Identifying Novel Functional Protein Complexes as Therapeutic Targets in Breast Cancer Therapy Through a Network Science-Based Approach

RICCARDO CARANGELO

Univeristy of Padua (ID 2057432), Network Science,
riccardo.carangelo@studenti.unipd.it

Compiled January 18, 2024

Breast cancer has always been a worldwide leading cause of mortality, affecting millions of individuals and their families. Despite all the advancements in our capability of an early detection and in treatment options, this disease shows significant challenges due to its heterogeneous nature, involving various subtypes and stages. The need to identify new ways for intervention and therapy is increased by the limitations of current treatments for some specific cases. This study wants to show how a network science approach directed towards biological systems, combined with a wise usage of the existing datasets, can be useful for developing tools that exploit the interactions that are present in the human cells. In this way it can be possible to bring simple and potentially usable results for proposing new study directions in therapeutic research. The application of this tool could help, in this specific case, breast cancer research. By using a high-confidence interaction data from the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) database, a scale-free network model was built, processed, and evaluated for extracting specific metrics. The obtained model was then further refined by incorporating gene expression data from the Human Protein Atlas (HPA), specifically targeting proteins with significant overexpression in breast cancer. Then, communities are extracted and used for performing a functional enrichment analysis using the Gene Ontology graph, revealing functionally coherent subgroups of proteins. Finally, the study employed the Pharos database to identify under-researched protein complexes and functionally related proteins that serve as potential therapeutic targets.

1. INTRODUCTION

Breast cancer, according to the National Cancer Institute definition and characterization, is a cancer type that develops from breast tissue and that includes a large variety of conditions, stages, and subtypes. Nowadays, breast cancer is a global health crisis that affects millions of individuals and places a significant burden on healthcare systems worldwide, since, despite the advancements in early detection methods and treatment modalities, the disease remains a leading cause of cancer-related mortality, with a global affection frequency of 2.2 million people and 685,000 global deaths (data for the year 2020, brought by Sung et al., 2021). This heterogeneous nature can be one of the most challenging aspects of breast cancer, complicating treatment strategies, since therapies that are effective for one subtype may not be necessarily effective for another one. This situation can be further exacerbated by its multifactorial etiology, like other cancer cases, involving genetic, environmental, and lifestyle factors (Allison & Sledge, 2014).

A. Network Science usefulness

The use of network science methods has provided powerful frameworks for understanding complex systems in many fields (Lewis, 2011), including biological systems, in which network science allows for the modeling and analysis of intricate interactions among various cellular components, such as proteins (Barabasi & Oltvai 2004). Protein-protein interaction (PPI) networks offer invaluable insights into cellular functions and mechanisms. By analyzing these networks, researchers can identify key proteins that have a role of hubs, and can be, therefore, involved in critical functions inside the cellular metabolism (Barabasi & Oltvai 2004). This is the context in which this study was performed, trying to leverage the capabilities of network science to investigate the human proteome with the primary objective of identifying novel components inside hubs of protein complexes that could serve as new potential therapeutic targets in breast cancer treatment, both from a functional and a structural point of view.

B. Project Outline

In order to achieve this objective, the study employed a varied approach. Initially, a scale-free network model of the human protein interactome was constructed using high-confidence interaction data from the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) database. This model was processed and analyzed to extract several information about the network and for determining the coefficients of the scale-free property. Then, gene expression data was introduced, selecting breast cancer-specific expressed genes, through the Human Protein Atlas (HPA). Subsequently, hubs-related breast cancer genes were selected and split into different communities within the network are identified and analyzed using modularity metrics, followed by a functional enrichment analysis based on the Gene Ontology (GO) dataset. Finally, the enriched proteins were compared to the Pharos database to identify under-researched protein complexes that are also hubs and that are important in breast cancer metabolic processes. This output can be used for discovering new potentially useful research directions in important related-by-function groups of proteins.

2. MATERIALS

This section provides a more in-depth presentation and characterization of the materials used in this study, in terms of datasets, code tools and libraries.

A. Datasets

For building the human protein interactome network, the STRING database (version 12.0), a well-known resource for protein-protein interactions, was used. In particular, two distinct datasets were fetched from the STRING database:

1. Physical subnetwork (`interaction.txt.gz`) file, which contains the interactome and the combined scores for the interaction accuracy (in this dataset the edges indicate that the proteins are part of a physical complex);

2. Mapping File (`mapping.txt.gz`), used to associate STRING IDs with gene names.

For fetching another mapping file (`PTHR18.0_human`), useful for mapping each gene ID to its respective UniProt ID, PANTHER db was used. PANTHER db is a comprehensive database that has been designed to classify proteins (and their respective genes) in order to facilitate high-throughput analysis.

The Human Protein Atlas and the Ensembl service (HPA, version 23.0 and Ensembl version 109) provided an updated pathology information file (`pathology.tsv.zip`), that shows protein expression data in various tissues and pathological conditions. Specifically, the HPA dataset shows staining profiles for various proteins in human tissues based on immunohistochemistry in tumor cases, using tissue micro-arrays and a log-rank *p*-value for a Kaplan-Meier analysis of correlation, performed to look at mRNA expression levels in relation to patient survival. The tab-separated file includes Ensembl gene identifier (*Gene*), gene name (*Gene name*), tumor name (*Cancer*), the number of patients annotated for different staining levels (*High*, *Medium*, *Low* and *Not detected*) and log-rank *p*-values for patient survival and mRNA correlation, with the following meanings:

- "Prognostic – favorable": a lower *p*-value indicates that a greater gene expression is correlated with a higher survival probability (i.e., favorable);
- "Unprognostic – favorable": a lower *p*-value indicates that there is not a significative correlation between gene expression and survival, but, if present, is favorable;
- "Prognostic – unfavorable": a lower *p*-value indicates that a greater gene expression is correlated with a lower survival probability (i.e., unfavorable);
- "Unprognostic – unfavorable": a lower *p*-value indicates that there is not a significative correlation between gene expression and survival, but, if present, is unfavorable.

The Gene Ontology (GO) dataset was used to download the `go-basic.obo` and `goa_human.gaf.gz` files from the Gene Ontology consortium, in order to provide structured information about the functions of genes and gene products. The GO dataset is a standardized vocabulary system used to describe the functions of genes and gene products across different databases.

