Heliyon 10 (2024) e27278

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Review article

Human protein–protein interaction networks: A topological comparison review

Rodrigo Henrique Ramos [a,b,*], Cynthia de Oliveira Lage Ferreira [a], Adenilso Simao [a]

[a] University of São Paulo, São Carlos, SP, Brazil
[b] Federal Institute of São Paulo, São Carlos, SP, Brazil

ARTICLE INFO

Dataset link: https://github.com/RodrigoHenriqueRamos/Human-Protein-Protein-Interaction-Networks-A-Topological-Comparison-Review

Keywords:
Protein–protein interaction networks
PPIN
Complex systems
Centrality measures
Network topology

ABSTRACT

Protein-Protein Interaction Networks aim to model the interactome, providing a powerful tool for understanding the complex relationships governing cellular processes. These networks have numerous applications, including functional enrichment, discovering cancer driver genes, identifying drug targets, and more. Various databases make protein-protein networks available for many species, including *Homo sapiens*. This work topologically compares four *Homo sapiens* networks using a coarse-to-fine approach, comparing global characteristics, sub-network topology, specific nodes centrality, and interaction significance. Results show that the four human protein networks share many common protein-encoding genes and some global measures, but significantly differ in the interactions and neighbourhood. Small sub-networks from cancer pathways performed better than the whole networks, indicating an improved topological consistency in functional pathways. The centrality analysis shows that the same genes play different roles in different networks. We discuss how studies and analyses that rely on protein-protein networks for humans should consider their similarities and distinctions.

1. Introduction

Complex networks are characterized by intricate structures and involve multiple nodes interacting in potentially unknown ways. These networks are subject to internal and external factors that can lead to self-organization and emergent phenomena [1].

The creation of Protein-Protein Interactions Networks (PPINs) became possible thanks to the advancements in large-scale methods to identify the functional relationship between genes. These methods include gene expression correlations, protein-protein interactions, text mining associations, and others [2]. Protein interactions play a crucial role in numerous biological processes, and analyzing their network can reveal unexpected biology, shedding light on complex pathways and potential therapeutic targets [3,4]. With the mapping of most humans' genes, genomics has made progress in identifying genetic variations and their correlation with diseases. However, understanding how these variations interfere with complex biological functions with thousands of interactions remains a significant challenge [4].

Researchers have utilized the analysis of PPINs and their functional modules to study complex diseases such as cancer and diabetes [5]. This has resulted in the development of computational methods aimed at improving our understanding of these diseases,

* Corresponding author at: Federal Institute of São Paulo, São Carlos, SP, Brazil.
E-mail addresses: ramos@ifsp.edu.br, rodrigohenrique.ramos@usp.br (R.H. Ramos).

https://doi.org/10.1016/j.heliyon.2024.e27278
Received 23 October 2023; Received in revised form 26 February 2024; Accepted 27 February 2024

Available online 1 March 2024

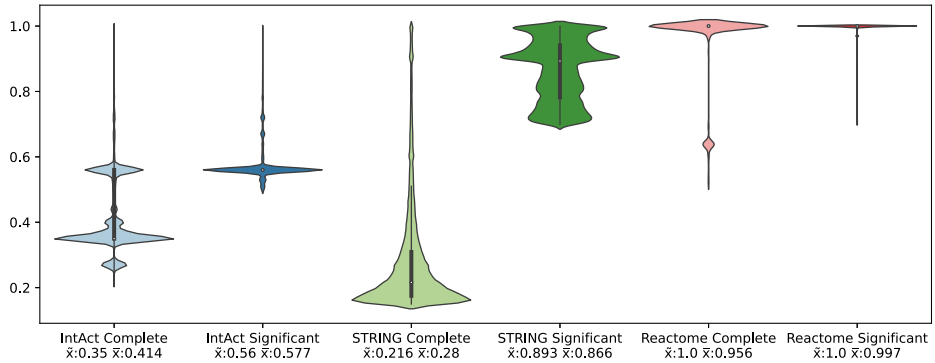

Fig. 1. Edge score distribution for all networks with scores. The median $\tilde{x}$ and mean $\overline{x}$ are closer in significant networks, normalizing the distribution. STRING shows the biggest difference between the significant and complete versions, since most of the edges score in the complete network are below 40%.

including identifying driver mutations in cancer [6]. Recent studies show a wide application of PPINs, such as drug repurposing [7–9], disease regulation and progression [10], virus mechanisms [11], and multi-omics interpretation [12].

Motivated by the broad applicability and use of PPINs and the existence of several databases that provide PPINs that aim to model the human interactome, this study explores the similarities and topological differences between PPINs following a coarse-to-fine approach. We analyzed global metrics, cancer sub-networks, interaction significance, intersections between nodes and edges, and the centrality of nodes. The results reveal that PPINs have similar global features, particularly in cancer sub-networks, but differ in measures related to the neighbourhood and mainly in the intersection of edges.

This work is organized as follows. The next section describes the databases and the pre-processing made considering the interaction significance. Afterwards, we analyze the nodes and edge intersection. The fourth and fifth section compares the PPINs and cancer sub-networks using global measures. After that, we explore the centrality role of three sets of nodes in the four PPINs. Finally, we present the results and the concluding remarks.

2. Databases and pre-processing

The combination of scientific literature curation and computational methods has led to the creation of numerous protein interaction databases [13]. Although there were doubts about the quality of interaction data at the beginning, significant advances in approaches over the past decade have greatly improved the data quality. Additionally, the standardization of data and curation practices across databases has enhanced the confidence and availability of interaction data [4]. The International Molecular Exchange consortium[1] (IMEx) strives to standardize curation rules related to interaction identification, while also establishing a standard data format, site search interface, and free access through the Creative Commons Attribution License.

