RESEARCH

Introduction to Focus Areas in Bioinformatics Project Week 10

Sina Glöckner[*], Christina Kirschbaum, Swetha Rose Maliyakal Sebastian and Gokul Thothathri

[*]Correspondence:
sina.gloeckner@fu-berlin.de
Institute for Informatics, Freie Universität Berlin, Takustr. 9, Berlin, DE
Full list of author information is available at the end of the article

Abstract

Goal of the project: This project aimed to analyze the protein interaction network of *Homo sapiens* by comparing it to an Erdos-Rényi network and a Lattice network as well as investigating its small world properties.

Main results of the project: We were able to successfully build a network and analyze its properties.

Personal key learnings:
Sina & Christina: We learned about utilizing NetworkX to analyze and visualize protein interaction networks.
Gokul: Understood about small world networks and NetworkX tool
Swetha:Understood gene ontology enrichment and modularity-maximisation for community detection

Estimation of the time:
Sina: 8h
Christina: 8h
Swetha: 6h
Gokul: 7h

Project evaluation: 2

Number of words: 1550

1 Scientific Background

BioGRID (Biological General Repository for Interaction Datasets) is an open-access database dedicated to the curation and archival preservation of protein, genetic, and chemical interactions for all major model organism species as well as humans [1]. Chemical–protein interactions for human therapeutic targets collected from the DrugBank database and manually curated bioactive compounds reported in the literature are captured by BioGRID [1]. BioGRID's genome-wide CRISPR/Cas9-based screens that show gene-phenotype and gene-gene interactions include a new dedicated feature. The biological network concept has been utilized to help with drug target selection, drug resistance, and off-target effects interpretation, and it is the foundation for targeted therapies and personalized medicine [1].

Proteins are molecules made up of amino acids that are linked together in a certain sequence to perform various activities within the organism. These molecules can catalyze chemical reactions, act as signals, or form part of the cell's structure [2]. Protein-Protein interactions (PPIs) are physical interactions between proteins that occur as a result of selective molecular docking in a biological environment. The interactome is a full map of protein interactions that can occur in a living organism

[2]. Proteins can connect in polyadic interactions as well as in pairs, forming multi-protein complexes, which are persistent interactions between many proteins [3].

PPI-data can be invaluable in translational contexts [4], by explaining mechanisms of infection spread and discovering novel targets, such as dependency factors, in addition to their role in elucidating a mechanistic understanding of fundamental biological processes from the molecular to the evolutionary scales [5]. The importance of automated PPI analysis is highlighted by the intricacy of protein-protein interactions, as well as the amount and noise of PPI data. A protein-protein interaction network (PPIN) is the typical way of expressing PPI data for computational analysis [6]. A graph is made up of nodes (vertices) and pairs of nodes that have been recognized (called edges, links, etc).

2 Modeling and Methods

Our goal was to investigate and analyze a protein interaction network for *Homo sapiens*.

For this purpose, we utilized data from the BioGRID [7] database. The chosen organism database provides in-depth information about protein interactions for 81 species. After downloading and unpacking the files, the file on humans was shown to be the biggest with a size of 523 MB, followed by data on the budding yeast *Saccharomyces cerevisiae* (423 MB). Initially, BioGRID only consisted of data on the budding yeast, but in later releases, more species were added. That includes protein interactions of *Escherichia coli* (93 MB) and *Schizosaccharomyces pombe* (47 MB).

To gain a general understanding, of the data, we started by investigating the Human Herpesvirus 6B. As Figure 1 shows, this data includes seven proteins that interact in four different ways. Each protein and each interaction is uniquely identified with a BioGRID ID. Additionally, information on the organism or the genes for the proteins is given. We focused on the protein identifiers since they are the most needed to build a PPIN.

