Published in final edited form as:
J Mol Biol. 2019 April 05; 431(8): 1650–1670. doi:10.1016/j.jmb.2019.03.008.

The balancing act of intrinsically disordered proteins: enabling functional diversity while minimizing promiscuity

Mauricio Macossay-Castillo[1,2,*], **Giulio Marvelli**[1,2], **Mainak Guharoy**[1,2], **Aashish Jain**[3], **Daisuke Kihara**[3,4], **Peter Tompa**[1,2,5], and **Shoshana J. Wodak**[1,*]

[1]VIB-VUB Center for Structural Biology, Vlaams Instituut voor Biotechnologie, Pleinlaan 2, 1050 Brussels, Belgium

[2]Structural Biology Brussels, Department of Bioengineering Sciences, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium

[3]Department of Computer Science, Purdue University, USA

[4]Department of Biological Sciences, Purdue University, Hockmeyer Structural Biology Building, 249 S. Martin Jischke Dr West Lafayette, IN 47907, USA

[5]Institute of Enzymology, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Magyar tudosok korutja 2, 1117 Budapest, Hungary

Abstract

Intrinsically disordered proteins (IDPs) or regions (IDRs) perform diverse cellular functions, but are also prone to forming promiscuous and potentially deleterious interactions. We investigate the extent to which the properties of, and content in, IDRs have adapted to enable functional diversity while limiting interference from promiscuous interactions in the crowded cellular environment. Information on protein sequences, their predicted intrinsic disorder and 3D structure contents, is related to data on protein cellular concentrations, gene co-expression, and protein-protein interactions (PPI) in the well-studied yeast *S. cerevisiae*. Results reveal that both the protein IDR content and the frequency of 'sticky' amino acids in IDRs (those more frequently involved in protein interfaces) decrease with increasing protein cellular concentration. This implies that the IDR content and the amino acid composition of IDRs experience negative selection as the protein concentration increases. In the *S. cerevisiae* PPI network, the higher a protein's IDR content, the more frequently it interacts with IDR-containing partners, and the more functionally diverse the partners are. Employing a clustering analysis of Gene Ontology (GO) terms we newly identify ~600 putative multifunctional proteins in *S. cerevisiae*. Strikingly, these proteins are enriched in IDRs and contribute significantly to all the observed trends. In particular, IDRs of multi-functional proteins feature more sticky amino acids than IDRs of their non-multifunctional counterparts, or

[*]Corresponding authors: Mauricio Macossay-Castillo, mauricio.macossay.castillo@vub.be, Shoshana J. Wodak, Shoshana.wodak@gmail.com.

Competing Interests
All the authors declare to have no competing interests.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

the surfaces of structured yeast proteins. This property likely affords sufficient binding affinity for the functional interactions, commonly mediated by short IDR segments, thereby counterbalancing the loss in overall IDR conformational entropy upon binding.

Introduction

To carry out their function, proteins tend to associate with other macromolecules (other proteins, nucleic acids, etc.) [1] as well as with small molecule ligands. These functional associations take place in the crowded cellular environment, where they have to compete with promiscuous non-functional binding events that may be detrimental to fitness because they sequester interaction partners [2,3].

There is mounting evidence that to mitigate the interference from non-functional interactions, nature acts at several levels. These include, adapting the amino acid sequence to modulate protein solubility and intrinsic stability [4,5], temporal, condition-dependent regulation of gene expression [6,7] or protein abundance and half-life [8]. Other means of reducing such interference involve post translational modifications (PTMs) [9,10], binding to various protein adaptors [11], or the sequestration of proteins within specialized cellular compartments [7,12].

Of particular interest in this context are intrinsically disordered proteins (IDPs) and proteins with intrinsically disordered regions (IDRs). IDPs/IDRs and their properties have been extensively reviewed [13,14]. These are polypeptides, or segments thereof, which are essentially devoid of stable secondary or tertiary structures, when in isolation. These segments are described as ensembles of conformations that interconvert on a range of timescales [15,16]. Nonetheless, a significant fraction of IDPs and IDRs have some regions with residual transient secondary structures. These regions often mediate function through association with other proteins (mostly to structured domains in these proteins), and can undergo disorder to order transitions as part of the recognition process [17].

Functional associations of IDPs/IDRs tend to be mediated by peptide recognitions motifs, the so-called short linear motifs (SLiMS), which are relatively well conserved in evolution [15,18,19]. While interactions mediated by individual motifs tend to be quite weak, interactions with several motifs of the same IDR may act cooperatively, thereby increasing the range of observed affinities [20,21], and enabling interactions with multiple partner proteins [22–25]. These properties empower IDPs/IDRs to carry out a diverse range of important functions, more particularly in regulatory [26,27] and signaling processes [15,28]. Intrinsic disorder has also been suggested to play a role in promoting the assembly of protein complexes [29].

The same properties, however, may also prompt disordered proteins to form promiscuous interactions with other proteins, especially at higher protein concentrations, causing for example, deleterious dosage sensitivity in yeast [2]. It has also been observed that proteins with a high degree of intrinsic disorder are often hubs in protein-protein interaction (PPI) networks of human and other model organisms [22,24,25,30,31], with hubs being defined as the subset of protein nodes that form interactions with many other proteins, typically more than ten. While many of these interactions may be functional in nature [32], the fact that they may also include non-functional interactions needs to be considered [33].

The cellular abundance of a protein is a crucial parameter governing its association with other cellular components, since this association depends on the standard free energy of the binding reaction and on the concentrations of the interacting components [34]. Previous analyses have shown that the abundance and residence time of IDPs in the cells of yeast *S. pombe* and human were tightly regulated by fine-tuning the rates of translation, protein degradation and transcript clearance [35]. Another mechanism of regulating availability of IDPs was suggested to involve synthesizing them at the site where they are required, by localization of the corresponding mRNA [36].

The main question addressed in the present study is if and to what extent the chemistry and other properties of IDPs and IDRs have adapted to mitigate their risk of engaging in promiscuous interactions, while enabling their diverse functional roles. To address this question, we focus on proteins of the yeast *S. cerevisiae*, one of the most extensively studied organisms in terms of protein function, proteome-scale PPI data [33,37], gene expression [38], and protein abundance [39]. We relate information on protein sequences, their predicted intrinsic disorder content and 3D structures to cellular protein abundance, and associate these relationships with protein function. In addition, we analyze patterns of the physical interactions of these proteins in a high-confidence (HC) PPI network of yeast soluble proteins [40]. These patterns are correlated to the intrinsic disorder content and abundance level of the corresponding proteins, as well as to the functional similarity and gene co-expression of their interacting partners. Importantly, we complement the analysis by identifying multi-functional proteins on the basis of their Gene Ontology (GO) annotations [41], and evaluate the relative contribution of protein multi-functionality versus its degree of intrinsic disorder to the observed trends.

Results of our analysis yield new insights into how the IDR content of proteins and the amino acid composition of IDRs have adapted to enable IDR-containing proteins to perform diverse functional roles, while reducing interference from promiscuous interactions in the crowded cellular environment.

Results and Discussion

Intrinsic disorder content in relation to protein abundance and functional annotations Figure 1 plots the fraction of proteins encoded by the *S. cerevisiae* genome containing a high degree of intrinsic disorder as a function of their abundance in the cell. Proteins with a high degree of disorder are defined as those with ≥ 30% of their residues predicted to be in IDRs of ≥20 consecutive residues along the chain. Disordered residues were predicted using the IUPred software [42,43], but essentially the same results were obtained using other disorder prediction methods (see Materials and Methods). Data on protein abundance were obtained from the PaxDb database [39]. These data are expressed in parts per million (ppm), a quantity linearly related to the protein copy number in cells [44].

The plot of Figure 1 shows a clear trend for the fraction of proteins displaying high intrinsic disorder to decrease with increasing protein abundance. This trend translates into a strong inverse correlation between the fraction of highly disordered proteins and their abundance (r_S = −0.76, with r_S being the Spearman's rank correlation coefficient), which is however weakly significant (*p* = 0.02), likely due to the dip in the fraction of highly disordered proteins in the lowest abundance range. Indeed, the correlation becomes stronger and highly statistically significant (r_S = −0.94 *p* = 2e-16) when ignoring proteins with abundance below 8 ppm. The latter tend to include membrane proteins, for which cytoplasmic abundance is underestimated [12], and various cellular proteins that are not reliably detected by current proteomics methods [45,46] (see Materials and Methods for details). A highly statistically significant but lower correlation coefficient is obtained (r_S = −0.11, *p* = 4e-15) when taking into account the fraction of all the residues in a protein that are predicted to be intrinsically disordered, over the entire range of protein abundances. Taken together, these results indicate that IDRs, especially IDRs ≥ 20 residues, are selected against as protein abundance increases.

Both the abundance and the degree of putative intrinsic disorder of a protein were shown to correlate with protein function to some extent [47]. To re-examine these correlations in our dataset, we computed the enrichment in the Biological process (BP) Gene Ontology (GO) terms in *S. cerevisiae* proteins in the highest and lowest one third of the abundance range, respectively. This was repeated for subsets in each of these two categories, comprising proteins of high- and low disorder, featuring ≥30% and <10% of their residues in IDRs of ≥20 residues, respectively, with results displayed in Figure 2.

The first two columns of the heatmap in Figure 2 clearly indicate that most BP terms related to cellular metabolism, ribosome biogenesis, translation and protein folding are overrepresented in high-abundance proteins, and are underrepresented in proteins of low abundance. This trend is reversed, with an overrepresentation in low-abundance proteins, and an under-representation in highly abundant ones, of processes related to the cell cycle, chromosome segregation, transcription, cellular component morphogenesis, and signal transduction.

Interestingly, these same trends are significantly strengthened in the subsets of proteins with the highest and lowest disorder content in each of the 2 abundance categories (last two heatmap columns in Figure 2). For example, the overrepresentation of metabolic processes, ribosome biogenesis, and translation and protein folding, is more pronounced in high-abundance proteins with low IDR content, than in all high-abundance proteins. The same applies to the underrepresentation of the same processes in low-abundance proteins with high IDR content. Likewise, the highlighted reverse trend for processes such as chromosome segregation, transcription and cell cycle processes is also significantly more pronounced when IDR content is considered.

These various trends are in good agreement with those reported in a previous study examining the relation between proteins with high IDR content in UniProt [48] and their biological process annotations [49]. The new insight provided by our analysis is that the relation of IDR content to protein function also depends on protein abundance, an aspect likely reflecting a more complex level of regulation and adaptation in evolution.

Residue properties of IDRs in relation to protein abundance

In the following, we investigate the relationships of various sequence properties of the IDRs with the abundance levels of the corresponding proteins. The main properties examined are the average stickiness of IDR residues, with stickiness representing an interaction propensity scale derived from structural data (see reference [4] and Figure 3d), and the average aggregation propensity of residues in IDR regions, as measured by different aggregation propensity scales (see Materials and Methods for details). These properties were computed only for IDRs of ≥20 consecutive residues along the polypeptide.

The most statistically significant correlation is found between the average stickiness value of residues in IDRs and protein abundance. From the scatter plot of Figure 3a we see that the average stickiness of IDRs decreases with protein abundance ($r_S = -0.31$ $p = 6e\text{-}59$). Interestingly, this trend is very similar to that reported previously for the average stickiness values of surface (solvent-accessible) residues of globular proteins in *S. cerevisiae* and other model organisms [4].

Retrieving the atomic coordinates of globular soluble yeast proteins, totaling 452 structures, from a recent version of the PDB (October 2017), and computing the average stickiness values of their surface residues, we confirm the previous reported [4] negative correlation between these values and protein abundance ($r_S = -0.26$, $p = 2e\text{-}8$) (Figure 3b). Using our more recent larger dataset of yeast protein structures, we also reproduce the reported negative correlation [4] between the average stickiness of residues in the protein interior and protein abundance ($r_S = -0.22$, $p = 2e\text{-}6$) (Supplementary Figure S1). But unlike in reference [4], we observe a weak positive correlation between protein size and, respectively, the stickiness of both interior and surface residues (Supplementary Figure S2a,b). The latter positive correlation may reflect the greater tolerance of larger proteins for surface residues to evolve towards optimizing interactions with binding partners at the cost of compromising intrinsic stability, owing to the smaller surface to volume ratio of larger proteins [50].

Finally, we note that the average residue hydrophobicity of IDRs, measured using the Kyte and Doollitile scale [51] (Supplementary Figure S3a), was also negatively correlated with protein abundance, albeit to a lesser degree ($r_S = -0.12$, $p = 4e\text{-}10$) than the average stickiness.

Following reference [4] we interpret the decrease of average IDR stickiness with protein abundance as reflecting the fact that proteins tend to limit promiscuous interactions by reducing the stickiness of residues in readily accessible, solvent exposed regions: the IDRs of at least 20 residues, analyzed here, and the surface residues of structured proteins and domains.

Somewhat unexpectedly, however, in proteins of all abundance levels, and particularly in less abundant proteins, the average stickiness of residues in IDRs is shifted towards higher values (more sticky) relative to surface residues of globular proteins ($p = 1e\text{-}10$ to $2e\text{-}4$), as witnessed by the distributions in Figure 3c. To gain further insight into the origins of this shift, we examine the amino acid propensities in the two types of regions respectively (Figure 4a), using the full yeast proteome as reference. We find that both IDRs and surface residues of globular proteins are depleted in sticky amino acids, especially in the aromatics W, Y, and F, and in C, and somewhat enriched in the least sticky charged amino acids (Figure 3d). Interestingly, the depletion in the four sticky amino acids is more pronounced in IDRs, whereas surface residues are enriched in the least sticky charged amino acids (E, D, K) (Figure 4a). To understand how these trends contribute to the higher stickiness of IDRs relative to surface residues, one must take into account amino acid frequencies in the proteome (Figure 4b). We see, indeed, that the more sticky amino acids (W, Y, F, C) are the least abundant amino acids overall, whereas the less sticky charged amino acids (E, D, K, R) are among the most abundant amino acids. Thus, the higher average stickiness of IDRs relative to surface residues is the consequence of the protein surface being enriched in less sticky and highly abundant charged amino acids. For the full list of amino acid frequencies and propensities see Supplementary Table S2.