In this study, we have chosen four human PPINs databases: HINT [14], IntAct [15], Reactome [16], and STRING [17]. These databases were chosen based on their frequent updates and citation rates. While IntAct and STRING have PPINs for multiple organisms, we have limited the networks to only *Homo sapiens*. Different databases may have varying methods of linking two proteins, but they assign a confidence score (i.e., edge weight) to each connection, which ranges from zero to one. This score represents the likelihood that the association is accurate [17]. HINT (High-quality INTeractomes) is an exception to this rule since its interactions come from a consensus between databases, keeping only significant interactions. Thus, its edges have no weight.

We followed HINT's method to enhance the quality of IntAct, Reactome, and STRING by creating significant versions. This was done by eliminating edges with low scores. A comparison of the edge score distribution before and after the removal is shown in Fig. 1. The complete versions of IntAct and STRING have the majority of their interactions with a confidence score below 50%, with STRING having a median of 28%. Reactome, on the other hand, has very few low-significant interactions, with a median of 96%. To create significant networks for Reactome and STRING, we removed interactions with scores below 70%. For IntAct, since there are barely any interactions with scores above 60%, we removed edges with a confidence score of 50% or less.

Fig. 1 shows how different the PPINs are when considering the significance of their interactions. Table 1 shows another important difference between the PPINs: the number of nodes and edges. STRING Complete has almost six million interactions; however, as shown in Fig. 1, most of these interactions have very low confidence scores. IntAct Complete has the second highest number interactions, but is 18 times lower than STRING Complete. This shows how important it is to remove low-confidence interactions, especially when comparing the PPINs.

After removing low-confidence scores, we eliminate self-loop edges and keep only the largest connected component. This pre-processing step is necessary because some network analysis measures require a single connected component and no self-loops. Reactome has no self-loop, and STRING has only one. HINT has 4% of its edges as self-loops, while IntAct has 1%. The largest connected component in each network has similar sizes, on average keeping 97.75% of its nodes. Table 2 shows the number of nodes and edges after the pre-processing steps.

[1] http://www.imexconsortium.org/.

Table 1
Number of nodes and edges in complete and significant networks. In the complete networks, the nodes range from 12,184 to 20,807, and the edges from 257,629 to 5,968,680.

Network	Nodes	Edges
HINT	15,386	119,494
IntAct Complete	20,807	328,288
IntAct Significant	13,437	91,701
Reactome Complete	13,953	257,629
Reactome Significant	12,184	228,699
STRING Complete	19,382	5,968,680
STRING Significant	16,812	252,953

Table 2
Final networks. The networks present a more evenly distribution among nodes and edges while maintaining only significative interactions.

Network	Nodes	Edges
HINT	14,763	114,588
IntAct	13,268	90,446
Reactome	11,873	228,447
STRING	16,582	252,801

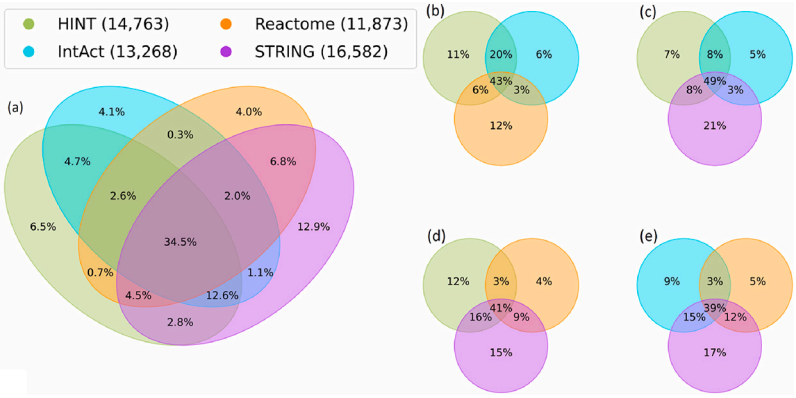

Fig. 2. Nodes intersection. All networks nodes intersection (a). Trio of networks nodes intersections (b), (c), (d), and (e). The networks have overlapping protein names while maintaining a significant amount of exclusive proteins.

All the analyses made from now on consider the networks present in Table 2. The final networks used in this work narrow the great difference in the number of nodes and edges among the original networks, making possible a fair comparison while keeping only interaction more likely to be true.

3. Nodes and edges intersections

PPINs are designed to represent the interactions within a cell. Even though these networks are regularly updated, the information they contain is still not comprehensive [5]. Various methods, such as link prediction, can be used to determine how two proteins interact with each other [18]. This section delves into the intersections that exist in the nodes and edges of the four PPINs.

Fig. 2 shows the nodes' intersection between the four networks and the intersection in groups of three. The union of all networks' nodes, Fig. 2 (a), corresponds to 21,475 unique nodes. From this union, 34.5% (7,408 nodes) are present in the four networks. Non-intersecting areas in Fig. 2 (a) show the percentage of unique nodes in each network. STRING harbours 2,770 unique nodes (12.9%), while 1,460 (6.8%) of the nodes are present only in STRING and Reactome. Fig. 2 (b) reveals that without STRING, HINT and IntAct share 20% of their nodes, the highest value when comparing groups of three. Without Reactome, Fig. 2 (c) shows the greatest intersection in groups of three. Fig. 2 (d) and Fig. 2 (e) present a similar behaviour, denoting a small difference when excluding HINT or IntAct.

The disparity in the networks' number of nodes impacts the percentual intersection analysis. To address this issue, Table 3 provides the percentage of nodes from each network (lines) included in the other networks (columns). HINT has 78% of its nodes included in IntAct and 79% in STRING. Even though the percentage is similar, STRING has more nodes than IntAct, indicating a

Table 3
Network nodes contained in other networks.

