Dominating Balanced Protein Interaction Networks in Cancer

Rianna Patricia S. Cruz[1,2], Hannah Mae C. Magno[1]
Joshua Dizon[2], Henry N. Adorna[1]
ha@dcs.upd.edu.ph
[1]Department of Computer Science (Algorithms & Complexity)
University of the Philippines Diliman
Diliman 1101 Quezon City, Philippines
[2] Philippine Genome Center
University of the Philippines Diliman
Diliman 1101 Quezon City, Philippines

ABSTRACT

As available proteomic data grows, so does our need for computational methods to process such data for practical applications — such as drug and therapeutic development. This is critical particularly in cancer treatments, where multiple mutations may obscure driver proteins and pathways to target for potential treatments.

To identify these driver proteins and pathways, we explore cancer networks' minimum connected dominating sets (MCDS), a set of topologically significant nodes of a network. We build on existing heuristic algorithms to find driver proteins of selected cancer networks via their MCDS.

From sets of known cancer driver proteins ($n = [8, 10]$) and essential proteins ($n = [991, 1415]$) of breast, ovarian, and pancreatic cancer, we generated protein interaction networks for each selected cancer, using balanced and directed graphs to model regulatory function.

We identified each interaction networks' driver proteins ($n = [40, 100]$) from their MCDS and validated each against sets of positive control driver proteins derived by other methods. From these driver protein sets, we performed pathway analysis to identify pathways enriched by these proteins. We then verified whether these proteins had a documented association with cancer.

Our driver proteins had measures of centrality (betweenness, degree centrality) higher than those of positive control proteins of the same cancer networks. This confirms their topological significance in their respective networks.

Pathway analysis identified over 300 pathways enriched with statistical significance. A survey on these pathways found that $79 - 80\%$ of these pathways are linked to cancer. They were also almost twice as likely to have a documented association with cancer than those not enriched by our identified driver proteins.

We not only identify specific potential driver proteins in cancer networks but also validate the potential of minimum connected dominating set-finding algorithms to identify driver proteins in protein regulatory networks. We validate the potential of balanced signed directed graphs in modeling regulatory functions of protein interaction networks.

KEYWORDS

signed graphs, balanced graphs, minimum connected dominating set, protein-protein interactions, cancer

1 INTRODUCTION

Omic research is at the core of the search for drug or therapeutic targets for diseases such as cancer [7] [6]. Protein-protein interaction (PPI) data is particularly promising, given its crucial role in the central dogma of molecular biology [7]. Biological regulatory networks interact via PPI and influence levels of expression of genes in a genome. These levels of gene expression affect risk or predisposition to certain diseases [4], such as cancer. Developing algorithmic approaches to identifying driver proteins in such networks is thus necessary, as these may point to proteins to target in treating these diseases. [8].

If we represent gene regulatory networks as graphs, then driver proteins may be driver nodes of such graphs. In this study, we identify such driver proteins. Milenkovic et al. [13] suggested that these proteins have high connectivity, degree, and centrality within their network, implying that topology-based methods have potential in identifying proteins in a network.

Various studies suggest that minimum dominating set-based methods for determining significant nodes of PPI networks may identify cancer-related or virus-targeted genes [23] [24] [14]. Bantang, Urog, and Adorna [11] proposed a heuristics-based approach to identifying the dominating tree of a scale-free network. Dizon et al. [9] built on this method to develop an algorithm to find the minimum connected dominating set of a protein-protein interaction network. Biones et al. [25] adapted this method to identify driver proteins considering gene regulatory function in the human PPI network.

In this paper, we present a method to model protein-protein networks regulatory function using balanced, signed, directed graphs, and an MCDS-based approach to identifying driver proteins in cancer networks. We subject the results of these methods to multiple validation tests to assess their viability in identifying driver proteins of gene regulatory networks.

2 PRELIMINARIES

2.1 Biological preliminaries

Definition 1 (Genes and Proteins). *Genes are chains of DNA that code instructions to assemble proteins. Proteins build and regulate various body functions. In this paper, we use the terms 'genes' and 'proteins' interchangeably.*

Definition 2 (Protein-protein interactions (PPI)). *Protein-protein interactions (PPI) are the physical interactions between proteins in a living organism [20].*

Definition 3 (Protein regulatory networks). *Proteins interact in various ways. In this paper, we explore up-regulation, down-regulation, and complex-forming interactions.*

A protein up-regulates another protein when it increases a cell's sensitivity to this other protein B, while it down-regulates a protein when it decreases this sensitivity. Groups of proteins form complexes when they interact with each other at the same time and location.

When groups of proteins interact in the above ways they form gene regulatory networks [3].

We provide a more detailed explanation of the above interactions in the Appendix.

Definition 4 (Cancer driver proteins, driver genes). *Cancer driver genes are genes whose mutations cause malignant or tumor growth. In graphical terms, driver proteins are driver nodes in protein interaction networks.*

Definition 5 (Biological Pathways). *A biological pathway is a series of molecular interactions in a cell. These either result in a change or a product in that cell.*

Definition 6 (Pathway analysis, pathway enrichment analysis). *Pathway analysis is a statistical method that identifies biological pathways that are enriched in a gene (protein) set more than would be expected by chance (with $p - value < 0.05$). This statistical method maps genotypes to their probable phenotypic manifestations.*

Different sets of genes may identify the same or similar sets of biological pathways with statistical significance.

2.2 Pharmaceutical preliminaries

Definition 7 (Target proteins, drug targets, or protein targets). *Target proteins are molecules associated with particular diseases in an organism. Pharmaceutical researchers aim to target these proteins with drugs to produce a desired therapeutic effect.*

In this paper, we seek target proteins for human breast, ovarian, and pancreatic cancer.

Cancer driver genes may not necessarily be viable drug targets and might not yet have drugs targeting them for therapeutic effect. Nonetheless, identifying new driver proteins for cancer may suggest novel targets for drug development.

We refer to these as 'drug targets' or 'known drug targets' if some approved drug claims to target these proteins, while we may refer to these as 'driver proteins' otherwise.

2.3 Mathematical preliminaries

2.3.1 General preliminaries.

Definition 8 (Path). *A path p is a sequence of alternating vertices and edges $v_0, e_1, v_1, e_2, v_2, ..., e_k, v_k$ in G where $e_i = v_{i-1}, v_i$, such that:*

- *All edges are distinct ($i.e.$ given $e_i, e_{\in}p, \forall i \neq j, e_i \neq e_j$)*
- *All vertices are distinct ($i.e.$ given $v_i, v_j \in p, \forall i \neq j, v_i \neq v_j$)*
- *The path starts and ends on distinct nodes ($i.e.$ $v_0 \neq v_k$)*

Figures 1 and 2 are examples of paths.

Definition 9 (Shortest path tree). *A shortest-path tree is a spanning tree T of G, rooted at vertex v, such that the path distance from root v to any other vertex u in T is the shortest path distance from v to any u in G. T may span G only if G is connected.*

If G is disconnected, then we take T that spans vertices that are reachable by v.