The GO dataset is organized into three main categories:

1. Molecular Function (MF): which describes the specific activities of gene products at the molecular level;

2. Cellular Component (CC): which describes the locations, at the levels of subcellular structures and macromolecular complexes, where a gene product performs its function;

3. Biological Process (BP): which describes broader biological goals or pathways that are accomplished by multiple molecular activities.

The GO database is structured as a directed acyclic graph (DAG), in which nodes represent GO terms and edges represent relationships between these terms. This structure is particularly suitable for representing hierarchical relationships of the Ontology. In particular, each of the three categories described above is the root of the GO DAG. Furthermore, there are several kinds of relationship in the GO network, while the most common one is the "is-a" relationship, which denotes a subclass relationship. For instance, "DNA repair" is-a "DNA metabolic process," means that "DNA repair" is a specific type of "DNA metabolic process." Other relationships include "part-of" and "regulates." By descending this network, from the root to its leaves, it is possible to meet more and more specific nodes. Leaf nodes are, thus, the most specific GO terms.

In the scope of this project, the "go-basic.obo" is a version of the Gene Ontology that is filtered to exclude certain advanced or specialized types of relationships and terms. It is provided in the OBO (Open Biomedical Ontologies) format, which is a text-based format used to represent ontologies. The `goa_human.gaf.gz` file is part of the Gene Ontology Annotation (GOA) and provides annotations for human genes, linking them to specific GO terms. This is a GAF file, i.e., a "Gene Association Format" file, which is a standard format for representing annotations in GO.

Swiss-Prot is a component of the Universal Protein Resource (UniProt) and is one of the most widely used protein sequence databases in the world, since it is the most important curated database of proteins, that ensures a high level of accuracy and reliability. Swiss-Prot was used to fetch the `uniprot_sprot.fasta.gz` file from the UniProt database, so as to get all annotated proteins to be used for the enrichment analysis.

Finally, Pharos is a comprehensive and integrated knowledgebase that focuses on the "druggability" of the human genome,

with the primary goal of aiding researchers in identifying potential therapeutic targets based on various criteria. This database was used to get the `TCRDv6.1.0_ALLexp.csv` file, which provides information related to target development levels (TDL). In particular, to categorize proteins based on their research status, the Pharos TDL defines four categories:

1. Tdark, this category is reserved for the most enigmatic proteins. These are targets about which virtually nothing is known. They represent the unknown territories of the human genome, offering both challenges and opportunities for researchers;

2. Tbio, these are targets that, while biologically significant, do not have known developed drugs or small molecule interactions. They represent proteins that are understood to some extent but haven't been extensively explored in the context of drug interactions;

3. Tchem, these targets have at least one ChEMBL compound with an activity cutoff of less than 30 nm;

4. Tclin, these category represents proteins with at least one approved drug.

B. Code and Libraries

For performing all the analysis, Python 3 was used in a Jupyter Notebook environment, including several libraries, as general purpose libraries like NumPy and Pandas for scientific computing and array data structure manipulation. Python standard libraries like Collections, GZip, ZipFile, and Shutil, were used for file compression and management. Then, TQDM, and Matplotlib were useful for plots and workflow visualization. NetworkX, SciPy and SciKit Learn were used for graph manipulation, metrics calculation and plotting. Finally, GOATOOLS was used for parsing the Gene Ontology structure and for performing the enrichment analysis.

Two Python functions were defined for this specific project:

1. A `robustness_analysis` function, for evaluating the graph robustness under different node removal strategies and metrics.

2. A `subgraph_of_nodes_and_neighbors` function, for recursively generating sub-graphs from a given set of nodes and their neighbors up to a specified-by-user depth.

3. METHODS AND RESULTS

This section is dedicated to show the methods used and the results obtained. All the process is split into small sub-sections.

A. Graph Processing

The first step was to download the STRING physical subnetwork dataset. In this case, only proteins with an interaction accuracy score of at least 0.9 (highest confidence) were selected, in order to build a "core" network of guaranteed-by-annotation protein interactions. In this process 16060 rows were removed from the original dataset (corresponding to 17.87 % of the total rows and each row representing a relation between two proteins).

Using the two mapping files introduced above, it was possible to get a triple identifier for each node, consisting in its STRING ID (i.e., the STRING internal identifier for each node protein), its UniProt ID (i.e., the name of the interacting domain protein) and its gene ID (i.e., the name of the gene coding for such protein). Using NetworkX, a graph object was made. The nature of the interactions among these proteins brought to an undirected graph, which was expected to be a scale-free (this aspect will be tackled later in this paper). This first graph has 8706 nodes and 44931 edges. However, the intrinsic relationships' pattern and the pre-processing performed during the STRING import, brought to several connected small components, shown in table 1.

All such small, connected components represent groups of protein domains whose physical interaction is proved by solid annotations (since the accuracy of the connections was requested to be above 0.900). However, there are not still reliable studies to place these groups in the more general context of the human protein interactome, which is why it is possible to observe them as isolated groups, after the STRING filtering process that pruned the network to remove all low-accuracy interactions. These small components lose their meaning in the macro-context of the interactome holistic study and, therefore, were removed, in order to only leave the biggest connected component of 7167 nodes and 43180 edges. After this further pruning step, 633 nodes were removed, corresponding to 8.83 % of the previous 8706-nodes graph.

With this final pre-processing step, it was possible to get a final core protein interactome human network, containing the most studied interacting proteins, with the highest interaction accuracy and a single big, connected component. Even before any further analysis, it is easy to speculate that the most important and well-studied human master regulators could appear in such a network.

B. Metrics Calculation and Inference

Now, it was possible to perform some network-specific statistics, to better know the network obtained above. In particular, the following metrics were evaluated: a density of 0.0017, an average degree of 12.0497, an average clustering coefficient of 0.4845, and an approximated betweenness centrality of 0.0016.

All this information was obtained using NetworkX, however, other metrics were calculated, using another method. Because of the slowness of NetworkX, due to the large size of the graph, it was necessary to take another calculation method. As suggested by the low density of 0.0017 just seen above, it was indeed observed that the adjacency matrix corresponding to this graph has a very high sparsity. This property can be a great advantage to make calculations faster and, therefore, the remaining metrics, were calculating by extrapolating the adjacency matrix and exploiting its high sparsity property, which is about 99.83 %. By using SciPy, it was possible to firstly extract the CSR (Compressed Sparse Row) format, which provided a new way of storing sparse matrices efficiently, using the following components:

1. a value array, containing all non-zero elements in the matrix;

2. a row pointer array, containing the starting index for each row in the value array;

3. a column index array, containing the column index for each element in the value array.

