44. Estrada E, Rodrguez-Velazquez J (2005) Subgraph centrality in complex networks. Phys Rev E 71: 056103.
45. Estrada E, Hatano N (2007) Statistical-mechanical approach to subgraph centrality in complex networks. Chemical Physics Letters 439: 247–251.
46. Pržulj N, Corneil DG, Jurisica I (2004) Modeling interactome: Scale-free or geometric? Bioinformatics 20: 3508–3515.
47. Pržulj N (2007) Biological network comparison using graphlet degree distribution. Bioinformatics 23: e177–e183.
48. Milo R, Shen-Orr SS, Itzkovitz S, Kashtan N, Chklovskii D, et al. (2002) Network motifs: simple building blocks of complex networks. Science 298: 824–827.
49. Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393: 440–442.
50. Guerrero C, Milenković T, Pržulj N, Jones JJ, Kaiser P, et al. (2008) Characterization of the yeast proteasome interaction network by qtax-based tag-team mass spectrometry and protein interaction network analysis. Proc Natl Acad Sci U S A 105: 13333–13338.
51. Memišević V, Milenković T, Pržulj N (2010) Complementarity of network and sequence structure in homologous proteins. Journal of Integrative Bioinformatics 7: 135.
52. Kuchaiev O, Milenković T, Memišević V, Hayes W, Pržulj N (2010) Topological network alignment uncovers biological function and phylogeny. Journal of the Royal Society Interface 7: 1341–1354.

53. Milenković T, Ng W, Hayes W, Pržulj N (2010) Optimal network alignment with graphlet degree vectors. Cancer Informatics 9: 121–137.
54. Stojmenovic I, Seddigh M, Zunic J (2002) Dominating sets and neighbor elimination-based broadcasting algorithms in wireless networks. IEEE Transactions on Parallel and Distributed Systems 13: 14–25.
55. Houmaidi M, Bassiouni M (2003) Dominating set algorithms for sparse placement of full and limited wavelength converters in wdm optical networks. Journal of Optical Networking 2.
56. Gao B, Yang Y, Ma H (2005) A new distributed approximation algorithm for constructing minimum connected dominating set in wireless ad hoc networks. International Journal of Comunication Systems. pp 743–762.
57. Rai M, Verma S, Tapaswi S (2009) A power aware minimum connected dominating set for wireless sensor networks. Journal of Networks 4.
58. Wu J, Li H (1999) On calculating connected dominating set for efficient routing in ad hoc wireless network. Proceedings od 3rd ACM International Workshop on Discrete Algorithms and Methods for Mobile Computing and Communications 18: 7–14.
59. Cooper C, Klasing R, Zito M (2005) Lower bounds and algorithms for dominating sets in web graphs. Internet Mathematics 2: 275–300.
60. Junker BH, Schreiber F (2008) Analysis of Biological Networks (Wiley Series in Bioinformatics) Wiley-Interscience.
61. Duckworth W, Wormald N (2006) On the independent domination number of random regular graphs. Combinatorics, Probability and Computing 15: 513–522.
62. de Magalhaes JP, Budovsky A, Lehmann G, Costa IJ, Fraifeld YV, et al. (2009) The Human Ageing Genomic Resources: online databases and tools for biogerontologists. Aging Cell 1: 65–72.
63. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, et al. (2004) A census of human cancer genes. Nature Reviews Cancer 4: 177–183.
64. Safran M, Solomon I, Shmueli O, Lapidot M, Shen-Orr S, et al. (2002) GeneCards 2002: towards a complete, object-oriented, human gene compendium. Bioinformatics 18: 1542–1543.
65. Kanehisa M, Goto S (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research 28: 27–30.
66. Hamosh A, Scott AF, Amberger J, Bocchini C, Valle D, et al. (2002) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Research 30: 52–55.
67. Fu W, Sanders-Beer B, Katz K, Maglott D, Pruitt K, et al. (2009) Human immunodeficiency virus type 1, human protein interaction database at ncbi. Nucleic Acids Research 37: D417–22.
68. Bandyopadhyay S, Chiang Cy, Srivastava J, Gersten M, White S, et al. (2010) A human MAP kinase interactome. Nature Methods 7: 801–805.
69. Kandasamy K, Mohan SS, Raju R, Keerthikumar S, Kumar GSS, et al. (2010) NetPath: a public resource of curated signal transduction pathways. Genome biology 11: R3+.
70. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, et al. (2006) Drugbank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Research 34: D668–D672.
71. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genetics 25: 25–29.