2.3.2 Signed graph preliminaries.

Definition 10 (Sign of a path). *The sign S of a path p is the product of the signs of its edges. This is*

$$S(p) = \prod_{(u,v)\in p} \sigma(u, v)$$

.

If we let m be the cardinality of a path's positive edges, and n be the cardinality of the same path's negative edges, then we may simplify this :

$$S(p) = (1^m)(-1^n) \qquad \textit{note identify property of 1} \qquad (1)$$

$$S(p) = -1^n \qquad (2)$$

Figure 1: The sign of the path is 1

Figure 2: The sign of the path is -1

Definition 11 (Balanced path). *Path p is balanced if its sign is positive ($i.e. S(p) > 0$). Otherwise, it is unbalanced. For example, Figure 1 shows a balanced path, while Figure 2 shows an unbalanced path.*

Definition 12 (Balanced signed directed graph). *A signed directed graph is balanced if the product of all its paths' signs is positive ($i.e. \Pi_{p_i \in G}(S(p_i)) > 0$). Otherwise, the graph is unbalanced. For example, Figure 3 presents a balanced signed directed graph rooted at node 1. The product of all paths ($e.g. \{1, 2\}$, $\{1, 3, 5\}$, $\{1, 4, 5\}$, $\{1, 4, 6, 7\}$) is positive.*

2.3.3 Dominating set preliminaries.

Definition 13 (Dominating set). *A dominating set is a subset D of the vertex set V of G where every vertex in V not in D is adjacent to at least one vertex in D. Figure 4 and Figure 5 show examples of dominating sets in directed graphs.*

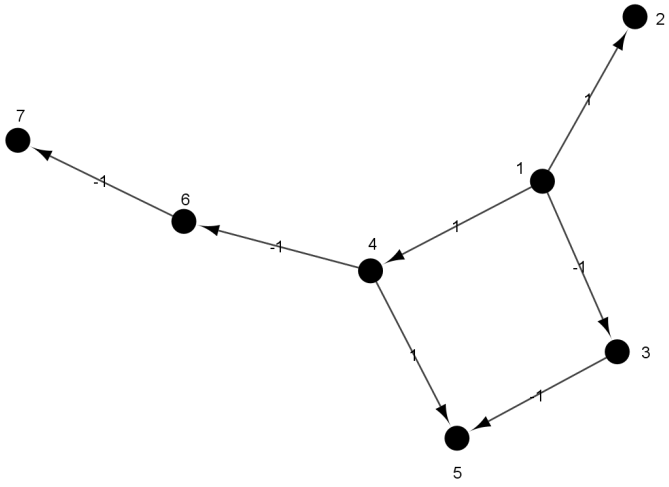

Figure 3: A signed directed graph

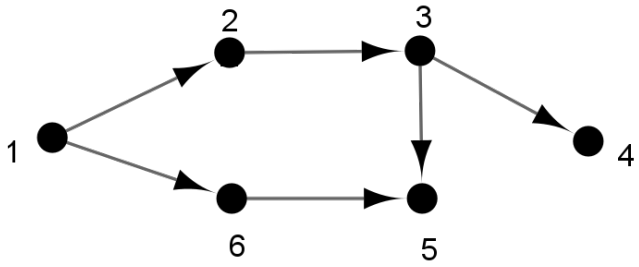

Figure 4: $\{1, 3\}$ is a dominating set of this graph.

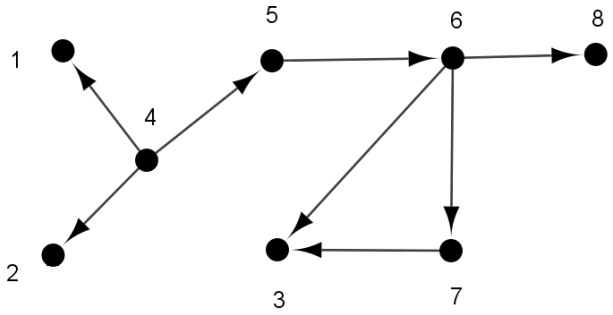

Figure 5: $\{4, 6\}$ is a dominating set of this graph

Definition 14 (Connected Dominating Set (CDS)). *The connected dominating set is a dominating set D of nodes that are connected in G.*

For example, $\{1, 2, 3, 4\}$ is a connected dominating set in Figure 4, while $\{4, 5, 6, 7\}$ or $\{4, 5, 6, 3\}$ is a connected dominating set in Figure 5.

Definition 15 (Minimum Connected Dominating Set (MCDS)). *An MCDS is a connected dominating set of G with the smallest size. For example, $\{2, 3\}$ is the MCDS in Figure 4, while $\{4, 5, 6\}$ is the MCDS in Figure 5.*

2.3.4 Measures of centrality.

Definition 16 (Measures of centrality). *Measures of centrality of a node v in G empirically reflect how topologically significant v is in G.*

Definition 17 (Degree). *The degree of a node v ($deg(v)$) is its number of edges. In directed graphs, we may distinguish in degree (edges coming into v) and out-degree (edges coming from v).*

Definition 18 (Degree centrality). *Degree centrality is the number of edges incident to a node. In undirected graphs, this is $deg(v)$. In directed graphs, this is the sum of the in-degree and out-degree of v.*

Definition 19 (Betweenness centrality). *Betweenness centrality of v in G is the percent of shortest paths in G that pass v.*

Definition 20 (Clustering coefficient). *The clustering coefficient of v in G is the percent of triangles in G that contain v.*

3 ALGORITHM TO BALANCE NETWORK

3.1 Requirements and rationale

3.1.1 Use of signed graphs. To model gene regulatory networks, we use signed directed graphs because signed edges may model their regulatory interactions. Negative edges may model down-regulation, while positive edges may model other interaction types.

3.1.2 Use of shortest-path tree. We use a shortest-path tree to model our gene regulatory network considering Occam's razor, which suggests that if we are given a multiplicity of solutions, we must choose the simplest solution. With gene regulatory networks where multiple paths may exist from any protein to another, this leads us to select the shortest path between these two proteins.

3.1.3 Need for root node. Our algorithm requires a root node (or a set of root nodes) to model the regulatory function of known driver proteins.

3.2 Input

A signed directed graph $G(V, E, \sigma)$
A root $v_s \in V$

3.3 Output

A balanced, shortest path tree T of G, rooted at v_s

3.4 Setup

At the i-th iteration of our algorithm, we define the following:

- Let E_i be the set of edges to keep from our original signed graph, where $E_0 = E$.