This higher stickiness of IDRs is interesting. It may reflect the need to afford a minimum binding affinity for the short linear motifs (SLiMS) of IDRs that commonly engage in interactions with binding partners. Such minimum binding affinity would partially compensate for the overall loss in conformational entropy upon binding, and may therefore relate to the more multi-functional nature of these proteins in comparison to their structured counterparts, an aspect examined further along this study. Also noteworthy is the enrichment in the more abundant S and T amino acids in IDRs relative to surface residues (Figure 4a). These very marginally sticky residues commonly undergo phosphorylation. Their enrichment in IDRs is hence consistent with the frequent role of IDRs in signaling processes [15,28].

Lastly, we find that there is virtually no overlap between putative IDRs and aggregation prone segments identified using two different computational methods: TANGO, an algorithm for predicting aggregation prone regions in unfolded polypeptide chains [52] and PASTA2.0, a predictor of residues likely to be part of amyloid fibrils [53]. Less than 1% of the residues in IDRs overlapped with aggregation prone residues predicted by both methods, indicating that unlike globular proteins [52–54], IDRs are essentially devoid of such segments. Recent findings on intracellular liquid-like protein compartments, which often lead to the formation of fibrillar aggregates involving IDRs, suggest indeed that these aggregation phenomena are distinct from those occurring in the unfolded state of structured proteins [55].

Relating protein properties to the pattern of their physical and functional interaction in the cell

Having gathered evidence that structural disorder has evolved properties that promote its capacity to form functional interactions with other proteins, we proceed to investigate how several of the protein properties analyzed above, relate to the patterns of physical and functional interactions proteins form in the cell. These patterns are investigated in the high confidence (HC) PPI network of *S. cerevisiae* soluble proteins, built from data obtained by two different groups using affinity purification and mass spectrometry (AP-MS) techniques [40]. This network, thereafter denoted as the Collins network, contains 1622 proteins, representing only a fraction (~25%) of all yeast proteins. These proteins engage in 9070 reliably detected 'interactions' (see Materials and Methods, and [40] for details). The detected 'interactions' represent in fact co-complex associations, of which only a fraction corresponds to direct physical contacts [33].

For each protein in the network, we retrieved their abundance values and several other properties of interest. The functional implications of the interactions was obtained by associating each protein node to its functional annotations, retrieved from the Gene Ontology (GO) resource [56,57], as well as its mRNA expression profiles measured in a set of different conditions (COXPRESdb [38]). This information was further analyzed to quantify the degree of functional diversity among the interacting proteins.

Relating protein abundance and intrinsic disorder to the number of interaction partners—The number of interactions formed by a protein in a PPI network depends on several factors. When present at higher concentrations, a protein may engage in promiscuous interactions with multiple partners [58,59] and is generally more readily detected by experimental methods [33,37]. Interactions with multiple partners may also be a distinctive property of proteins that carry out multiple functions. It is therefore not surprising to find that the number of interaction partners (node degree) of a protein in the Collins yeast network, is significantly positively correlated with its abundance ($r_S = 0.18$, $p = 5.2\text{e-}13$) (Figure 5a). As a result, hubs in the network, defined as proteins (nodes) connected to at least 10 other proteins, are enriched in abundant proteins (Figure 5b).

Next, we investigate the relation between the number of interaction partners of a protein and its intrinsic disorder level, with the latter defined as described above. Depicting the histograms of predicted protein disorder and the corresponding node degree distributions side by side (Figure 6a), there appears to be no correlation between the two parameters ($r_S = -0.02$, $p = 0.44$). This lack of correlation confirms earlier findings by Schnell et al. [60], but disagrees with a several other studies according to which network hubs tend to be more intrinsically disordered than non-hub proteins [22,24,25,61].

Being fully aware that PPI networks of the same organism detected by different experimental methods or derived from different databases may differ substantially (for review see [33,37,62–64]), we hypothesized that differences between the yeast PPI networks used in different studies are the main reason for these contradictory observations. To test this hypothesis, we computed histograms displaying the fractions of proteins featuring increasing levels of disorder in hubs and 'end' proteins respectively. This was done for proteins in the PPI network used here, and for those in a recent version (March 2018) of the multi-validated (MV) PPI network of yeast, downloaded from BioGRID (interactions supported by at least 2 publications) [65]. We considered this literature-curated network, hereafter denoted as BioGRID, because earlier versions of this network were used in two of the above mentioned earlier studies [22,32].

Our results show striking differences of the corresponding histograms (Figure 6b,c). In the Collins network, hub proteins tend to be less disordered (albeit not significantly) than ‘end’ proteins (average $p \geq 0.05$). On the other hand, hubs in the BioGRID network are significantly more disordered than ‘end’ proteins (maximum $p = 1e\text{-}5$), in excellent agreement with previous observations (Figure S1 of reference [22]). These opposing trends cannot be attributed to a bias in protein disorder content of the respective networks, since both networks are similarly enriched in IDP/IDRs relative to the proteome as a whole (Supplementary Figure S4).

Reproducing the same trends as in reference [22] with an up-to-date version of the BioGRID network, therefore, suggests that other differences between the BioGRID and Collins networks may be at play. In the Materials and Methods section, we present evidence that the BioGRID and Collins networks differ substantially in a number of important aspects. We show, in particular, that the Collins network is significantly less noisy than the BioGRID network, despite the fact that the latter includes PPI data from multiple detection methods, in agreement with previous finding [33], thereby justifying the use of the Collins network throughout our analysis.

Relating protein intrinsic disorder to the properties of its interaction partners—Having demonstrated that the level of intrinsic disorder of a protein is not significantly correlated with the number of its interaction partners in the PPI network, we now examine the relation of protein disorder with various properties of its partners. In particular, we considered the disorder level of the interacting partners, their gene co-expression relationship and their functional similarity based on GO terms. The latter two properties were evaluated for protein pairs, and averaged over all pairs of the direct interaction partners of a given protein node (see Materials and Methods for details).

Our analysis reveals that more disordered protein nodes tend to interact with more disordered partners ($r_S = 0.31$ $p = 2e\text{-}37$) (Figure 7b,c). On the other hand, statistically significant negative correlations are detected between the disorder level of the protein node and both the average pairwise semantic similarity of the Biological Process GO annotations ($r_S = -0.13$ $p = 4e\text{-}4$), and the average pairwise correlation coefficient of the gene expression profiles ($r_S = -0.15$, $p = 3e\text{-}7$) of its partners (Figure 7b,d,e).

To further scrutinize the latter findings, we examined properties of hub proteins and their interaction partners. Hubs interact with a larger number of partners, yielding more robust statistics for trends in various partner properties. We segregated hubs into two groups: hubs with weakly co-expressed partners (average pairwise Pearson correlation coefficient $<|0.5|$), and hubs with highly co-expressed partners (average pairwise Pearson correlation coefficient ≥ 0.5) (see Materials and methods). In line with the correlations described above, we find that, on average, hubs with weakly co-expressed partners contain nearly twice as many disordered residues than those interacting with highly co-expressed partners (19% versus 11%, $p = 5e\text{-}6$) (Figure 8a). The weakly co-expressed partners of the first category of hubs are also more disordered than the highly co-expressed partners of the second category (26% of disordered residues versus 17%, $p = 4e\text{-}41$)) (Figure 8b). Interestingly, the same weakly and highly co-expressed partners of the corresponding two hub categories display, on average, significantly different abundance levels (*p* = 3e-35), and half-lives (*p* = 6e-5) (Figure 8c,d). Average values for all the parameter distributions are listed in the Table of Figure 8. These various trends hold when segregating hubs into those with low-to-medium, and high abundance levels, respectively (Supplementary Figure S5).

In summary, the picture emerging from these observations is that more disordered proteins tend to interact with more disordered partners. The interaction partners of these proteins are also more weakly co-expressed and functionally diverse, as well as less abundant and shorter lived. These trends are consistent with previous findings, where more disordered network hubs were suggested to play distinct roles in diverse signaling cascades [32] (Supplementary Figure S6). In the next section, we explore an alternative explanation that takes the multifunctional character of the protein nodes into account.

Contribution of multifunctional proteins to the observed trends

While the low functional similarity and weak co-expression of interaction partners are often linked to noise (promiscuous interactions) in the corresponding PPI data [33,37,66–68], they may also reflect a genuine functional behavior, namely of proteins performing multiple cellular functions. How to define such proteins remains however, a subject of much debate. Some authors focus on proteins performing unrelated molecular functions (moonlighting) [69–72], others integrate information from functional (molecular and cellular) annotations, and/or protein interaction data [41,69]. These different definitions and approaches often yield small and poorly overlapping sets of multifunctional proteins for the same organism (for data on yeast, for example, see Supplementary Section I).

To evaluate the contribution of protein multifunctionality to the observed trends, we identified putative multifunctional proteins in the *S. cerevisiae* proteome. To this end, the GO functional annotations of each yeast protein (including multidomain proteins) were clustered using a metric reflecting the similarity between these annotations [41]. A protein was classified as multi-functional if it is annotated with GO terms belonging to at least 2 clusters (see Materials and Methods, and Supplementary Figures S7 and S8 for details). The analysis was performed for both the GO Biological Process (BP) and Molecular Function (MF) terms.

By applying this procedure to the full *S. cervecisiae* proteome (6437 proteins), we identified a total of 595 putative multifunctional proteins (MFPs) on the basis of the GO BP terms, and 423 MFPs on the basis of the GO MF terms. These two sets are essentially distinct as only 74 proteins are common to both. For the purpose of the present analysis, we considered mainly the MFPs based on GO BP terms, as we were primarily interested in multiple functions of a protein, most likely to affect the number and types of interactions it forms in the cell.

Results show that the identified putative MFPs are enriched in structural disorder (proteins with >30% of their residues in IDRs ≥20 residues). As shown in Table 1, the full set of 595 MF proteins is 44% enriched in IDRs, relative to the entire proteome. The subset of MFPs that is part of the PPI network is 37% enriched in IDRs relative to all proteins in the network, whereas MFP hubs are most highly enriched in IDRs (65%), relative to non-MFP hubs in the network. This enrichment is all the more significant considering that MFPs tend to be slightly more abundant than the entire yeast proteome (Supplementary Figure S9a), a trend that should in principle lower the proportion of IDRs (Figure 1). The observed enrichment therefore agrees with the view that protein intrinsic disorder may promote more diverse functional roles in the cell [72].

Furthermore, we find that the IDRs of the predicted MFPs display on average higher stickiness values than their non-functional complement, when considered in the context of both the whole yeast proteome and the interactome (Table 2). This trend is consistent with the hypothesis presented above that higher stickiness of IDRs relative to surface residues of globular proteins may be linked to the thermodynamics of the IDR mediated functional interactions.

Next we investigate the potential influence of intrinsic disorder and multifunctionality on the properties of their interaction partners. To tease out these influences, we analyzed various properties of the interaction partners of MFPs and their non-MFPs counterpart with the same extreme intrinsic disorder levels: respectively, those with ≥30% and those with <10% of their residues in disordered regions.

The results summarized in Figure 9 (and the Table within) show that among highly disordered proteins, the subset of MFPs interacts with more disordered partners on average (31% versus 25% disordered residues in partners of MFPs versus non-MFPs, $p = 4e\text{-}3$). These partners also tend to display a lower level of functional diversity (average semantic similarity of 0.40 for MFPs versus 0.52 for non-MFPs, $p = 4e\text{-}3$), and weaker gene co-expression (0.33 for MFPs versus 0.42 for non-MFPs, $p = 1e\text{-}3$) than their equally disordered non-MFP counterparts (Figure 9a-c). On the other hand, among the set of most highly structured (least disordered) proteins, the interaction partners of MFPs display on average a similar level of disorder and functional diversity as non-MFPs, while nonetheless maintaining a similar lower level of gene-co-expression as in the highly disordered set (Supplementary Figure S10). Importantly, consistent with higher intrinsic disorder in MFPs, 20% of the proteins in the highly disordered set are predicted to be multifunctional, whereas this fraction is only 13% for the more structured set.

Lastly we note that somewhat surprisingly, MFPs interact on average with fewer partners in the yeast PPI network than their non-MFP counterparts (Supplementary Figure S9b). Considering that MFPs represent a small fraction of the proteins in the network, this observation may reflect the general bias in the PPI data towards more abundant proteins, which are more likely to form promiscuous interactions [33,37].

Thus, taken together, these observations indicate that multifunctionality contributes significantly to the elevated stickiness of IDRs in comparison to the surfaces of globular proteins. Multifunctionality also increases significantly the propensity of disordered proteins to interact with more disordered and functionally diverse partners.

Concluding remarks

In this study, we evaluated how intrinsic disorder of proteins and the properties of disordered regions have adapted to mitigate the risk of engaging in non-functional interactions, while at the same time enabling the corresponding proteins to perform diverse cellular functions. To this end, we used information on the protein sequences, 3D structures, as well as genome-scale data on cellular protein concentrations, gene coexpression, and protein-protein interactions in one of the most thoroughly studied model organism, the yeast *S. cerevisiae.*

We found that the fraction of highly disordered proteins decreases significantly with increasing protein abundance, in agreement with previous findings [35]. This trend suggests that protein intrinsic disorder is selected against as protein abundance increases, in line with the observation that disorder is associated with deleterious dosage sensitivity in yeast [2].

A significant new finding is that IDRs of highly abundant proteins have adapted their amino acid composition in order to minimize promiscuous interactions with other proteins, in a way similar to surface residues of globular proteins [4]. This adaptation has lowered the frequency of ‘sticky’ amino acids found more frequently in protein-protein interfaces, and raised the frequency of polar and charged amino acids, which tend to remain solvated even for proteins present at high concentrations in the cell. We find however, that IDPs/IDRs of ≥20 contiguous disordered residues maintain a somewhat higher average stickiness than surface residues at all protein abundance levels. This is explained by the lower frequency of less sticky polar and charged amino acids in IDRs than on the surface of globular proteins.

