Research Article

Links between critical proteins drive the controllability of protein interaction networks

Stefan Wuchty[1,2,3], *Toni Boltz*[1] *and Hande Küçük-McGinty*[1]

[1] Department of Computer Science, University of Miami, Coral Gables, FL, USA
[2] Center of Computational Sciences, University of Miami, Coral Gables, FL, USA
[3] Sylvester Comprehensive Cancer Center, University of Miami, Miami, FL, USA

Focusing on the interactomes of *Homo sapiens*, *Saccharomyces cerevisiae*, and *Escherichia coli*, we investigated interactions between controlling proteins. In particular, we determined critical, intermittent, and redundant proteins based on their tendency to participate in minimum dominating sets. Independently of the organisms considered, we found that interactions that involved critical nodes had the most prominent effects on the topology of their corresponding networks. Furthermore, we observed that phosphorylation and regulatory events were considerably enriched when the corresponding transcription factors and kinases were critical proteins, while such interactions were depleted when they were redundant proteins. Moreover, interactions involving critical proteins were enriched with essential genes, disease genes, and drug targets, suggesting that such characteristics may be key for the detection of novel drug targets as well as assess their efficacy.

Received: February 8, 2017
Revised: March 23, 2017
Accepted: April 7, 2017

Keywords:
Critical proteins / Network control / Protein–protein interactions

Additional supporting information may be found in the online version of this article at the publisher's web-site

1 Introduction

The focus of modern network research has shifted to the determination of nodes that allow the control of an entire network. Controllability is a classical notion in control theory, which quantifies our ability to steer a dynamical system from any initial state to any final state in the state space. However, controllability of networks is rather a function of their topology. Liu et al. introduced a maximum matching approach to find a subset of control nodes in directed networks [1]. Specifically, the analysis of a directed interaction network between human signaling proteins revealed that such control nodes appear to be disease genes and drug targets [2]. In undirected protein interaction networks, controlling genes are represented by a minimum dominating set (MDSet) that "covers" all proteins in the underlying interaction network. Nacher and Akutsu suggested an optimization procedure to determine MDSets of nodes that provide control of undirected networks [3]. In a previous analysis, we showed that MDSet proteins were enriched with essential genes, disease genes as well as appeared in regulatory interactions [4, 5]. As a caveat, such analyses of MDSets only focused on one control configuration. In fact, many MDSets of the same size may exist, suggesting that critical/intermittent proteins always/sporadically appear in any control configuration (MDSet). In a recent contribution, an algorithm was proposed that allowed the determination of the roles of nodes in every configuration of MDSets [6, 7]. In particular, critical nodes that were defined as appearing in every configuration of MDSets were found to be preferably essential and appeared predominately expressed in different human tissues.

While the previously mentioned analyses shed a light on the node-specific biological characteristics of a small set of network controllers, we aimed to take the next step by

Correspondence: Dr. Stefan Wuchty, Department of Computer Science, University of Miami, 1365 Memorial Drive, room 310D, Coral Gables, FL 33146, USA
E-mail: wuchtys@cs.miami.edu
Fax: +1-305-284-4122

Abbreviations: GWAS, genome wide association; **ILP**, integer linear programming; **MDSet**, minimum dominating set

Significance of the study

In our analysis, we investigated the role of proteins that control the underlying networks of protein interactions in *H. sapiens, S. cerevisiae,* and *E. coli*. Determining all control configurations, we found that a subset of proteins that participate all configurations. Independent of the organisms considered, interactions that involve such critical proteins had the most prominent effects on the underlying network topology. From a functional perspective, we found that regulatory interactions and phosphorylation events appear to be controlled by critical proteins, when the corresponding transcription factors and kinases were critical. Furthermore, we observed that interactions with critical proteins were enriched with essential genes, disease genes, and drug targets. As a consequence, we conclude that characteristics of critical proteins are independent of the underlying organism and may be key for the detection of novel drug targets and their efficacy.

primarily investigating interactions between such nodes. Focusing on the currently best-investigated interactomes of *Homo sapiens*, *Saccharomyces cerevisiae,* and *Escherichia coli*, we determined critical, intermittent, and redundant proteins. While critical proteins appeared to be highly central, bottlenecks preferably were interactions that involved at least one critical protein. Furthermore, we found that such interactions that involved critical proteins were more functionally incoherent. As for a different level of cellular organization, critical proteins occurred in interactions that connected different protein complexes and essential genes and were enriched in regulatory interactions. As for human diseases, we found that critical proteins were interacting when they were involved in the same diseases. Furthermore, interacting critical proteins were enriched with drug targets, suggesting that controlling proteins appeared central in the domination of disease-related networks.

2 Materials and methods

2.1 Determination of critical intermittent and redundant nodes

We defined a set $S \subseteq V$ of nodes in a network $G = (V, E)$ as an MDSet if every node $v \in V$is either an element of S or adjacent to an element of S. In a binary integer linear programming (ILP), we assigned a binary variable $x_v = 1$ when a protein $v \in V$ that participates in interactions E in a protein interaction network G is an element of the MDSet, and $x_v = 0$ otherwise. The smallest set of MDSet nodes is obtained by $min \sum_{v \in V} x_v$, subject to the constraint $x_v + \sum_{w \in \Gamma(v)} x_w \geq 1$, where $\Gamma(v)$ is the set of interaction partners of protein v. However, many optimal solutions exist that provide MDSets of the same size. Such characteristics suggest the existence of subsets of nodes that always (critical nodes), never (redundant nodes), and sporadically appear in MDSets (intermittent nodes). To find such subsets, our objective is to determine if $v \in MDSet$ always appears in the MDSet of any solution. For each $v \in MDSet$, we create an ILP as before and assume that $x_v = 0$ (i.e. not participating in the MDS). After solving the ILP, we determine the size of the corresponding MDSet N_v that we obtained with $x_v = 0$. If $N_v > N$, v is a critical node, and intermittent otherwise. For all nodes that did not participate in the original MDSet, $v \notin MDSet$, we need to check if they always appear outside MDSets. For $v \notin MDSet$, we create an ILP as before and assume that $x_v = 1$ (i.e. participating in the MDSet). After solving the ILP, we determine the size of the corresponding MDSet N_v that we obtained with $x_v = 1$. If $N_v > N$, v is a redundant node, and intermittent otherwise [6,7]. To solve these ILP problems, we utilized a branch-and-bound algorithm [8] as implemented by the *lpSolve* library.