Table 1: The paths in the tree and their corresponding signs

j	Target node	$p_{0,j}$	Signs of edges in $p_{0,j}$	Number of negative edges	(P)
-	1	1 is not reachable from 0.			
0	2	$0 \to 8 \to 7 \to 2$	-1, -1, 1	2	1
1	3	$0 \to 8 \to 7 \to 2 \to 3$	-1, -1, 1, -1	**3**	**-1**
2	4	$0 \to 4$	1	0	1
3	5	$0 \to 8 \to 9 \to 5$	-1, 1, -1	2	1
4	6	$0 \to 4 \to 6$	1, -1	**1**	**-1**
5	7	$0 \to 8 \to 7$	-1, -1	2	1
6	8	$0 \to 8$	-1	**1**	**-1**
7	9	$0 \to 8 \to 9$	-1, 1	**1**	**-1**

- Let V_i be the set of nodes reachable from v_s, considering all pruned edges
- Let G_i be the connected component of all nodes reachable from v_s, considering E_i i.e. $G_i(V_i, E_i)$
- T_i is the shortest path tree of G_i, rooted at v_s.
- Let P_i be the set of all shortest paths in T_i. P_i contains paths $p_{i,j}$.
- Let Q_i be the set of edges we must prune to balance paths in P_i. Q_i start each run as an empty set. Throughout the iteration, it is populated with negative edges that unbalance T_i.

3.5 Algorithm Definition

(1) Find the shortest path tree T_i of G given E_i.
(2) Find P_i from T_i. Let $Q_i = \emptyset$
(3) For all paths $p_{i,j} \in P_i$, if the sign of $p_{i,j}$ is -1 (if n is odd in $(p_{i,j}) = -1^n$), then add to Q_i the negative edge that is furthest from v_s in path $p_{i,j}$.
(4) After iterating through all $p_{i,j} \in P_i$, remove from G all edges in Q_i such that $E_{i+1} = E_i - Q_i$.
(5) Let G_{i+1} be the connected component to v_s.
(6) If Q_i is not an empty set, then find T_{i+1}, considering E_{i+1}, and repeat from step 2.

3.6 Sample run of the algorithm

In Figure 6, we present a sample signed graph. We set node 0 as our root. At $i = 0$, we identify the shortest path tree T_0 of G_0 from our root to every other node in G_0. Because there are paths in T_0 with a negative sign, T_i is unbalanced. The paths in this tree and their corresponding signs are found in Table 1. We set $Q_0 = \emptyset$.

The sign of $p_{0,1}$ is negative. To balance this branch of our tree, we prune the negative edge in this path furthest from our root node. In $p_{0,1}$, this edge is $2 \to 3$. We add this edge to Q_0.

In $p_{0,4}$, edge $4 \to 6$ is the negative edge furthest from our root. We add $4 \to 6$ to Q_0 as well. We repeat this for all paths $p_{0,j}$ with a negative sign.

Processing all such negative paths for $i = 0$, gives us is $Q_0 = \{2 \to 3, 4 \to 6, 0 \to 8, 8 \to 9\}$.

To balance our shortest-path tree, we remove all edges in Q_0 from E_0. E_1 then becomes $E_1 = E_0 - Q_0$. Expanding this gives us $E - \{2 \to 3, 4 \to 6, 0 \to 8, 8 \to 9\}$.

After we remove these edges, node 3 becomes unreachable from our root node. Thus $V_1 = V_0 - \{3\}$. We then set G_1 as the subgraph of G with reduced edges E_1 and nodes V_1.

From G_1, we find T_1, and repeat the above steps until T_{i+1} is balanced. This occurs when we may have no shortest paths to balance. This occurs when we may add no more edges to Q_i to prune.

Figure 7 presents the results of this algorithm on our graph. The shortest path tree of this graph is balanced because it contains no unbalanced paths. The MCDS of the network is shown in Figure 8.

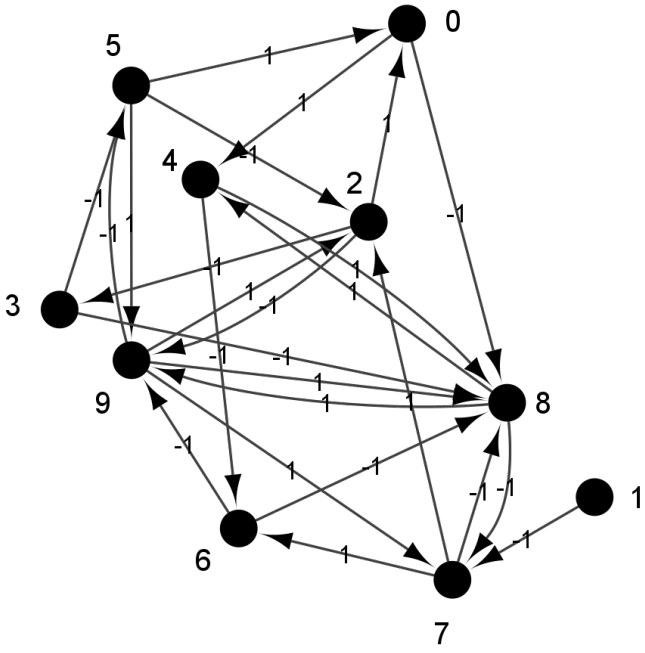

Figure 6: A signed directed graph

3.7 Proof of correctness

3.7.1 Proof of balance within a single path. Let G be a signed directed graph
Let v_s be some root node in G (i.e. $v_s \in V$)
Let p be any path that exists from v_s to a target node $v_t \in V$ reachable from v_s
Let n be the number of negative edges along with such single path p

Recall that the sign of a path is $S(p) = -1^n$, where n is the number of negative edges along with p.

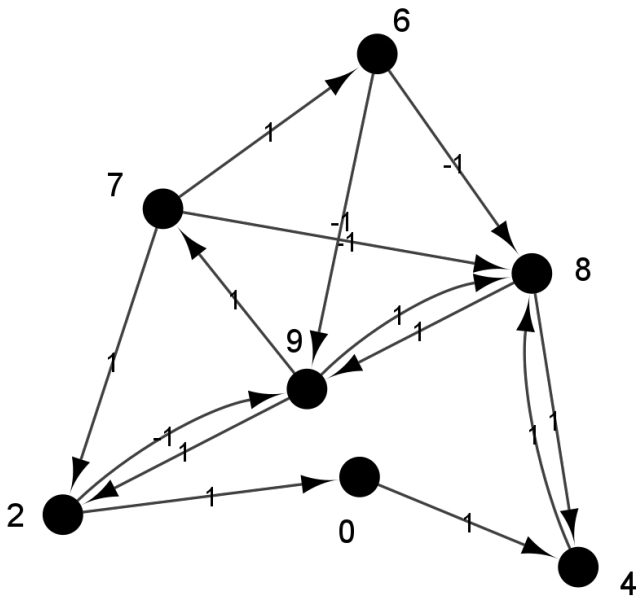

Figure 7: Pruned and balanced directed signed graph

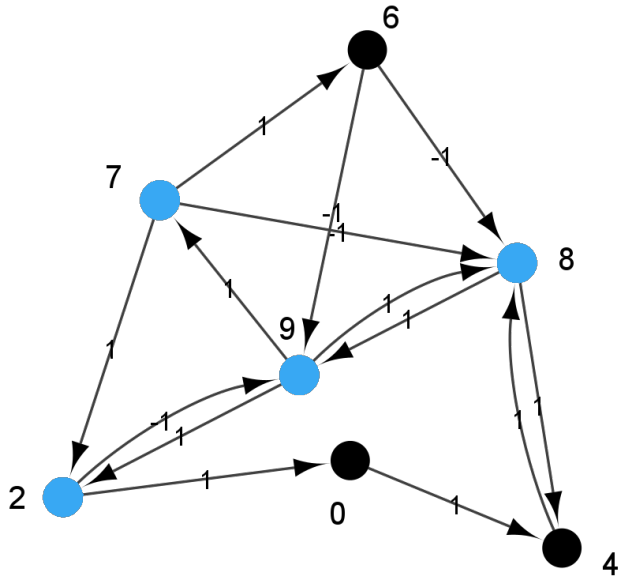

Figure 8: MCDS of the graph

Base case.