	HINT	IntAct	Reactome	STRING	Average
HINT		78%	61%	79%	73%
IntAct	88%		64%	81%	77%
Reactome	76%	71%		86%	78%
STRING	70%	65%	62%		65%

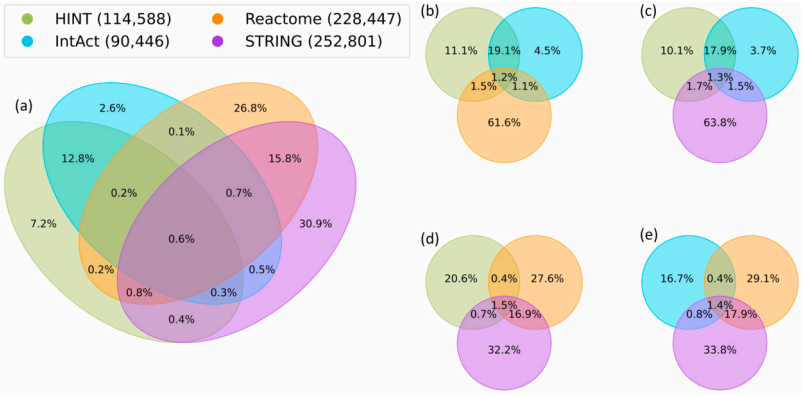

Fig. 3. Edges (protein interactions) intersection. All networks edges intersection (a). Trio of networks edges intersections (b), (c), (d), and (e). Only 0.6% of the protein interactions are shared with all the PPINs, while the majority of interactions are unique or shared in pairs of networks.

Table 4
Network edges contained in other networks.

	HINT	IntAct	Reactome	STRING	Average
HINT		62%	8%	10%	27%
IntAct	78%		9%	12%	33%
Reactome	4%	4%		40%	16%
STRING	4%	4%	36%		15%

greater intersection between HINT and IntAct than with HINT and STRING. The average column shows that around 3/4 of the nodes in each network are contained in other networks. Overall, Table 3 shows that PPINs are still incomplete and in development. While there is some overlap between the networks, it still needs to be completed.

We replicate the analysis made with nodes using edges (the interactions between proteins). The results show significant differences. Fig. 3 (a) reveals that only 0.6% of the edges are shared among all networks. The in trio analyses also show a mutual overlap ranging from 1.2%, Fig. 3 (b), to 1.5%, Fig. 3 (d). Without STRING, Fig. 3 (b), Reactome keeps 61.6% unique edges. While without Reactome, Fig. 3 (c), STRING keeps 63.8% of unique edges. Fig. 3 (d) and Fig. 3 (e) present a similar behaviour, denoting a small difference when excluding HINT or IntAct. The number of edges in the networks has a greater difference than the number of nodes, which can affect the percentage analysis. Table 4 shows a clear intersection pattern between HINT and IntAct, as well as Reactome and STRING.

The previous results show a similarity between the nodes present in the networks but a significant variation in overlapping edges. We selected 10 popular genes [19] to analyze the edges between them and further compare the interactions overlap. There are 45 possible edge combinations. We present in Table 5 only the 21 edges that appear in at least one of the networks. The total number of edges impacts the chance that an edge exists in a network, but STRING has 11% more edges than Reactome while having twice as many “matches”. The 10 selected nodes are considered hubs, as shown in Section 6, and the probability of edges between hubs is related to assortativity, a subject we will address in Fig. 9.

Table 5 reinforces the finding that, albeit the networks share a considerable number of nodes, the interaction between them varies greatly. Just the edge EGDR-AKT1 is present in all networks, while 11 edges are only in STRING.

4. Global networks measures

In this section, we present six global measures to explore and compare the networks as a whole. Firstly, we analyze the degree distribution and characterize the networks as scale-free using the Power Law Package [20].

A key difference between random and real-world networks is their degree distribution. Real-world networks tend to have a scale-free distribution where few nodes are highly connected (hubs), while most nodes have few neighbours [21]. Thus, the mean and variance cannot capture the distribution behaviour. While not all real-world networks are scale-free, the ones that are follow a degree distribution probability similar to a power law. This means that, in scale-free networks, it is more likely to find small-degree nodes

Table 5
Edges between 10 popular genes. The table presents only pairs of genes interacting in at least one network.

Gene Pair	APOE ESR1	APOE IL6	EGFR AKT1	EGFR ESR1	EGFR IL6	EGFR VEGFA	ESR1 AKT1	IL6 AKT1	TNF AKT1	TNF APOE	TNF EGFR
HINT			X								
IntAct			X	X							
Reactome	X		X	X	X	X	X				
STRING		X	X	X	X	X	X	X	X	X	X
Gene Pair	TNF IL6	TNF VEGFA	TP53 AKT1	TP53 EGFR	TP53 ESR1	TP53 IL6	TP53 TNF	TP53 VEGFA	VEGFA AKT1	VEGFA IL6	
HINT			X								
IntAct											
Reactome			X	X				X	X		
STRING	X	X	X	X	X	X	X	X	X	X	

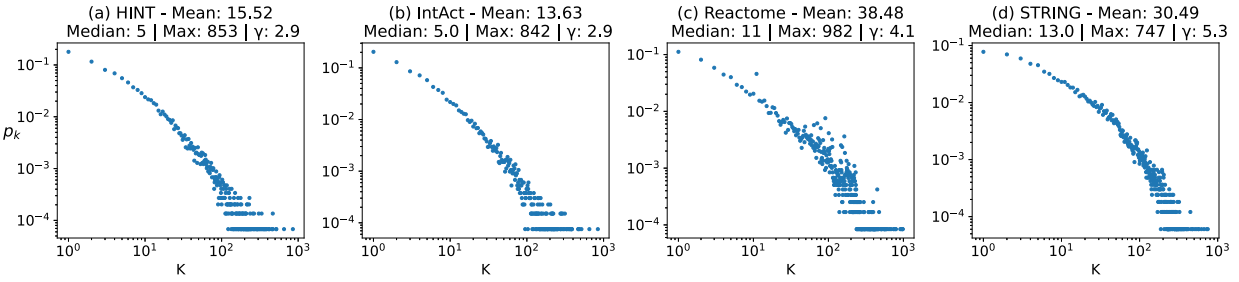

Fig. 4. Scale free characterization. The four PPINs share a similar scale free degree distribution, especially the pairs (a) HINT & (b) IntAct and (c) Reactome & (d) STRING.