2.2 Protein–protein interactions

We collected a total of 10 531 interactions between 2492 proteins that were experimentally determined using a yeast-two-hybrid and co-complex approaches in *E. coli* [9,10]. As for *S. cerevisiae* we collected 22 243 interactions between 4467 yeast proteins that were determined by large-scale yeast-two hybrid and co-complex approaches from the HINT database [11]. In the same way, we assembled a network of 28 627 high-quality protein interactions between 8495 human proteins from the HINT database as well [11].

2.3 Regulatory interactions

We collected 95 722 links between 209 human transcription factor and 8910 human genes from the TRANSFAC [12] database as provided by mSigDB [13]. As for regulatory interactions in yeast, we utilized 48 082 regulatory interactions between 183 transcription factors and 6403 genes from the YEASTRACT database [14]. Specifically, such regulatory interactions were indicated if a binding site of given transcription factor appeared in the promoter of the underlying genes. Furthermore, we utilized the RegulonDB database [15], totaling 4442 interactions between 187 transcription factors and 1638 target genes.

2.4 Phosphorylation events

As for phosphorylation events in human, we obtained 7346 interactions between 357 kinases and 2181 human proteins from the kinaseNetworkX database [16]. Such links represent a kinase-specific phosphorylation site in a given protein. Furthermore, we collected 3466 experimentally determined phosphorylation events between 128 kinases and 4400 substrates in yeast [17].

2.5 Disease genes in *H. sapiens*

We collected 496 oncogenes and 876 tumor-suppressor genes from the CancerGenes database [18] which collects such information from the literature.

Collecting data from the HPIDB database, we used 697 human proteins that were targeted by the Hepatitis C virus, as well as 255 targets of the Herpes simplex virus, 1272 targets of HIV-1, 396 targets of the Influenza A virus, and 317 targets of the Vaccinia virus [19]. In total, we assembled a set of 2359 proteins that are targeted by viruses. As for a set of genes that are required for the viral infection process, we used 262 genes that were required by the Hepatitis C virus to infect a human host cell [20]. As for the Herpes simplex virus we collected 358 such genes [21]. Furthermore, we utilized 917 such genes of HIV-1 [22–24], 1101 genes of the Vaccinia virus [25] and 1251 genes of the Influenza A virus [26–29]. In total, we assembled a set of 3752 human genes that are required for viral infection processes.

We utilized a large set of disease-associated genes from the DisGeNet database [30], containing more than 429 036 associations, between 17 381 genes and 15 093 diseases, disorders, and clinical or abnormal human phenotypes. In particular, we utilized a set of 7690 genes that were associated to environmental chemicals and corresponding diseases as of the Comparative Toxicogenomics Database [31]. Furthermore, we used 2315 genes that were associated to genetic disorders as of the OMIM database [32] and 2113 genes that were determined through genome wide association (GWAS) as of the NHGRI GWAS catalogue [33]. We augmented this GWAS dataset with genes that were obtained from literature mining analysis, allowing us to obtain a total of 7803 genes. Moreover, we collected a set of 2639 genes that were associated with rare diseases, as of the Orphanet database (www.orpha.net). Furthermore, we utilized 2921 genes that were determined by the Clinvar database to clinical disease phenotypes through their variations [34].

2.6 Drug targets

We collected information about 2474 human and 369 *E. coli* drug targets from the Drugbank database [35].

2.7 Essential genes

As a source of information about essential genes, we collected 712 essential genes in *E. coli* and 1110 essential genes in *S. cerevisiae* from the DEG database [36], and obtained 2708 essential genes in *H. sapiens* from the online gene essentiality database [37].

2.8 Protein complexes in *E. coli, S. cerevisiae,* and *H. sapiens*

We utilized a set of 517 protein complexes in *E. coli* from a co-affinity purification study that was followed by MS analyses [10]. As for *S. cerevisiae*, we utilized 430 protein complexes compiled in [38], including the SGD Macromolecular Complex GO standard [39], the CYC2008 protein complex catalogue [40], and a set of manually curated protein complexes. Furthermore, we utilized 1843 protein complexes in *H. sapiens* from the CORUM database [41].

2.9 Protein complex participation coefficient

For each protein that is involved in at least one protein complex, we defined the protein complex participation coefficient of a protein i as $P_\mathrm{i} = \sum_{\mathrm{s}=1}^{\mathrm{N}} (\frac{n_{\mathrm{i,s}}}{\sum_{\mathrm{s}=1}^{\mathrm{N}} n_{\mathrm{i,s}}})^2$, where $n_{\mathrm{i,s}}$ is the number of links that protein i had to proteins in complex s out of N total complexes. If a protein predominantly interacted with partners of the same complex, P tended to 1 and vice versa [42].

2.10 Enrichment analysis

As for the enrichment of given functional, regulatory, and topological interactions between critical, intermittent, and redundant proteins, we counted the number of pairs that are connected by such links, N_A. As a null model, we shuffled the annotation of proteins, randomly assigning such labels, generating nonoverlapping random sets of critical, intermittent, and redundant sets of proteins. Furthermore, our concept allowed us to keep the number of random critical, intermittent, and redundant proteins constant. Based on such random sets, we analogously counted the corresponding random number of interactions between random critical, intermittent, and redundant proteins, $N_{\mathrm{r,A}}$, and defined the enrichment of such interactions as $E_\mathrm{A} = \mathrm{lg}_2(\frac{N_\mathrm{A}}{N_{\mathrm{r,A}}})$. Analogously, we determine the enrichment of genes in a set A (e.g. disease gene set).

2.11 Betweeness centrality

As a global measure of its centrality, we calculated an edges betweenness, indicating an interactions appearance in shortest paths through the whole network. In particular, we defined betweenness centrality c_{B} of an edge e as $c_{\mathrm{B}}(e) = \sum_{\mathrm{s}\neq \mathrm{t}\in \mathrm{V}} \frac{\sigma_{\mathrm{st}}(e)}{\sigma_{\mathrm{st}}}$, where σ_{st} is the number of shortest paths between proteins s and t while $\sigma_{\mathrm{st}}(e)$ is the number of shortest paths running through e. Analogously, we determined the betweeness centrality of node v as $c_{\mathrm{B}}(v) = \sum_{\mathrm{s}\neq \mathrm{t}\neq v\in \mathrm{V}} \frac{\sigma_{\mathrm{st}}(v)}{\sigma_{\mathrm{st}}}$.

