The Topology of the Growing Human Interactome Data

Vuk Janjić[1] and Nataša Pržulj[1,*]

[1]Department of Computing, Imperial College London, London, SW7 2RH, United Kingdom

Summary

We have long moved past the one-gene–one-function concept originally proposed by Beadle and Tatum back in 1941; but the full understanding of genotype–phenotype relations still largely relies on the analysis of static, snapshot-like, interaction data sets. Here, we look at what global patterns can be uncovered if we simply trace back the human interactome network over the last decade of protein-protein interaction (PPI) screening. We take a purely topological approach and find that as the human interactome is getting denser, it is not only gaining in structure (in terms of now being better fit by structured network models than before), but also there are patterns in the way in which it is growing: (a) newly added proteins tend to get linked to existing proteins in the interactome that are not know to interact; and (b) new proteins tend to link to already well connected proteins. Moreover, the alignment between human and yeast interactomes spanning over 40% of yeast's proteins — that are involved in regulation of transcription, RNA splicing and other cell-cycle-related processes — suggests the existence of a part of the interactome which remains topologically and functionally unaffected through evolution. Furthermore, we find a small sub-network, specific to the "core" of the human interactome and involved in regulation of transcription and cancer development, whose wiring has not changed within the human interactome over the last 10 years of interacome data acquisition. Finally, we introduce a generalisation of the clustering coefficient of a network as a new measure called the cycle coefficient, and use it to show that PPI networks of human and model organisms are wired in a tight way which forbids the occurrence large cycles.

1 Introduction

The first high-throughput human protein-protein interaction (PPI) network data started appearing shortly after the publication of the first reference sequence of the human genome, over a decade ago. Subsequently, the torrent of '-omics' data has shifted systems biology research from hypothesis-based data analysis to data-inspired hypothesis generation; but this revolution introduced by the post-genomics era has not often translated directly into new therapeutic developments [1]. The dawn of the 21st century began with the achievement of the first big goal towards understanding the underlying mechanisms of life — the genome sequencing. The next big challenge that lies ahead is obtaining complete interactomes for a number of species, including human [2, 3].

*To whom correspondence should be addressed. Email: natasha@imperial.ac.uk

Just over a decade ago, all available individually reported protein-protein interactions were in the range of hundreds. Since the introduction of high-throughput screening, including yeast two-hybrid [4] and affinity purification [5], these new technologies have been generating thousands of interactions each year, causing an explosion in available molecular data. Nevertheless, current high-throughput PPI data are still of low coverage and have high rates of false positives [6, 7, 8].

As we have been building a rich foundation of molecular interaction data over the years, we have now reached a point of examining the contribution of interaction data to new biological insight. For instance, a recent large-scale data fusion study (integrating 11 sources of human molecular data) showed that it is the set of genetic interactions which has the most influence on human diseases, despite its small size and sparseness compared to all other available data [9]. On the other hand, proteins do not work in isolation to carry out almost all biological process, but together, by being wired into a complex network whose properties have given us important insights over the years. The availability and quality of these data could impact the conclusions of scientific analyses.

To make sure that data generation is as unbiased as possible and that PPI network collection progress is on the right track, it was recently shown that different interaction detection biotechnologies produce consistent PPI topology, i.e., that PPI topology is largely independent of biotechnology used for generating it [10]. Currently, we are at a unique point in time where we can observe PPI network growth as new PPI data become available. In this study, we examine how human PPI network acquisition has been progressing thus far, and whether its topological properties changed over time in a way that could provide interesting insights into the principles that underlie the complex protein interaction machinery.

We examine the human PPI network at its current stage on its path to data collection completion, and study how its topology has been changing over the past 10 years: we analyse how newly added proteins get wired into the existing PPI network and if parts of the human and yeast interactomes gain in topological and functional similarity. We also find a core sub-network in the human PPI network, which has been topologically conserved since the earliest versions of the interactome and find that it is involved in some of the most important biological processes related to the cell's development and progression, as well as disease formation.

It is well established that conserved PPIs exist between species, even as distant as yeast and human; this is used for transferring functional annotation of proteins from one species to another, reconstructing phylogenetic relationships between different species (including viruses), predicting and validating new interactions between proteins, finding cross-species conserved motifs that correspond to specific cellular machinery (such as amino acid phosphorylation, DNA replication initiation, protein folding, regulation of cell cycle progression) [11, 12, 13, 14, 15, 16, 17, 18, 19, 20]. Conversely, here we examine how addition of new PPI data affects the large conserved human–yeast regions of their interactome networks, both topologically and functionally.

Finally, we introduce a new network topology measure, which we call the *cycle coefficient*. It is a generalisation of the clustering coefficient and represents the likelihood of any two nodes with a common interactor to be connected through cycles within the network. Using this measure,

we find that large cycles are not common in the interactomes, i.e., that PPI networks are tightly wired with short cycles between their proteins.

2 Methods

2.1 Data

We obtained all protein-protein interaction (PPI) data sets from the Center for Cancer Systems Biology (CCSB) Interactome Database[1]. We used the following yeast-two-hybrid (Y2H) human interactomes: HI-2005, the data set from Rual et al. (2005) [21]; HI-2011, the union of Yu el al. (2011) [8], Vankatesan et al. (2009) [6] and Rual et al. (2005) [21] data sets; HI-2013, the latest human PPI data set (Marc Vidal's pre-publication human PPI data available at `http://interactome.dfci.harvard.edu/H_sapiens/index.php`).