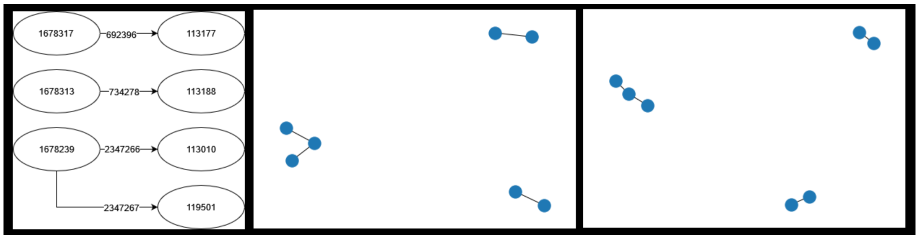


Figure 1: **Small PIN example** This figure shows the protein network of the Human Herpesvirus. On the left is a sketch of all protein IDs that are connected due to interactions. The pictures in the middle and on the right show different ways to draw the network with the spring layout in NetworkX.

With the help of the python package `NetworkX` [8], a small example network was constructed and visualized (Figure 1 middle and right).

The data can include self-loops and parallel edges, to prevent them from influencing the computation, they were removed. The package itself does not allow parallel edges and provides a function to recognize and delete self-loops.

With these prerequisites fulfilled we analyzed the human network. First, we found the largest component, which is the biggest connected set of nodes. Then, the degree distribution was visualized in a histogram and compared to an Erdös-Rényi (ER) network [9].

To further understand the network, multiple centrality measures were utilized to find hubs. The degree centrality C_D signifies the number of edges originating from one node. In the case of PPIN, that is the number of interactions a protein is a part of. The number of edges is limited by the number of nodes, which results in Equation 1 for the centrality of one node p_k.

$$C_D(p_k) = \frac{\sum_{i=1}^{n} a(p_i, p_k)}{n-1} \tag{1}$$

$$\text{with } a(p_i, p_j) = \begin{cases} 1 & p_i \text{ and } p_j \text{ are connected} \\ 0 & \text{else} \end{cases} \tag{2}$$

Additionally, we used the eigenvector centrality [10]. This measurement gives each node a prestige score, the higher the score of one node, the more other high-scoring nodes are its neighbors. Both methods are common in social network analysis [11] as well as PINs [12, 13]. The correlation between both methods was calculated and visualized, as well as the best scoring proteins.

Lastly, we compared the network to an ER network and a lattice network of the same size. For this purpose, we calculated the clustering coefficients and the average path lengths. To reduce computation time, the average path length was calculated on a subsample of the original data set consisting of 1000 nodes.

3 Results

3.1 Human Protein Interaction Network

The whole PPIN for Homo Sapiens was built from its BioGRID file. The network consists of 27,274 nodes, while the largest connected component holds 27,269 nodes and 732,760 edges. Therefore, approximately 99.98% of the network is included in the largest connected component. This is visualized in Figure 7.

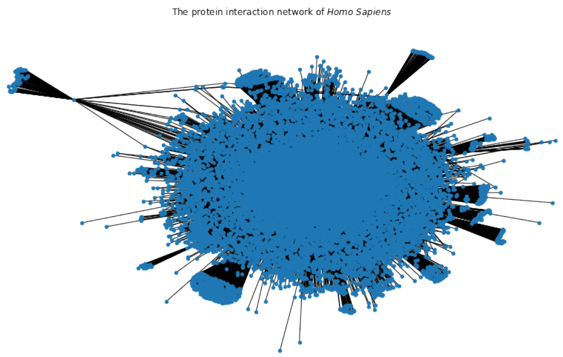


Figure 2: **Human Protein Interaction Network.** The figure shows the largest component of Protein Interaction Network for the Homo Sapiens.

Further calculations were made with the largest connected component instead of the whole graph. For the component of the network, a mean degree of 54 was computed. The degree distribution can be seen in the histogram Figure 3. In an Erdös–Rényi network, the number of edges per node depends on the chosen p. For a fitting p, the degree distribution could be similar to our network.