3 Results

Considering protein–protein interactions, we aimed at the elucidation of proteins that are important for the controllability of the underlying network. In particular, networks are dominated by MDSets that can be determined by an integer linear programming (ILP), allowing the determination of the smallest set of nodes where each non-MDSet node is adjacent to a node in the MDSet. However, many different configurations of MDSets exist that have the same cardinality. Such an assumption implies sets of nodes that always, partially, or never participate in MDSets. In particular, we defined proteins as critical if they always participated in the MDSet of a given configuration (Fig. 1A). Furthermore, we considered redundant nodes that never appeared in MDSets while intermittent nodes sporadically occurred in MDSets. Applying an algorithm that allowed us to determine such sets of nodes [6, 7], we considered protein–protein interaction networks of *H. sapiens, S. cerevisiae,* and *E. coli* that were experimentally determined through high-throughput yeast two-hybrid and co-complexing approaches. In the table of Fig. 1B, we observed that the percentage of critical nodes was roughly <10%, while intermittent nodes constituted 25–30% of all proteins. Investigating their degree distributions, we observed that sets of such proteins featured fat tails (Supporting Information Fig. S1). Repeating such an analysis with protein interaction networks that were composed of binary and co-complex interactions, respectively, we observed similar results (Supporting Information Fig. S2). Furthermore, the mean degree of critical proteins far exceeded the corresponding values of intermittent and redundant proteins that were close to the mean degree of all proteins in the underlying interaction networks. Such observations were consistent with corresponding values in binary (i.e. yeast two-hybrid) and co-complex interaction networks in all organisms (Supporting Information Fig. S2).

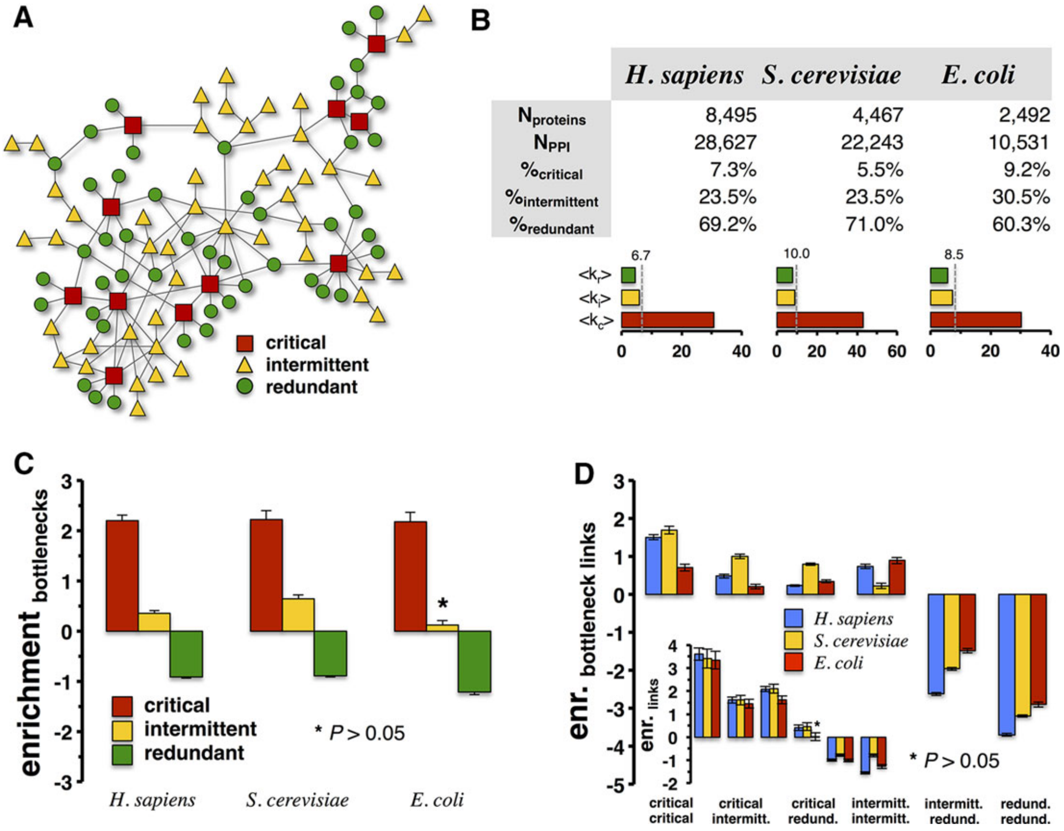
3.1 Topological characteristics

As for other topological characteristics, we calculated the betweeness centrality of all nodes in the underlying network. Choosing the top 20% of proteins with highest betweeness centrality as a set of bottleneck nodes, we calculated the enrichment of such proteins in sets of critical, intermittent, and redundant proteins. Given all proteins in the underlying interaction network, we sampled sets of proteins, by randomly shuffling their labels, generating nonoverlapping, random sets of critical, intermittent, and redundant proteins. Specifically, we observed that critical proteins in all organisms were strongly enriched with bottlenecks ($p < 10^{-4}$). Albeit weaker, intermittent proteins were enriched with bottleneck nodes as well, while critical proteins hardly were bottlenecks ($p < 10^{-4}$, Fig. 1C). The enrichment with bottlenecks suggests that critical and intermittent proteins have disruptive effects on the topology of the underlying networks. To measure a protein's impact on an interaction network's resilience, we performed a robustness analysis by sorting all critical proteins according to their degree in the underlying interaction networks. Starting with the most connected we gradually cut proteins and calculated the number of connected components after each deletion step. We found that the successive deletion of critical proteins had a higher impact on network robustness by producing more connected components while removing more interactions in *E. coli, S. cerevisiae,* and *H. sapiens* (Supporting Information Fig. S3). In comparison, we considered sets of equal size of most connected intermittent and redundant proteins and observed a sizeable but still weaker impact of most connected intermittent proteins. In turn, the removal of redundant nodes hardly had an impact on network stability.

As for topological characteristics that revolved around interactions, we counted the number of interactions in the underlying interactions networks that appeared between critical, intermittent, and redundant proteins. Randomly sampling sets of critical, intermittent, and redundant nodes, we determined the number of interactions between these three sets of proteins and found that interactions that involved critical nodes were significantly enriched (inset, Fig. 1D). In turn, interactions that involved redundant nodes mostly appeared depleted ($p < 10^{-4}$). Furthermore, we determined the betweeness centrality of all interactions and defined the top 20% of edges with highest betweeness as bottleneck edges. Determining the number of bottleneck interactions between proteins that belong to different sets, we randomly sampled sets of critical, intermittent, and redundant proteins. In Fig. 1D, we observed that bottleneck nodes are mostly enriched between critical proteins, while interactions that involved intermittent and redundant proteins were largely devoid of bottleneck links ($p < 10^{-4}$).