(1) $n = 0$. There are no negative edges in p. It is balanced as is.

(2) $n = 1$. There is 1 negative edge in p. Prune this negative edge. Let $\hat{p}$ be the path that remains reachable from v_s. Path $\hat{p}$ now has $\hat{n} = 1 - 1 = 0$ negative edges. Thus, $\hat{p}$ is balanced.

General cases.

(1) The path p is balanced. This implies that path p has an even number of negative edges. We can express this number of negative edges as $n = 2k$, for some $k \geq 0$.
(2) The path p is unbalanced. This implies that path p has an odd number of negative edges. We can express this number of negative edges as $n = 2k + 1$, for some $k \geq 0$.

Inductive hypothesis. Let n be the number of negative edges in some path p, where $n > 1$.

(1) If n is even, then we can express it as $n = 2k$ for some $k \geq 0$. The sign of path p is then $S(p) = (-1)^{2k} = ((-1)^2)^k = 1^k = 1$, which is positive. Path p is thus balanced as-is and does not require pruning.
(2) If n is odd, then we can express it as $n = 2k + 1$ for some $k \geq 0$. The sign of path p is then:

$$S(p) = (-1)^{2k+1} \tag{3}$$
$$= (-1)^{2k} * -1 \tag{4}$$
$$= ((-1)^2)^k * -1 \tag{5}$$
$$= 1^k * -1 \quad \text{note identify property of 1} \tag{6}$$
$$S(p) = -1 \tag{7}$$

$S(p)$ is negative. To balance this negative path, we prune the negative edge furthest from our root v_s.

Let $\hat{p}$ be the path that remains reachable from v_s after pruning. Since we removed 1 negative edge, now $\hat{n} = (2k+1)-1 = 2k$, which is even. As proved above, the sign of $\hat{p}$ is thus positive and balanced.

Inductive step. Let $m+1$ be the number of negative edges in some p, where $m > 1$.

(1) If m is even, then we can express it as $m = 2k$ for some $k \geq 0$. Given this m, then $m + 1 = 2k + 1$, which is odd.

We then prune the edge as defined above, to produce path $\hat{p}$. Path $\hat{p}$ will have $m+1-1$ negative edges, where $(m+1)-1 = (2k+1)-1 = 2k$, which is even.

Thus, the sign of the path will be positive, as proved above. Path $\hat{p}$ is then balanced after pruning.

(2) If m is odd, then we can express it as $m = 2k + 1$, for some $k \geq 0$. If $m = 2k+1$, then $m+1 = (2k+1)+1 = 2k+2 = 2(k+1)$, which is even.

Thus, the sign of the path would be positive. Thus p is balanced without pruning.

Therefore, for $n \geq 0$, the above algorithm will produce a balanced path.

3.7.2 Proof of balance throughout the graph. Let G be a signed directed graph. Let v_s be some source node in G (i.e. $v_s \in V$).

Recall, at the i-th iteration of our algorithm, we have:

(1) Let T_i be the shortest path tree of G rooted at v_s
(2) Let P_i the set of paths in T_i, rooted at v_s
(3) Let Q_i be the set of edges to prune to balance paths in P_i.

Base cases. The following are base cases for our proof of balance:

(1) If $|Q_i| = 0$ then P_i contains no paths with a negative sign. The graph is balanced at the i-th iteration of the algorithm.
(2) If $|Q_i| = 1$ then one path p in P_i has a negative sign. Path p must then have an odd number of negative edges. In this case, if we prune p as illustrated above, then the resulting $\hat{p}$ is balanced. Thus, $|Q_{i+1}| = |Q_i| - 1 = 1 - 1 = 0$, and the graph is balanced at step $i + 1$.

Proof of balance when $|Q_i| > 1$. If we remove a single edge in T_i, only the following are possible:
We remove an edge in T_i that makes at least one node unreachable from v_s. Therefore, $|V_{i+1}| < |V_i|$.
We remove an edge in T_i but no nodes become unreachable from v_s. Therefore, $|V_{i+1}| = |V_i|$. Thus, in either case, $|V_{i+1}| \leq |V_i|$.

There are only as many shortest paths in P_i as there are vertices reachable from v_s in T_i. Because of this, P_i is bound by $|P_i| \leq |V_i|$. As the number of vertices is bound by a decreasing interval, then the number of paths is similarly bound by $|P_{i+1}| \leq |P_i|$. In other words, the number of shortest paths in T_i cannot increase.

Note that we may prune at most one edge per path in P_i. Paths in P_i thus map to edges in Q_i by some one-to-one mapping. Therefore, the size of Q_i is bound by $|Q_i| \leq |P_i|$. Thus the number of unbalanced paths in T_i also cannot increase in size. The number of edges to prune from these shortest paths also cannot increase and is bound by $|Q_{i+1}| \leq |Q_i|$.

At step i, if T_i is unbalanced, then there must be at least j unbalanced paths in P_i, where $j \geq 1$. We may convert each of these unbalanced paths $p_{i,j}$ in P_i into a balanced path $\hat{p_{i,j}}$. To balance all $p_{i,j}$ unbalanced paths in P_i, we must prune j negative edges in Q_i.

If we prune at least 1 unbalanced edge in Q_{i-1} at each i step, then the remaining number of unbalanced paths in T_i converges to:

$$|Q_i| \leq |Q_{i-1}| - 1 \quad \text{since j} \geq 1 \tag{8}$$
$$\leq |Q_{i-2}| - 1 - 1 = |Q_{i-2}| - 2 \tag{9}$$
$$\ldots \tag{10}$$
$$\leq |Q_0| - 1_1 - 1_2 - 1_3 \ldots - 1_i \quad \text{Prune} \geq 1 \text{ edge at each } i \text{ step} \tag{11}$$
$$\leq |Q_0| - i \tag{12}$$
$$|Q_{lim_{i\to\infty}}| \leq |Q_0| - lim_{i\to\infty} \quad \text{Note that } Q_0 \text{ is a finite set. } |Q_0| \text{ is finite.} \tag{13}$$
$$|Q_{lim_{i\to\infty}}| \leq 0 \tag{14}$$

This shows that at each step i, a decreasing interval bounds the size of set Q_{i+1}. This eventually converges to 0, where there are only balanced paths in T_i. Thus our algorithm terminates in i (finite number) iterations with balanced shortest path tree T_i.