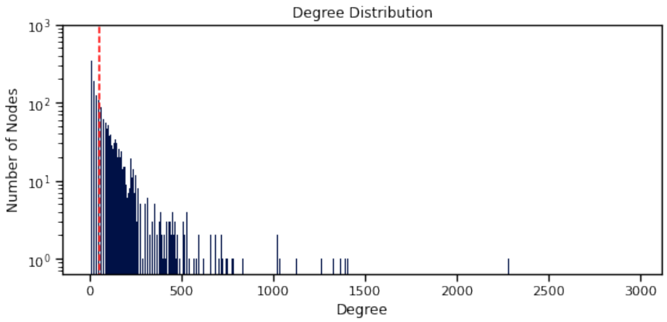


Figure 3: **Degree Distribution of the Human PPIN.** The histogram shows the degree distribution, with the red line marking the mean degree.

The centrality was measured with two different methods, the degree centrality, and the eigenvector centrality. For the degree centrality, the values ranged from $1.09e^{-1}$ to $3.70e^{-5}$, while the eigenvector centrality reached values from $1.05e^{-1}$ to $2.12e^{-8}$. Additionally, the correlation for the two centrality measures was computed, resulting in a coefficient of 0.9182, and visualized in Figure 4.

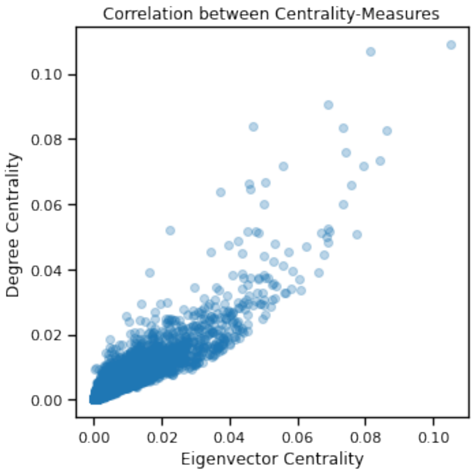


Figure 4: **Correlation of the Centrality Measures.** The figure shows the correlation between degree and eigenvector centrality.

Finally, the clustering coefficient and the average path length were calculated for the Homo Sapiens PPIN as well as an Erdös–Rényi with $p = 0.002$ and a Lattice network with similar size and density. Because the Lattice networks have a fixed number of edges, the number of nodes and edges could not be adjusted as well as with Erdös–Rényi.

The Human Protein Interaction Network had a clustering coefficient of 0.1064, the Erdös–Rényi network of 0.0020, and the Lattice network of 0.4026. For the average path length, the Human Protein Interaction Network resulted in a path length of 5, the Erdös–Rényi network of 9, and the Lattice network of 34. A complete summary of the used networks and the results can be found in Table 1. The Human Protein Interaction network is better than a random network like Erdös-Rényi, but it is not fully connected like a Lattice network.

Table 1: Results for the network of the Homo Sapiens and the Erdös–Rényi and Lattice networks.

Network	Number of Nodes	Number of Edges	Clustering Coefficient	Average Path Length
Human	27269	732760	0.1064	4.53
ER	27269	743393	0.0020	9.83
Lattice	33366	99365	0.4026	33.83

3.2 Small World Network

Each shortest path between vertices in a network has a nonlinear influence on the overall system, minimizing the distance not just between the two vertices it connects, but also between other connected vertices to the parent vertex, and so forth. The essential consequence is that the change to a smaller world is almost unnoticeable at the local scale. The example model in the study for the spread of an infectious disease is a purposefully simplified test scenario. Two key conclusions can be drawn: For smaller p, the infection's critical infectiousness, or the point at which half of the population is infected, the so-called half life, drops steadily. If a single outbreak is enough to infect the entire global population, the time period $T(p)$ will be similar to the $L(p)$ curve. [14]

3.3 Community detection in protein interaction networks

3.3.1 Modularity-Maximization

For community analysis, `NetworkX` includes a modularity-maximization option. This was used to calculate community structure in the Homo sapiens Protein Interaction Network. A graph depicting the size distribution of the communities was obtained, as seen in Figure 5.