3.2 Molecular functions

Utilizing GO functions, we determined semantic similarities (3, 4) of all protein interactions in *H. sapiens,* focusing on terms of the biological processes, molecular function, and cellular component ontologies. Specifically, we obtained values between 0 and 1, where values close to 1 indicate highest similarity of GO terms and vice versa [43]. In Fig. 2A, we found that interactions appear to be more heterogeneous when critical proteins were involved. In turn, we observed a shift toward more functional homogeneity when redundant nodes were involved. While such observations were confirmed in yeast, we found inconclusive results in *E. coli* (Supporting Information Fig. S4).



	H. sapiens	*S. cerevisiae*	*E. coli*
$N_{proteins}$	8,495	4,467	2,492
N_{PPI}	28,627	22,243	10,531
$\%_{critical}$	7.3%	5.5%	9.2%
$\%_{intermittent}$	23.5%	23.5%	30.5%
$\%_{redundant}$	69.2%	71.0%	60.3%

Figure 1. (A) In a toy network, we illustrate the concept of critical, intermittent, and redundant nodes. (B) In the table, we present statistics of protein interaction networks of *H. sapiens*, *S. cerevisiae*, and *E. coli* and of their corresponding critical, intermittent, and redundant proteins. Notably, we found that critical proteins were strongly connected, while degrees of intermittent and redundant nodes revolved around the mean degree of all proteins in the underlying interaction networks (dashed lines). In (C), we defined the top 20% of proteins with the highest node betweeness as a set of bottleneck proteins. Randomly sampling sets of critical, intermittent, and critical proteins 10 000 times, we found that critical nodes were strongly enriched with bottlenecks. While intermittent nodes were moderately enriched, we also found a significant depletion of redundant nodes in the underlying set of bottleneck proteins ($p < 10^{-4}$). (D) Randomly sampling sets of critical, intermittent, and critical proteins, we observed that interactions involving critical proteins were enriched ($p < 10^{-4}$). In turn, interactions that involved redundant proteins were depleted. In the main plot, we defined sets of the top 20% of interactions with highest edge betweeness as bottleneck links in the underlying protein interaction networks. Furthermore, we observed that bottleneck interactions were significantly enriched if they involved a critical protein while we found the opposite when redundant proteins were involved ($p < 10^{-4}$).

3.3 Protein complexes

Moving to a higher level of cellular organization, we calculated the complex participation coefficients of proteins, a value that indicates a protein's tendency to interact with different complexes through their interactions. The complex participation coefficient tends toward 1 if the given protein predominantly interacts with proteins in the same complex and vice versa. In particular, we utilized a set of 517 protein complexes in *E. coli* [10], 409 protein complexes in *S. cerevisiae* [38] and 1843 protein complexes in *H. sapiens* [41]. In the inset of Fig. 2B, we observed that human critical proteins had a significantly lower participation than intermittent and redundant proteins, suggesting that critical proteins reached into many different protein complexes through their interactions. Critical, intermittent, and redundant proteins in yeast and *E. coli* showed similar patterns (Supporting Information Fig. S5). In Fig. 2B, we calculated the number of interactions that appeared within and between human complexes. Randomly sampling critical, intermittent, and redundant proteins, we observed that interactions that involved human critical proteins preferably appeared between complexes. We also found a significant but weaker enrichment signal, indicating that interactions with critical nodes appeared within complexes as well. Notably, interactions between and within complexes were depleted when intermittent and redundant nodes were involved. We observed similar results, when we considered complexes in *E. coli* and *S. cerevisiae* (Supporting Information Fig. S6).

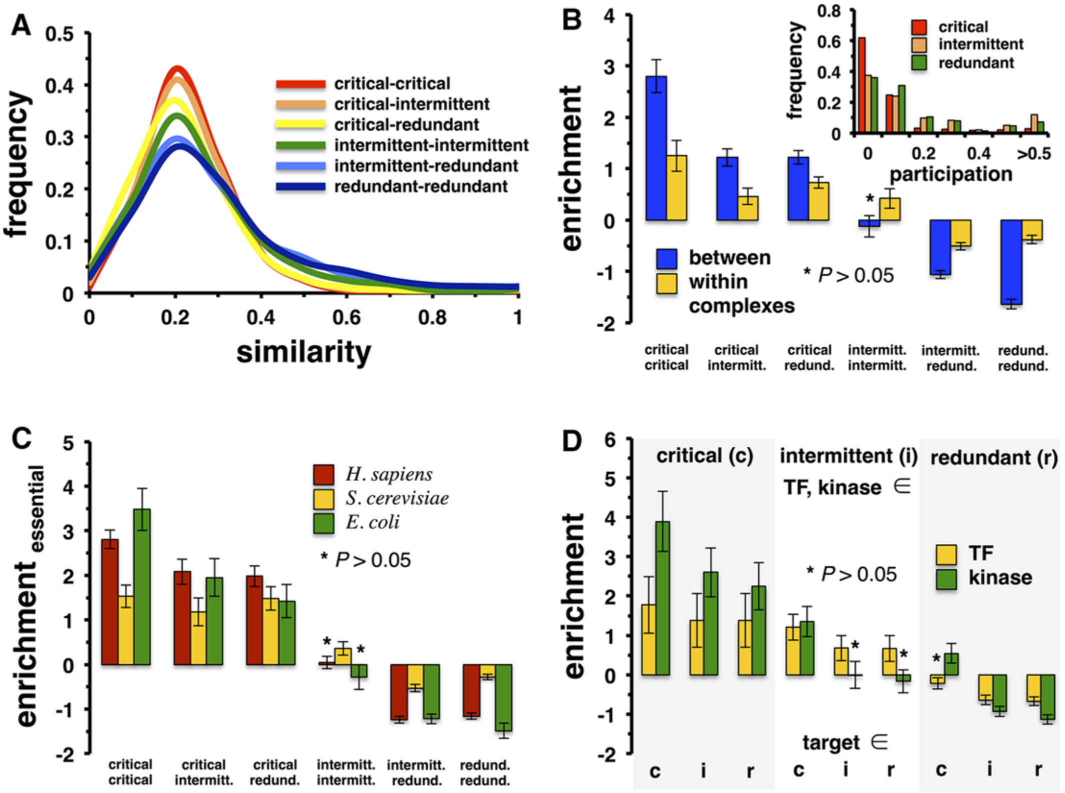


Figure 2. (A) Using human protein interactions, we calculated the functional similarity of interacting proteins. In particular, we observed that interactions between critical proteins appeared to be less functionally coherent compared to interactions between redundant proteins. (B) Using more than 1800 experimentally verified protein complexes in *H. sapiens*, we calculated the participation coefficient of each protein, using the underlying human protein interaction network. Notably, we found that critical proteins participated in more different protein complexes through their interactions than their intermittent and redundant counterparts (inset). Using protein interactions, we counted the number of interactions that appeared within and/or between protein complexes. Randomizing sets of critical, intermittent, and redundant proteins, we observed that interactions that had at least one critical protein preferably appeared between and within complexes ($p < 10^{-4}$). In (C), we observed that interactions between essential proteins preferably appeared between critical proteins in all organisms ($p < 10^{-4}$) while appeared depleted involving redundant proteins. In (D), we determined the enrichment of human phosphorylation events and transcriptional, regulatory interactions in a human protein interaction network. Randomly sampling critical, intermittent, and redundant proteins, we clearly observed that such interactions preferably occurred when at least kinases or transcription factors were critical or intermittent proteins ($p < 10^{-4}$).