3.7.3 Practical relaxation of balance. Complete balance in these graphs eliminates all negative (down-regulating) interactions in our model. This harms rather than improves the accuracy of our model. For practicality with real data sets, we relax our algorithm to balance only the shortest path branches to leaf nodes. We then remove only the furthest down-regulating edges from our root node and maintain, rather than closer down-regulating edges which may ultimately be balanced further along a path.

4 ALGORITHM TO FIND THE MINIMUM CONNECTED DOMINATING SET OF NETWORK

We use a topology-based MCDS-finding algorithm defined by Dizon et al. [11] [9] [25].

Given graph $G(V, E)$, let U be a set of visited nodes, and W be a set of covered nodes. At the start of our algorithm, U and W are empty sets.

4.1 Algorithm definition

(1) Identify node v of highest degree in G.
(2) Initialize the MCDS as the set containing only v.
(3) Initialize a set of visited nodes U as a set containing only v.
(4) Identify the neighborhood of v and add this neighborhood to both set U and W.
(5) Find w in W which covers the most unvisited nodes and w to the MCDS.
(6) Add the neighborhood of w to U.
(7) Subtract w from W, but add the neighbors of w to W.
(8) If $U = V$, then all nodes have been visited and the MCDS is found. Otherwise, set v as w and repeat from step 4.

5 DATA

5.1 Protein-protein interaction (PPI) data

We use PPI data from the Signaling Network Open Resource (SIGNOR) database [17]. For this study, we limit our analysis to interactions between proteins, complexes, and protein families that up-regulate, down-regulate, or form complexes.

5.2 Modeling PPI data

To model these networks, we set proteins as nodes and their interactions as edges in a signed directed graph, where interaction type (up-regulation, down-regulation) defines edge sign.

5.3 Test data sets

5.3.1 A priori known cancer networks. Kanhaiya et al. [10] identified driver protein networks of breast, ovarian, and pancreatic cancer based on these networks' structural controllability.

To accomplish this, they determined essential protein networks associated with these cancers ($|V|$ = 900 to 1600 nodes, $|E|$ = 1500

Table 2: Network properties at each step of our test runs

Network	Root nodes	Original node count	Original edge count	Node count after balancing	Edge count after balancing	Driver proteins found	Percent of nodes removed to balance	Percent of MCDS in Pruned
a priori known breast	breast driver proteins	1415	2435	550	1085	100	61.13%	18.18%
a priori known ovarian	ovarian driver proteins	1047	1579	294	458	44	71.92%	14.97%
a priori known pancreatic	pancreatic driver proteins	991	1484	241	367	51	75.68%	21.16%
naive network	breast driver proteins	5681	16414	2628	8977	599	53.74%	22.79%
naive network	ovarian driver proteins	5681	16414	2627	9064	595	53.76%	22.6%
naive network	pancreatic driver proteins	5681	16414	2621	8902	596	53.86%	0.227394124

to 2500 edges). In our paper, we refer to these as 'a priori known networks'.

From each of these, they identified minimum node sets with full controllability of their entire network ($|V| = 130$ to 170 nodes). In this paper, we compare these node sets to the MCDS we find from the same networks.

5.3.2 Naive networks. Besides our a priori known networks, we test our algorithm on naive networks. These are networks are not informed with prior knowledge of PPI regulatory functions. Our naive networks contain all proteins in the SIGNOR dataset ($|V| = 3051$, with $|E| = 10892$ edges).

5.3.3 Driver proteins as root nodes. From their set of driver proteins, Kanhaiya et al. [10] identified US Food and Drug Administration approved drug targets for each cancer. We use these drug targets as root nodes for our algorithm.

5.4 Validation data sets

5.4.1 Positive control data sets from cancer literature. We validate our results against positive control driver proteins from Bailey et al [1]. This paper identified 299 unique cancer driver genes across 33 cancer types from a consensus of driver protein-finding algorithms. From these, we use proteins related to our selected cancers ($|V| =$ 239 proteins total).

We further validate our results against driver protein sets derived by several other algorithmic methods. These sets are from Nikzainal et al. [15] for breast cancer ($|V| = 93$), Ryland et al. [21] for ovarian cancer ($|V| = 15$), and Biankin et al. [2] for pancreatic cancer ($|V| = 16$).

5.4.2 Pathway database. We use the Reactome database [5] — a curated and peer-reviewed pathway database — to perform pathway analysis.

6 METHODS

6.1 Representing PPI cancer networks with signed directed graphs

We represent cancer networks as signed directed graphs $G(V, E, \sigma)$ [16]. In this graph, we set nodes as proteins and the interactions between them as edges. Interaction type (up-regulation, down-regulation) defines edge sign. Only down-regulating edges receive a negative sign, while other interaction types receive a positive sign.

6.2 Finding driver proteins of cancer networks

For each of our test cancer networks, we find driver proteins as follows:

(1) Balance the cancer network with our relaxed balancing algorithm; then
(2) Find the MCDS of the balanced network as defined above.

6.3 Experimental setup

For each cancer, we apply the above algorithms on the following networks:

(1) A naive network of all proteins in SIGNOR dataset
(2) An a priori defined cancer network

We use driver proteins from Kanhaiya et al. [10] ($|V| = 9$ to 11 nodes) as roots for our algorithm.

With 2 (balanced and naive) networks each for 3 cancers (breast, ovarian, and pancreatic), we produced 6 sets of driver proteins.

6.4 Validating our driver protein sets

We validate our driver protein sets:

(1) Compare the measures of centrality of our protein sets with those of controlling proteins from [10].
(2) Identify the consensus of our protein sets with validation data sets from cancer literature [15] [21] [2].
(3) Identify pathways enriched by our protein sets with statistical significance.
(4) Identity in which pathways enriched by statistical significance have been associated with cancer.

6.5 Pathway analysis and pathway association with cancer

We identify the pathways enriched in our geneset with statistical significance using the Reactome Pathway database. To identify which pathways have a documented association with cancer, we performed a systematic review of cancer literature using the United States National Library of Medicine's PubMed literature database. This database hosts over 30 million citations and abstracts of life science and biomedical literature. We quantify a pathway's association with cancer as the number of PubMED-curated publications linking such pathways to cancer, tumors, metastasis, or oncogenes.

7 RESULTS AND DISCUSSIONS

7.1 MCDS of naive networks do not identify cancer-specific driver proteins

Table 2 presents the properties of our networks throughout each test run. Naive networks resulted in driver protein sets of similar size. The variance between driver protein set size of naive networks is only 4.33, while that of a priori networks is 931. All naive networks shared 580 nodes (96.83 – 97.48% of these sets) despite these sets having distinct root sets.

We suspect that the large number of edges in naive networks caused this limited specificity across networks. Having more edges implies that more paths would exist between any pair of nodes. In this case, pruning an edge is less likely to cause a disconnect in the graph to any target node, as there would likely be an alternate shortest path towards this target node. The search space for our MCDS algorithm does not gain specificity from a balancing step in dense graphs.