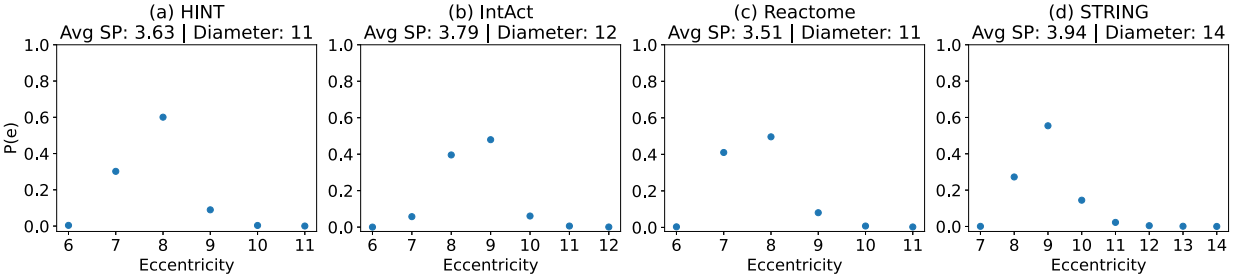

Fig. 5. Small world behaviour: the X-axis shows the eccentricity value, and the Y-axis shows the percentage of nodes with such value. The four PPINs show a small world behaviour, even with thousands of nodes, the average shortest path is smaller than 4.

than high-degree nodes. Figs. 4 (a), (b), (c), and (d), shows the degree distribution probability in log-log scale. After the network name, there is information about the degree distribution.

The average degree is more than double the median, showing that most nodes have a small degree and few hubs are responsible for increasing the average value. The value Max shows the degree of the biggest hub. Even though STRING, Fig. 4 (d), has the most edges, it has the smaller Max value. The last information, γ shows the approximate exponent of a power-law distribution that best represents the degree distribution of each network [20]. Although the PPINs have varying numbers of nodes and edges and differ in their intersections, they share similarities in their degree distribution. The plots demonstrate a similar trend, with Reactome, Fig. 4 (c), appearing slightly less defined in the middle. The plots' scales are also nearly identical. These observations classify all four PPINs as scale-free networks, where the majority of nodes have a small degree, and only a few hubs have a much higher degree than the average.

One of the consequences expected from scale-free networks is also being considered a small world. The term "small-world" is used to describe networks that, despite having thousands of nodes, have a relatively small diameter and eccentricity distribution. The eccentricity of a node refers to the maximum shortest path from that node to all other nodes, while the diameter is the maximum eccentricity found in the network. Figs. 5 (a), (b), (c), and (d), displays the eccentricity probability distribution, diameter and average shortest path of the PPINs.

The behaviour of the four networks regarding the small world phenomena is quite similar overall. The average shortest path (Avg SP) ranges from 3.51 to 3.94, indicating that, on average, information takes less than four "steps" to move through the network. Except for STRING, Fig. 5 (d), the diameter is also similar, ranging from 11 to 12. STRING has 16,582 nodes, with the shortest path between its two farthest nodes being 14. Looking at the probability distribution for eccentricity, we can see that in STRING,

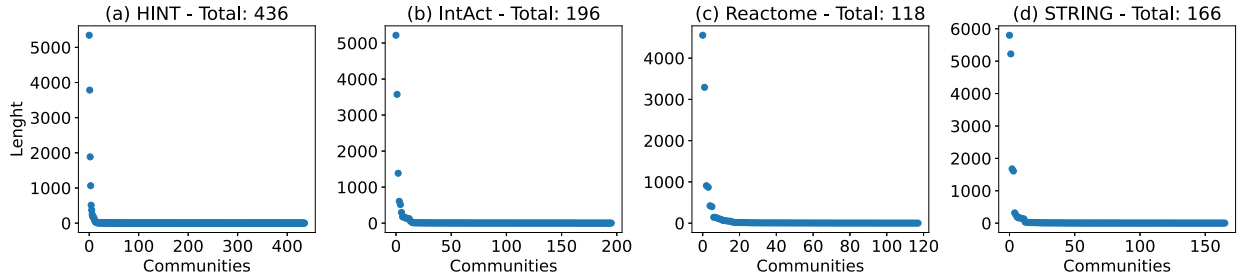

Fig. 6. Communities: the X-axis shows the number of communities found, and the Y-axis shows the number of nodes in each community. The community length distribution size is similar, with few large and many small communities.

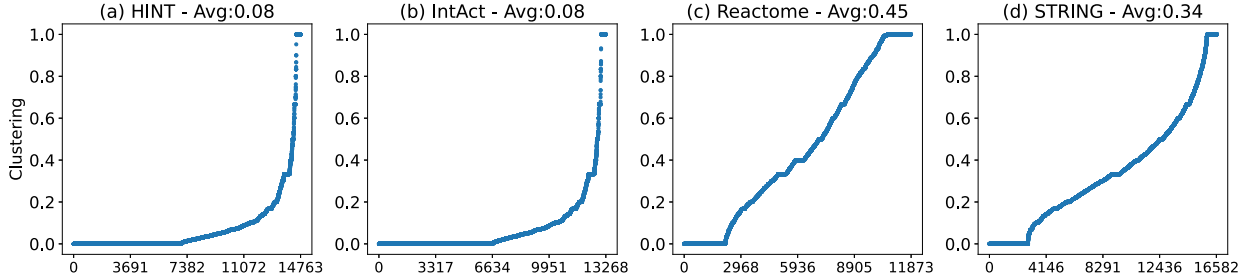

Fig. 7. Clustering. The pair (a) HINT & (b) IntAct have almost equal distribution. (c) Reactome & (e) STRING are similar, while being distinct from (a) HINT & (b) IntAct.