3.4 Essential proteins

Given that critical proteins were found enriched with essential proteins [6], we hypothesized that essential proteins appeared predominately between critical proteins. Collecting essential genes in *E. coli* and *S. cerevisiae* from the DEG database [36], and in *H. sapiens* from the online gene essentiality database [37], we determined the number of interactions between essential proteins. Randomizing critical, intermittent, and redundant protein sets, we found that interactions between essential proteins were indeed enriched with critical proteins in all organisms (Fig. 2C).

3.5 Regulatory interactions

Assuming that critical, intermittent, and redundant proteins may significantly contribute to cellular control processes, we hypothesized that transcription factors and their target genes may significantly appear in sets of critical and intermittent proteins. Specifically, we used 95 722 regulatory interactions between 209 human transcription factors and 8910 target genes from the TRANSFAC database [12, 13]. Assuming that the same logic applies to phosphorylation events we collected 7346 interactions between 357 kinases and 2181 human proteins from the kinaseNetworkX database [16]. Specifically, we counted the times a pair of transcription factors and a given target gene appeared between critical, intermittent, and redundant proteins. Randomly sampling critical, intermittent, and redundant proteins, we observed that regulatory interactions and phosphorylation events were significantly enriched when corresponding transcription factors and kinases were critical proteins ($p < 10^{-4}$, Fig. 2D). In turn, such interactions appeared generally depleted when transcription factors and kinases were redundant proteins.

In Supporting Information Fig. S7, we observed similar, albeit weaker results when we considered such regulatory interactions in *S. cerevisiae* and *E. coli*.

3.6 Human diseases

To investigate the role of critical, intermittent, and redundant proteins in human diseases, we collected sets of disease genes from the DisGeNet database [30].

In the inset of Fig. 3A, we determined the enrichment of genes that were associated to a disease. We observed that critical and intermittent proteins were enriched in such sets, while redundant genes were significantly depleted. In the main plot of Fig. 3A, we counted the number of diseases a given gene is involved in and observed that critical proteins were increasingly enriched among proteins that appeared in numerous diseases. In Supporting Information Fig. S8A, we refined our analysis and focused on proteins that were associated to genetic diseases [32]. We found that critical and intermittent proteins were enriched in such sets, while redundant genes were significantly depleted. Such patterns were confirmed focusing on sets of genes that were associated to environmental chemicals and corresponding diseases [31], rare diseases and clinical disease phenotypes through their variations [34] ($p < 10^{-4}$). Utilizing experimentally verified disease genes that were determined through GWASs [33], we surprisingly found no significant enrichment/depletion signals. As for interactions between disease genes, we observed that such interactions were enriched when critical proteins were involved. Such a result was confirmed when we considered interactions between proteins that were involved in at least one common disease (Fig. 3B). Analogously, we refined our analysis considering different sources of diseases, allowing us to find similar results in Supporting Information Fig. S8B.

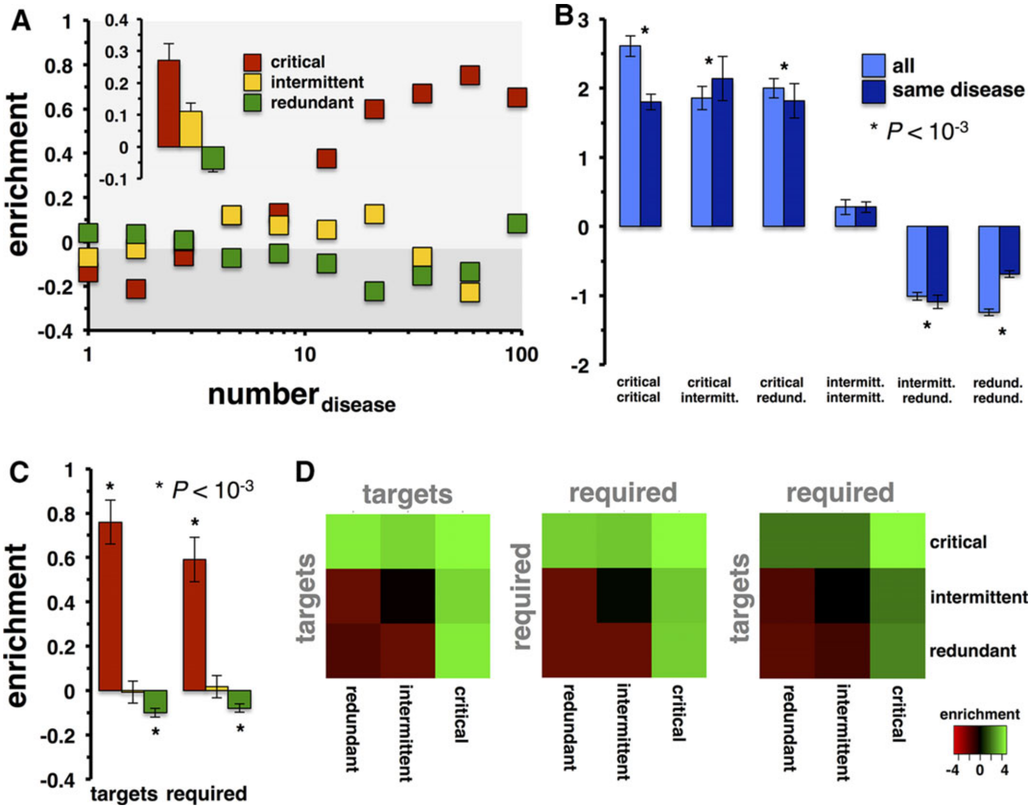


Figure 3. (A) Utilizing disease-specific data, we observed that critical and intermittent proteins in a human protein interaction network were significantly enriched with disease-associated genes (inset, $p < 10^{-4}$). Determining the number of diseases a given protein was involved in, we observed that critical proteins appeared to be enriched among highly disease involved proteins. In (B), we counted the number of interactions between disease relevant proteins. Randomizing sets of critical, intermittent, and redundant genes, we found that interactions between disease genes favor critical proteins. Furthermore, we observed a similar result when we considered interactions between proteins that shared at least one disease ($p < 10^{-4}$). (C) As for human viral targets and genes that were required for the viral infection, we found that both sets were enriched with critical proteins ($p < 10^{-4}$). (D) Randomly sampling critical, intermittent, and redundant protein in a human protein interactions network, we calculated the enrichment of interacting pairs of proteins that were targeted or were required for a viral infection. Most strikingly, we observed that interactions between human proteins were strongly enriched with genes that were targeted and required by human viruses.