This suggests that MCDS-based driver protein-finding algorithms on balanced naive networks do not find cancer-specific driver proteins, despite being balanced to cancer-specific proteins. These proteins may be driver proteins for the human proteome in general rather than for specific cancers.

Based on their variance in set size and composition, we found greater specificity in driver proteins found with a priori known cancer networks. We focus on driver proteins found from these networks in the following sections.

7.2 MCDS-derived driver proteins are more topologically significant than control groups

We performed an independent t-test on our driver protein sets' measures of centrality against those of Kanhaiya's driver protein networks, to assess the significance of their difference. Table 3 presents the truncated results of this test. From this, we determined that the means for all measures except the clustering coefficient are significantly different.

The median values of our measures are higher than Kanhaiya's across all measures. From this, we claim that our driver proteins are more topologically significant than those of Kanhaiya, though they cluster similarly. Figures 9 and 10 show the distributions of these measures across breast networks. These are consistent across all cancers.

Among all networks, a priori known cancer networks scored the highest median range of all measures of centrality except clustering coefficient (Figure **??**). A priori known networks also had the highest betweenness centrality among networks tested.

In a study on S. cerevisiae protein interaction networks [12], Dirk observed that high betweenness centrality may identify global regulators of a network. They explain these proteins regulate their respective networks because their position in multiple shortest paths allows them to monitor communication between vertices via these shortest paths.

Koschützki and Schreiber also found that essential proteins have higher mean measures of centrality than non-essential proteins. Protein network global regulators and essential proteins may show potential as drug targets for further study.

Table 3: Independent T-test on measures of centrality of our driver proteins and those of Kanhaiya

Network	Measure of centrality	P-value	Means are significantly different?
breast	Betweenness	$2.49x10^{-6}$	yes
	Clustering coefficient	0.15	no
	Total degree/degree centrality	$2.69x10^{-7}$	yes
ovarian	Betweenness	$7.13x10^{-4}$	yes
	Clustering coefficient	$3.85x10^{-4}$	yes
	Total degree/degree centrality	$5.42x10^{-4}$	yes
pancreatic	Betweenness	$2.45x10^{-4}$	yes
	Clustering coefficient	0.23	no
	Total degree/degree centrality	$6.06x10^{-4}$	yes

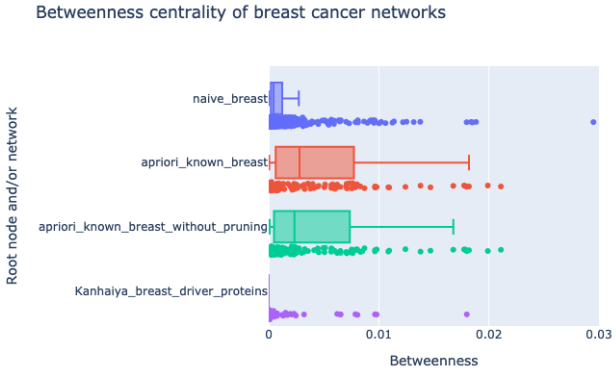

Figure 9: Betweenness centrality of breast cancer networks

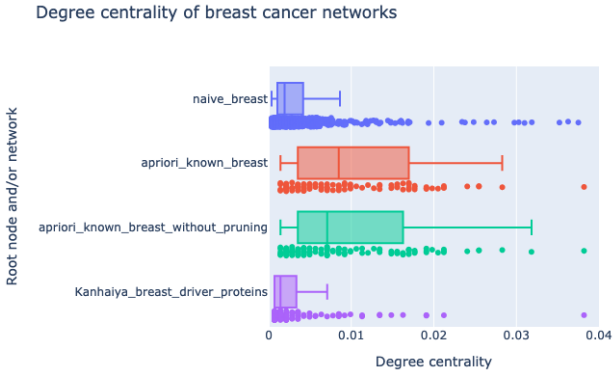

Figure 10: Degree centrality of breast cancer networks

7.3 MCDS-derived driver proteins identify known driver proteins

We validate our driver proteins sets against driver protein sets from Bailey et al [1], a comprehensive list of cancer driver proteins. Figure 11 shows the percent of Bailey proteins in our test results.

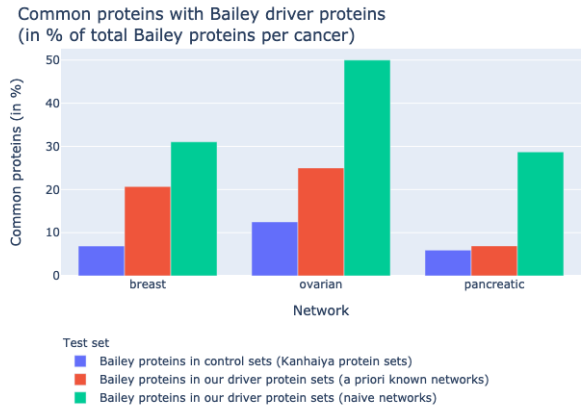

Figure 11: Common proteins with Bailey cancer driver proteins.

Naive network's MCDS identifies the largest percent of Bailey et al oncogenes. This shows the potential of MCDS-based methods in identifying driver proteins in the human proteome. Though we have shown that these MCDS do not identify cancer-specific proteins, the large percent of proteins captured by our naive network's MCDS suggest that MCDS-based methods may still show potential in finding oncogenes in naive PPI networks.

7.4 MCDS-based methods may complement other driver protein-finding methods to find consensus driver protein sets

Figure 12 shows a Venn diagram of driver protein sets from our test runs and validation sets.

Our results for a priori known networks defined more specific sets of driver proteins when complemented by other driver protein-finding algorithms. The consensus of multiple driver-protein finding methods can be used to determine the most significant driver proteins from large sets of driver proteins.

Below we can see that the consensus (intersection) size of our results is comparable in size to those between other methods. Similar Venn diagrams for ovarian and pancreatic cancer may be found in the appendix.

7.5 MCDS-derived driver protein sets enrich pathways linked to cancer

We performed pathway enrichment analysis on our driver protein sets for our networks. The list of pathways enriched with statistical significance for each network are Reactome pathways from Cytoscape.

Our driver protein sets identified proteins enriched for biological processes such as gene transcription, signal transduction, immunity,

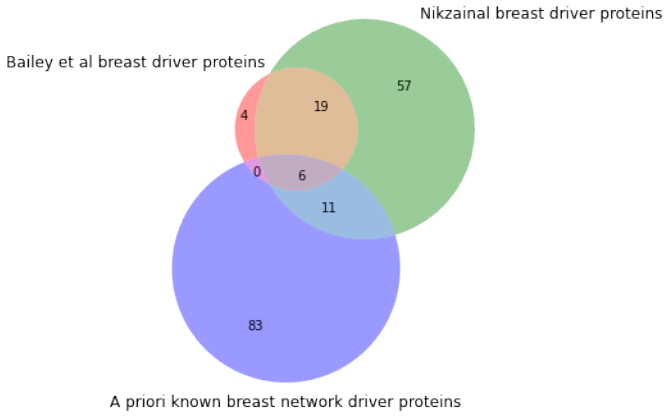

Figure 12: Venn diagram of our driver proteins for breast cancer, and those of Bailey et al, Nikzainal

and apoptosis. Enriched pathways related to breast cancer included pathways for growth signaling and cell cycle regulation.