We also used Y2H protein-protein interaction (PPI) networks of model organisms to compare the human interactome with interactomes of other species (also downloaded from CCSB Interactome Database). The PPI network of *A. thaliana* (AI-1) is constructed from data published by the Arabidopsis Interactome Mapping Consortium (2011) [22]. To increase coverage, we constructed the worm, yeast and fly PPI networks as unions of multiple data sets: the worm (*C. elegans*) PPI network (WI-2) is constructed as the union of WI-2004 [23] and WI-2007 [24] data; the yeast (*S. cerevisiae*) PPI network (YI-2) is constructed as the union of data provided by Yu et al. (2008) [7], Ito et al. (2001) [25] and Uetz et al. (2000) [26]; and the fly (*D. melanogaster*) PPI network (FI-2) is constructed as the union of data by Stanyon et al. (2004, Finley Lab) [27], Formstecher et al. (2005, Hybrigenics) [28] and Giot et al. (2003, CuraGen) [29]. Table 1 summarises the basic network properties of all used data sets.

Table 1: Basic network properties of the analysed PPI networks. The column labels are as follows: $|N|$, number of nodes; $|E|$, number of edges; CC, clustering coefficient; APL, avarage path length; ANN, average number of neighbours; d, diameter; and r, radius.

PPI Network	$\|N\|$	$\|E\|$	CC	APL	ANN	d	r
HI-2005	1,523	2,549	0.033	4.35	3.77	12	6
HI-2011	2,163	3,718	0.027	4.57	3.73	12	6
HI-2013	4,228	13,427	0.054	4.06	6.52	11	6
AI-1	2,634	5,529	0.050	4.75	4.48	16	8
WI-2	2,235	3,232	0.023	5.29	3.13	15	8
YI-2	1,966	2,705	0.056	5.61	3.05	14	8
FI-2	8,023	27,795	0.011	4.28	6.99	10	6

2.2 Graphlets and graphlet degree distribution agreement

Graphlets are small, connected, induced sub-graphs of a large network [30, 31] that describe the wiring patterns around nodes in the network [32]: all 30 graphlets with 2 to 5 nodes and

[1] `http://interactome.dfci.harvard.edu/`

all 73 symmetries in them (called *orbits*) are illustrated in Figure 3 (graphlets are labelled G_0 to G_{29}, while orbits are labelled from 0 to 72). Previously, orbits were used to provide a more constraining measure of a node's position within a network [32]. Here, we use orbits to show exactly how the new proteins were added and how they impacted the topology of the human interactome.

Graphlet degree distribution (GDD) agreement is thus far the most sensitive measure which shows how similar the structure of two networks is. It works by generalising the degree distribution so that instead of comparing only the degree distributions of two networks, it also compares how similar the two networks are in terms of sub-structures such as triangles or squares (see [31] and [32] for extensive details).

2.3 Random network models

We construct random model networks, with the same number of nodes and edges as the original data, coming from five network models most commonly used for modelling PPI networks: Erdös-Rényi random graphs (ER), Erdös-Rényi random graphs with the same degree distribution as the data (ER-DD), Geometric random graphs (constructed using 3-dimensional Euclidean space, denoted by GEO), Scale Free Barabasi-Albert type networks (SF) and stickiness-index based networks (STICKY) (described in [31] and [33]). To increase confidence, for each of these five random network models corresponding to each of the three data networks (HI-2005, HI-2011 and HI-2013), we generate 30 network instances. This produces 450 random model networks (30 instances for 5 models for each of the 3 data networks; 150 model instances per data network). To see which model fits the data, we measure the similarity between the three human PPI networks and each of the 150 generated networks by computing the GDD agreement between them (see section 2.2 for introduction to GDD). We compute the average and standard deviation of the GDD agreement between the data network and all of the 30 generated instances of one model, and we do so for each of the five random models. We report average and standard deviations of GDD agreement between data and model networks for each of the five models.

2.4 Network alignment

Aligning networks is a process of mapping nodes of one network onto the nodes of another with the goal of maximising the number of aligned edges between the aligned nodes. The problem is computationally intractable due to the underlying sub-graph isomorphism problem that is NP-complete [34]. Hence, approximate solutions are sought. Analogous to sequence alignment, network alignment algorithms can be local and global. There exists a number of network alignment algorithms [11, 12, 13, 14, 15, 16, 17, 18, 19]. The topological quality of alignment is usually measured by *edge correctness (EC)*, which is the percentage of edges of the smaller network that are correctly aligned to edges of the larger network [17].

Since in Section 3.3 we look at the similarity in "wiring" between the human PPI network and PPI networks of model organisms, we use a network alignment algorithm that would align

interactomes based purely on topology. Hence, we use MI-GRAAL [19] restricted to using only topological similarities to find similar nodes to align.

2.5 Cycle coefficient

We generalise the notion of the clustering coefficient to *cycle coefficient* (defined below) and compare cycle coefficients of the PPI data and random network models. Note that the standard clustering coefficient can be interpreted as the likelihood of two nodes being neighbours (connected by a path of length 1, i.e., an edge) given that they share a common neighbour. We generalise this as follows.

Definition — A *cycle coefficient* of order k of a node v, denoted as $C_k(v)$, is a fraction of all pairs of neighbours of v connected by some path of length $\leq k-2$ not going through v. By definition, we put $C_k(v) = 0$ for any k, if the degree of node $v \leq 1$. Analogous to the standard average clustering coefficient of a network, the *average cycle coefficient* of a network is the average of cycle coefficients of all nodes in the network. In other words, it is the likelihood of two nodes being connected by some path of length $\leq k-2$, given that they share a common neighbour. Equivalently, the cycle coefficient of order k of network G, $C_k(G)$, is the fraction of all node pairs in the network that belong to some cycle of length $\leq k$, given that they share a common neighbour (note the importance of $\leq$ here and also see the example below). Thus, $C_3(G)$ is the standard clustering coefficient of network G.