As for human viral targets, we compiled a set of human host proteins that were targeted by HIV, Influenza, Vaccinia, Herpes, HPV, and Hepatitis virus as well as genes that were required for the infection process of the corresponding viruses. In a previous analysis, we found that viral targets and required genes closely clustered together in a human protein interaction network [44]. As a consequence, we hypothesized that interactions between targets and required genes may favor the presence of critical proteins. Indeed, Fig. 3C suggested that critical and intermittent proteins were enriched with targeted and required genes ($p < 10^{-4}$). In the heatmaps in Fig. 3D, we determined the enrichment of interacting pairs of critical, intermittent, and redundant genes that were targeted or required by viruses. Notably, we observed that interactions between targets and required genes were mostly carried by critical proteins. In particular, we found that interacting critical proteins were enriched with interacting targets and required genes ($p < 10^{-4}$). In Supporting Information Fig. S9, we repeated such an analysis with onco- and tumor-suppressor genes, suggesting that critical proteins are enriched with such tumor-related genes. Furthermore, we observed that interacting critical proteins significantly appeared between onco- and tumor-suppressor genes.

3.7 Drug targets

Assuming that critical proteins are preferably enriched with disease genes, we hypothesized that such proteins may be prime drug targets as well. Collecting drug targets in human and *E. coli* from the Drugbank database [35], we determined the number of drugs a given protein was associated with. In Fig. 4A, we observed that critical human proteins appeared to be enriched among proteins that are targets to numerous drugs. In Fig. 4B, we determined the enrichment of critical, intermittent, and redundant genes among human and *E. coli* drug targets, clearly suggesting that critical proteins were prime drug targets in both organisms. As for characteristics that revolved around interactions between drug targets, we found that interactions between drug targets were enriched with critical proteins in both human and *E. coli* (Fig. 4C).

4 Discussion

In conclusion, we determined critical, intermittent, and redundant proteins in the currently best investigated interactomes of *H. sapiens*, *S. cerevisiae,* and *E. coli*. Still the question remains as to how to interpret the role of such sets of proteins for the control of an undirected protein interaction network. The notion of control is immediately transparent in directed biological interactions, allowing the flow of biological information. For example, biological signals are mediated by kinases that phosphorylate a substrate protein to activate a downstream signaling cascade. In turn, the controllability of an undirected protein interaction network is more of a topological nature. Assuming that protein interaction networks mediate undirected biological information flow as well, critical and intermittent proteins would represent the smallest set that “covers” all proteins in the underlying interaction network as critical/intermittent proteins always/sporadically appear in any control configuration (MDSet). As a consequence, controllability of undirected protein interactions appears to be rather a question of the proteins topological reachability in a network than exerting some type of control through a biological mechanism.

Such an interpretation strongly suggests that critical and intermittent proteins should have a high number of interaction partners and placed in topologically central positions. Indeed, we observed that critical proteins appeared as the most connected nodes in the underlying networks as their mean degree was far higher than the mean degree of all nodes. However, we also stress that the sets of intermittent and redundant genes harbor hubs as their corresponding

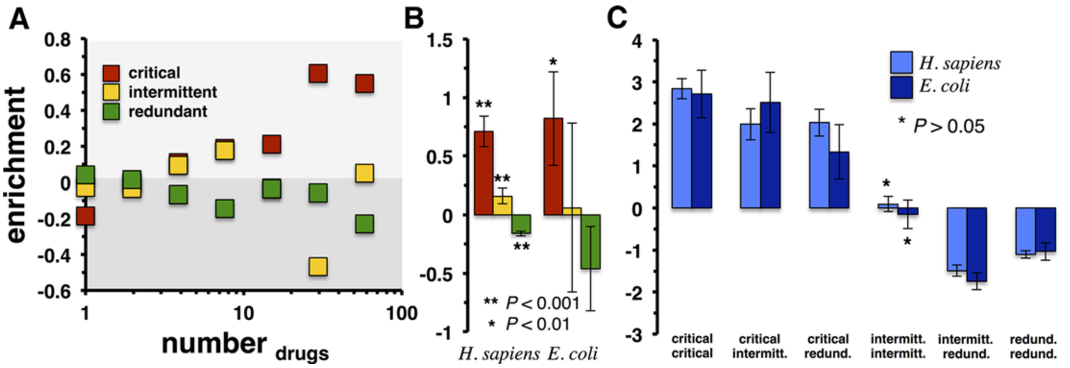


Figure 4. (A) Determining the number of drugs that work on a given protein, we observed that critical proteins appeared to be enriched among proteins that are frequent drug targets. In (B), we determined the enrichment of critical, intermittent, and redundant genes among human and *E. coli* drug targets. In both organisms, we observed that critical proteins were prime drug targets. In (C), we determined the enrichment of interactions between proteins that are drug targets. In both organisms, we found that interactions between drug targets were enriched with critical proteins ($p < 10^{-4}$).

degree distributions have fat tails as well. Yet, their mean degree revolves around the mean degree of all nodes in the underlying networks. The topological meaning of such nodes is immediately obvious when we consider their penchant to appear in sets of bottleneck nodes and interactions. Given that bottlenecks are defined as the topologically most central nodes and links, critical nodes appear to be most enriched with bottleneck nodes. Such an observation may be rooted in the extraordinarily high degree of critical nodes, as they indicate central placement in the underlying networks. In turn, intermittent proteins appeared enriched as well while we observed the opposite for redundant nodes, suggesting that degree alone is neither a criterion to be a critical or intermittent node nor indicative of centrality. In fact, the occurrence of critical and intermittent nodes is rather a question of the determination of the lowest number of strategically placed nodes. As a corollary, the topological relevance of critical and intermittent nodes is further emphasized by the observation that they appear enriched in bottleneck interactions. Notably, we observed that interactions that involved a critical node always appeared enriched with bottleneck links. However, the same only applies to interactions between intermittent nodes, indicating the topological limits of intermittent nodes.