It is easy to notice that this is a memory efficient format, that is also ideal for make computation faster, since operations on zero elements (which are the large majority in sparse matrices) can be

skipped entirely. SciPy offers a specific function for calculating all the distances using the CSR format, by exploiting the Dijkstra algorithm. In this way, using the matrix of the distances, it was possible to calculate the average distance, which is 2.7713, and the diameter, which is 18.

In general, as mentioned above, considering the low density and high sparsity, it can be deduced that the graph is very sparse, so there is only a small fraction of possible connections between its nodes. Also, the high average clustering coefficient in combination with the low average betweenness centrality suggest the presence of several protein communities that are highly interconnected, but possess a number of bridge nodes, which link the various clusters together. These bridge nodes are represented by proteins with a high betweenness centrality and could play a crucial role in the transmission of information between pairs of different clusters. Furthermore, the discrepancy between the average distance and the diameter can be considered remarkable. In fact, while most proteins can communicate rapidly, there are some pairs of proteins that are exceptionally far apart in the network. This last fact could suggest the existence of very specific signaling pathways travelling in different clusters.

By incorporating all this information, it is possible to conclude that the metrics describe a graph that represents a system in which there are several subsystems of proteins that perform specific functions and are highly interconnected. However, the communication or interaction among these clusters is mediated by a small number of key proteins. These bridge proteins could play a crucial role in coordinating or regulating functions between different cellular subsystems. In other words, all these metrics confirm the typical structure of a protein interactome graph, that appears, indeed, as a complex and highly integrated biological system and, gives a proof of a well-built model that seems to well-capture the characterizing elements of the human protein network.

C. Degree Distribution Analysis

The analysis of the degree distribution was performed to understand the network topology. A well-built human protein interaction network should show a scale-free topology for some given $k_{\min}$. In Figure 1 it is possible to see a degree distribution with a very big number of nodes that show a small number of connections, while a very small amount of nodes that show a very high number of connections. This trend reveals a very fast decay, in terms of frequency, as the degree increases. There is, therefore, a small number of proteins that physically interact with a huge number of other proteins, while a large amount of proteins interacts with a small number of other proteins. The proteins with the highest number of interactions are the best candidate when looking for master regulators and usually have a fundamental role in the network, in terms of controlling and regulating cell metabolic processes.

This first qualitative analysis can provide fundamental suggestions about the topology of the network. However, it is necessary to go deeper. Many real-world networks, including biological, technological, and social ones, show, indeed, a degree distribution that follows a power law (feature that is also suggested by the qualitative analysis of the degree distribution presented above), which is described by the following probability distribution:

$$\mathbb{P}[k] = Ck^{-\gamma} \sim k^{-\gamma}$$

Where k is the degree variable of the network, while C is a coefficient determined by the total normalization condition, and γ is called the exponent parameter. Passing to the logarithm:

$$\log \mathbb{P}[k] = \log Ck^{-\gamma} \sim \log k^{-\gamma} = -\gamma \log k$$

So, it is possible to simply interpret this relation as a line, by putting $Y = -\gamma X$ where $Y = \log \mathbb{P}[k]$ and $X = \log k$. In other words, by plotting the log-log scale of the degree distribution, the value of γ can be found by finding the minus coefficient of the linear interpolation line. So, in this interpretation $-\gamma$ is basically the slope of the shown line. This fact implies the presence of a linear relationship between the degree distribution against the degree frequency on a log-log scale (as it is possible to see in Figure 2).

It is also possible to see the linear interpolation in the Complementary Cumulative Distribution Function (CCDF) in Figure 3, which shows the log-log plot of the CCDF (calculated as 1 – CDF, i.e., the Cumulative Distribution Function) against the degree. This is another common way to visualize and analyze the degree distribution that shows the probability that a randomly chosen node has a degree greater than k.

The interpolation line gives a $\gamma = 2.1092$, with a $k_{\min}$ of 26, chosen by maximizing the γ value on the set of all the degrees. In this way it was possible to calculate the C coefficient which is 41.1603.

So, by considering $k_{\min} = 26$, it is possible to show that the network presents a scale-free topology. In this context, it is important to perform a proper analysis to identify the hubs of the network, which will be the central part of this study from now on.

D. Hubs Identification

A fundamental part of this investigation wants to focus on identifying the hub nodes, i.e., the nodes that hold the greatest number of connections. By building a degree dataset for each node and by sorting such dataset by degree in a descending way, it was possible to identify the top 10 most connected nodes, shown in the table 2.

In this table it is possible to see genes that code for proteins associated with the ubiquitin system, such as RPS27A, UBA52, UBC, and UBB. Their role in the ubiquitination process is fundamental to cellular homeostasis. By ensuring timely protein turnover and eliminating damaged or unnecessary proteins, the ubiquitin-proteasome system acts as a guardian of protein quality and regulator of protein levels. Given its foundational role in cell function, it is entirely logical for ubiquitin-related proteins to be deeply interconnected within the network.

Moreover, the prominence of ribosomal proteins, including RPS27A, UBA52, RPS23, RPS6, and RPS16, underscores the central place that protein synthesis occupies in cellular biology. Ribosomes, the cellular factories where genetic information is translated into functional proteins, have several interactions. Their extensive array of interactions reflects the myriad regulatory and structural proteins they engage with, establishing them as central hubs in the network.

Additionally, the central role of histone proteins, such as H3C13, H4C6, and H3C12, must be considered. Responsible for the task of packaging vast stretches of DNA within the compact cell nucleus, histones also play a fundamental role in gene regulation, modulating the accessibility of DNA to transcription machinery and acting as gatekeepers of gene expression. This dual role, both as structural scaffolds and regulatory entities,

necessitates extensive interactions, reinforcing their position as integral hubs in the network.

In order to select actual hubs from the degree-sorted nodes table, a 90-th percentile degree threshold was set for each node. All the nodes with a degree greater or equal to this threshold were selected to be hubs. In this way a degree threshold of 31 was set and 724 hubs were found. The hubs of this scale-free network are critical nodes, playing a basic role in cell metabolism, while making the network very robust to random attacks, but also fragile to specific attacks, as it will be seen in the next section.