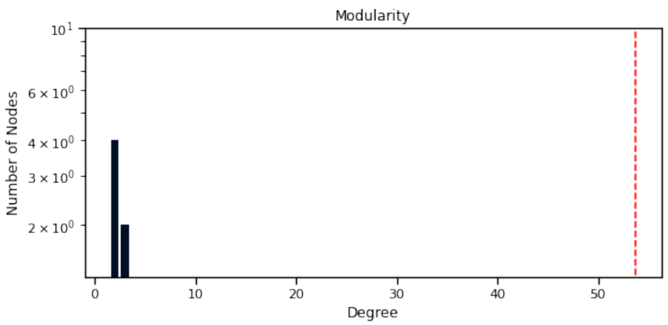
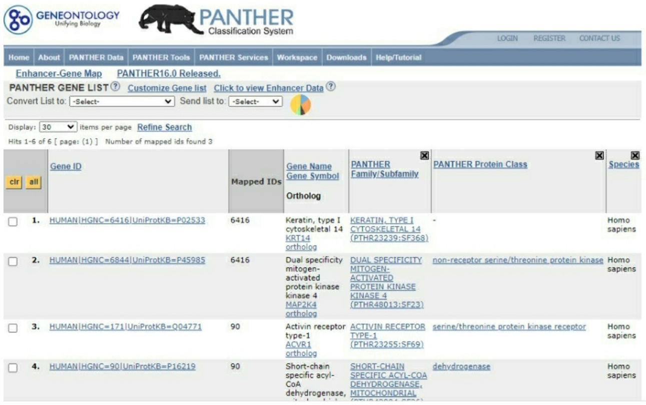


Figure 5: **Modularity Graph** The figure shows the number of nodes against the degree for the Homo Sapiens.

3.3.2 Gene ontology enrichment

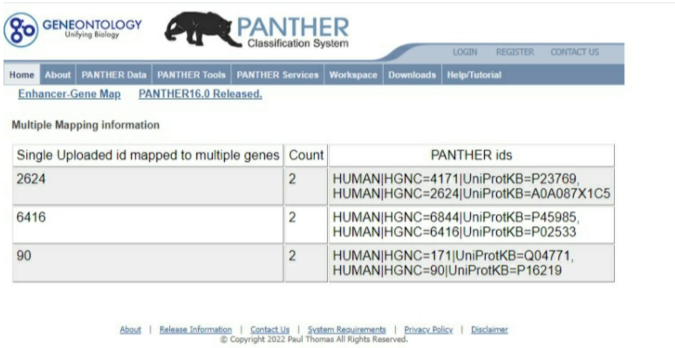
The number of genes tagged to individual categories in the Gene Ontology has a heavy-tailed distribution15. However, little research has been done into how these underlying annotation qualities could affect the outcomes of functional analysis tools. The ontology is a systematic tree-like structure that contains concepts (referred to as GO terms) and their interconnections. The GO annotation is a database of all genes that have been labeled and are associated with ontology concepts that describe those genes. The communities identified in the modularity graph are explored to find the relevant GO terms associated with the genes. The results of a Gene Ontology Enrichment are uniquely mapped gene information and the multiple mapped gene IDs for the given gene community input.



Hits 1-6 of 6 [page: (1)] Number of mapped ids found 3

		Gene ID	Mapped IDs	Gene Name Gene Symbol Ortholog	PANTHER Family/Subfamily	PANTHER Protein Class	Species
□	1.	HUMAN|HGNC=6416|UniProtKB=P02533	6416	Keratin, type I cytoskeletal 14 KRT14 ortholog	KERATIN, TYPE I CYTOSKELETAL 14 (PTHR23239:SF368)	-	Homo sapiens
□	2.	HUMAN|HGNC=6844|UniProtKB=P45985	6416	Dual specificity mitogen-activated protein kinase kinase 4 MAP2K4 ortholog	DUAL SPECIFICITY MITOGEN-ACTIVATED PROTEIN KINASE KINASE 4 (PTHR48013:SF23)	non-receptor serine/threonine protein kinase	Homo sapiens
□	3.	HUMAN|HGNC=171|UniProtKB=Q04771	90	Activin receptor type-1 ACVR1 ortholog	ACTIVIN RECEPTOR TYPE-1 (PTHR23255:SF69)	serine/threonine protein kinase receptor	Homo sapiens
□	4.	HUMAN|HGNC=90|UniProtKB=P16219	90	Short-chain specific acyl-CoA dehydrogenase,	SHORT-CHAIN SPECIFIC ACYL-COA DEHYDROGENASE, MITOCHONDRIAL	dehydrogenase	Homo sapiens