Assuming that critical, intermittent, and redundant proteins and their immediate interactions indeed play a governing, biological role in the underlying protein interaction networks, we expected their involvement in biological processes that control a cell. Indeed, we found that critical proteins may exert their governance by reaching into more protein complexes than intermittent and redundant proteins, largely connecting complexes through their interactions. Considering protein functions between interacting proteins, we observed a gradual change from heterogeneous to more homogeneous functions when we considered interactions between critical proteins to interactions between redundant proteins. While this pattern largely holds, we observed significant differences between yeast, human, and *E. coli*. In particular, interactions with critical proteins appeared between more functionally heterogeneous proteins in yeast and human, while differences appear rather blurred in *E. coli*. Such an observation may be rooted in the quality of the underlying protein interaction data as the yeast and the human interactome is far better investigated and curated than the *E. coli* counterpart. In contrast to the human and yeast interactome, the vast majority of *E. coli* interactions were determined with a co-complex approach. Compared to binary interactions that have been found with yeast two-hybrid techniques co-complex approaches tend to strongly connect functionally homogeneous proteins as well as provide denser networks, potentially increasing the percentage of critical and intermittent proteins (Fig. 1B).

Furthermore, we found that interactions that involve critical proteins are enriched with essential genes. Such a results is a corollary to previous results, indicating that essential proteins tend to cluster together in protein interaction networks of different organisms [45]. Furthermore, critical proteins have already been observed to be essential [6]. However, our results indicate that such an observations were largely a matter of centrally placed critical proteins, as essential proteins preferably interact when one is a critical protein. On a functional level, we previously found that transcription factors and protein kinases were significantly present among MDSet proteins [4]. As a corollary, pairs of proteins that involved at least one critical protein were preferably connected by regulatory interactions and phosphorylation events. Notably, enrichment signals of such links that mediated biological control were strongest when a transcription factor or kinase was a critical protein.

Another strong indication that critical, intermittent, and redundant proteins and their immediate interactions indeed play a governing, biological role in the underlying protein interaction networks came from our observations that critical nodes appeared to be strongly enriched with disease genes. Such a result confirms previous reports that network controllers tend to be disease-related genes in a directed human signaling network [2]. While this characteristic applied for different subsets of disease genes, we surprisingly found the absence of any enrichment among disease genes that were curated from GWAS studies. Such an observation may be rooted in the fact that the involvement of correlations of genomic variations and disease may be less well experimentally curated than other biomedical information. As a corollary, we found that disease genes tend to interact, when interactions between two disease genes involved at least one critical protein. Notably, such a result applied to all sources of disease information, suggesting that interactions of critical proteins could guide the search for disease relevant genes. As for infectious diseases, we found that critical proteins were enriched with proteins that were targeted as well as required by human viruses. While we recently observed that such sets of virus relevant genes tend to cluster together [44], we found that interactions between targeted and required genes preferably occurred when both proteins were critical in the human underlying interaction network. As a final indication of the relevance of critical proteins for the controllability of a network, we found that critical proteins in human and *E. coli* are preferable drug targets while drug targets interacted when they involved at least one critical protein in both organisms. As a consequence, we concluded that such a characteristic could be potentially harnessed to find other disease genes and drug targets as well as assess the efficacy of novel drugs.

Our observations also may dependent on the inherent, yet inevitable bias toward interactions that involve heavily investigated proteins [46] as well as rates of false positives and negatives. Furthermore, estimates indicate that current interactomes are far from being complete. In our previous work, we showed that the sizes of MDSets are relatively stable, when we accounted for the presence of false interactions [4]. Given that critical proteins are proteins that always appear in MDSets of any configuration, we expect that sets of critical proteins may be relatively robust in the presence of an inherent bias as well as noise. Although we cannot rule out that sets of critical, intermittent, and redundant proteins will not change once biases are removed and interactomes approach completion, we assume that the biological and topological characteristics of such protein sets will not dramatically change, as our observations seem to be largely driven by central topological placement of the underlying critical proteins.

The authors have declared no conflict of interest.

5 References

[1] Liu, Y. Y., Slotine, J. J., Barabasi, A. L., Controllability of complex networks. *Nature* 2011, *473*, 167–173.

[2] Vinayagam, A., Gibson, T. E., Lee, H. J., Yilmazel, B. et al., Controllability analysis of the directed human protein interaction network identifies disease genes and drug targets. *Proc. Natl. Acad. Sci. USA* 2016, *113*, 4976–4981.

[3] Nacher, J., Akutsu, T., Dominating scale-free networks with variable scaling exponent: heterogeneous networks are not difficult to control. *New J. Phys.* 2012, *14*, 073005.

[4] Wuchty, S., Controllability in protein interaction networks. *Proc. Natl. Acad. Sci. USA* 2014, *111*, 7156–7160.

[5] Khuri, S., Wuchty, S., Essentiality and centrality in protein interaction networks revisited. *BMC Bioinform.* 2015, *16*, 109.

[6] Ishitsuka, M., Akutsu, T., Nacher, J. C., Critical controllability in proteome-wide protein interaction network integrating transcriptome. *Sci. Rep.* 2016, *6*, 23541.

[7] Nacher, J. C., Akutsu, T., Analysis of critical and redundant nodes in controlling directed and undirected complex networks using dominating sets. *J. Compl. Networks* 2014, *2*, 394–412.

[8] Land, A. H., Doig, A. G., An automatic method of solving discrete programming-problems. *Econometrica* 1960, *28*, 497–520.

[9] Rajagopala, S. V., Sikorski, P., Kumar, A., Mosca, R. et al., The binary protein-protein interaction landscape of *Escherichia coli*. *Nat. Biotechnol.* 2014, *32*, 285–290.

[10] Hu, P., Janga, S. C., Babu, M., Diaz-Mejia, J. J. et al., Global functional atlas of *Escherichia coli* encompassing previously uncharacterized proteins. *PLoS Biol.* 2009, *7*, e96.

[11] Das, J., Mohammed, J., Yu, H., Genome-scale analysis of interaction dynamics reveals organization of biological networks. *Bioinformatics* 2012, *28*, 1873–1878.

[12] Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I. et al., TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* 2006, *34*, D108–D110.

[13] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S. et al., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* 2005, *102*, 15545–15550.