E. Robustness Analysis

Robustness, in this context, refers to the ability of the network to maintain its general structural properties in case of perturbations. These perturbations can be random (such as the random failure of nodes) or targeted, like a deliberate attack on the most specific nodes. The robustness analysis presented here is designed to evaluate how the human protein interaction network built before responds to these perturbations. Since the network has a scale-free topology, it is expected to be robust to random attacks, while weak to targeted attack with a preference for hub nodes (Figure 4).

The robustness analysis consists in simulating the removal of nodes from the network and observe the subsequent changes in its structural properties, using specific metrics. Two distinct strategies for node removal were employed:

1. Random removal, in which, nodes are selected at random and removed from the network, simulating random failures in the network (like random non-silent mutations that make a gene not-working in terms of transcription);

2. Targeted removal, in which the removal strategy involves the deliberate removal of the hubs, simulating focused failures (like what happens in many important diseases).

The metrics used are the following:

1. Size of the largest connected component (relative to the original network size), in this case, the larger this component is, the more robust the network is considered;

2. Density value, which evaluates the relative density of the network after node removal using the density metric, which is defined as the ratio of the number of actual connections to the number of possible connections in a network.

The robustness analysis was conducted over a range of fractions (from 0 to 0.99), with a resolution of 50 points (Figure 4). In this way, it was possible to evaluate the network's response to several levels of node removal, from minimal to almost complete removal. The results of this analysis are predictably in line with what should be observed in a scale-free graph, strongly emphasizing the great structural importance of hubs, which are crucial in preserving the overall topology of the network, ensuring its functionality and effectiveness. Furthermore, the analysis highlights the resilience that this specific kind of network topology possesses against random attacks on its nodes.

Hub proteins have a primary role in controlling and preserving the cell functions, centralizing the upstream functions' control of the metabolic processes. The focus on hubs and their closest neighbors (to be able to take into accounts entire protein complexes or interacting protein pairs) is important in terms of studying the effect of specific illnesses.

F. Breast Cancer Genes Selection and Enrichment

The HPA cancer updated dataset was used to select genes that are specifically related to breast cancer and that can be considered significant. To do this, only prognostic and unfavorable cases were selected, using a p-value threshold of 5 % and considering only breast cancer patients with at least one record of high staining levels. This method allowed the selection of highly breast cancer-correlated genes in terms of differential positive expression and unfavorable prognosis. The group of highly correlated genes was then extracted from the interactome graph. Then, direct neighbors of each breast cancer-related gene were also included. These neighbors represent proteins or complexes that physically interact with the breast cancer-related genes. In this sub-graph, only hubs were selected for an enrichment analysis (Figure 5).

In order to facilitate the enrichment analysis and to filter out most isolated proteins, the sub-graph was divided into communities using the label propagation algorithm. The Label Propagation Algorithm (LPA) is a community detection method in which each node in the network is firstly assigned a unique label. The algorithm then proceeds iteratively, adopting for each node the label that most of their neighbors currently have, with connections broken arbitrarily. This process of label adoption and updating continues until convergence, where nodes no longer change their labels. At the end of the algorithm, nodes with the same label are considered to be part of the same community.

In this way, the top 5 most populated communities were selected for the enrichment process using Gene Ontology, getting in this way hub proteins of the network, that are highly expressed in breast cancer in case of unfavorable prognosis and that also tend to form communities. The aim was to work on proteins that are important for the network and for the disease therapy, and that are also related to each other. The enrichment step was performed using the top 5 communities mentioned above against all the Swiss-Prot proteins, that constitute a highly reliable baseline for making comparison. The enrichment was propagated to parents along the Gene Ontology DAG, proceeding backward until the root of the graph. Furthermore, the edges considered, are "is_a" edges, a fundamental and straightforward relationship in the GO hierarchy that denotes a subclass relationship between two GO terms. The used method to filter the most significant proteins is a False Discovery Rate (FDR) correction using the Benjamini-Hochberg procedure with a 5 % p-value threshold (Benjamini & Hochberg, 1995). The analysis was performed five times, one time for each of the five communities. The results of this analysis led to small groups of proteins that are particularly linked to each other because of the presence of some specific function they perform. This new functional way of grouping unfavorable breast-related hub proteins was the baseline for selecting new potential therapeutical targets both in terms of proteins and functions.

G. Less Studied Proteins Selection

Now that groups of high-interest proteins were found, it is possible to select the proteins that are less studied, in order to suggest new potential directions that could help both basic research and biomedical study. For this final part, the Pharos notation is used.

For the purpose of quantifying the level of research on each significantly enriched protein, a simple inverse metric was developed. In such "research score" system, the more unknown a protein is, the higher its score is, resulting in the following logic:

1. Tdark proteins, being the least studied, were given a score

of 1;

2. Tbio proteins received a score of 0.5, indicating a moderate level of understanding;

3. Tchem and Tclin proteins, being more researched and known, were assigned scores of 0.25 and 0, respectively.

These points are protein-specific, but, in order to assign a final score to each function (i.e., each small group of enriched proteins), the final score of each group was computed as the average of the individual scores of the proteins contained in such group. The underlying logic was to gauge the overall research status of the entire protein group, once again promoting a holistic approach rather than focusing on individual proteins.

Final reliable results combine the enriched groups of proteins with a FDR correction *p*-value threshold that was less than 5 % (for getting only the significant enrichments) with the groups of proteins with an average research score greater than 0.75 (for getting the most unknown functions, with their respective groups of proteins). The table 3 provide the top 10 interesting results of this research.

By excluding groups with single proteins (i.e., the first three trivial groups, containing just one protein), it is possible to notice several groups with 3 proteins and a last one with 71 proteins. The function that links these proteins is reported in the first columns, while the last column shows the score that highlights the lack of studies for these groups and, therefore, for the function represented by these groups.

4. CONCLUSIONS

As already mentioned above several times, by using the pipeline presented in this paper, it was possible to get a list of cellular functions (with their respective group of proteins), here summed up for a quick recap:

1. each cellular function is significant for the selected group of proteins (FDR GO significance);

2. all the physical protein interactions among proteins have a likelihood of at least 90 % (STRING accuracy score) and represent interactions patterns or protein complexes;

3. all the functional groups of proteins are constituted by hubs of the scale-free core human protein interactome network (network analysis results);

4. all the functions and their proteins are highly correlated, in terms of expression, with human breast cancer in experimental cases with non-favorable prognosis (HPA significance score);

5. all the functions and their proteins are unknown both from the biomedical and the basic research points of view (Pharos-derived metric).