Example 1 — We compute $C_3(v)$, $C_4(v)$ and $C_5(v)$ of node v in the graph presented in Figure 1A. Node v has 4 neighbours and therefore there are 6 possible pairs of its neighbours. None of the neighbours of v are connected by an edge and therefore, the standard clustering coefficient of this node is $C_3(v) = \frac{0}{6} = 0$. Next, one pair of neighbours of v is connected by a path of length 2 not passing through v and therefore, $C_4(v) = \frac{1}{6}$. To calculate $C_5(v)$, notice that there are two pairs of neighbours of v that are connected by paths of length ≤ 3 and therefore $C_5(v) = \frac{2}{6} = \frac{1}{3}$.

Example 2 — We compute $C_3(u)$, $C_4(u)$ and $C_5(u)$ of node u in the graph presented in Figure 1B. Node u has 4 neighbours and therefore there are 6 possible pairs of its neighbours. Its $C_3(u) = 0$, $C_4(u) = 0$ and $\forall k \geq 5$ its $C_k(u) = \frac{1}{6}$. Therefore even though node u in Figure 1B is a member of many more cycles of length 5 than node v in Figure 1A, its $C_5(u)$ is less than $C_5(v)$ of node v in the Figure 1A. This is because the cycle coefficient of order k is concerned with the presence or absence of the cycles between node pairs, not the number of cycles.

Note that the cycle coefficient is always less than 1. Also, $\forall\ k_1 < k_2$, the average cycle coefficients of network G satisfy $C_{k_1}(G) \leq C_{k_2}(G)$. In particular, the cycle coefficient is always greater than or equal to the clustering coefficient.

3 Results

We focus on human protein-protein interaction (PPI) data made available over the last 10 years through yeast two-hybrid (Y2H) high-throughput interaction detection technology. We find

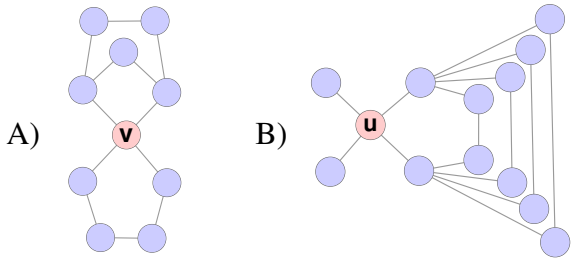

Figure 1: Cycle coefficient computation example.

that, from 2005 to 2013, the average number of neighbours increased and both the average path length and the diameter decreased, suggesting that the human PPI network is becoming more compact and less sparse (Table 1). Also, the degree distributions of HI-2005, HI-2011 and HI-2013 roughly follow a power-law (Figure 2), possibly meaning that Y2H screening of human protein interactions has been progressing in a consistent manner. A non-quantitative change in the degree distribution (e.g. a shift away from the power-law distribution to, say, a random distribution) would indicate a major change in the global topological properties of the human PPI network. As this is not the case, and the distributions differ only in the number of proteins having a certain degree, this suggests that new screening experiments are adding proteins and interactions to the human PPI network in a topologically consistent way.

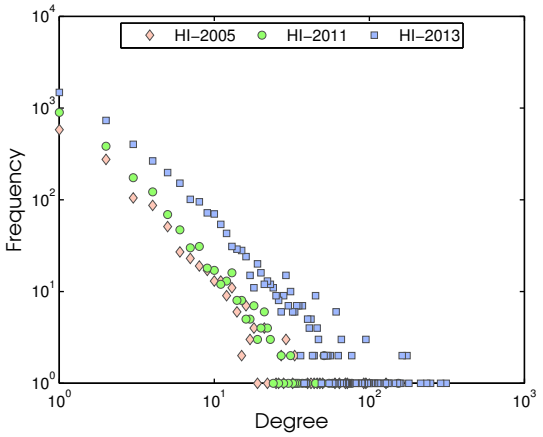

Figure 2: Degree distributions for HI-2005, HI-2011 and HI-2013 (log-log scale).

3.1 Changes in network topology: new proteins as mediators

The HI-2013 data set has 2,525 newly added proteins compared to the previous HI-2011 interactome version; (note that 460 proteins from HI-2011 are absent from HI-2013). We asked what the impact of the addition of these 2,525 proteins on HI-2013 network topology was and how the new proteins got "wired" into the human PPI network. To answer this, we analyse HI-2013 using graphlets (for details see Methods, section 2.2).

The degree distribution of the 2,525 newly added proteins again follows a power-law, with 1,233 of them being linked to only one other protein in HI-2013, 473 of them being linked to

two other proteins, 235 to three, etc., but there are 16 new proteins with over 70 interacting partners in HI-2013 (four of which have over 200 interacting partners, and one of them has 313 interacting partners – KRT40 is the most connected protein of HI-2013). We see this in detail in Figure 3, where we colored red all orbits i such that when we count all proteins in HI-2013 with non-empty i^{th} orbit, we find that the newly added proteins contribute over 50% to these counts.