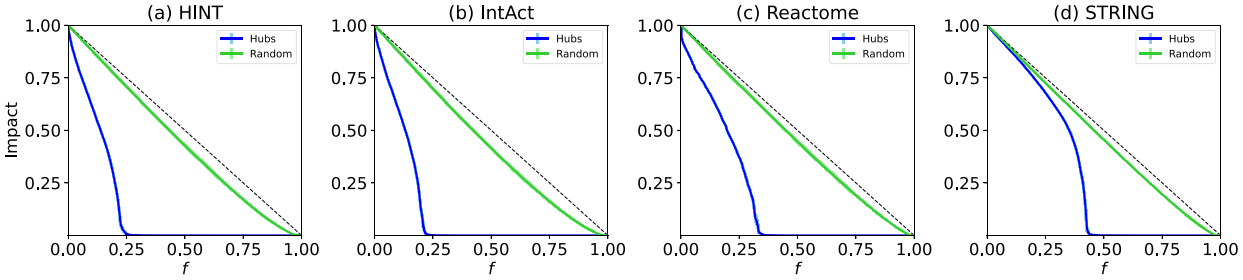

Fig. 8. Network resilience. The PPINs show a behaviour expected from scale free networks: resilient to random removal and fragile to hub removal. (d) STRING is the most resilient due to its degree assortativity.

almost 60% of nodes have an eccentricity of 9, and less than 10% have an eccentricity greater than 12, indicating that the network's diameter is a scenario for a few nodes.

Groups of proteins work together to accomplish complex tasks [4]. These groups form clusters and communities, with larger clusters often associated with functional modules, such as pathways [2]. The study of PPIN and its functional modules has been utilized to investigate complex diseases like cancer and diabetes [3,5]. Figs. 6 (a), (b), (c), and (d), shows the total number of communities found in the PPINs, along with their length distribution.

Just as it happened with the small world analysis, the PPINs show an overall similar behaviour in communities. The total number of communities in Fig. 6 (a) IntAct, Fig. 6 (b) Reactome, and Fig. 6 (d) STRING ranges from 118 to 196. All networks have very few large modules, with the vast majority having less than 100 nodes. Fig. 6 (b) HINT, for example, has 5,344 nodes in its largest community, while 425 communities have less than 100 nodes. Since the communities are made from interconnections within the nodes, we also present the clustering distribution for the fours PPINs in Figs. 7 (a), (b), (c), and (d).

Fig. 7 (a) HINT and Fig. 7 (b) IntAct share the same average clustering and almost identical clustering distribution, with few nodes having 100% clustering and nearly half having zero clustering. Fig. 7 (c) Reactome has the most interconnections between nodes. While the PPINs share similarities in terms of communities, there are noticeable differences in node clustering.

In Figs. 8 (a), (b), (c), and (d), the resilience of the networks are examined through the removal of nodes via random and hub attacks. Random attacks test for failures that occur unintentional, whereas hub removal simulates planned attacks on important nodes.

On the graph, the X-axis (f) shows the fraction of nodes removed, while the Y-axis shows the impact on the largest connected component. We conducted 30 random removal executions and 10 hub removal executions. A vertical line positioned over the impact line represents the standard deviation between these executions, indicating the variance in the results. In all cases, the

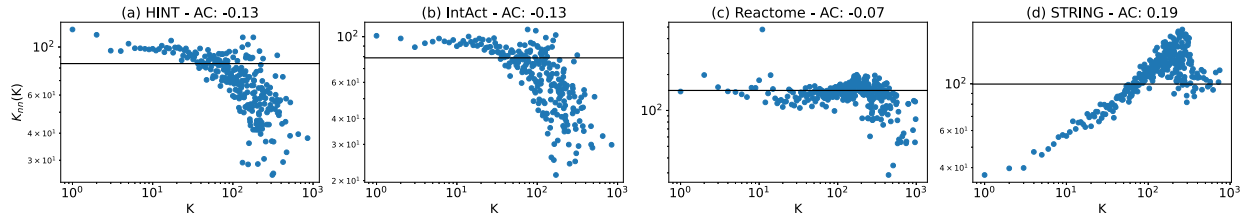

Fig. 9. Assortativity: the X-axis represents all nodes with the same degree K, and the Y-axis shows the average neighbours' degree of nodes of these nodes. The PPINs do not follow a linear distribution. HINT & IntAct follows the same trend, while Reactome and STRING have unique comportment.

Table 6
Topological analysis for cancer pathways networks.

	LCC (%)	AvgSP	AvgC	ρ	Com
Colorectal HINT	84%	3.32	0.20	0.06	6
Colorectal IntAct	76%	4.05	0.21	0.05	8
Colorectal Reactome	100%	2.14	0.61	0.19	4
Colorectal STRING	100%	1.97	0.62	0.22	4
Pancreatic HINT	80%	3.01	0.26	0.08	6
Pancreatic IntAct	68%	4.05	0.16	0.06	7
Pancreatic Reactome	100%	1.97	0.57	0.22	4
Pancreatic STRING	100%	1.96	0.61	0.22	4
Glioma HINT	76%	2.60	0.25	0.09	6
Glioma IntAct	73%	3.35	0.22	0.07	6
Glioma Reactome	97%	1.92	0.67	0.27	4
Glioma STRING	100%	2.09	0.68	0.24	4

variance is minimal. The dotted black line indicates zero impact, where removal does not break the network in more than one connected component. The four PPINs are very resilient to random removal and weak to hubs removal, as expected from scale free-networks [22]. Out of the four, Fig. 8 (a) HINT and Fig. 8 (b) IntAct are the most vulnerable to hub attacks, as they become completely dysfunctional after 25% of their hubs are removed. Fig. 8 (c) Reactome is more robust than Fig. 8 (a) HINT and Fig. 8 (b) IntAct, as it only reaches zero on the Y-axis after 37% of its hubs are removed. On the other hand, Fig. 8 (d) STRING shows remarkable resilience to hub attacks, with only around 45% of its hubs removal leading to a zero Y-axis value. The density and clustering of the networks play a critical role in their resilience, but compared to Reactome, STRING has lower values in both metrics. We attribute STRING's high resilience to its degree assortativity, which enables hubs to connect to other hubs and create many redundant paths within the network's core.