We propose that the trend towards higher stickiness of IDRs likely reflects the need to provide sufficient binding affinity for the SLiMs-mediated functional interactions to counterbalance the large loss of conformational entropy that generally accompanies these interactions. A good example is the binding of the trans-activator domain (TAD) of p53 to the nuclear coactivator-binding domain (NCBD) of CREB-binding protein (CBP), largely driven by intermolecular hydrophobic contacts involving 2 short segments of the disordered portion of p53 [73].

Investigating the pattern of interactions made by IDPs/IDRs in the context of the high confidence *S. cerevisiae* PPI network yielded further important insight into the balancing act of protein disorder, enabling it to engage in diverse functional interactions while minimizing promiscuity. Overall, proteins with a higher level of intrinsic disorder were found to interact more frequently with more highly disordered partners. These partners also tend to be functionally more diverse, as evident by the reduced similarity of their GO functional annotations and the weaker correlation of their gene expression profiles.

Most importantly still, our study makes significant headway towards assessing the respective contributions of functional versus promiscuous interactions to the observed trends. GO functional annotations were used to identify ~600 putative MFPs in *S. cerevisiae*, and these proteins were mapped to and analyzed in the context of the yeast PPI network. Interestingly, we found that multifunctional proteins, which represent ~20% of the highly disordered proteins in the PPI network (Table 1), make a significant contribution to the trends described above. They are enriched in IDRs (by 44–65%, Table 1), their interaction partners tend to be more disordered (*p* = 4e-3) and the functional similarity of these partners is significantly lower (*p* = 4e-3 for their partners GO-BP semantic similarity annotations; *p* = 1e-3 for their partners gene co-expression) (Figure 9 a-c), in comparison to their non-MFPs counterpart.

Moreover, the IDRs of MFPs display higher stickiness values than IDRs of their non-multifunctional complement (Table 2), lending support to the idea that the somewhat higher average stickiness of IDRs reflects the thermodynamic requirements of forming specific interactions with cognate proteins, mediated by SLiMs. Indeed, our yeast MFPs do feature, on average, a somewhat higher coverage of SLiMs predicted by the ANCHOR algorithm [74] than non-MFPs (between 32 and 34% for MFPs in the whole proteome and Collins interactome, and 28% for the non-MFPs in both datasets respectively, see Supplementary Figure S11).

Nevertheless, it is reasonable to assume that a fraction of the interactions formed by both structured proteins and IDPs/IDRs is likely non-functional under certain conditions, but serve as a reservoir for future adaptation to altered conditions and environmental contexts [3,75]. Evidence has, indeed, been accumulating that IDPs/IDRs are frequently involved in intracellular liquid-like compartments, where they form both functional and spurious interactions [55,76]. More data on such cellular assemblies, and on the functional interactions involving IDPs/IDRs in specific systems [77] are clearly needed to better evaluate the functional contexts of the trends uncovered in our analysis.

It also remains to be seen if the trends observed here for yeast proteins are shared by other organisms. The study of Levy et al. [4] showed that the stickiness of surface residues in structured human proteins is more weakly anti-correlated to protein abundance than in yeast. This may be attributed to the fact that protein abundance is ill defined in human cells, due to the variety of cell types, and to a higher degree of compartmentalization of proteins in human cells, which relieves some of the pressure to decrease their propensity to engage in promiscuous associations. This may also be the case for IDPs/IDRs, as the latter can be translated into a spatially localized fashion within mammalian cells [36].

Materials and Methods

Datasets

Data on protein sequences abundance levels and half-life—Data on *S. cerevisiae* protein sequences were taken as provided by the Protein Abundance Database (PaxDb) version 4.0 (March 2016) [39].

Data on protein abundance levels in yeast *S. cerevisiae*, expressed as parts per million (ppm), a quantity linearly related to the protein copy number in the cell [44], were obtained from PaxDb). We used the integrated organism-level abundance data provided by PaxDB for a total of 6437 yeast proteins. From these data, 1457 proteins with very low abundance levels (<2ppm) were discarded from our analysis, as they were found to have higher variability among individual datasets deposited in PaxDB, and therefore considered as less reliable (see Supplementary Figure S12).

Data on the steady-state half-life of yeast proteins expressed in minutes were obtained from [78].

Gene co-expression data—Data on the Pearson correlation coefficient of mRNA expression profiles of *S. cerevisiae* gene pairs, taken to represent the co-expression levels of the corresponding gene, were extracted from the COXPRESdb [38]. The mRNA profiles were computed from mRNA expression data (microarrays) measured under a large set of different conditions and retrieved from the ArrayExpress database <https://www.ebi.ac.uk/arrayexpress>. The expression data were normalized using the robust multi-array average (RMA) method [79].

Protein-protein interaction data—Data on genome-wide protein-protein interactions (PPI) in the yeast *S. cerevisiae* were taken from the study of Collins et al. [40]. This study derived a high confidence (HC) PPI network involving soluble yeast proteins, starting from the raw data obtained by affinity purification and mass spectrometry analysis (AP-MS) in the two independent high throughput studies by Krogan et al.[66] and Gavin et al.[80].

The HC ‘Collins’ yeast PPI network was derived using the Protein Enrichment (PE) statistical scoring scheme to filter out spurious interactions. Applying a PE threshold of ≥3.19 [40], the ‘Collins’ PPI network contains 9070 high confidence co-complex associations, commonly denoted as ‘interactions’ among 1622 distinct proteins (1600 protein with abundance levels > 2ppm). The estimated TP/FP ratio for this dataset is ~30, where TP (true positive) are detected interactions from the positive examples, and FP (false positive) are detected interactions from the negative examples. The PPI dataset with the associated PE scores was downloaded from <http://interactome-cmp.ucsf.edu/>.

In addition, we examined *S. cerevisiae* PPI data retrieved from the March 2018 release of the BioGRID database [65]. These data include protein-protein interactions annotated from the scientific literature, detected by both high- and low- throughput methods of variable stringencies and includes both co-complex and binary interactions. We used the multi-validated (MV) dataset of BioGRID. In this dataset, the reliability criteria for a given interactions is based on the number of experimental systems and publications where the physical interaction between the proteins was detected.

Data on structures of yeast globular proteins—A dataset of yeast globular proteins was assembled from the protein data bank (RCSB-PDB) [81] (Sep. 2017). We selected entries comprising x-ray structures determined at 3Å resolution or better, which contain only proteins (no DNA or RNA) from *S. cerevisiae*, displaying ≤ 90% sequence identity. For these structures, we downloaded both the atomic coordinates and biological assembly assignments. Structures representing membrane proteins (annotated intra- or trans-membrane regions in UniProt [48] or predicted by TOPCONS [82]), those lacking UniProt [48] identifiers or protein abundance information in PaxDb [39], or identified as probable errors in QSbio [83] (low and very-low confidence), were discarded. This set was further reduced to a total of 452 structures by retaining entries containing only a single protein chain (e.g. hetero-complexes were not considered). All subsequent calculations were performed on the biologically meaningful assembly state of the protein (Biological Unit).

Gene functional annotations, enrichment analysis, and functional similarity

Functional annotations for yeast proteins were downloaded from the Gene Ontology [56] (GO) (July 2016). Unless otherwise specified, only GO terms based on evidence codes IDA (Inferred from Direct Assay) and IPI (Inferred from Physical Interaction) from respectively, the Biological Process and Molecular Function ontologies were considered for the present analysis. The GO enrichment analysis was performed using the BiNGO [84] plugin for Cytoscape [85].

The extent to which two interacting proteins carry out similar functions was evaluated using the *semantic similarity measure*. The specific measure was that of Wang et al. [86] implemented in the R-package GOSemSim [87]. This measure evaluates the similarity of two GO terms based on both the locations of these terms in the GO graph and their relations to their ancestor terms.

Comparing the Collins and BioGRID PPI networks

To justify the use of the Collins PPI network for the bulk of our analysis we compared various properties of interacting protein pairs in respectively, the MV BioGRID and Collins yeast PPI networks, with those of a random network having the same node degree distribution as the Collins network (generated as described below).

The evaluated properties are derived from independent data types routinely used to gauge the quality and biological relevance of experimentally derived PPI networks. For individual interacting protein pairs in each network we computed the GO BP (Biological Process) semantic similarity score [86], the gene co-expression Pearson correlation coefficient, and the ratio of the protein abundance levels [38,39]. These values were then averaged over all the interacting pairs in each of the networks.

First we note that while 42% of the proteins of the larger MV BioGRID network are shared with the HC Collins network, only 14% of the PPI in MV BioGRID are also part of the Collins network (Supplementary Figure S13a,b). As a result the proteins defined as ‘hubs’ (protein nodes connected to >10 interaction partners in the network) and those defined as ‘ends’ (proteins connected to only 1 partner) differ substantially (Supplementary Figure S13c,d).

Next our analysis shows (Figure 10a) that the distribution of the Pearson correlation coefficient of the mRNA expression profiles of interacting pairs for both the HC Collins and MV BioGRID networks, are significantly different from the distribution for the random network. On the other hand, the distribution for the HC Collins network is bimodal, with a significant fraction of interacting pairs (~57%, 4387 out of 7636 pairs with a coexpression correlation value) displaying highly correlated expression profiles (≥0.5), shifting it further towards higher values relative to the random distribution than the distribution computed for the MV BioGRID network. We find furthermore, that the HC Collins network features a larger fraction of functionally similar interacting pairs (SemSim values ≥0.6), and a smaller fraction of functionally diverse pairs (≤0.4) than the MV BioGRID network, and that the SemSim distributions of both networks depart significantly from random (Figure 10b). In addition, the distributions of the pairwise protein abundance ratio and the half-life ratio of interacting protein pairs are more significantly shifted towards lower values in the HC Collins network than for the MV BioGRID network, although the distributions for both networks differ significantly from random (Figure 10c,d).

On the basis of this multipronged comparison, we conclude that the HC Collins network is of higher quality than the MV BioGRID network, confirming earlier findings that the HC Collins network is one of the highest quality interaction networks of the soluble *S. cerevisiae* proteome and therefore is the network of choice for the present analysis. Many of its interactions were indeed confirmed by subsequent publications, allowing the generation of CYC2008, a widely used updated version of annotated yeast protein complexes[88].

Generation of random networks

Ten different random networks were created, which preserve the number of protein nodes and the node degree distribution of the Collins PPI network [40]. Starting from the Collins network, binary interactions (network edges) were randomly shuffled using the *sample_degseq* function in the *igraph* package for R <http://www.igraph.org/r>, avoiding repeating interactions and self-loops. Next, the identities of the protein nodes were randomly shuffled using the *sample* function without replacement in the *base* package of R <http://www.r-project.org>. The resulting networks were pruned to eliminate the few interactions that were also in the Collins network, yielding essentially identical random networks shown in the Supplementary Figure S14.

Residue properties and amino acid propensities

The *stickiness* scale of Levy and Teichmann [4] (values were taken from the Supplementary Material of ref [4]), was used to evaluate the propensity of a protein residue to be part of an interaction interface with another protein. In addition, we used the Kyte-Doolittle amino acid hydrophobicity scale [51]. Surface residues were defined as those with a relative solvent accessible surface area (RSASA) of at least 20, with $RSASA=SASA_{protein} / SASA_{free}$, where $SASA_{protein}$ is the accessible surface area of the residue in the protein structure, and $SASA_{free}$ is the accessible surface area of the corresponding amino acid free in solution. The accessible surface area values were computed using FreeSASA [89].

The amino acid composition of intrinsically disordered regions (IDRs) and surface residues of globular proteins, respectively, was evaluated using a propensity score. The latter was derived by calculating the fractional occurrence (f) of each individual amino acid with respect to the total number of residues in each dataset (IDRs or surface residues in the dataset of globular proteins). Amino acid propensities (p) were then calculated using $p = \log_{10} (f_A / f_B)$, where f_A is the fractional contribution of an amino acid for a given residue subset (IDRs or protein surfaces) and f_B is its average fraction computed for the full yeast proteome.

Prediction of intrinsically disordered and aggregation-prone residues

Residues likely to be part of IDRs were predicted from the amino acid sequence by the IUPred software [43], using the long-window option. To decrease the level of noise, we considered only segments of at least 20 consecutive residues predicted to be disordered.

Additionally, we analyzed predicted IDRs from the meta-predictor MobiDB-lite [90] and observed the same trends as with IDRs inferred by IUPred. These are: 1) the negative correlation between the fraction of highly disordered proteins and protein abundance levels gets stronger when filtering out proteins with an abundance of <2 and <8 ppm ($r_S = -0.26$ $p = 0.48$ and $r_S = -0.84$ $p = 2e\text{-}3$, respectively), 2) At the protein abundance cutoff of 2ppm, there is a negative correlation between protein abundance and IDR stickiness ($r_S = -0.29$ $p = 1e\text{-}47$) and 3) and a significant, but weak negative correlation between protein abundance and IDR hydrophobicity ($r_S = -0.07$ p = 5e-4). Moreover, we observed a very high correlation between the amino acid propensities in IDRs predicted by both software tools ($r_S = 0.98$ p = 6e-6) using the whole yeast proteome as reference.

The propensity of amino acid residues to be aggregation prone, was evaluated from the amino acid sequence using respectively the TANGO [52] and PASTA2.0 [53] software. TANGO, which predicts aggregation prone regions in unfolded polypeptide chains, was applied using the default parameters for pH, temperature, and ionic strength (7, 298°K and 0.1 respectively). An aggregation prone region was defined as a segment of at least five consecutive residues with a β-aggregation propensity of minimum 5%.

PASTA2.0, a predictor of residues likely to be part of amyloid fibrils [53], was applied using parameters optimized to achieve 90% specificity. The file containing the “best pairs” was used to retrieve the putative amyloid forming regions.

Identifying putative multifunctional proteins

Putative multifunctional proteins in the genome of *S. cerevisiae* were identified on the basis of publically available functional annotations, using a slight modification of the procedure of Khan et al. (2014) [41].