In our driver protein set for our a priori known breast cancer network, we identified breast cancer-linked pathways for MAPK targets and nuclear events mediated by MAP kinases ($p - value = 2.37x10^{-8}$). This driver protein set also enriched pathways for other breast cancer-related pathways such as VEGFR2 mediated cell proliferation, oxidative stress-induced senescence, apoptotic cleavage of cellular proteins, PLCG1 events in ERBB2 signaling, among others [19].

According to Smolle et al. [22], the EGFR receptor protein is over-expressed in 30% to 98% ovarian cancer cases are present in the pathway for signaling by Receptor Tyrosine Kinases.

In our driver proteins for our a priori known ovarian cancer network, the ovarian cancer-linked pathway for signaling by Receptor Tyrosine Kinases was enriched ($p - value = 6.0x10^{-15}$). In our driver protein sets for pancreatic cancer networks, we identified pathways for MAPK/MAPK3 signaling and Negative regulation of the PI3K/AKT network, which are among the most regularly activated signaling pathways in pancreatic cancer [18].

7.6 Majority of pathways enriched in MCDS-based driver protein set have documented link with cancer

We comprehensively surveyed each pathway enriched in our driver protein sets for their documented association with cancer. To quantify a pathway's association with cancer, we measured the number of publications that linked that pathway to cancer.

Percent of pathways with a modest amount of literature (< 100 publications) associated with cancer is comparable across pathway sets. A larger percent of pathways enriched in our driver protein sets have a fair amount of literature (100 publications) linking these pathways to cancer, compared to the same percentage of all pathways. Pathways enriched in our driver protein sets also have a smaller percent of pathways with no documented association with cancer, compared to the same percentage of all pathways.

Table 4: Percentage of pathways enriched in a priori networks with documented link to cancer

Pathway set from driver proteins set	Association documented in <100 publications	Association documented in >100 publications	No documented association
pathways enriched in breast	43.57%	36.12%	20.32%
pathways enriched in ovarian	42.26%	38.49%	19.24%
pathways enriched in pancreatic	41.51%	38.23%	20.26%
pathways not enriched in breast	50.08%	19.90%	30.18%
pathways not enriched in ovarian	49.77%	19.91%	30.48%
pathways not enriched in pancreatic	50.16%	20.09%	29.91%
All pathways	42.55%	32.46%	24.98%

Our driver protein sets also enriched double the percent of pathways with a fair amount of literature linking these pathways to cancer. Our driver protein sets had 38-38% of pathways enriched with a fair amount of literature linking such pathways to cancers, while only 19.9-20% of pathways not enriched by our driver had a fair amount of literature linking these pathways to cancer. Over 10% fewer pathways had no documented association with cancer in pathways enriched by our driver protein sets.

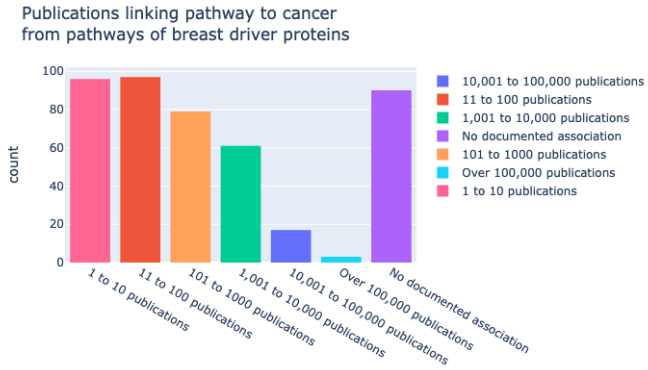

Figure 13: Publication linking pathway to cancer from pathways of breast driver proteins

8 CONCLUSION AND RECOMMENDATIONS

We developed a targeted approach to identifying driver proteins of cancer regulatory networks.

In this approach, we model protein-protein interaction networks as balanced signed directed graphs, rooted in a small ($|V| < 10$) set of root nodes. We then identify driver proteins from as the minimum connected dominating set of this balanced graph. The balancing of networks aims to model the regulatory behavior of gene regulatory networks.

From networks we modeled, we found that resulting driver proteins sets had higher measures of centrality or topological properties than sets of known driver proteins. We observed that the results from a priori defined target disease networks had higher measures of centrality than those from networks with no prior knowledge of their target diseases.

Networks with no prior knowledge (naive networks) may be dense and would not gain specificity from balancing as much as less dense, a priori defined disease target networks may.

Pathway enrichment analysis on our set of driver proteins confirms high enrichment for several pathways linked with specific cancers. A comprehensive literature survey of pathways enriched by our driver proteins showed that a larger percentage of these pathways have a well-documented link with cancer, compared to the same percentage across all pathways. Our driver proteins' pathways also had a smaller percentage of pathways with no link to cancer, compared to the same percentage of all pathways.

Extensions for this study include exploring other cancer networks and cancer-related pathways. Our method may also be used to complement and identify consensus among drug-target databases of these cancer networks. Our algorithm can be further improved by exploring alternate approaches to modeling regulatory behavior.

ACKNOWLEDGEMENTS

REFERENCES

[1] Matthew H Bailey, Collin Tokheim, Eduard Porta-Pardo, Sohini Sengupta, Denis Bertrand, Amila Weerasinghe, Antonio Colaprico, Michael C Wendl, Jaegil Kim, Brendan Reardon, et al. 2018. Comprehensive characterization of cancer driver genes and mutations. *Cell* 173, 2 (2018), 371–385.

[2] Andrew V Biankin, Nicola Waddell, Karin S Kassahn, Marie-Claude Gingras, Lakshmi B Muthuswamy, Amber L Johns, David K Miller, Peter J Wilson, Ann-Marie Patch, Jianmin Wu, et al. 2012. Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. *Nature* 491, 7424 (2012), 399–405.

[3] E. Davidson and M. Levin. 2005. Gene regulatory networks. *Proceedings of the National Academy of Sciences* 102, 14 (April 2005), 4935–4935. https://doi.org/10.1073/pnas.0502024102

[4] Valur Emilsson, Gudmar Thorleifsson, Bin Zhang, Amy S. Leonardson, Florian Zink, Jun Zhu, Sonia Carlson, Agnar Helgason, G. Bragi Walters, Steinunn Gunnarsdottir, Magali Mouy, Valgerdur Steinthorsdottir, Gudrun H. Eiriksdottir, Gyda Bjornsdottir, Inga Reynisdottir, Daniel Gudbjartsson, Anna Helgadottir, Aslaug Jonasdottir, Adalbjorg Jonasdottir, Unnur Styrkarsdottir, Solveig Gretarsdottir, Kristinn P. Magnusson, Hreinn Stefansson, Ragnheidur Fossdal, Kristleifur Kristjansson, Hjortur G. Gislason, Tryggvi Stefansson, Bjorn G. Leifsson, Unnur Thorsteinsdottir, John R. Lamb, Jeffrey R. Gulcher, Marc L. Reitman, Augustine Kong, Eric E. Schadt, and Kari Stefansson. 2008. Genetics of gene expression and its effect on disease. *Nature* 452, 7186 (March 2008), 423–428. https://doi.org/10.1038/nature06758