In our last global network measure, we explored degree assortativity. This measure evaluates how nodes connect with other nodes based on their degree. Figs. 9 (a), (b), (c), and (d), shows the assortativity coefficient (AC) after the network name. These values range from -0.13 to 0.19. Positive AC indicates that low degree nodes tend to connect with low degree nodes, and high degree nodes to connect with other high degree nodes. Negative AC indicates the opposite, where low degree nodes tend to connect with high degree.

The AC alone does not capture the actual assortativity behaviour found in the PPINs. Thus, Fig. 9 also present the assortativity distribution. In the plot, K (X-axis) represents all nodes with degree K, and $K_{nn}(K)$ (Y-axis) is the average neighbours' degree of nodes with degree K. Each sub-plot's solid black horizontal line represents neutral assortativity (AC equals zero). Fig. 9 (a) HINT and Fig. 9 (b) IntAct have the same AC and a similar distribution. In both cases, nodes with degree smaller than forty show a neutral assortativity. As the degree increases, the plot chances tendency, indicating that hubs are connected with small degree nodes. Fig. 9 (c) Reactome keeps a neutral assortativity until around degree three hundred. Only the significant hubs show negative assortativity. Overall, the three PPINs share similarities: neutral assortativity within small degree nodes and negative assortativity on hubs. Fig. 9 (d) STRING has the opposite behaviour. Small degree nodes are connected with small degree nodes, showing a positive assortativity trend until degree two hundred and fifty. After that, big hubs tend to connect with other similar or smaller hubs.

5. Cancer sub-networks

This section compares the PPINs by analyzing sub-networks extracted from cancer pathways. Pathways consist of interacting genes that contribute to specific biological functions. These pathways serve as the foundational components of a cell's complex system [23]. The KEGG database [24] makes available cancer pathways, i.e., lists of genes associated with specific types of cancer. Using pathways for colorectal, pancreatic, and glioma cancer, we create induced subgraphs from the PPINs, totalizing twelve cancer pathway networks. Table 6 presents these networks and five topological measures in order to explore how the same set of genes creates distinct sub-networks in different PPINs.

The column LCC (%) shows the percentage of nodes present in the largest connected component relative to genes in the KEGG's pathways. The cancer pathways for Colorectal, Pancreatic, and Glioma have respectively 86, 76, and 75 genes. It is a tiny set compared to thousands of nodes presented in the PPINs. Although small in size, the cancer pathways are biologically significant

(a) Colorectal

	HINT	IntAct	Reactome	STRING
HINT		45%	76%	83%
IntAct	71%		82%	86%
Reactome	16%	11%		85%
STRING	16%	10%	76%	

(b) Pancreatic

	HINT	IntAct	Reactome	STRING
HINT		37%	76%	74%
IntAct	70%		80%	93%
Reactome	17%	10%		77%
STRING	17%	12%	78%	

(c) Glioma

	HINT	IntAct	Reactome	STRING
HINT		51%	76%	80%
IntAct	76%		87%	91%
Reactome	15%	12%		74%
STRING	18%	14%	80%	

Fig. 10. Cancer pathways networks: edges contained in other networks.

and indicate important interactions within the cell. On average, the LCC contains 88% of the genes associated with the selected cancer types in the KEGG's database. Reactome and STRING have the greatest overlap, while IntAct has the least. The column AvgSP is the average shortest path, AvgC is the average clustering, ρ is network density, and Com is the number of communities. These four global measures show that the sub-networks from HINT and IntAct are similar. The same happens with the sub-networks from STRING and Reactome. The similarities follow the trend found in previous analyses, with the difference that cancer networks from STRING and Reactome are closer together than the actual PPINs. The AvgC is significantly higher in the cancer network than with PPINs, especially in HINT and IntAct, where global clustering is only 8%. This behaviour is expected, since pathways usually form modules with high clustering coefficient [3].

We replicate the analysis made in Table 4 with the cancer networks. Figs. 10 (a), (b), and (c), presents one table for each cancer type showing the percentual of edges from one network contained in the other networks. Compared to Table 4, we see a better overall intersection among the cancer network than with the whole networks. Across Figs. 10 (a), (b), and (c), HINT decreases its association with IntAct from 62% to an average of 44% and significantly increases with Reactome and STRING, jumping from 8% and 10% to an average of 76% and 79%. IntAct also showed a similar increase with Reactome and STRING, but without decreasing it with HINT. Reactome and STRING also increase their mutual association from 40% and 36% to an average of 79% and 78%.

The results show that small sub-networks from cancer pathways enhance the similarity between PPINs using traditional measures and significantly increase edge intersections. Pathways are main functional modules within cells, and these results indicate that although the entire PPINs exhibit differences, mainly in edges, the similarity increases in pathway networks.

6. Genes centrality role

In addition to conducting global and sub-network analyses, we also examine the topological role of three sets of nodes in the four PPINs. We have chosen 10 famous genes [19], 10 genes associated with type 2 diabetes [25], and 10 genes related to Alzheimer's disease [26]. Fig. 11 shows their percentile positions in three centrality measures, the first indicating connectivity and the other two neighbourhoods. The standard deviation is presented after the gene names, rounded to two decimal places.