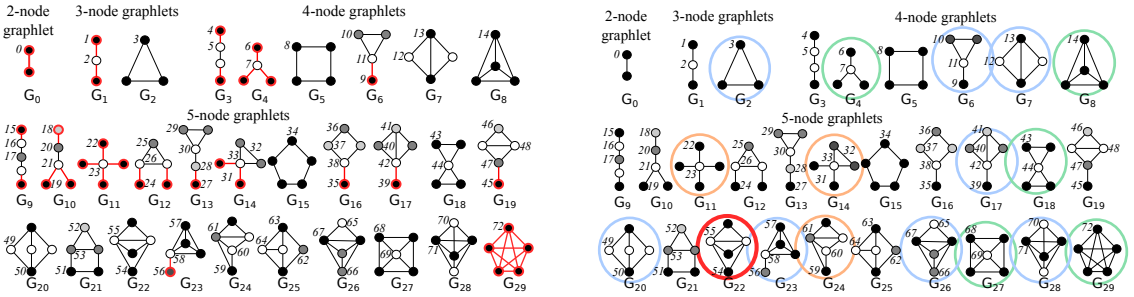

Figure 3: Left panel — Orbits coloured red are the most frequent ones in the newly added proteins (i.e., in the 2,525 proteins that exist in HI-2013, but not in HI-2011). Right panel — Presence of specific topological wirings (i.e., graphlets) in the full HI-2013 network. Circled graphlets are those for which a large portion of their counts in HI-2013 come from proteins added to HI-2013.

To gain further insight into the wiring patterns (i.e. topology) of HI-2013 caused by added proteins, we analyse graphlet counts (note, the higher the number of occurrences of a graphlet, the more prominent that wiring pattern within the network) and find several surprisingly specific topological wirings in HI-2013 caused by the 2,525 newly added proteins (Figure 3, right panel). We look for graphlets such that over 50% of their counts in HI-2013 come from at least one of the newly added proteins: in Figure 3 (right panel), if 50%–59% of graphlet G_i counts ($i \in \{0, 1, \ldots, 29\}$) in HI-2013 include at least one newly added protein, we circle graphlet G_i in blue; if 60%–69% of graphlet G_i counts in HI-2013 include at least one newly added protein, we circle graphlet G_i in green; if over 70% of graphlet G_i counts in HI-2013 include at least one newly added protein, we circle that graphlet in orange; finally, graphlet G_{22} is circled in red since 94% of its counts in HI-2013 include at least one newly added protein. What does this mean? For instance, graphlet G_{22} is present 2,454,886 times in HI-2013 and 2,302,331 of these counts include newly added proteins. This means that the newly added proteins are responsible for 94% of HI-2013 wiring patterns described by graphlet G_{22}, and being able to visualise the change in wiring in this way (i.e. by observing graphlet G_{22} and other prominent grphalets) facilitates the analysis of these newly introduced topological features.

Next, in order to distinguish the precise topological position of newly added proteins within, we look whether newly added proteins tend to touch G_{22} at orbit 54 or orbit 55. This will tell us the exact orbit within graphlet G_{22} which describes how the new proteins interact with the rest of the network. We find that about 94% of orbit 55 counts in HI-2013 come from the newly added proteins, while about 58% of orbit 54 counts come from them. Hence, one of the topological patterns by which newly added proteins got wired into HI-2013 is described by orbit 55, indicating that the new proteins tend to link to existing proteins in the interactome that were not interacting between themselves. Similarly, we look at orbits of graphlet G_{24} and find that about 78% of orbit 61 counts in HI-2013 come from the newly added proteins, about 64%

of orbit 60 counts come from the new proteins, and about 48% of orbit 59 counts come from the new proteins — again, this indicates that the newly added proteins are more likely to get linked to existing proteins that do not interact between themselves. We conclude the same when we analyse orbits of G_{14}. In G_{11}, new proteins contributed 73% to the count of orbit 23 and 53% to the count of orbit 22, again confirming the above conclusion. The same can be concluded from similar analyses of orbits of the remaining circled graphlets.

3.2 The human interactome is getting more complete

We have seen how newly added proteins and interactions tend to get wired into the human PPI network. Now, we look at the topology of the entire HI-2013 interactome and how it evolves as more PPI data becomes available. For each of the three human PPI networks (HI-2005, HI-2011 and HI-2013), we generate random network models of their size to see which one best fits the data (see Methods section 2.3 for details on constructing random models).

We see that the best fitting network model for the human interactome is the stickyness-index based model (STICKY), followed by the geometric model (GEO). This is consistent across all three instances of the human interactome data (Figure 4). However, we find that the relative rise in the fit of GEO over time (HI-2005 $\rightarrow$ HI-2013) is 6.31% and the relative drop in the fit of ER is 2.16%, pointing to the fact that the human interactome is getting more complete. This is because geometric graphs have already been shown to model well higher-confidence and more complete PPI data [30, 31, 10].

We also examine the fit of random network models to the interactomes of model organisms worm, plant, yeast and fly (Figure 5). YI-2 is best modelled by GEO (yeast is currently considered to have the most complete interactome), AT-1 by GEO and STICKY (Figure 5), and we have seen in Figure 4 that the topology of the human interactome has been approaching GEO over the years as well. The interactomes of worm and fly are currently best modelled by STICKY, with GEO being the second best fitting model.

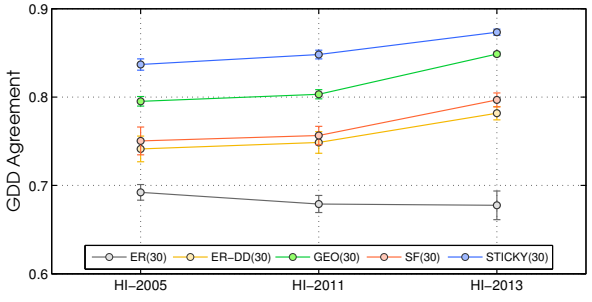

Figure 4: The fit of random network models (ER, ER-DD, GEO, SF, and STICKY) to the three versions of the human PPI networks.