[14] Abdulrehman, D., Monteiro, P. T., Teixeira, M. C., Mira, N. P. et al., YEASTRACT: providing a programmatic access to curated transcriptional regulatory associations in *Saccharomyces cerevisiae* through a web services interface. *Nucleic Acids Res.* 2011, *39*, D136–D140.

[15] Gama-Castro, S., Salgado, H., Santos-Zavaleta, A., Ledezma-Tejeida, D. et al., RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic Acids Res.* 2016, *44*, D133–D143.

[16] Cheng, F., Jia, P., Wang, Q., Zhao, Z., Quantitative network mapping of the human kinome interactome reveals new clues for rational kinase inhibitor discovery and individualized cancer therapy. *Oncotarget* 2014, *5*, 3697–3710.

[17] Sharifpoor, S., Nguyen Ba, A. N., Youn, J. Y., van Dyk, D. et al., A quantitative literature-curated gold standard for kinase-substrate pairs. *Genome Biol.* 2011, *12*, R39.

[18] Higgins, M. E., Claremont, M., Major, J. E., Sander, C., Lash, A. E., CancerGenes: a gene selection resource for cancer genome projects. *Nucleic Acids Res.* 2007, *35*, D721–D726.

[19] Kumar, R., Nanduri, B., HPIDB—a unified resource for host-pathogen interactions. *BMC Bioinform.* 2010, *11*(Suppl. 6), S16.

[20] Li, Q., Brass, A. L., Ng, A., Hu, Z. et al., A genome-wide genetic screen for host factors required for hepatitis C virus propagation. *Proc. Natl. Acad. Sci. USA* 2009, *106*, 16410–16415.

[21] Griffiths, S. J., Koegl, M., Boutell, C., Zenner, H. L. et al., A systematic analysis of host factors reveals a Med23-interferon-lambda regulatory axis against herpes simplex virus type 1 replication. *PLoS Pathog.* 2013, *9*, e1003514.

[22] Stegen, C., Yakova, Y., Henaff, D., Nadjar, J. et al., Analysis of virion-incorporated host proteins required for herpes simplex virus type 1 infection through a RNA interference screen. *PloS One* 2013, *8*, e53276.

[23] Brass, A. L., Dykxhoorn, D. M., Benita, Y., Yan, N. et al., Identification of host proteins required for HIV infection through a functional genomic screen. *Science* 2008, *319*, 921–926.

[24] Konig, R., Zhou, Y., Elleder, D., Diamond, T. L. et al., Global analysis of host-pathogen interactions that regulate early-stage HIV-1 replication. *Cell* 2008, *135*, 49–60.

[25] Sivan, G., Martin, S. E., Myers, T. G., Buehler, E. et al., Human genome-wide RNAi screen reveals a role for nuclear pore proteins in poxvirus morphogenesis. *Proc. Natl. Acad. Sci. USA* 2013, *110*, 3519–3524.

[26] Brass, A. L., Huang, I. C., Benita, Y., John, S. P. et al., The IFITM proteins mediate cellular resistance to influenza A H1N1 virus, West Nile virus, and dengue virus. *Cell* 2009, *139*, 1243–1254.

[27] Karlas, A., Machuy, N., Shin, Y., Pleissner, K. P. et al., Genome-wide RNAi screen identifies human host factors crucial for influenza virus replication. *Nature* 2010, *463*, 818–822.

[28] Konig, R., Stertz, S., Zhou, Y., Inoue, A. et al., Human host factors required for influenza virus replication. *Nature* 2010, *463*, 813–817.

[29] Shapira, S. D., Gat-Viks, I., Shum, B. O., Dricot, A. et al., A physical and regulatory map of host-influenza interactions reveals pathways in H1N1 infection. *Cell* 2009, *139*, 1255–1267.

[30] Queralt-Rosinach, N., Pinero, J., Bravo, A., Sanz, F., Furlong, L. I., DisGeNET-RDF: harnessing the innovative power of the Semantic Web to explore the genetic basis of diseases. *Bioinformatics* 2016, *32*, 2236–2238.

[31] Davis, A. P., Murphy, C. G., Johnson, R., Lay, J. M. et al., The comparative toxicogenomics database: update 2013. *Nucleic Acids Res.* 2013, *41*, D1104–D1114.

[32] Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., McKusick, V. A., Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* 2005, *33*, D514–D517.

[33] Welter, D., MacArthur, J., Morales, J., Burdett, T. et al., The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* 2014, *42*, D1001–D1006.

[34] Landrum, M. J., Lee, J. M., Benson, M., Brown, G. et al., ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* 2016, *44*, D862–D868.

[35] Wishart, D. S., Knox, C., Guo, A. C., Shrivastava, S. et al., DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* 2006, *34*, D668–D672.

[36] Luo, H., Lin, Y., Gao, F., Zhang, C. T., Zhang, R., DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements. *Nucleic Acids Res.* 2014, *42*, D574–D580.

[37] Chen, W. H., Minguez, P., Lercher, M. J., Bork, P., OGEE: an online gene essentiality database. *Nucleic Acids Res.* 2012, *40*, D901–D906.

[38] Baryshnikova, A., Costanzo, M., Kim, Y., Ding, H. et al., Quantitative analysis of fitness and genetic interactions in yeast on a genome scale. *Nat. Methods* 2010, *7*, 1017–1024.

[39] Cherry, J. M., Hong, E. L., Amundsen, C., Balakrishnan, R. et al., Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.* 2012, *40*, D700–D705.

[40] Pu, S., Wong, J., Turner, B., Cho, E., Wodak, S. J., Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Res.* 2009, *37*, 825–831.

[41] Ruepp, A., Waegele, B., Lechner, M., Brauner, B. et al., CORUM: the comprehensive resource of mammalian protein complexes—2009. *Nucleic Acids Res.* 2010, *38*, D497–D501.

[42] Wuchty, S., Siwo, G., Ferdig, M. T., Viral organization of human proteins. *PloS One* 2010, *5*, e11796.

[43] Yu, G., Li, F., Qin, Y., Bo, X. et al., GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* 2010, *26*, 976–978.

[44] Mariano, R., Khuri, S., Uetz, P., Wuchty, S., Local action with global impact: highly similar infection patterns of human viruses and bacteriophages. *mSystems* 2016, *1*, e00030–15.

[45] Wuchty, S., Uetz, P., Protein-protein interaction networks of *E. coli* and *S. cerevisiae* are similar. *Sci. Rep.* 2014, *4*, 7187.

[46] Rolland, T., Tasan, M., Charloteaux, B., Pevzner, S. J. et al., A proteome-scale map of the human interactome network. *Cell* 2014, *159*, 1212–1226.