In the first sub-plot, we present the percentile position in the degree distribution for the set of popular genes. The first gene, TP53, is a major hub in all PPINs. The percentile position for HINT, IntAct, Reactome and STRING, respectively, is 0.993, 0.994, 0.997, 0.999. Thus all points in the plot superpose. Something similar happens with the third gene, EGFR, and the last two genes, ESR1 and AKT1. The overlapping of points seen in these four genes indicates that, in terms of degree, these genes occupy the same topological role in the four PPINs. We would expect to find many overlappings in different measures and sets of genes if the PPINs were similar. However, Fig. 11 shows the opposite. Overall, the same gene has a spread percentile distribution among the PPINs. TP53, albeit a top hub in all PPINs, has a distinct neighbourhood in different PPINs. The percentile position for Average Neighbors Degree is 0.387 (HINT), 0.318 (IntAct), 0.415 (Reactome), and 0.767 (STRING). For clustering, the percentiles are 0.569 (HINT), 0.524 (IntAct), 0.195 (Reactome), and 0.173 (STRING). The genes TMEM106B, from Alzheimer, and TCF7L2, from Diabetes, also have a similar degree and different neighbourhoods.

7. Conclusion

Considering that distinct databases make available PPINs for human interactome, this study aimed to topologically compare four networks: HINT, IntAct, Reactome, and STRING. Our analysis comprehends interaction significance, nodes and edges intersections, global measures, sub-networks comparisons, and node centrality. This coarse-to-fine approach offers a comprehensive overview, showing that the PPINs share some global characteristics, especially in cancer sub-networks, while differing in measures associated with neighbourhood and mainly in edges intersection.

Globally, the networks are scale-free, and the degree distribution in all of them follows the same trend. They are also small worlds and share similarities in diameter, average shortest path, eccentricity and community size. In the clustering distribution, HINT and IntAct are practically identical, with distant values from Reactome and STRING, which are similar to each other. The resilience to random and hub attacks followed the expected behaviour from scale-free networks, with STRING being the most resistant. HINT and IntAct perform similarly in assortativity, with a neutral behaviour in small degree nodes and a decreasing assortativity as the degree

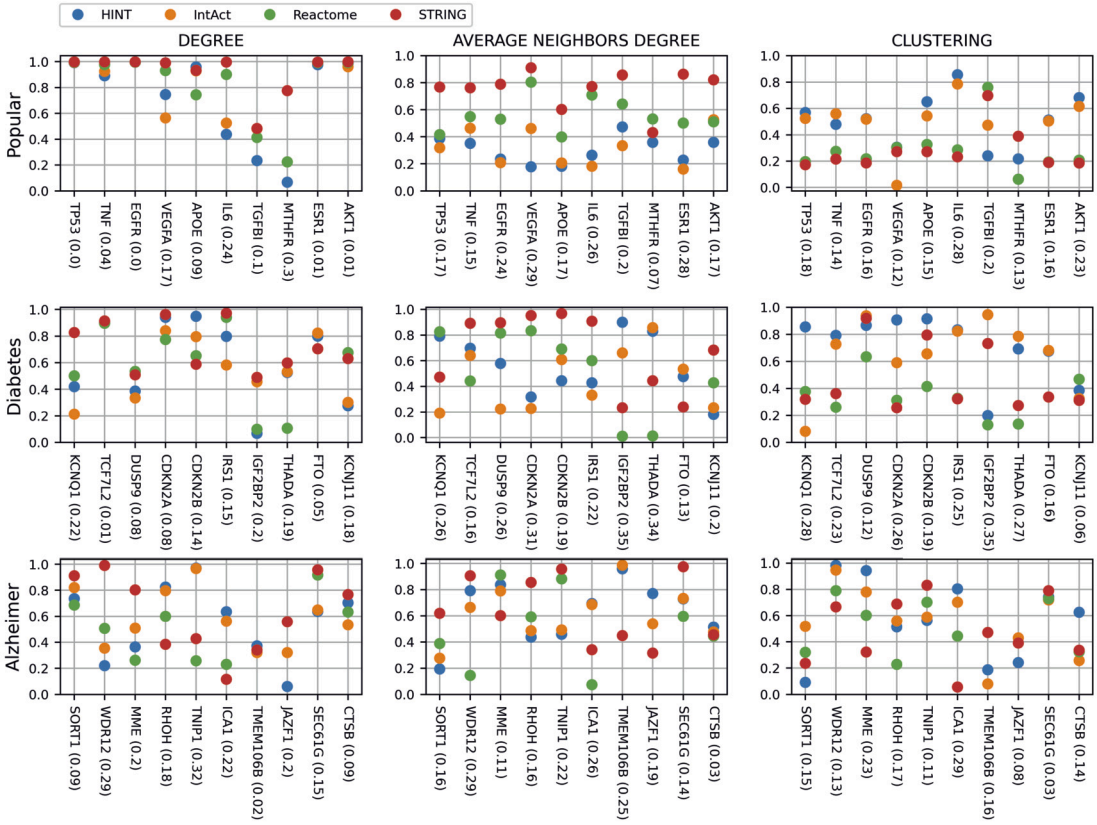

Fig. 11. Percentile position for individual genes.

increases. Reactome showed a negative trend only with the highest hubs, with the other nodes being neutral. STRING differs from the others PPINs, with a positive assortativity in small degree nodes and mixed in hubs.

The significant differences in PPINs lay in the number of edges, edges weight, edges intersection and neighbourhood-associated measures. Even with the removal of low-score edges, the average of edges from one network contained in others is 22.75%, and the average of nodes contained in other networks is 73.25%. This shows that PPINs harbour many common protein-encoding genes but not their interactions. Centrality measures for the same genes also perform differently in the PPINs, even if the degree is similar. Sub-networks from cancer pathways showed to be more alike than the whole PPINs, showing the PPINs maintain topological similarities in functional modules.