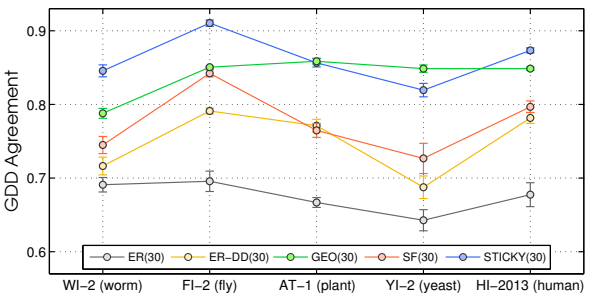

Figure 5: The fit of random network models (ER, ER-DD, GEO, SF, and STICKY) to the PPI networks of human and model organisms.

3.3 Alignment of interactomes across species

We perform a topological network alignment (see Methods, section 2.4) of the three human PPI networks with PPI networks of plant, worm, yeast and fly. The results show that as the human interactome is getting more complete over time, the correctly aligned network region with yeast and worm increase, while with the plant and fly decrease (Figure 6 and Table 2). The aligned sub-network of HI-2005 and YI-2 contains 783 proteins and 644 interactions with the largest connected component (LCC) in the alignment containing 318 proteins and 318 interactions (so it has only one cycle); the alignment of HI-2011 and YI-2 has 1,014 proteins and 866 interactions with its LCC containing 514 proteins and 514 interactions (again, almost a tree, i.e., it contains only one cycle); and the alignment of HI-2013 and YI-2 has 1,306 proteins and 1,152 interactions with its LCC containing 844 proteins and 845 interactions (again, almost a tree, with only two cycles).

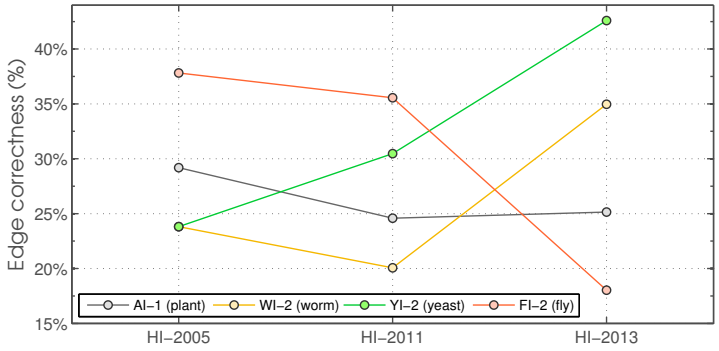

Figure 6: Alignment between human and non-human PPI networks. The y-axis shows the percentage of aligned interactions (edges) between PPI networks (numbers of correctly aligned interactions are given in Table 2).

Table 2: Alignment between human PPI networks and PPI networks of *A. thaliana*, *C. elegans*, *S. cerevisiae* and *D. melanogaster*: edge correctness values are given as percentages with respect to the perfect alignment of all edges of the smaller network to the edges of the bigger network. In brackets are the numbers of correctly aligned edges, meaning that if protein A in network 1 is aligned to protein A' in network 2 and protein B in network 1 is aligned to protein B' in network 2, then if AB is an edge in network 1, A'B' is an edge in network 2.

	AI-1 / plant	WI-2 / worm	YI-2 / yeast	FI-2 / fly
HI-2005	29.18% (744)	24.16% (616)	25.26% (644)	38.87% (991)
HI-2011	24.79% (922)	20.60% (766)	32.01% (866)	35.55% (1322)
HI-2013	23.53% (1301)	35.30% (1141)	42.59% (1152)	18.02% (2420)

We examine the biological function of yeast–human alignments by computing the enrichment of Gene Ontology[2] (GO) terms in each of the three human–yeast aligned sets of proteins and find that the enriched biological process (BP) and molecular function (MF) terms include regulation of transcription, DNA repair, cell cycle and apoptosis as some of the top statistically significantly enriched GO terms (p-value ≤ 0.01, all p-values were adjusted using Benjamini-Hochberg multiple hypothesis testing procedure). Also, we compute GO term overlap for each

[2] http://www.geneontology.org/

human–yeast aligned protein pair in each of the three alignments. Shared GO terms between a pair of inter-species aligned proteins are indicative of their similar biological functions, therefore providing a link between topological analysis and biological function. We find that in each of the three human–yeast alignments, over 40% of aligned proteins have at least one GO term in common (42.83% in HI-2005–yeast alignment, 44.77% in HI-2011–yeast alignment, and 40.59% in HI-2013–yeast alignment). This indicates that, since the alignments are obtained purely from the topology of the interactomes and since biological function is conserved (through both enriched and shared GO terms in the inter-species alignments), the human interactome has been growing so that its topology is reflective of biological function. We did not analyse the alignment of the human interactome with those of other model organisms because their interactomes are less complete and contain more noise than interactomes of yeast and human.

3.4 There are no "large cycles" in PPI networks

To examine the change of cycle content in the human interactome over time, as well as to compare the cycle content of human interactome and the interactomes of other organisms, we generalise the clustering coefficient, which corresponds to triangles (3-node cycles), to analogous coefficients that correspond to larger cycles (4-node, 5-node, 6-node cycles etc.; as described in section 2.5). The results are presented in Figure 7 (left panel): while HI-2005 and HI-2011 had similar content of cycles, HI-2013 has many more cycles; also, for all versions of the human interactome, there are practically no cycles with more than 7 proteins; similar holds for PPI networks of model organisms. This indicates that the interactomes are "tightly wired" in the way that forbids large cycles. Examining causes and implications of this is a subject of further research.