[5] Antonio Fabregat, Konstantinos Sidiropoulos, Guilherme Viteri, Oscar Forner, Pablo Marin-Garcia, Vicente Arnau, Peter D'Eustachio, Lincoln Stein, and Henning Hermjakob. 2017. Reactome pathway analysis: a high-performance in-memory approach. *BMC Bioinformatics* 18, 1 (March 2017). https://doi.org/10.1186/s12859-017-1559-2

[6] Margherita Francescatto, Marco Chierici, Setareh Rezvan Dezfooli, Alessandro Zandonà, Giuseppe Jurman, and Cesare Furlanello. 2018. Multi-omics integration for neuroblastoma clinical endpoint prediction. *Biology Direct* 13, 1 (Jan. 2018). https://doi.org/10.1186/s13062-018-0207-8

[7] Amanda L. Garner and Kim D. Janda. 2011. Protein-Protein Interactions and Cancer: Targeting the Central Dogma. *Current Topics in Medicinal Chemistry* 11, 3 (Feb. 2011), 258–280. https://doi.org/10.2174/156802611794072614

[8] Andrei A. Ivanov, Fadlo R. Khuri, and Haian Fu. 2013. Targeting protein–protein interactions as an anticancer strategy. *Trends in Pharmacological Sciences* 34, 7 (July 2013), 393–400. https://doi.org/10.1016/j.tips.2013.04.007

[9] S. Bera J. Dizon, J. Malolos and H. Adorna. 2019. Minimum Connected Dominating Sets on Protein-Protein Interaction Networks. In *MCDS on PPI networks*. Proceedings of 19th Philippine Computing Science Congress.

[10] Krishna Kanhaiya, Eugen Czeizler, Cristian Gratie, and Ion Petre. 2017. Controlling directed protein interaction networks in cancer. *Scientific reports* 7, 1 (2017), 1–12.

[11] J. Bantang K.J. Urog and H.N. Adorna. 2018. Dominating Tree Problem Heuristics for Scale-Free Networks. In *DTP*. Proceedings Workshop on Computation: Theory and Practice 2018.

[12] Dirk Koschützki and Falk Schreiber. 2008. Centrality analysis methods for biological networks and their application to gene regulatory networks. *Gene regulation and systems biology* 2 (2008), GRSB–S702.

[13] Tijana Milenković, Vesna Memišević, Anthony Bonato, and Nataša Pržulj. 2011. Dominating biological networks. *PloS one* 6, 8 (2011), e23016.

[14] Jose C. Nacher and Tatsuya Akutsu. 2016. Minimum dominating set-based methods for analyzing biological networks. *Methods* 102 (June 2016), 57–63. https://doi.org/10.1016/j.ymeth.2015.12.017

[15] Serena Nik-Zainal, Helen Davies, Johan Staaf, Manasa Ramakrishna, Dominik Glodzik, Xueqing Zou, Inigo Martincorena, Ludmil B Alexandrov, Sancha Martin, David C Wedge, et al. 2016. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* 534, 7605 (2016), 47–54.

[16] Le Ou-Yang, Dao-Qing Dai, and Xiao-Fei Zhang. 2015. Detecting protein complexes from signed protein-protein interaction networks. *IEEE/ACM transactions on computational biology and bioinformatics* 12, 6 (2015), 1333–1344.

[17] Livia Perfetto, Leonardo Briganti, Alberto Calderone, Andrea Cerquone Perpetuini, Marta Iannuccelli, Francesca Langone, Luana Licata, Milica Marinkovic, Anna Mattioni, Theodora Pavlidou, Daniele Peluso, Lucia Lisa Petrilli, Stefano Pirrò, Daniela Posca, Elena Santonico, Alessandra Silvestri, Filomena Spada, Luisa Castagnoli, and Gianni Cesareni. 2015. SIGNOR: a database of causal relationships between biological entities. *Nucleic Acids Research* 44, D1 (10 2015), D548–D554. https://doi.org/10.1093/nar/gkv1048 arXiv:https://academic.oup.com/nar/article-pdf/44/D1/D548/16661237/gkv1048.pdf

[18] Kishore Polireddy and Qi Chen. 2016. Cancer of the pancreas: molecular pathways and current advancement in treatment. *Journal of Cancer* 7, 11 (2016), 1497.

[19] David C Qian, Jinyoung Byun, Younghun Han, Casey S Greene, John K Field, Rayjean J Hung, Yonathan Brhane, John R Mclaughlin, Gordon Fehringer, Maria Teresa Landi, et al. 2015. Identification of shared and unique susceptibility pathways among cancers of the lung, breast, and prostate from genome-wide association studies and tissue-specific protein interactions. *Human molecular genetics* 24, 25 (2015), 7406–7420.

[20] Javier De Las Rivas and Celia Fontanillo. 2010. Protein–Protein Interactions Essentials: Key Concepts to Building and Analyzing Interactome Networks. *PLoS Computational Biology* 6, 6 (June 2010), e1000807. https://doi.org/10.1371/journal.pcbi.1000807

[21] Georgina L Ryland, Sally M Hunter, Maria A Doyle, Franco Caramia, Jason Li, Simone M Rowley, Michael Christie, Prue E Allan, Andrew N Stephens, David DL Bowtell, et al. 2015. Mutational landscape of mucinous ovarian carcinoma and its neoplastic precursors. *Genome medicine* 7, 1 (2015), 87.

[22] Elisabeth Smolle, Valentin Taucher, Martin Pichler, Edgar Petru, Sigurd Lax, and Johannes Haybaeck. 2013. Targeting signaling pathways in epithelial ovarian cancer. *International journal of molecular sciences* 14, 5 (2013), 9536–9555.

[23] Arunachalam Vinayagam, Travis E. Gibson, Ho-Joon Lee, Bahar Yilmazel, Charles Roesel, Yanhui Hu, Young Kwon, Amitabh Sharma, Yang-Yu Liu, Norbert Perrimon, and Albert-László Barabási. 2016. Controllability analysis of the directed human protein interaction network identifies disease genes and drug targets. *Proceedings of the National Academy of Sciences* 113, 18 (2016), 4976–4981. https://doi.org/10.1073/pnas.1603992113 arXiv:https://www.pnas.org/content/113/18/4976.full.pdf

[24] S. Wuchty. 2014. Controllability in protein interaction networks. *Proceedings of the National Academy of Sciences* 111, 19 (April 2014), 7156–7160. https://doi.org/10.1073/pnas.1311231111

[25] R. D. Jalandoni J. A. Dizon A. T. Young Y. L. Briones, M. R. Castro and H. N. Adorna. 2019. A directed minimum connected dominating set for protein-protein interaction networks. In *Directed MCDS for PPI networks*. Presented at Workshop on Computation: Theory and Practice.