In summary, protein annotations were retrieved from UniProtKB Swiss-Prot database <http://www.uniprot.org/>. GO terms [56,57] from all UniProtKB evidence codes were extracted. Next, the GO terms belonging to Molecular Function (MF) and Biological Process (BP) category were clustered separately. Clustering was done using single linkage clustering [91] based on the sematic similarity among GO terms, with similarity cutoff of 0.1 and 0.5. A protein was classified as multi-functional if it is annotated with GO terms a) belonging to 2 or more clusters with sematic similarity cutoff of 0.1 and b) belonging to 4 or more clusters with sematic similarity cutoff of 0.5 (Supplementary Figures S7 and S8). These cutoff values were empirically determined (2014) [41]. When only MF GO terms are available, the number of clusters based on BP terms was assigned as 0 (zero). Multi-domain proteins were included in the analysis. Ideally Multi-functional proteins identified by this approach should be validated on the basis of manual curation of the literature and other sources [41]. But this very time consuming process was therefore foregone in this study, but may be performed in the future, using recently developed text-mining techniques [92].

In the present study, we considered protein clusters derived on the basis of the GO BP terms. Results were essentially unchanged when considering the union of putative multifunctional proteins inferred using respectively the MF and BP GO annotations.

Statistical analyses

The statistical analyses and the corresponding graphical representations were performed using custom R scripts. To evaluate the relationships between variables we used the non-parametric Spearman's rank correlation coefficient (r_S) and its *p*-value as implemented in the *cor.test* function in R *Stats* package. To compare pairs of distributions of the same variable, the nonparametric Wilcoxon rank-sum (Mann-Whitney U) test was used. In case of multiple distribution comparison, the Bonferroni correction was applied. To compare the fraction of elements in different groups we used the two-sample test for equality of proportions with continuity correction (Pearson's chi-squared statistic).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Wim Vranken and Tamas Lazar for helpful discussions, and are particularly grateful to Emmanuel Levy for critical reading of the manuscript and helpful suggestions. MM-C acknowledges support from the Mexican National Council of Science and Technology (CONACYT) PhD Fellowship [215503]. This work was supported in part by the Odysseus Grant G.0029.12 from the Research Foundation Flanders (FWO) and OTKA grant K124670 to PT, and by a grant from the National Institute of General Medical Sciences of the National Institutes of Health (R01GM123055) and the National Science Foundation (DMS1614777, CMMI1825941) to DK.

Abbreviations

AP-MS	affinity purification and mass spectrometry
BP	biological process
FP	false positive
GO	gene ontology
HC	high confidence
IDP/R	intrinsically disordered protein/region
MF	molecular function
FP	multifunctional protein
MV	multi-validated
NCBD	nuclear coactivator binding domain
PaxDB	protein abundance database
PDB	protein data bank
PE	protein enrichment
PPI	protein-protein interaction
ppm	parts per million

PTM	post translational modification
RMA	robust multi-array average
r_S	Spearman’s rank correlation coefficient
RSASA	relative solvent accessible surface
SLiM	short linear motif
TP	true positive

References

[1]. Alberts B, The cell as a collection of protein machines: preparing the next generation of molecular biologists., Cell. 92 (1998) 291–4. http://www.ncbi.nlm.nih.gov/pubmed/9476889 (accessed August 28, 2018). [PubMed: 9476889]

[2]. Vavouri T, Semple JI, Garcia-Verdugo R, Lehner B, Intrinsic Protein Disorder and Interaction Promiscuity Are Widely Associated with Dosage Sensitivity, Cell. 138 (2009) 198–208. http://www.sciencedirect.com/science/article/pii/S0092867409004541 (accessed September 26, 2013). [PubMed: 19596244]

[3]. Levy ED, Landry CR, Michnick SW, How perfect can protein interactomes be?, Sci. Signal 2 (2009) pe11. doi:10.1126/scisignal.260pe11.

[4]. Levy ED, De S, Teichmann S. a., Cellular crowding imposes global constraints on the chemistry and evolution of proteomes, Proc. Natl. Acad. Sci 109 (2012). doi:10.1073/pnas.1209312109.

[5]. Jacob E, Unger R, Horovitz A, N-terminal domains in two-domain proteins are biased to be shorter and predicted to fold faster than their C-terminal counterparts, Cell Rep. 3 (2013) 1051–6. doi:10.1016/j.celrep.2013.03.032. [PubMed: 23602567]

[6]. Batada NN, Shepp LA, Siegmund DO, Stochastic model of protein-protein interaction: why signaling proteins need to be colocalized, Proc. Natl. Acad. Sci. U. S. A 101 (2004) 6445–9. doi: 10.1073/pnas.0401314101. [PubMed: 15096590]

[7]. Zhang J, Maslov S, Shakhnovich EI, Constraints imposed by non-functional protein–protein interactions on gene expression and proteome size, Mol. Syst. Biol 4 (2008) 210. doi:10.1038/msb.2008.48. [PubMed: 18682700]

[8]. Fishbain S, Inobe T, Israeli E, Chavali S, Yu H, Kago G, Babu MM, Matouschek A, Sequence composition of disordered regions fine-tunes protein half-life, Nat. Struct. Mol. Biol 22 (2015) 214–21. doi:10.1038/nsmb.2958. [PubMed: 25643324]

[9]. Duan G, Walther D, The Roles of Post-translational Modifications in the Context of Protein Interaction Networks, PLOS Comput. Biol 11 (2015) e1004049. doi:10.1371/journal.pcbi.1004049. [PubMed: 25692714]

[10]. Beltrao P, Trinidad JC, Fiedler D, Roguev A, Lim WA, Shokat KM, Burlingame AL, Krogan NJ, Evolution of phosphoregulation: comparison of phosphorylation patterns across yeast species., PLoS Biol. 7 (2009) e1000134. doi:10.1371/journal.pbio.1000134. [PubMed: 19547744]

[11]. Davey NE, Morgan DO, Building a Regulatory Network with Short Linear Sequence Motifs: Lessons from the Degrons of the Anaphase-Promoting Complex., Mol. Cell 64 (2016) 12–23. doi:10.1016/j.molcel.2016.09.006. [PubMed: 27716480]

[12]. Levy ED, Kowarzyk J, Michnick SW, High-resolution mapping of protein concentration reveals principles of proteome architecture and adaptation., Cell Rep. 7 (2014) 1333–40. doi:10.1016/j.celrep.2014.04.009. [PubMed: 24813894]

[13]. van der Lee R, Buljan M, Lang B, Weatheritt RJ, Daughdrill GW, Dunker AK, Fuxreiter M, Gough J, Gsponer J, Jones DT, Kim PM, Kriwacki RW, Oldfield CJ, Pappu RV, Tompa P, Uversky VN, Wright PE, Babu MM, Classification of Intrinsically Disordered Regions and Proteins, Chem. Rev 114 (2014) 6589–6631. doi:10.1021/cr400525m. [PubMed: 24773235]

[14]. Theillet F-X, Binolfi A, Frembgen-Kesner T, Hingorani K, Sarkar M, Kyne C, Li C, Crowley PB, Gierasch L, Pielak GJ, Elcock AH, Gershenson A, Selenko P, Physicochemical Properties of Cells and Their Effects on Intrinsically Disordered Proteins (IDPs), Chem. Rev 114 (2014) 6661–6714. doi:10.1021/cr400695p. [PubMed: 24901537]

[15]. Wright PE, Dyson HJ, Intrinsically disordered proteins in cellular signalling and regulation, Nat. Rev. Mol. Cell Biol 16 (2014) 18–29. doi:10.1038/nrm3920.

[16]. Varadi M, Tompa P, The Protein Ensemble Database, in: Adv. Exp. Med. Biol, 2015: pp. 335–349. doi:10.1007/978-3-319-20164-1_11.

[17]. Toth-Petroczy A, Palmedo P, Ingraham J, Hopf T, Berger B, Sander C, Marks D, Structured States of Disordered Proteins from Genomic Sequences, Cell. 167 (2016) 158–170.e12. doi:10.1016/j.cell.2016.09.010. [PubMed: 27662088]

[18]. Habchi J, Tompa P, Longhi S, Uversky VN, Introducing Protein Intrinsic Disorder, Chem. Rev 114 (2014) 6561–6588. doi:10.1021/cr400514h. [PubMed: 24739139]

[19]. Tompa P, Davey NE, Gibson TJ, Babu MM, A Million Peptide Motifs for the Molecular Biologist, Mol. Cell 55 (2014) 161–169. doi:10.1016/j.molcel.2014.05.032. [PubMed: 25038412]

[20]. Uversky VN, Dunker AK, The case for intrinsically disordered proteins playing contributory roles in molecular recognition without a stable 3D structure, F1000 Biol. Rep 5 (2013) 1. doi:10.3410/B5-1. [PubMed: 23361308]

[21]. Kiss R, Bozoky Z, Kovács D, Róna G, Friedrich P, Dvortsák P, Weisemann R, Tompa P, Perczel A, Calcium-induced tripartite binding of intrinsically disordered calpastatin to its cognate enzyme, calpain, FEBS Lett. 582 (2008) 2149–2154. doi:10.1016/j.febslet.2008.05.032. [PubMed: 18519038]

[22]. Haynes C, Oldfield CJ, Ji F, Klitgord N, Cusick M, Radivojac P, Uversky VN, Vidal M, Iakoucheva LM, Intrinsic Disorder Is a Common Feature of Hub Proteins from Four Eukaryotic Interactomes, (2006). https://dash.harvard.edu/handle/1/4853417 (accessed May 17, 2016).

[23]. Oldfield CJ, Meng J, Yang JY, Yang MQ, Uversky VN, Dunker AK, Flexible nets: disorder and induced fit in the associations of p53 and 14–3-3 with their partners., BMC Genomics. 9 Suppl 1 (2008) S1. doi:10.1186/1471-2164-9-S1-S1.

[24]. Patil A, Nakamura H, Disordered domains and high surface charge confer hubs with the ability to interact with multiple proteins in interaction networks, FEBS Lett. 580 (2006) 2041–2045. doi:10.1016/j.febslet.2006.03.003. [PubMed: 16542654]

[25]. Singh GP, Ganapathi M, Dash D, Role of intrinsic disorder in transient interactions of hub proteins, Proteins Struct. Funct. Bioinforma 66 (2006) 761–765. doi:10.1002/prot.21281.

[26]. Liu J, Perumal NB, Oldfield CJ, Su EW, Uversky VN, Dunker AK, Intrinsic Disorder in Transcription Factors, Biochemistry. 45 (2006) 6873–6888. doi:10.1021/bi0602718. [PubMed: 16734424]

[27]. Garcia-Pino A, Balasubramanian S, Wyns L, Gazit E, De Greve H, Magnuson RD, Charlier D, van Nuland NAJ, Loris R, Allostery and intrinsic disorder mediate transcription regulation by conditional cooperativity., Cell. 142 (2010) 101–11. doi:10.1016/j.cell.2010.05.039. [PubMed: 20603017]

[28]. Tompa P, Multisteric regulation by structural disorder in modular signaling proteins: an extension of the concept of allostery., Chem. Rev 114 (2014) 6715–32. doi:10.1021/cr4005082. [PubMed: 24533462]

[29]. Hegyi H, Schad E, Tompa P, Structural disorder promotes assembly of protein complexes., BMC Struct. Biol 7 (2007) 65. doi:10.1186/1472-6807-7-65. [PubMed: 17922903]

[30]. Han J-DJ, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, Dupuy D, Walhout AJM, Cusick ME, Roth FP, Vidal M, Evidence for dynamically organized modularity in the yeast protein–protein interaction network, Nature. 430 (2004) 88–93. doi:10.1038/nature02555. [PubMed: 15190252]

[31]. Vallabhajosyula RR, Chakravarti D, Lutfeali S, Ray A, Raval A, Identifying Hubs in Protein Interaction Networks, PLoS One. 4 (2009) e5344. doi:10.1371/journal.pone.0005344. [PubMed: 19399170]

[32]. Kim PM, Sboner A, Xia Y, Gerstein M, The role of disorder in interaction networks: a structural analysis., Mol. Syst. Biol 4 (2008) 179. doi:10.1038/msb.2008.16. [PubMed: 18364713]

[33]. Wodak SJ, Vlasblom J, Turinsky AL, Pu S, Protein-protein interaction networks: the puzzling riches., Curr. Opin. Struct. Biol 23 (2013) 941–53. doi:10.1016/j.sbi.2013.08.002. [PubMed: 24007795]

[34]. Janin J, Quantifying biological specificity: The statistical mechanics of molecular recognition, Proteins Struct. Funct. Genet 25 (1996) 438–445. doi:10.1002/prot.4. [PubMed: 8865339]

[35]. Gsponer J, Futschik ME, Teichmann SA, Babu MM, Tight regulation of unstructured proteins: from transcript synthesis to protein degradation., Science. 322 (2008) 1365–8. doi:10.1126/science.1163581. [PubMed: 19039133]

[36]. Weatheritt RJ, Gibson TJ, Babu MM, Asymmetric mRNA localization contributes to fidelity and sensitivity of spatially localized systems., Nat. Struct. Mol. Biol 21 (2014) 833–9. doi:10.1038/nsmb.2876. [PubMed: 25150862]

[37]. Pu S, Vlasblom J, Turinsky A, Marcon E, Phanse S, Trimble SS, Olsen J, Greenblatt J, Emili A, Wodak SJ, Extracting high confidence protein interactions from affinity purification data: At the crossroads, J. Proteomics 118 (2015) 63–80. doi:10.1016/j.jprot.2015.03.009. [PubMed: 25782749]

[38]. Okamura Y, Aoki Y, Obayashi T, Tadaka S, Ito S, Narise T, Kinoshita K, COXPRESdb in 2015: co-expression database for animal species by DNA-microarray and RNAseq-based expression data with multiple quality assessment systems, Nucleic Acids Res 43 (2015) D82–D86. doi:10.1093/nar/gku1163. [PubMed: 25392420]

[39]. Wang M, Herrmann CJ, Simonovic M, Szklarczyk D, von Mering C, Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines., Proteomics. (2015). doi:10.1002/pmic.201400441.