We compared cycle coefficients of the human interactome with those of model networks and found that the all three versions of the human interactome are much closer to STICKY and GEO than to ER graphs with respect to cycle coefficients (Figure 7, right panel).

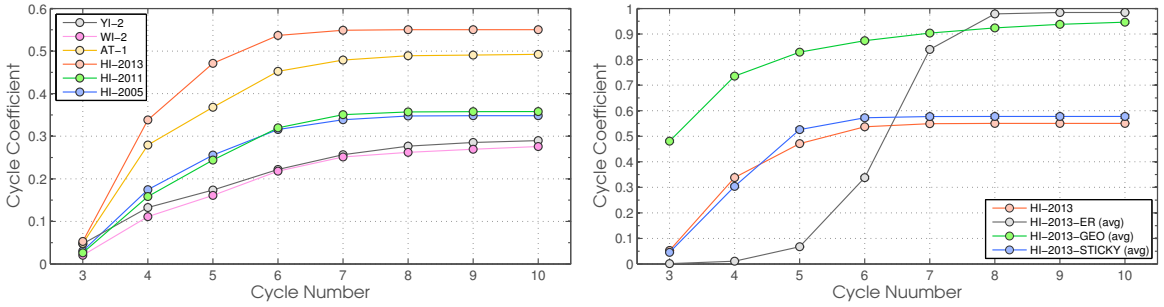

Figure 7: Left panel — Cycle coefficients for PPI networks. Numbers on the horizontal axis correspond to cycle size, i.e., 3 stands for the cycle coefficient for 3-node cycles, C_3 (this is the clustering coefficient), 4 for the cycle coefficient for 4-node cycles etc. The vertical axis gives the value of the i^{th} cycle coefficient of a network, for $i = 1, 2, 3, \ldots, 10$. Right panel — Cycle coefficients for HI-2013 and model networks.

3.5 The topologically conserved core of the human interactome

It was previously found that by applying k-core decomposition (see [35] and [36] for methodological details) to the human PPI network from BioGRID [3] and HPRD[4], a unique topological structure emerges — namely, the "core" sub-network of the human interactome — which is enriched in disease genes, drug targets and contains genes that are known to drive disease formation. Hence, we examine how this core changed through the three versions of the human interaction networks, HI-2005, HI-2011, and HI-2013.

We find that the core of the interactome gets larger and denser with time (Table 3), and that a part of it remains topologically conserved (i.e., not altered by addition of new proteins and interactions over time) across all three networks: it consists of 20 proteins and 59 interactions (Tables 3, 4 and Figure 8). This set of 20 proteins is statistically significantly enriched with GO terms of cellular localisation, in particular, with non-membrane-bounded organelles, which are known to govern cellular structure and morphology and include organelles such as ribosomes, the cytoskeleton and chromosomes. The 20 proteins are also enriched in coiled coil domains, which are usually present on transcription factors, proteins involved in cell proliferation and growth, regulation of gene expression, and HIV related proteins. Many of them are, indeed, involved in regulation of transcription and cancer development (Table 4). For example, keratin protein family is statistically significantly enriched in the core of HI-2013: tumor tissues have been shown to strongly express keratines [37, 38], and also solid epithelial tumors (both primary carcinomas and their metastases) have been shown to exclusively contain keratin intermediate-sized filaments [39, 40].

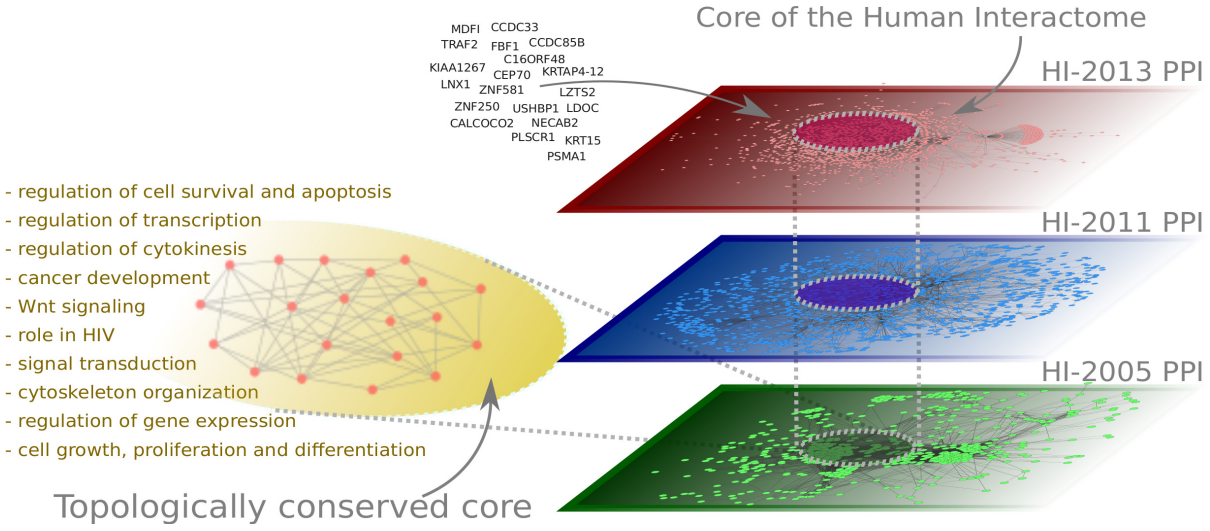

Figure 8: The core of the human interactome is conserved (i.e., remains topologically unchanged) as the interactome grows.