Just for providing an instance of preliminary analysis, it is possible, by looking at the table above, to analyze the functions that popped up from the pipeline. For instance, important groups of proteins that appear often are related with histone modifications. Histones are the protein components of chromatin and play a crucial role in packaging the DNA for regulating gene expression and keep the inner nucleus structure. Post-translational modifications of histones, such as mono-ubiquitination, are very important in dictating the transcriptional status of genes and the active protein profile of a cell in a specific moment. Aberrations in histone modifications can lead to dysregulated gene expression profiles, contributing to the multifaceted nature of cancer. Despite histones and their modifications are still widely studied in relation to cancer, this specific pipeline highlighted groups of genes related to such function that are still understudied and underknown. The same can be said about mitochondrial translation. So, this kind of approach can allow researches to find functional aspects of a research area that are still not treated, but it can also used to find out totally new functions that were ignored before.

The method presented in this study could be improved and adapted for other human pathologies or other organisms, such as, for example, the search for pathological patterns within model organisms. Furthermore, although in this study proteins and functions were considered on the basis of their minimum research level, it might also be interesting to use Tbio as target category, in order to provide known proteins and functions for which there are not yet interesting therapeutic compounds.

Finally, other features generated during the enrichment can be considered for selecting important functions and group of proteins; for instance, the depth of the function in the GO network can be an interesting metric of specificity for considering new potential research ways.

In conclusion, the possibility of carrying out such an analysis was due to the powerful tools provided by graph theory applied to molecular biology and structural biochemistry. All this work highlights how a holistic approach to biomedical sciences can allow high-throughput data to be used to bring scientific value, with potentially strong implications in biomedical and clinical research fields.

5. REFERENCES

Aleksander, S. A., Balhoff, J., Carbon, S., Cherry, J. M., Drabkin, H. J., Ebert, D., ... & Zarowiecki, M. (2023). The Gene Ontology knowledgebase in 2023. *Genetics*, 224(1), iyad031.

Allison, K. H., & Sledge, G. W. (2014). Heterogeneity and cancer. *Oncology*, 28(9), 772-772.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., ... & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1), 25-29.

Barabasi, A. L., & Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature reviews genetics*, 5(2), 101-113.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1), 289-300.

Kelleher, K. J., Sheils, T. K., Mathias, S. L., Yang, J. J., Metzger, V. T., Siramshetty, V. B., ... & Oprea, T. I. (2023). Pharos 2023: an integrated resource for the understudied human proteome. *Nucleic acids research*, 51(D1), D1405-D1416.

Lewis, T. G. (2011). Network science: Theory and applications. John Wiley & Sons.

Mi, H., Dong, Q., Muruganujan, A., Gaudet, P., Lewis, S., & Thomas, P. D. (2010). PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. *Nucleic acids research*, 38(suppl_1), D204-D210.

Martin, F. J., Amode, M. R., Aneja, A., Austine-Orimoloye, O., Azov, A. G., Barnes, I., ... & Flicek, P. (2023). Ensembl 2023. *Nucleic acids research*, 51(D1), D933-D941.

SEER Training Modules, Module Name. U. S. National Institutes of Health, National Cancer Institute. Day Month Year (of access) https://training.seer.cancer.gov/.

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 71(3), 209-249.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., ... & Von Mering, C. (2015). STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic acids research*, 43(D1), D447-D452.

Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., ... & Narechania, A. (2003). PANTHER: a library of protein families and subfamilies indexed by function. *Genome research*, 13(9), 2129-2141.

Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., ... & Pontén, F. (2015). Tissue-based map of the human proteome. *Science*. Retrieved from https://www.proteinatlas.org

UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Research*, 2023, 51.D1: D523-D531.

6. TABLES AND FIGURES

Table 1. Component size and numerosity for each isolated component.

Component size	Component numerosity
7167	1
17	1
16	2
12	4
11	1
10	3
9	1
8	6
7	7
6	10
5	22
4	33
3	127
2	306

Table 2. Degree and function for each gene.

Gene name	Protein name	Degree	Function
RPS27A	Ubiquitin-40S ribosomal protein S27a	255	Involved in protein synthesis and degradation processes
UBA52	Ubiquitin-60S ribosomal protein L40	235	Plays a role in protein translation and the ubiquitin-proteasome pathway
UBC	Polyubiquitin-C	172	A chain of ubiquitin molecules involved in marking proteins for degradation
UBB	Polyubiquitin-B	168	Another chain of ubiquitin that targets proteins for degradation
H3C13	Histone H3.2	158	A core component of the nucleosome, vital for DNA packaging in the nucleus
H4C6	Histone H4	147	Another core component of the nucleosome, essential for DNA packaging and accessibility
RPS23	40S ribosomal protein S23	139	A component of the ribosome, the cell's protein factory
RPS6	40S ribosomal protein S6	136	Plays a role in protein synthesis, especially during cell growth and proliferation
RPS16	Ribosomal protein S16	132	Another component of the ribosome involved in protein synthesis
H3C12	Histone H3.1	132	A variant of Histone H3, crucial for nucleosome structure and function

Table 3. Functional analysis results.