[40]. Collins SR, Kemmeren P, Zhao X-C, Greenblatt JF, Spencer F, Holstege FCP, Weissman JS, Krogan NJ, Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae., Mol. Cell. Proteomics 6 (2007) 439–50. doi:10.1074/mcp.M600381-MCP200. [PubMed: 17200106]

[41]. Khan I, Chen Y, Dong T, Hong X, Takeuchi R, Mori H, Kihara D, Genome-scale identification and characterization of moonlighting proteins, Biol. Direct 9 (2014) 30. doi:10.1186/s13062-014-0030-9. [PubMed: 25497125]

[42]. Dosztányi Z, Csizmók V, Tompa P, Simon I, The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins., J. Mol. Biol 347 (2005) 827–39. doi:10.1016/j.jmb.2005.01.071. [PubMed: 15769473]

[43]. Dosztányi Z, Csizmok V, Tompa P, Simon I, IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content., Bioinformatics. 21 (2005) 3433–4. doi:10.1093/bioinformatics/bti541. [PubMed: 15955779]

[44]. Levy ED, Michnick SW, Landry CR, Protein abundance is key to distinguish promiscuous from functional phosphorylation based on evolutionary information., Philos. Trans. R. Soc. Lond. B. Biol. Sci 367 (2012) 2594–606. doi:10.1098/rstb.2012.0078. [PubMed: 22889910]

[45]. Simicevic J, Schmid AW, Gilardoni PA, Zoller B, Raghav SK, Krier I, Gubelmann C, Lisacek F, Naef F, Moniatte M, Deplancke B, Absolute quantification of transcription factors during cellular differentiation using multiplexed targeted proteomics, Nat. Methods 10 (2013) 570–576. doi:10.1038/nmeth.2441. [PubMed: 23584187]

[46]. Keshishian H, Addona T, Burgess M, Kuhn E, Carr SA, Quantitative, multiplexed assays for low abundance proteins in plasma by targeted mass spectrometry and stable isotope dilution., Mol. Cell. Proteomics 6 (2007) 2212–29. doi:10.1074/mcp.M700354-MCP200. [PubMed: 17939991]

[47]. Xie H, Vucetic S, Iakoucheva LM, Oldfield CJ, Dunker AK, Uversky VN, Obradovic Z, Functional Anthology of Intrinsic Disorder. 1. Biological Processes and Functions of Proteins with Long Disordered Regions, J. Proteome Res 6 (2007) 1882–1898. doi:10.1021/pr060392u. [PubMed: 17391014]

[48]. The UniProt Consortium, UniProt: the universal protein knowledgebase, Nucleic Acids Res. 45 (2017) D158–D169. doi:10.1093/nar/gkw1099. [PubMed: 27899622]

[49]. Necci M, Piovesan D, Tosatto SCE, Large-scale analysis of intrinsic disorder flavors and associated functions in the protein sequence universe., Protein Sci. 25 (2016) 2164–2174. doi:10.1002/pro.3041. [PubMed: 27636733]

[50]. Jaramillo A, Wernisch L, Héry S, Wodak SJ, Folding free energy function selects native-like protein sequences in the core but not on the surface., Proc. Natl. Acad. Sci. U. S. A 99 (2002) 13554–9. doi:10.1073/pnas.212068599. [PubMed: 12368470]

[51]. Kyte J, Doolittle RF, A simple method for displaying the hydropathic character of a protein., J. Mol. Biol 157 (1982) 105–32. http://www.ncbi.nlm.nih.gov/pubmed/7108955 (accessed January 5, 2015). [PubMed: 7108955]

[52]. Fernandez-Escamilla A-M, Rousseau F, Schymkowitz J, Serrano L, Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins, Nat. Biotechnol 22 (2004) 1302–1306. doi:10.1038/nbt1012. [PubMed: 15361882]

[53]. Walsh I, Seno F, Tosatto SCE, Trovato A, PASTA 2.0: an improved server for protein aggregation prediction, Nucleic Acids Res. 42 (2014) W301–7. doi:10.1093/nar/gku399. [PubMed: 24848016]

[54]. Linding R, Schymkowitz J, Rousseau F, Diella F, Serrano L, A comparative study of the relationship between protein structure and beta-aggregation in globular and intrinsically disordered proteins., J. Mol. Biol 342 (2004) 345–53. doi:10.1016/j.jmb.2004.06.088. [PubMed: 15313629]

[55]. Protter DSW, Rao BS, Van Treeck B, Lin Y, Mizoue L, Rosen MK, Parker R, Intrinsically Disordered Regions Can Contribute Promiscuous Interactions to RNP Granule Assembly, Cell Rep. 22 (2018) 1401–1412. doi:10.1016/j.celrep.2018.01.036. [PubMed: 29425497]

[56]. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium., Nat. Genet 25 (2000) 25–9. doi:10.1038/75556. [PubMed: 10802651]

[57]. The Gene Ontology Consortium, Expansion of the Gene Ontology knowledgebase and resources, Nucleic Acids Res. 45 (2017) D331–D338. doi:10.1093/nar/gkw1108. [PubMed: 27899567]

[58]. McGuffee SR, Elcock AH, Diffusion, crowding & protein stability in a dynamic molecular model of the bacterial cytoplasm., PLoS Comput. Biol 6 (2010) e1000694. doi:10.1371/journal.pcbi.1000694. [PubMed: 20221255]

[59]. Frembgen-Kesner T, Elcock AH, Computer Simulations of the Bacterial Cytoplasm., Biophys. Rev 5 (2013) 109–119. doi:10.1007/s12551-013-0110-6. [PubMed: 23914257]

[60]. Schnell S, Fortunato S, Roy S, Is the intrinsic disorder of proteins the cause of the scale-free architecture of protein–protein interaction networks?, Proteomics. 7 (2007) 961–964. doi:10.1002/pmic.200600455. [PubMed: 17285562]

[61]. Dosztányi Z, Chen J, Dunker AK, Simon I, Tompa P, Disorder and sequence repeats in hub proteins and their implications for network evolution., J. Proteome Res 5 (2006) 2985–95. doi:10.1021/pr060171o. [PubMed: 17081050]

[62]. Yu H, Braun P, Yildirim MA, Lemmens I, Venkatesan K, Sahalie J, HirozaneKishikawa T, Gebreab F, Li N, Simonis N, Hao T, Rual J-F, Dricot A, Vazquez A, Murray RR, Simon C, Tardivo L, Tam S, Svrzikapa N, Fan C, de Smet A-S, Motyl A, Hudson ME, Park J, Xin X, Cusick ME, Moore T, Boone C, Snyder M, Roth FP, Barabasi A-L, Tavernier J, Hill DE, Vidal M, High-Quality Binary Protein Interaction Map of the Yeast Interactome Network, Science (80-.). 322 (2008) 104–110. doi:10.1126/science.1158684.

[63]. Jensen LJ, Bork P, Biochemistry. Not comparable, but complementary., Science. 322 (2008) 56–7. doi:10.1126/science.1164801. [PubMed: 18832636]

[64]. Goll J, Uetz P, The elusive yeast interactome, Genome Biol. 7 (2006) 223. doi:10.1186/gb-2006-7-6-223. [PubMed: 16938899]

[65]. Chatr-Aryamontri A, Oughtred R, Boucher L, Rust J, Chang C, Kolas NK, O’Donnell L, Oster S, Theesfeld C, Sellam A, Stark C, Breitkreutz BJ, Dolinski K, Tyers M, The BioGRID interaction database: 2017 update, Nucleic Acids Res. 45 (2017) D369–D379. doi:10.1093/nar/gkw1102. [PubMed: 27980099]

[66]. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, Punna T, Peregrín-Alvarez JM, Shales M, Zhang X, Davey M, Robinson MD, Paccanaro A, Bray JE, Sheung A, Beattie B, Richards DP, Canadien V, Lalev A, Mena F, Wong P, Starostine A,

Canete MM, Vlasblom J, Wu S, Orsi C, Collins SR, Chandran S, Haw R, Rilstone JJ, Gandi K, Thompson NJ, Musso G, St Onge P, Ghanny S, Lam MHY, Butland G, Altaf-Ul AM, Kanaya S, Shilatifard A, O'Shea E, Weissman JS, Ingles CJ, Hughes TR, Parkinson J, Gerstein M, Wodak SJ, Emili A, Greenblatt JF, Global landscape of protein complexes in the yeast Saccharomyces cerevisiae, Nature. 440 (2006) 637–643. doi:10.1038/nature04670. [PubMed: 16554755]

[67]. Babu M, Díaz-Mejía JJ, Vlasblom J, Gagarinova A, Phanse S, Graham C, Yousif F, Ding H, Xiong X, Nazarians-Armavil A, Alamgir M, Ali M, Pogoutse O, Pe'er A, Arnold R, Michaut M, Parkinson J, Golshani A, Whitfield C, Wodak SJ, MorenoHagelsieb G, Greenblatt JF, Emili A, Genetic interaction maps in Escherichia coli reveal functional crosstalk among cell envelope biogenesis pathways., PLoS Genet. 7 (2011) e1002377. doi:10.1371/journal.pgen.1002377. [PubMed: 22125496]

[68]. Wodak SJ, Vlasblom J, Pu S, High-throughput analyses and curation of protein interactions in yeast., Methods Mol. Biol 759 (2011) 381–406. doi:10.1007/9781-61779-173-4_22. [PubMed: 21863499]

[69]. Chapple CE, Robisson B, Spinelli L, Guien C, Becker E, Brun C, Extreme multifunctional proteins identified from a human protein interaction network, Nat. Commun 6 (2015) 7412. doi: 10.1038/ncomms8412. [PubMed: 26054620]

[70]. Huberts DHEW, van der Klei IJ, Moonlighting proteins: an intriguing mode of multitasking., Biochim. Biophys. Acta 1803 (2010) 520–5. doi:10.1016/j.bbamcr.2010.01.022. [PubMed: 20144902]

[71]. Mani M, Chen C, Amblee V, Liu H, Mathur T, Zwicke G, Zabad S, Patel B, Thakkar J, Jeffery CJ, MoonProt: a database for proteins that are known to moonlight, Nucleic Acids Res. 43 (2015) D277–D282. doi:10.1093/nar/gku954. [PubMed: 25324305]

[72]. Tompa P, Szász C, Buday L, Structural disorder throws new light on moonlighting., Trends Biochem. Sci 30 (2005) 484–9. doi:10.1016/j.tibs.2005.07.008. [PubMed: 16054818]

[73]. Lee CW, Martinez-Yamout MA, Dyson HJ, Wright PE, Structure of the p53 Transactivation Domain in Complex with the Nuclear Receptor Coactivator Binding Domain of CREB Binding Protein, Biochemistry. 49 (2010) 9964–9971. doi:10.1021/bi1012996. [PubMed: 20961098]

[74]. Mészáros B, Simon I, Dosztányi Z, Prediction of protein binding regions in disordered proteins., PLoS Comput. Biol 5 (2009) e1000376. doi:10.1371/journal.pcbi.1000376. [PubMed: 19412530]

[75]. Heo M, Maslov S, Shakhnovich E, Topology of protein interaction network shapes protein abundances and strengths of their functional and nonspecific interactions., Proc. Natl. Acad. Sci. U. S. A 108 (2011) 4258–63. doi:10.1073/pnas.1009392108. [PubMed: 21368118]

[76]. Maharana S, Wang J, Papadopoulos DK, Richter D, Pozniakovsky A, Poser I, Bickle M, Rizk S, Guillén-Boixet J, Franzmann TM, Jahnel M, Marrone L, Chang Y-T, Sterneckert J, Tomancak P, Hyman AA, Alberti S, RNA buffers the phase separation behavior of prion-like RNA binding proteins., Science. 360 (2018) 918–921. doi:10.1126/science.aar7366. [PubMed: 29650702]

[77]. Contreras-Martos S, Piai A, Kosol S, Varadi M, Bekesi A, Lebrun P, Volkov AN, Gevaert K, Pierattelli R, Felli IC, Tompa P, Linking functions: an additional role for an intrinsically disordered linker domain in the transcriptional coactivator CBP, Sci. Rep 7 (2017) 4676. doi: 10.1038/s41598-017-04611-x. [PubMed: 28680062]

[78]. Christiano R, Nagaraj N, Fröhlich F, Walther TC, Global Proteome Turnover Analyses of the Yeasts S. cerevisiae and S. pombe, Cell Rep. 9 (2014) 1959–1965. doi:10.1016/j.celrep.2014.10.065. [PubMed: 25466257]

[79]. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP, Exploration, normalization, and summaries of high density oligonucleotide array probe level data, Biostatistics. 4 (2003) 249–264. doi:10.1093/biostatistics/4.2.249. [PubMed: 12925520]

[80]. Gavin A-C, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dümpelfeld B, Edelmann A, Heurtier M-A, Hoffman V, Hoefert C, Klein K, Hudak M, Michon A-M, Schelder M, Schirle M, Remor M, Rudi T, Hooper S, Bauer A, Bouwmeester T, Casari G, Drewes G, Neubauer G, Rick JM, Kuster B, Bork P, Russell RB, Superti-Furga G, Proteome survey reveals modularity of the yeast cell machinery, Nature. 440 (2006) 631–636. doi:10.1038/nature04532. [PubMed: 16429126]

[81]. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE, The Protein Data Bank., Nucleic Acids Res. 28 (2000) 235–42. http://www.ncbi.nlm.nih.gov/pubmed/10592235 (accessed October 11, 2017). [PubMed: 10592235]

[82]. Bernsel A, Viklund H, Hennerdal A, Elofsson A, TOPCONS: consensus prediction of membrane protein topology, Nucleic Acids Res. 37 (2009) W465–W468. doi:10.1093/nar/gkp363. [PubMed: 19429891]

[83]. Dey S, Ritchie DW, Levy ED, a., Nat. Methods 15 (2018) 67–72. doi:10.1038/nmeth.4510. [PubMed: 29155427]

[84]. Maere S, Heymans K, Kuiper M, BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks, Bioinformatics. 21 (2005) 3448–3449. doi: 10.1093/bioinformatics/bti551. [PubMed: 15972284]

[85]. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T, Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks, Genome Res. 13 (2003) 2498–2504. doi:10.1101/gr.1239303. [PubMed: 14597658]

[86]. Wang JZ, Du Z, Payattakool R, Yu PS, Chen C-F, A new method to measure the semantic similarity of GO terms., Bioinformatics. 23 (2007) 1274–81. doi:10.1093/bioinformatics/btm087. [PubMed: 17344234]

[87]. Yu G, Li F, Qin Y, Bo X, Wu Y, Wang S, GOSemSim: an R package for measuring semantic similarity among GO terms and gene products., Bioinformatics. 26 (2010) 976–8. doi:10.1093/bioinformatics/btq064. [PubMed: 20179076]

[88]. Pu S, Wong J, Turner B, Cho E, Wodak SJ, Up-to-date catalogues of yeast protein complexes, Nucleic Acids Res. 37 (2009) 825–831. doi:10.1093/nar/gkn1005. [PubMed: 19095691]

[89]. Mitternacht S, FreeSASA: An open source C library for solvent accessible surface area calculations., F1000Research. 5 (2016) 189. doi:10.12688/f1000research.7931.1. [PubMed: 26973785]

[90]. Necci M, Piovesan D, Dosztányi Z, Tosatto SCE, MobiDB-lite: fast and highly specific consensus prediction of intrinsic disorder in proteins, Bioinformatics. 33 (2017) 1402–1404. doi: 10.1093/bioinformatics/btx015. [PubMed: 28453683]

[91]. Gower JC, Ross GJS, Minimum Spanning Trees and Single Linkage Cluster Analysis, Appl. Stat 18 (1969) 54. doi:10.2307/2346439.