[3] http://www.thebiogrid.org
[4] http://www.hprd.org

Table 3: Network properties of core sub-networks of HI-2005, HI-2011, HI-2013, and the intersection of the core sub-networks (i.e., the conserved core sub-graph). The column labels are as follows: $|N|$, number of nodes; $|E|$, number of edges; CC, clustering coefficient; APL, avarage path length; ANN, average number of neighbours; d, diameter; and r, radius.

	$\|N\|$	$\|E\|$	CC	APL	ANN	d	r
HI-2005-Core	62	284	0.109	2.14	9.16	4	3
HI-2011-Core	91	449	0.105	2.26	9.86	4	3
HI-2013-Core	138	1722	0.210	1.94	24.95	3	2
Intersection	20	59	0.240	1.80	5.90	3	2

Table 4: The biological function of the 20 proteins from the conserved core across the three versions of the human interactome.

	Gene name	Function
1	PSMA1	proteasome subunit
2	LZTS2	regulation of cytokinesis and Wnt signaling
3	MDFI	regulation of transcription and Wnt signaling
4	ZNF250	potential regulator of transcription
5	TRAF2	regulation of cell survival and apoptosis
6	LNX1	signal transduction and potential role in tumorgenesis
7	CALCOCO2	cytoskeleton organization
8	PLSCR1	cell proliferation and differentiation
9	ZNF581	regulation of transcription
10	KRT15	keratin (intermediate filament) protein family (KAP)
11	KRTAP4-12	keratin (intermediate filament) protein family (KAP)
12	CCDC33	coiled-coil domain protein
13	FBF1	keratin cell polarity
14	CEP70	mitotic spindle organization
15	LDOC1	role in cancer development
16	CCDC85B	repressor of transcription; cell growth
17	KIAA1267	regulation of transcription
18	NECAB2	binding partner of adenosine A2A receptor
19	USHBP1	unknown
20	C16ORF48	unknown

4 Discussion

We performed a network analysis of the human protein-protein interaction (PPI) data sets published over the past decade to study the topological evolution of the human interactome. We found that the human PPI network is becoming more compact and less sparse. It is interesting that while mostly the same proteins were hubs in HI-2005 and HI-2011, that is not the case in HI-2013, where other proteins took over the role of hubs — mainly keratin-type proteins. Note that some proteins that were hubs in HI-2011 lost edges in HI-2013. This may be due to the fact that HI-2011 has 460 proteins that HI-2013 does not have. So perhaps a union of HI-2011 and HI-2013 would provide a more complete version of the interactome.

We found a topological pattern of interactome growth process: the newly screened proteins and interactions tend to link existing proteins in the interactome that were not interacting between themselves; and these newly added topological features are contributing to the more complete topology of the interactome. For instance, we find that the human interactome is losing the random, Erdös-Rényi-like structure, while at the same time gaining a more geometric (GEO) structure, which is characteristic to well studied and more complete interactomes, such as that of baker's yeast. Also, by aligning the human interactome to the interactomes of model organisms, we see an increasing functional and topological overlap with the yeast's interactome, and a divergence from the plant's interactome.

This led us to search for a conserved part of the interactome, which we found: we identified a topologically and functionally conserved "core sub-structure" (preserved from the 2005 version of the interactome), enriched in biological processes related to transcription and cancer development.

We conclude that the human interactome is evolving in principled, non-random ways. The search for mechanisms driving its data acquisition is far from complete, however, as more data becomes available we are beginning to gain insights into global trends which govern the growth of the "protein interactome space".

Acknowledgements

This work was supported by the European Research Council (ERC) Starting Independent Researcher Grant 278212, the National Science Foundation (NSF) Cyber-Enabled Discovery and Innovation (CDI) OIA-1028394, the Serbian Ministry of Education and Science Project III44006, and ARRS project J1-5454. We thank Dr. Oleksii Kuchaiev for help with cycle coefficient work. The authors declare that there is no conflict of interests regarding the publication of this article.

References

[1] M. Vidal, M. E. Cusick and A.-L. Barabasi. Interactome networks and human disease. *Cell*, 144(6):986–998, 2011.

[2] M. Vidal. A unifying view of 21st century systems biology. *FEBS Letters*, 583(24):3891–3894, 2009.

[3] W. Kelly and M. Stumpf. Protein–protein interactions: from global to local analyses. *Current Opinion in Biotechnology*, 19(4):396–403, 2008.

[4] S. Fields and O. Song. A novel genetic system to detect protein protein interactions. *Nature*, 340:245–246, 1989.

[5] G. Rigaut, A. Shevchenko, B. Rutz, M. Wilm, M. Mann and B. Séraphin. A generic protein purification method for protein complex characterization and proteome exploration. *Nature Biotechnology*, 17(10):1030–1032, 1999.

[6] K. Venkatesan, J. Rual, A. Vazquez et al. An empirical framework for binary interactome mapping. *Nature Methods*, 6(1):83–90, 2009.

[7] H. Yu, P. Braun, M. Yıldırım et al. High-quality binary protein interaction map of the yeast interactome network. *Science*, 322(5898):104–110, 2008.

[8] H. Yu, L. Tardivo, S. Tam et al. Next-generation sequencing to generate interactome datasets. *Nature Methods*, 8(6):478–480, 2011.

[9] M. Žitnik, V. Janjić, C. Larminie, B. Zupan and N. Pržulj. Discovering disease-disease associations by fusing systems-level molecular data. *Scientific Reports*, 3, 2013.

[10] V. Janjić, R. Sharan and N. Pržulj. Modelling the yeast interactome. *Scientific Reports*, 4:4273, 2014.

[11] J. Berg and M. Lassig. Local graph alignment and motif search in biological networks. *Proceedings of the National Academy of Sciences of USA*, 101:14689–14694, 2004.