Function	Number of proteins	Score
prenylation	1	1.00
protein prenylation	1	1.00
inactivation of paternal X chromosome by genomic imprinting	1	1.00
histone H2A-K119 monoubiquitination	3	0.83
protein monoubiquitination	3	0.83
histone ubiquitination	3	0.83
histone monoubiquitination	3	0.83
histone H2A ubiquitination	3	0.83
histone H2A monoubiquitination	3	0.83
mitochondrial translation	71	0.80

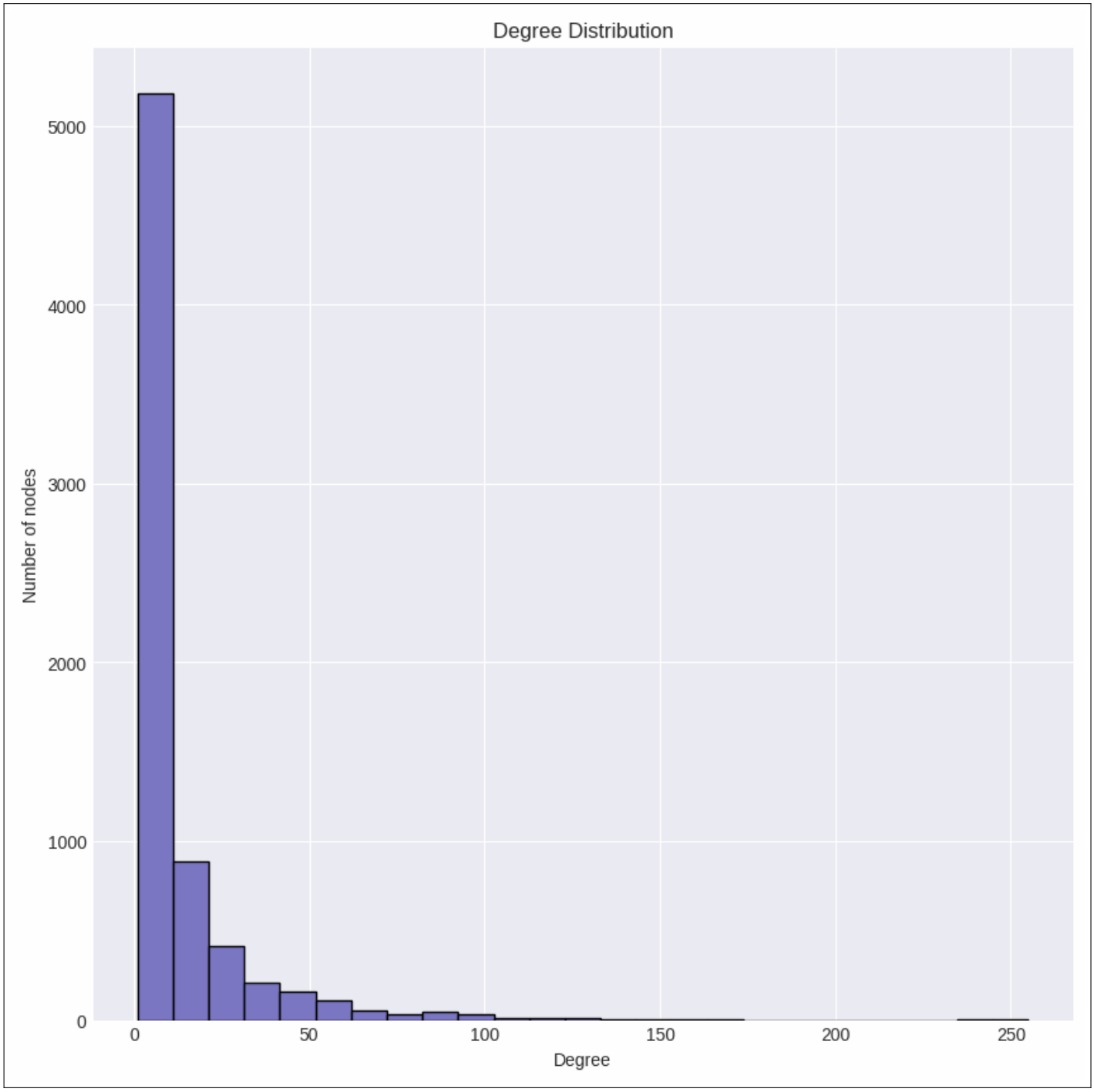

Fig. 1. This plot indicates the distribution histogram of the degree and offers an immediate view of how the degrees are distributed across the network, showing a steep decay of the degree frequency.