The PPINs are evolving in an effort to model the human cell interactome, and albeit they share common nodes and some global characteristics, the interactions and associated measures differ. PPINs are used in many studies and computational analyses, like cancer driver genes discovery. Considering the results we present in this paper, we hypothesise that changing the PPIN may alter the result of previous findings. For future works, we propose to investigate this hypothesis and create a consensus PPIN that encompasses different PPINs.

CRediT authorship contribution statement

Rodrigo Henrique Ramos: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Data curation, Conceptualization. **Cynthia de Oliveira Lage Ferreira:** Writing – review & editing, Writing – original draft, Supervision, Conceptualization. **Adenilso Simao:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Rodrigo Henrique Ramos reports financial support and equipment, drugs, or supplies were provided by São Paulo Research Foundation (FAPESP). Rodrigo Henrique Ramos reports financial support and equipment, drugs, or supplies were provided by Center for Mathematical Sciences Applied to Industry (CeMEAI). Rodrigo Henrique Ramos reports financial support was provided by Brazilian National Research and Technology Council (CNPq). Rodrigo Henrique Ramos reports financial support was provided by Brazilian Federal Foundation for Support and Evaluation of Graduate Education (CAPES). If there are other authors, they declare that they

have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

All the code and data used in this work is available at: https://github.com/RodrigoHenriqueRamos/Human-Protein-Protein-Interaction-Networks-A-Topological-Comparison-Review.

Acknowledgements

The authors acknowledge the financial support received from the Federal Institute of Sao Paulo (IFSP), the University of São Paulo (USP), the São Paulo Research Foundation (FAPESP), the Center for Mathematical Sciences Applied to Industry (CeMEAI), the Brazilian National Research and Technology Council (CNPq), and the Brazilian Federal Foundation for Support and Evaluation of Graduate Education (CAPES).

References

[1] Albert Réka, Albert-László Barabási, Statistical mechanics of complex networks, Rev. Mod. Phys. 74 (1) (2002) 47.
[2] P. Porras, Network analysis of protein interaction data: an introduction, https://doi.org/10.6019/tol.networks_t.2016.00001.1, 2016.
[3] Gavin C.K.W. Koh, et al., Analyzing protein–protein interaction networks, J. Proteome Res. 11 (4) (2012) 2014–2031.
[4] Kasper Lage, Protein–protein interactions and genetic diseases: the interactome, Biochim. Biophys. Acta, Mol. Basis Dis. 1842 (10) (2014) 1971–1980.
[5] Nahid Safari-Alighiarloo, et al., Protein-protein interaction networks (PPI) and complex diseases, Gastroenterol. Hepatol. Bed Bench 7 (1) (2014) 17.
[6] Jorge Francisco Cutigi, Adriane Feijo Evangelista, Adenilso Simao, Approaches for the identification of driver mutations in cancer: a tutorial from a computational perspective, J. Bioinform. Comput. Biol. 18 (03) (2020) 2050016.
[7] Trang T.T. Truong, et al., Repurposing drugs via network analysis: opportunities for psychiatric disorders, Pharmaceutics 14 (7) (2022) 1464.
[8] Yadi Zhou, et al., Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2, Cell Discov. 6 (1) (2020) 14.
[9] Feixiong Cheng, István A. Kovács, Albert-László Barabási, Network-based prediction of drug combinations, Nat. Commun. 10 (1) (2019) 1197.
[10] Gihanna Galindez, et al., Network-based approaches for modeling disease regulation and progression, Comput. Struct. Biotechnol. J. (2022).
[11] David E. Gordon, et al., Comparative host-coronavirus protein interaction networks reveal pan-viral disease mechanisms, Science 370 (6521) (2020) eabe9403.
[12] Antoine Bodein, et al., Interpretation of network-based integration from multi-omics longitudinal data, Nucleic Acids Res. 50 (5) (2022) e27.
[13] Damian Szklarczyk, Lars Juhl Jensen, Protein-protein interaction databases, in: Protein-Protein Interactions, Springer, 2015, pp. 39–56.
[14] Jishnu Das, Haiyuan Yu, HINT: high-quality protein interactomes and their applications in understanding human disease, BMC Syst. Biol. 6 (1) (2012) 1–12.
[15] Sandra Orchard, et al., The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases, Nucleic Acids Res. 42 (D1) (2014) D358–D363.
[16] Guanming Wu, Xin Feng, Lincoln Stein, A human functional protein interaction network and its application to cancer data analysis, Genome Biol. 11 (5) (2010) 1–23.
[17] Damian Szklarczyk, et al., The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest, Nucleic Acids Res. 51 (D1) (2023) D638–D646.
[18] Ho Yin Yuen, Jesper Jansson, Better link prediction for protein-protein interaction networks, in: 2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE), IEEE, 2020, pp. 53–60.
[19] Elie Dolgin, The most popular genes in the human genome, Nature 551 (7681) (2017) 427–432.
[20] Jeff Alstott, Ed Bullmore, Dietmar Plenz, Powerlaw: a Python package for analysis of heavy-tailed distributions, PLoS ONE 9 (1) (2014) e85777.
[21] Albert-László Barabási, Network Science, Cambridge University Press, Cambridge, 2015.
[22] Réka Albert, Hawoong Jeong, Albert-László Barabási, Error and attack tolerance of complex networks, Nature 406 (6794) (2000) 378–382.
[23] Bijay Jassal, et al., The reactome pathway knowledgebase, Nucleic Acids Res. 48 (D1) (2020) D498–D503.
[24] Minoru Kanehisa, Susumu Goto, KEGG: Kyoto encyclopedia of genes and genomes, Nucleic Acids Res. 28 (1) (2000) 27–30.
[25] Ali Omar, Genetics of type 2 diabetes, World J. Diabetes 4 (4) (2013) 114.
[26] Céline Bellenguez, et al., New insights into the genetic etiology of Alzheimer's disease and related dementias, Nat. Genet. 54 (4) (2022) 412–436.