[12] B. P. Kelley, Y. Bingbing, F. Lewitter, R. Sharan, B. R. Stockwell and T. Ideker. PathBLAST: a tool for alignment of protein interaction networks. *Nucleic Acids Research*, 32(Web Server issue):W83–W88, 2004.

[13] R. Sharan, S. Suthram, R. M. Kelley, T. Kuhn, S. McCuine, P. Uetz, T. Sittler, R. M. Karp and T. Ideker. Conserved patterns of protein interaction in multiple species. *Proceedings of the National Academy of Sciences of USA*, 102:1974–1979, 2005.

[14] Z. Liang, M. Xu, M. Teng and L. Niu. Comparison of protein interaction networks reveals species conservation and divergence. *BMC Bioinformatics*, 7(1):457, 2006.

[15] M. Koyuturk, Y. Kim, U. Topkara, S. Subramaniam, W. Szpankowski and A. Grama. Pairwise alignment of protein interaction networks. *Journal of Computational Biology*, 13(2):182–199, 2006.

[16] J. Flannick, A. Novak, S. Balaji, H. Harley and S. Batzglou. Graemlin general and robust alignment of multiple large interaction networks. *Genome Research*, 16(9):1169–1181, 2006.

[17] R. Singh, J. Xu and B. Berger. Pairwise global alignment of protein interaction networks by matching neighborhood topology. In *RECOMB*, pages 16–31. Springer, 2007.

[18] T. Milenkovic, W. Leong Ng, W. Hayes and N. Pržulj. Optimal network alignment with graphlet degree vectors. *Cancer Informatics*, 9:121–137, 2010.

[19] O. Kuchaiev and N. Przulj. Integrative network alignment reveals large regions of global network similarity in yeast and human. *Bioinformatics*, 27(10):1390–1396, 2011.

[20] M. Ostaszewski, S. Eifes and A. del Sol. Evolutionary conservation and network structure characterize genes of phenotypic relevance for mitosis in human. *PloS ONE*, 7(5):e36488, 2012.

[21] J. Rual, K. Venkatesan, T. Hao et al. Towards a proteome-scale map of the human protein–protein interaction network. *Nature*, 437(7062):1173–1178, 2005.

[22] A. I. M. Consortium. Evidence for network evolution in an arabidopsis interactome map. *Science*, 333(6042):601–607, 2011.

[23] S. Li, C. Armstrong, N. Bertin et al. A map of the interactome network of the metazoan c. elegans. *Science Signalling*, 303(5657):540, 2004.

[24] N. Simonis, J. Rual, A. Carvunis et al. Empirically controlled mapping of the caenorhabditis elegans protein-protein interactome network. *Nature Methods*, 6(1):47–54, 2008.

[25] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori and Y. Sakaki. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences*, 98(8):4569–4574, 2001.

[26] P. Uetz, L. Giot, G. Cagney et al. A comprehensive analysis of protein–protein interactions in saccharomyces cerevisiae. *Nature*, 403(6770):623–627, 2000.

[27] C. Stanyon, G. Liu, B. Mangiola, N. Patel, L. Giot, B. Kuang, H. Zhang, J. Zhong, R. Finley Jr et al. A drosophila protein-interaction map centered on cell-cycle regulators. *Genome Biology*, 5(12):R96, 2004.

[28] E. Formstecher, S. Aresta, V. Collura et al. Protein interaction mapping: a drosophila case study. *Genome Research*, 15(3):376–384, 2005.

[29] L. Giot, J. Bader, C. Brouwer et al. A protein interaction map of drosophila melanogaster. *Science*, 302(5651):1727–1736, 2003.

[30] N. Pržulj, D. G. Corneil and I. Jurisica. Modeling interactome: Scale-free or geometric? *Bioinformatics*, 20(18):3508–3515, 2004.

[31] N. Pržulj. Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23(2):e177–e183, 2007.

[32] T. Milenković and N. Pržulj. Uncovering biological network function via graphlet degree signatures. *Cancer Informatics*, 6:257–273, 2008.

[33] N. Pržulj and D. J. Higham. Modelling protein-protein interaction networks via a stickiness index. *Journal of the Royal Society Interface*, 2006.

[34] S. Cook. The complexity of theorem-proving procedures. In *Proceedings of the 3rd Annual ACM Symposium on Theory of Computing: 1971; New York*, pages 151–158. 1971.

[35] V. Batagelj and M. Zaversnik. An O(m) algorithm for cores decomposition of networks. *Symposium A Quarterly Journal In Modern Foreign Literatures*, cs.DS/0310(m):1–10, 2003.

[36] V. Janjić and N. Pržulj. The core diseasome. *Molecular BioSystems*, 8:2614–2625, 2012.

[37] A. Markey, E. Lane, D. Macdonald and I. LEIGH. Keratin expression in basal cell carcinomas. *British Journal of Dermatology*, 126(2):154–160, 1992.

[38] R. Moll, W. Franke, D. Schiller, B. Geiger, R. Krepler et al. The catalog of human cytokeratins: patterns of expression in normal epithelia, tumors and cultured cells. *Cell*, 31(1):11–24, 1982.

[39] F. Ramaekers, D. Haag, A. Kant, O. Moesker, P. Jap and G. Vooijs. Coexpression of keratin-and vimentin-type intermediate filaments in human metastatic carcinoma cells. *Proceedings of the National Academy of Sciences*, 80(9):2618–2622, 1983.

[40] M. Hendrix, E. Seftor, Y. Chu, K. Trevor and R. Seftor. Role of intermediate filaments in migration, invasion and metastasis. *Cancer and Metastasis Reviews*, 15(4):507–525, 1996.