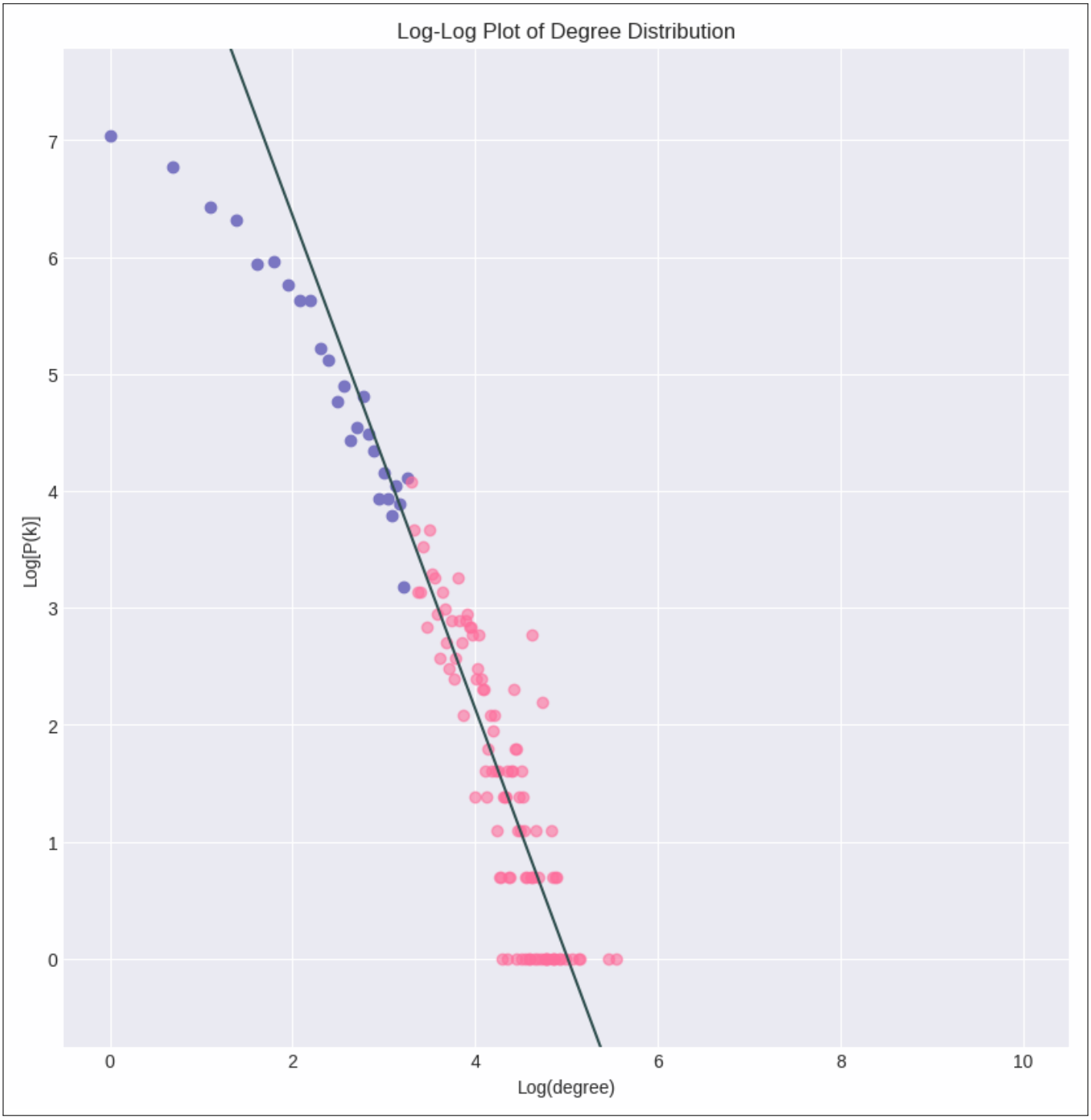

Fig. 2. Degree distribution scatter plot on a log-log scale. The line was calculated by using the linear regression interpolation with minimum k of 26 (determined by maximizing the γ value on the set of all the degrees). Purple points are under the minimum k, while pink points are above the minimum k.

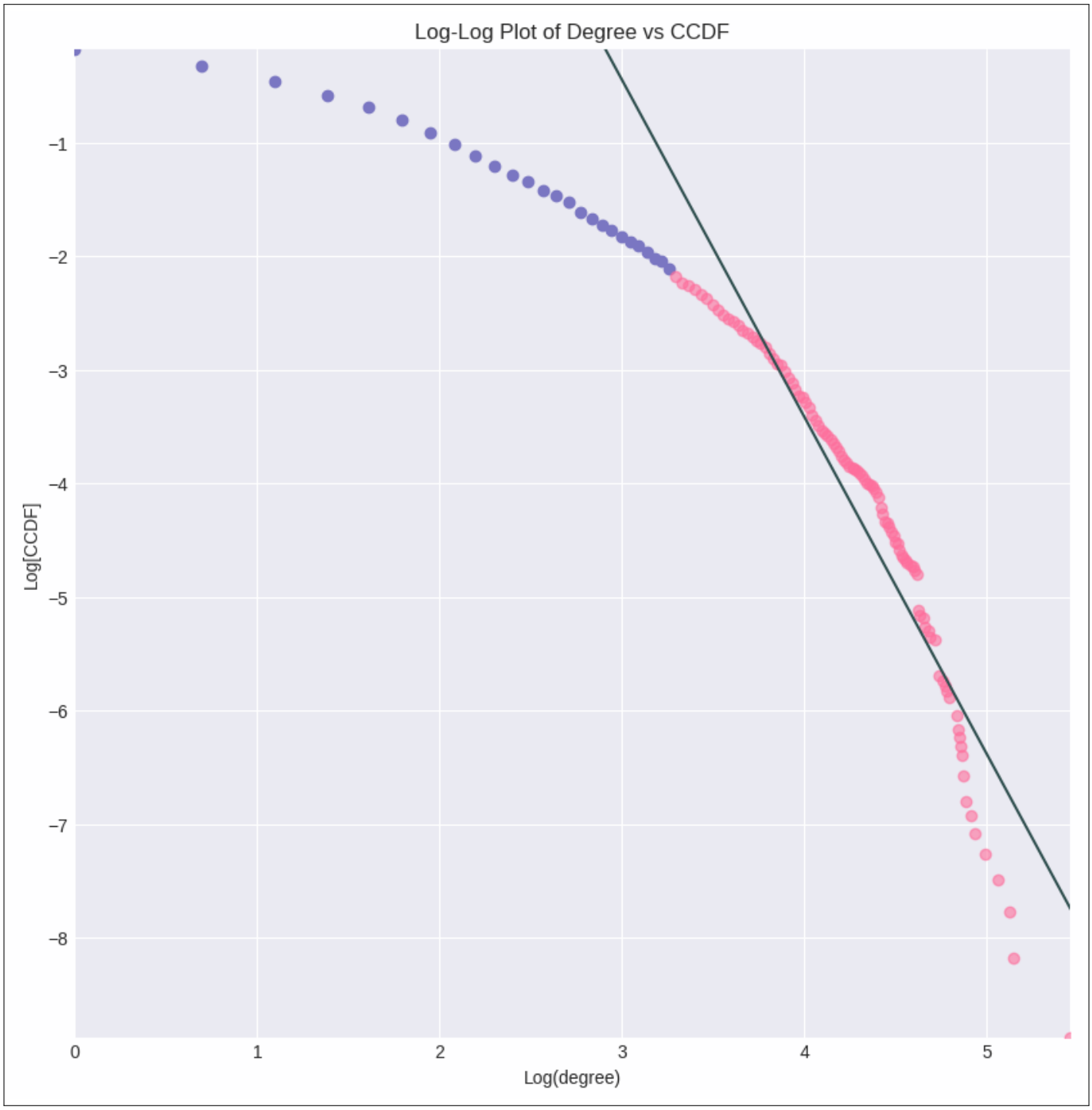

Fig. 3. Log-log plot of the Complementary Cumulative Distribution Function (CCDF) against the degree distribution. The line here derives by the linear interpolation line calculated, considering the minimum k of 26. Purple points are under the minimum k, while pink points are above the minimum k.