[92]. Jain A, Gali H, Kihara D, Identification of Moonlighting Proteins in Genomes Using Text Mining Techniques., Proteomics. (2018) e1800083. doi:10.1002/pmic.201800083.

[93]. Benjamini Y, Yekutieli D, The Control of the False Discovery Rate in Multiple Testing under Dependency, Ann. Stat 29 (2001) 1165–1188. doi:10.2307/2674075.

Highlights:

- Intrinsically disordered proteins are prone to forming weak promiscuous interactions
- We set out to determine how nature mitigates interference from such interactions
- We study protein disorder level, sequence properties and cellular context in yeast
- Protein disorder level and sequence properties have evolved to minimize promiscuity
- Protein disorder and sequence properties also evolved to enable functional diversity

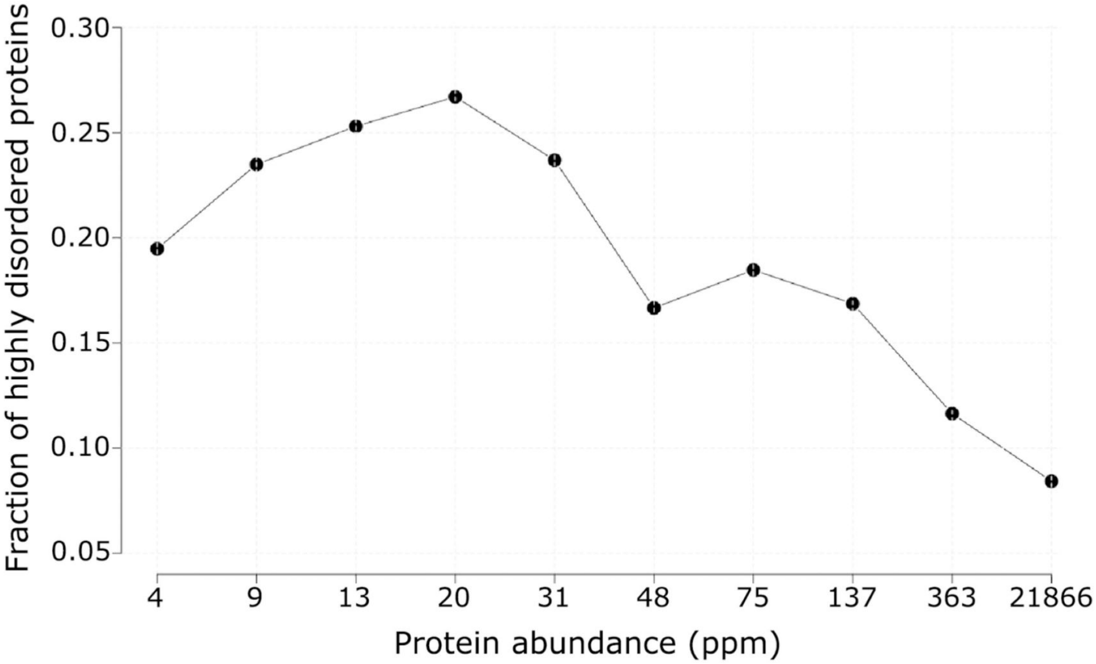


Figure 1.

Fraction of *S. cerevisiae* proteins with a high degree of putative intrinsic disorder, as a function of their cellular abundance.

The graph reveals a strong negative correlation between the fraction of highly disordered proteins and the corresponding protein abundance level (Spearman correlation $r_S = -0.76$, pvalue = 0.02). A stronger and highly significant correlation is obtained when applying a somewhat higher abundance threshold and a lower but significant correlation when taking into account the fraction of all the disordered residues in a protein (see text). Proteins featuring a high degree of intrinsic disorder are defined as those containing ≥30% of their residues in IDRs of at least 20 consecutive residues. Proteins in the yeast proteome were divided according to their abundance into ten equally populated bins (498 proteins each). Numbers along the horizontal axis represent the rounded upper limit in the protein abundance range of each bin.

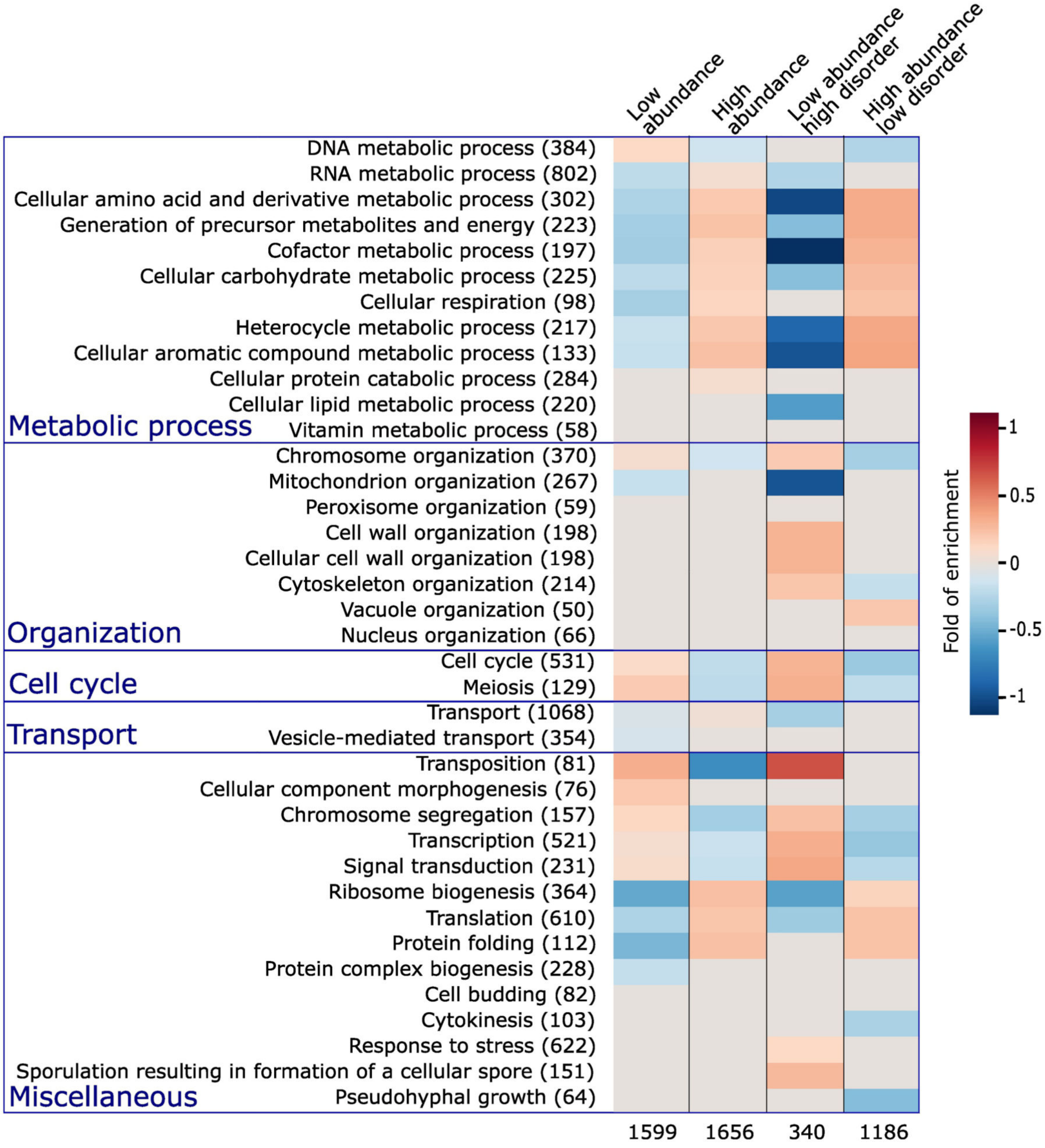


Figure 2.

Enrichment in Gene Ontology terms for proteins with different abundance levels and intrinsic disorder contents.

The first column lists the biological process (BP) terms of the Gene Ontology [56,57] for which statistically significant enrichments were obtained. The number of *S. cerevisiae* proteins with the corresponding BP term in the reference dataset is indicated in parentheses. The last 4 columns represent the heat maps of the computed statistically significant enrichment levels for each of the listed BP terms among proteins in the following 4 categories respectively. Proteins among the 33.3 percentile most abundant and least abundant proteins, respectively (columns 2 and 3); the subset of proteins from each of the previous categories with high (≥ 30%) and low (<10%) IDR content, respectively (columns 4 and 5).

The number of proteins in each of the 4 categories is listed at the bottom. The Biological Process (BP) Gene ontology (GO) terms are those from the GO slim yeast ontology in the BiNGO [84] application for Cytoscape [85] with all possible evidence codes. Enrichments were computed using as reference, all *S. cerevisiae* proteins with available GO annotation and abundance levels of at least 2 ppm (4900 proteins in total). The statistical significance of the computed enrichment was evaluated using the hypergeometric test with Benjamin and Hochberg false discovery rate (FDR) multiple testing correction [93].

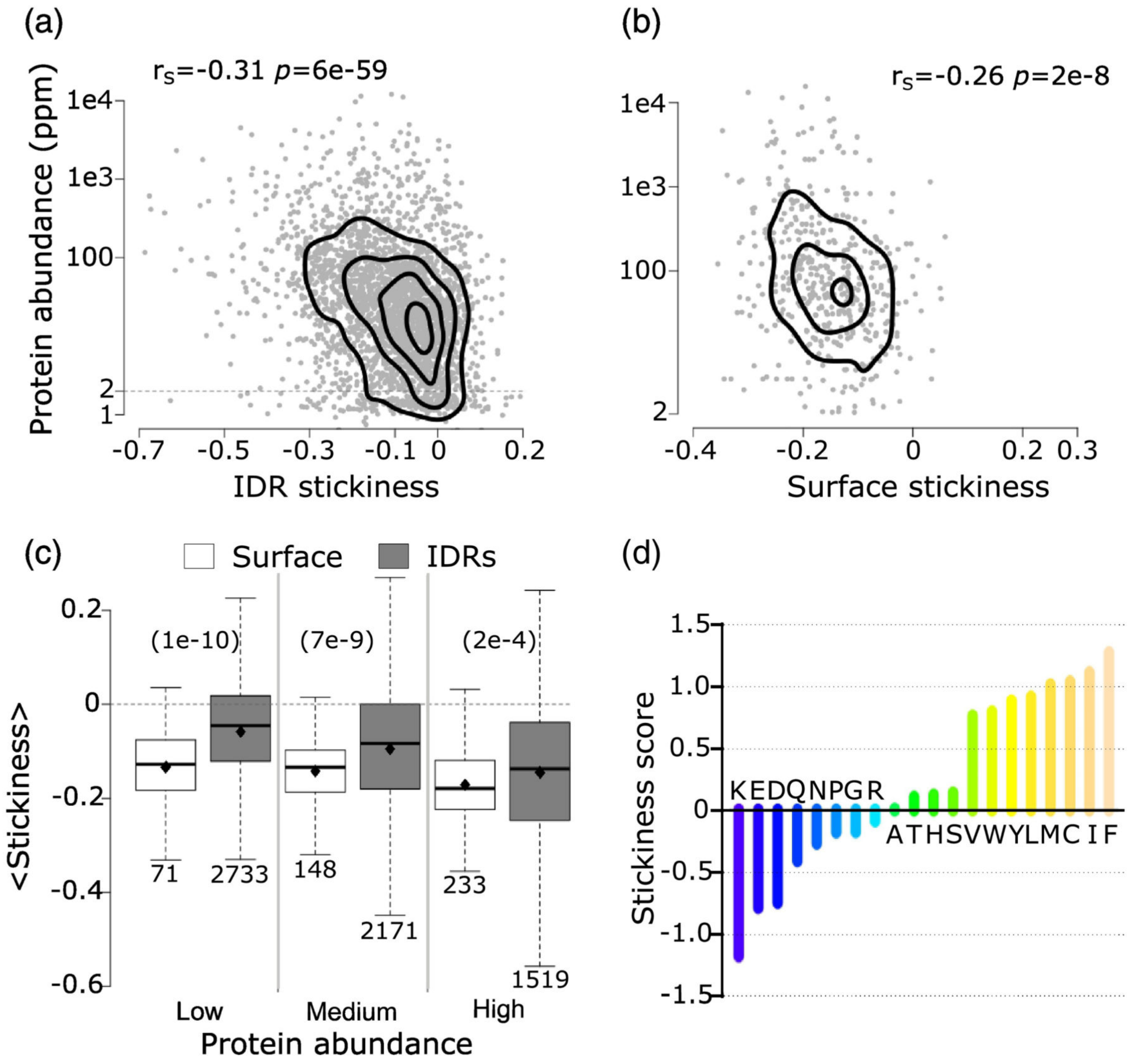


Figure 3.