Figure 6: **Uniquely Mapped information** The figure shows Uniquely Mapped IDs for Homo sapiens



Multiple Mapping information

Single Uploaded id mapped to multiple genes	Count	PANTHER ids
2624	2	HUMAN|HGNC=4171|UniProtKB=P23769, HUMAN|HGNC=2624|UniProtKB=A0A087X1C5
6416	2	HUMAN|HGNC=6844|UniProtKB=P45985, HUMAN|HGNC=6416|UniProtKB=P02533
90	2	HUMAN|HGNC=171|UniProtKB=Q04771, HUMAN|HGNC=90|UniProtKB=P16219

About | Release Information | Contact Us | System Requirements | Privacy Policy | Disclaimer
© Copyright 2022 Paul Thomas All Rights Reserved.

Figure 7: **Multiple Mapping Information** The figure shows the Multiple Mapped Information for Homo sapiens.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

Gokul Thothathri and Swetha Rose Maliyakal Sebastian constructed and analyzed the small world network and wrote the Scientific Background and the Results for this part. Sina Glöckner and Christina Kirschbaum constructed and analyzed the human protein interaction network and wrote Methods and Modeling as well as Results for this part.

References

1. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: A general repository for interaction datasets. Nucleic acids research. 2006 01;34:D535–9.
2. Palacio A, Pastor O. Towards a Shared, Conceptual Model-Based Understanding of Proteins and Their Interactions. IEEE Access. 2021 05;PP:1–1.
3. Klimm F, Deane C, Reinert G. Hypergraphs for predicting essential genes using multiprotein complex data. Journal of Complex Networks. 2021 05;9.
4. Sozdinler M. ProLiVis: Protein-Protein Interaction Literature Visualization System. 2021 11;.
5. Dohrmann J, Puchin J, Singh R. Global multiple protein-protein interaction network alignment by combining pairwise network alignments. BMC Bioinformatics. 2015 09;16:S11.
6. Wang L, Xuexia M, Rui N, Zhang Z, Zhang J, Cai J. MultiCapsNet: A General Framework for Data Integration and Interpretable Classification. Frontiers in Genetics. 2021 11;12.
7. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. Nucleic Acids Research. 2006 Jan;34(suppl_1):D535–D539. Available from: https://doi.org/10.1093/nar/gkj109.
8. Proceedings of the Python in Science Conference (SciPy): Exploring Network Structure, Dynamics, and Function using NetworkX;. Available from: http://conference.scipy.org/proceedings/SciPy2008/paper_2/.
9. Erdös P, Rényi A. On the evolution of random graphs. In: The Structure and Dynamics of Networks. Princeton University Press; 2011. p. 38–82. Available from: https://www.degruyter.com/document/doi/10.1515/9781400841356.38/html.
10. Schoenberg IJ. Publications of Edmund Landau. In: Turán P, editor. Number Theory and Analysis: A Collection of Papers in Honor of Edmund Landau (1877–1938). Boston, MA: Springer US; 1969. p. 335–355. Available from: https://doi.org/10.1007/978-1-4615-4819-5_23.
11. Bródka P, Skibicki K, Kazienko P, Musiał K. A degree centrality in multi-layered social network. In: 2011 International Conference on Computational Aspects of Social Networks (CASoN); 2011. p. 237–242.
12. Jeong H, Mason SP, Barabási AL, Oltvai ZN. Lethality and centrality in protein networks. Nature. 2001 May;411(6833):41–42. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 6833 Primary_atype: Research Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/35075138.
13. Wang P, Yu X, Lü J. Identification and Evolution of Structurally Dominant Nodes in Protein-Protein Interaction Networks. IEEE Transactions on Biomedical Circuits and Systems. 2014 Feb;8(1):87–97. Conference Name: IEEE Transactions on Biomedical Circuits and Systems.
14. Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks. Nature. 1998 Jun;393(6684):440–442. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 6684 Primary_atype: Research Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/30918.