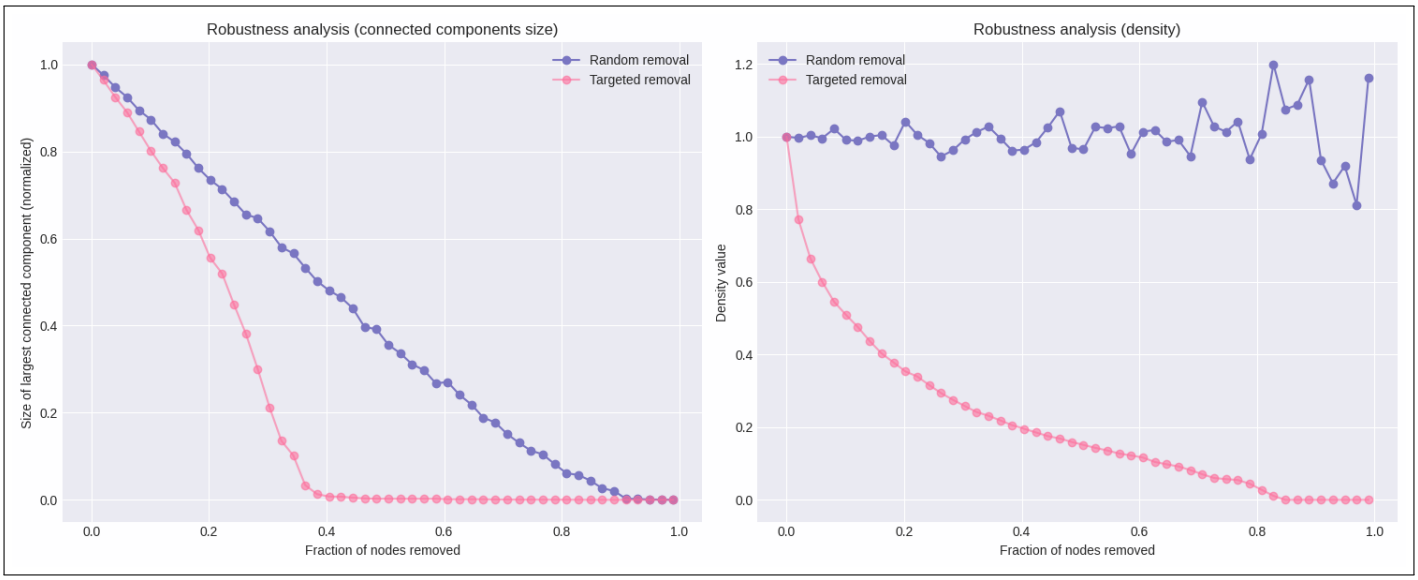

Fig. 4. Random attacks (purple plot) vs targeted attacks (pink plot). On the left, the robustness was derived from measuring the size of the largest connected component during the attack process; on the right, the robustness was derived from measuring the density during the attack process.

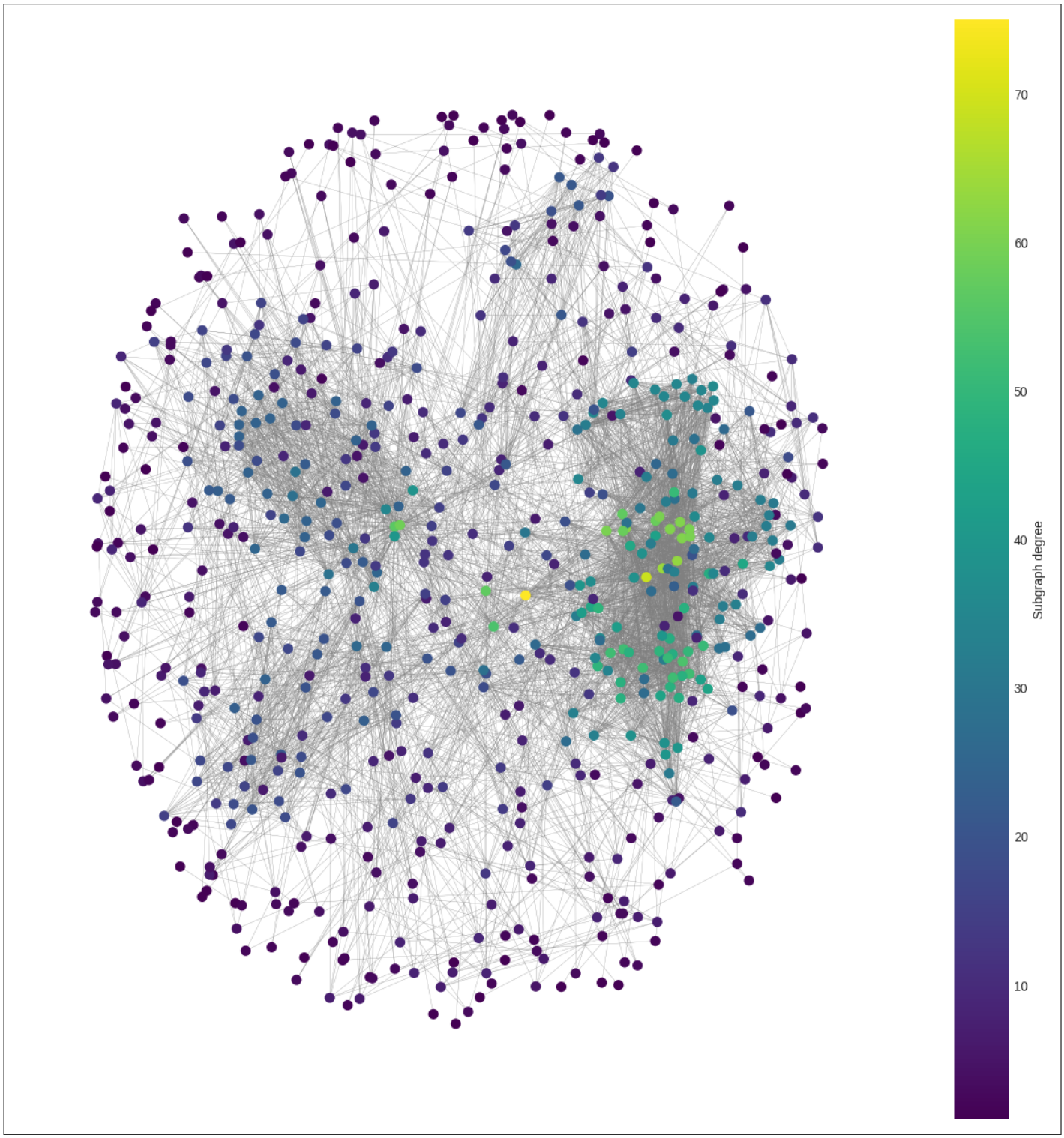

Fig. 5. Sub-graph of the hub nodes (with their immediate neighbors) selected for the enrichment analysis, with a gradient coloration to show the degrees in the sub-graph. This core network represents a set of highly important (hubs) proteins that are also highly correlated (in terms of expression) to the presence of breast cancer.