Average residue stickiness versus protein abundance.

a) Scatter plot of the average residue stickiness in IDRs ≥20 residues of 2874 (2566 ≥2 ppm) *S. cerevisiae* proteins, as a function of protein abundance. b) Scatter plot of average stickiness of surface residues in 452 yeast globular proteins retrieved from the PDB, as a function of protein abundance. The Spearman's correlation coefficients (r_S) and the corresponding *p*values are given at the top of each panel. For both datasets, correlations were calculated using proteins with an abundance value of at least 2ppm. c) Boxplots of average residue stickiness on the surface of globular proteins (white) and in IDRs ≥20 residues (gray) for proteins in three abundance ranges: low (2 – 15.5ppm), medium (15.6 – 64.7ppm) and high (64.8 – 21866). The corresponding average stickiness values (diamond in each boxplot) are: −0.13, 0.06, −0.14, −0.10, −0.17 and −0.15. *p*-values (computed using the Wilcoxon rank-sum test) between pairs of distributions in the three abundance ranges are given at the top of the panel in parentheses. The number of data points in each distribution (n) is shown below each boxplot. The horizontal line in each boxplot indicates the median value. Outliers are not depicted. d) Bar plot of the stickiness score of amino acids in [4]. Protein abundance values were obtained from PaxDb [39] (see Material and Methods).

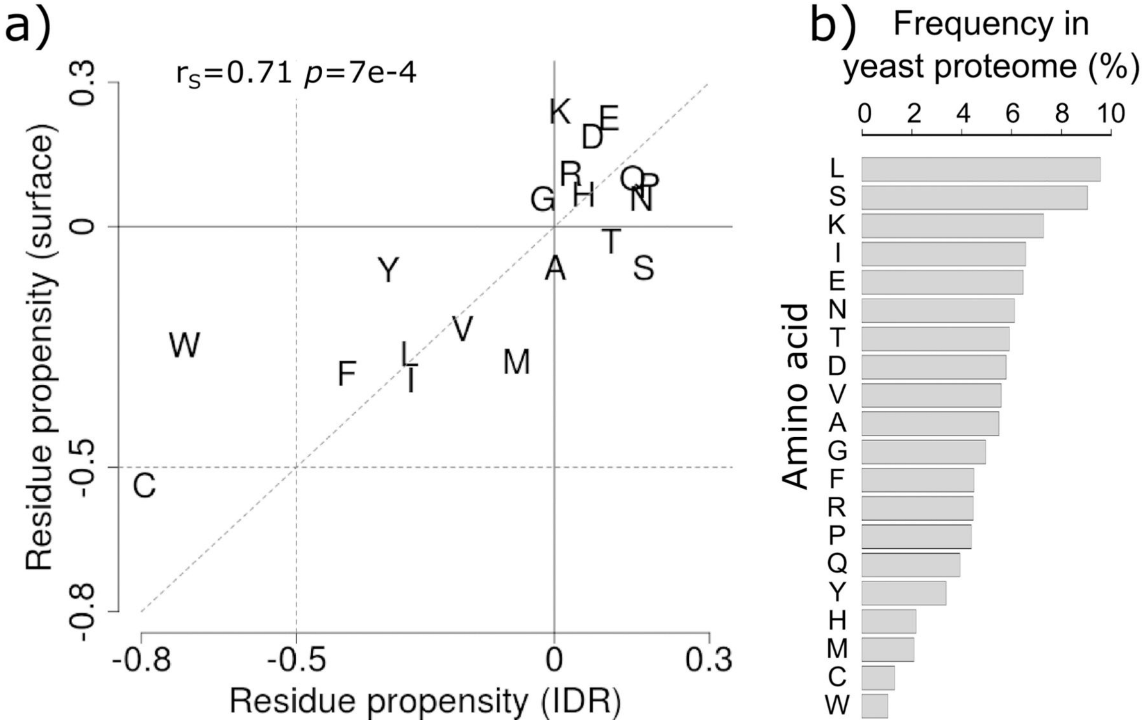


Figure 4.

Comparing amino acid propensities of IDRs to those on surfaces of globular proteins.

a) Amino acid propensities in IDRs ≥20 residues (horizontal axis) versus those on the surfaces of globular proteins/domains (vertical axis) using the yeast proteome as reference (see Materials and Methods). Amino acids are represented by the one-letter code. The Spearman's correlation coefficient between the two sets of propensities and the corresponding *p*-value are given at the top of the graph. b) Frequencies of amino acids (%) in the yeast proteome.

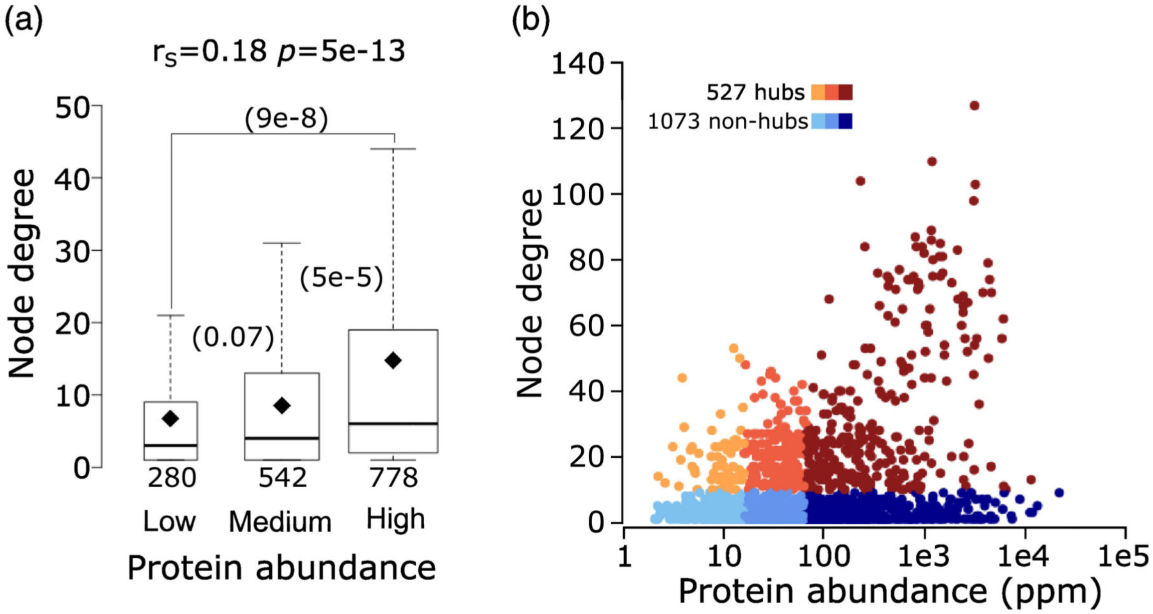


Figure 5.

Relating protein abundance to the number of interaction partners (node degree) in the yeast interactome.

a) Boxplots of the number of interaction partners (node degree) for proteins in three ranges of abundance: low (2 – 15.5ppm), medium (15.6 – 64.7ppm) and high (64.8 – 21866ppm). *P*-values (computed using the Wilcoxon rank-sum test) between pairs of distributions in the three abundance ranges are given in parentheses. The number of data points in each distribution (n) is shown below each boxplot. The horizontal line in each boxplot indicates the median value and the diamond indicates the mean value. Outliers are not depicted in the panel. b) Scatter plot of protein abundance versus node degree. Hub proteins (those with ≥10 interaction partners) are colored from orange to brown and non-hubs (<10 interaction partners) are colored from light to dark blue, according to the 3 protein abundance ranges given in (a). Hubs represent 22%, 30% and 39% respectively, of proteins of low-, medium- and high abundance in the HC Collins interactome.

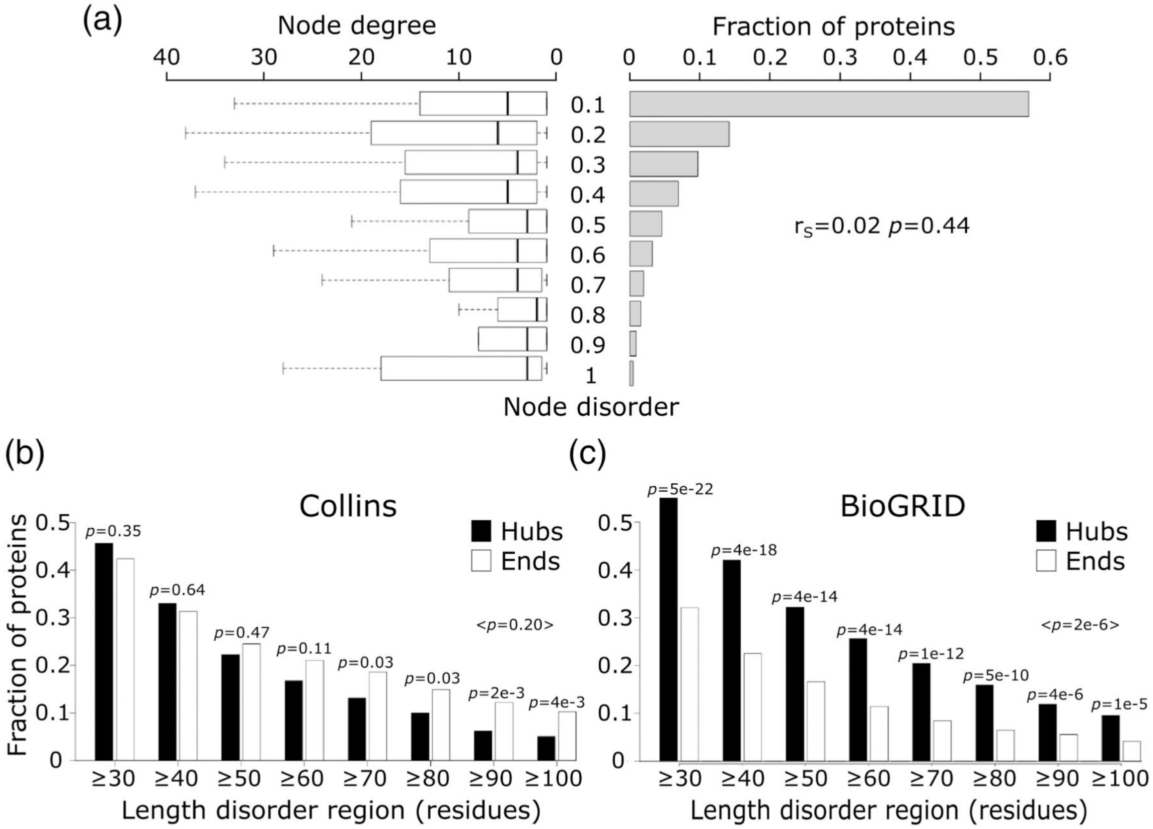


Figure 6.

Node degree of proteins with different levels of intrinsic disorder.

a) Depicted on the right hand side is the histogram of the intrinsic disorder ranges (fraction of disordered residues in IDRs ≥20 residues long) of protein nodes in the Collins yeast PPI network [40]. Shown on the left are boxplots representing the distributions of the number of interaction partners (node degree) for individual disorder ranges of disorder. The Spearman's correlation coefficient (r_S), and its corresponding *p*-value are given at the right hand side of the panel. b) The fraction of hub (≥10 partners) and 'end' (1 partner) proteins at different IDR length cutoffs in the HC yeast PPI network of Collins. c) The same as (b) but for proteins in a recent version of the BioGRID [65] interactome. In both panels, the *p*-value between the fraction of hub and end proteins was calculated using the two-sample test for equality of proportions with continuity correction (Pearson's chi-squared statistic).

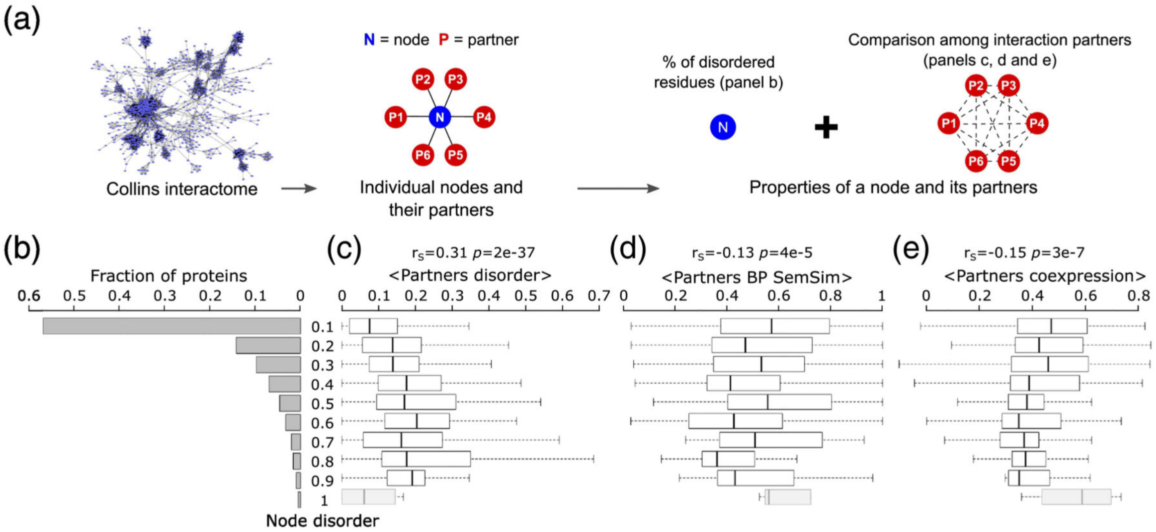
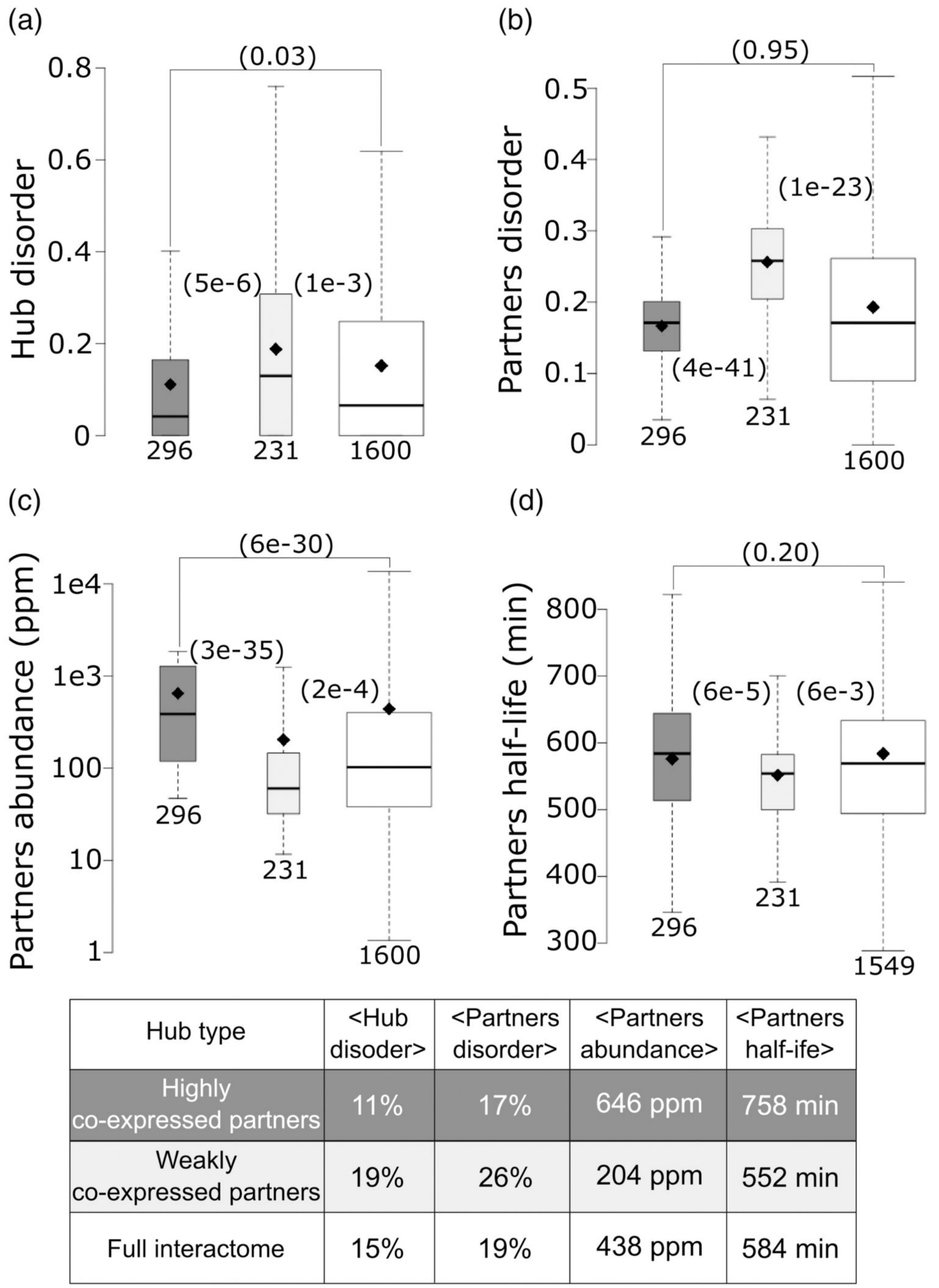


Figure 7.

Relating intrinsic disorder of protein nodes to the properties of their interaction partners.

a) Schematic representation of the source of the values depicted in the following panels. b) histogram of the intrinsic disorder levels of protein nodes of the HC Collins yeast PPI network [40], depicted in Fig. 6a. Shown to the right are box plots representing the distributions of various properties of the interaction partners of protein nodes in the disorder ranges depicted on the far-left histogram: c) Average disorder (fraction of disordered residues in IDRs≥ 20 residues long) of interaction partners, d) average functional similarity (semantic similarity of their biological process gene ontology annotation, BP SemSim) of interactions partners and e) average co-expression (Pearson correlation coefficient of the mRNA expression profiles [38]) of interaction partners. The Spearman's correlation coefficient (r_S), and its corresponding *p*-value are given at the top of each panel. Nodes with the highest disorder level comprising <10 proteins (greyed box plots), were not considered in the analysis.

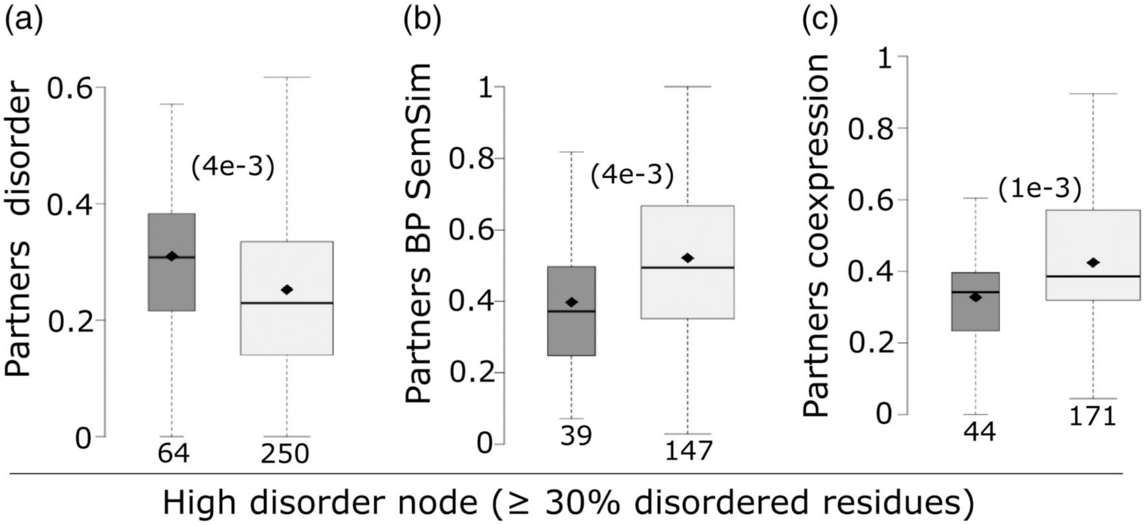


Hub type	<Hub disoder>	<Partners disorder>	<Partners abundance>	<Partners half-ife>
Highly co-expressed partners	11%	17%	646 ppm	758 min
Weakly co-expressed partners	19%	26%	204 ppm	552 min
Full interactome	15%	19%	438 ppm	584 min

Figure 8.

Contrasting the properties of protein nodes and their interaction partners in hubs with highly and weakly co-expressed interaction partners.

a) Box plots depicting the average disorder level of: hubs with highly co-expressed partners (average partners' mRNA expression profiles Pearson correlation coefficient ≥0.5; dark gray), hubs with weakly co-expressed partners (average partners' mRNA expression profiles Pearson correlation coefficient <|0.5|; light gray), all protein nodes in the HC Collins interactome (white), b-d) Boxplots depicting respectively the distributions of: disorder content, abundance levels (ppm) and half-life (min) of the interaction partners for the same 3 categories of protein nodes as in (a). Abundance values are from PaxDb [39] and yeast protein half-lives are taken from reference [78]. *p*-values (computed using the Wilcoxon rank-sum test) between pairs of distributions in the three groups are given in parentheses. The number of data points in each distribution (n) is shown below each boxplot. The horizontal line in each boxplot indicates the median value and the diamond indicates the mean value. Outliers are not depicted. The Table at the bottom lists the average values of the corresponding displayed distributions.



Proteins	<Partners disorder>	<Partners BP SemSim>	<Partners co-expression>
Multifunctional (MFP)	31%	0.40	0.33
Non-multifunctional (non-MFP)	25%	0.52	0.42

Figure 9.

Relations of high intrinsic disorder content and the multifunctional nature of protein nodes to the properties of their interaction partners.

Panels a-c) compare the distributions of the disorder levels, functional similarity (BP SemSim), and the mRNA co-expression levels, respectively, of interaction partners in the two groups of proteins with a high disorder content (≥30% residues in IDRs ≥20 residues): putative multi-functional protein nodes of the Collins PPI network [40] (dark grey) and nonmulti-functional proteins nodes (light grey). *p*-values (computed using the Wilcoxon ranksum test) between the pair of distributions are given in parentheses. The number of data points in each distribution is shown below each boxplot. The horizontal line in each boxplot indicates the median value and the diamond indicates the mean value. Outliers are not depicted. The Table at the bottom lists the average values of the corresponding displayed distributions.

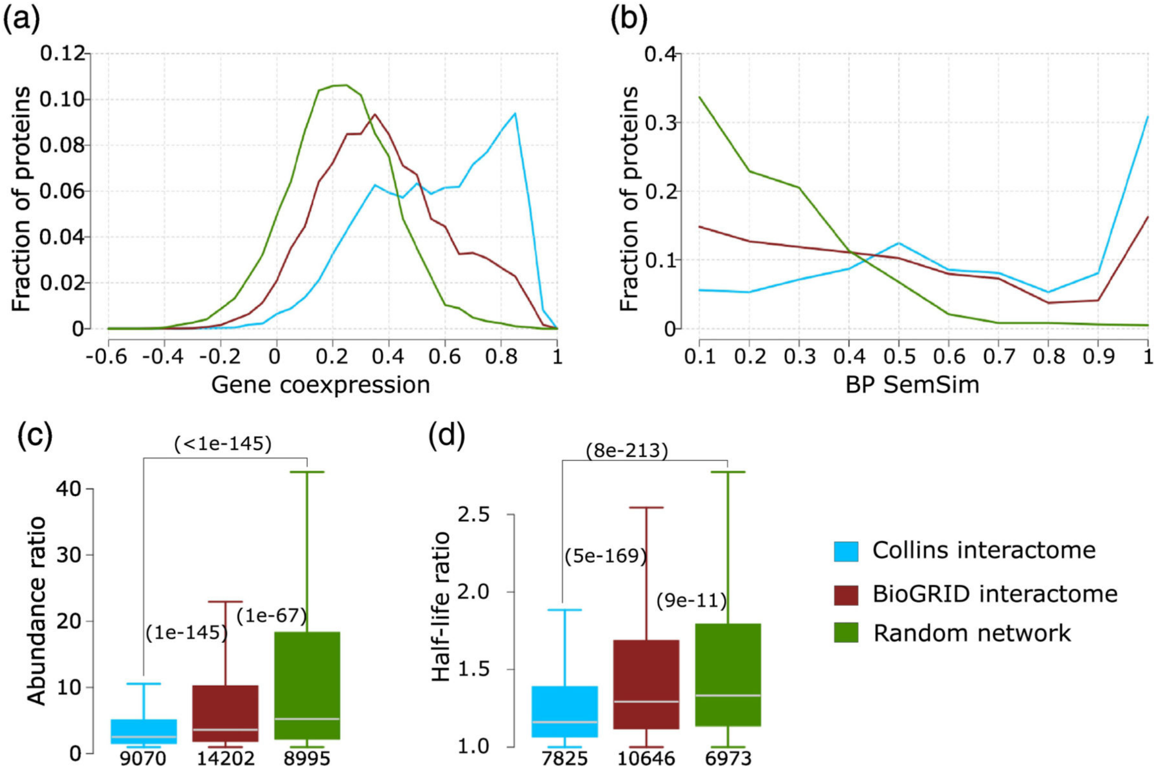


Figure 10.

Comparing the Collins and BioGRID interactomes of *S. cerevisiae*.

The HC interactome of Collins [40] and the multi-validated BioGRID interactome [65] (March 2018) are compared against each other and to a random PPI network (see Material and Methods for detail). Four properties of interacting proteins pairs in the network are compared: a) Pearson correlation coefficient of mRNA expression profiles (gene co-expression), b) the semantic similarity of the biological process GO annotation (BP SemSim), c) the ratio of protein abundance levels and d) the ratio of protein half-life. Panels c, d) *p*-values (computed using the Wilcoxon rank-sum test) between pairs of distributions in the three networks are given in parentheses. The number of data points in each distribution is shown below each boxplot. Data on protein abundance are from PaxDb [39], and those on half-life are from reference [78].

Table 1.

Intrinsic disorder enrichment in multifunctional proteins (MFP) of *S. cerevisiae*

Dataset	Proteins	Highly disordered proteins	Fraction highly disordered proteins	p^*	Enrichment
Whole proteome	6437	1055	0.16	5e-5	1.44
MFPs	595	137	0.23		
Interactome	1622	314	0.19	0.02	1.37
MFPs in interactome	246	64	0.26		
Hub proteins	530	92	0.17	0.08	1.65
MFP hubs	58	16	0.28		

MFPs: multifunctional proteins

Highly disordered proteins are those with ≥30% disordered residues

[*] *P*-values from the two-sample test for equality of proportions (Pearson's chi-squared statistic)

Table 2.

Average stickiness scores in IDRs of MFP and non-MFP in *S. cerevisiae*

	Dataset	Proteins	<Stickiness>	SD	*p*[*]
Whole proteome	MFPs	367	−0.0947	0.1151	0.04
	NMFPs	2507	−0.1095	0.1195	
Collins interactome	MFPs	160	−0.1144	0.1129	6e-4
	NMFPs	771	−0.1474	0.1210	

MFPs: multifunctional proteins

NMFPs: non-multifunctional proteins

<Stickiness>: average IDR stickiness per protein. More negative values correspond to lower stickiness

SD: standard deviation

[*] *P*-values from the two-sample test for equality of proportions (Pearson's chi-squared statistic)