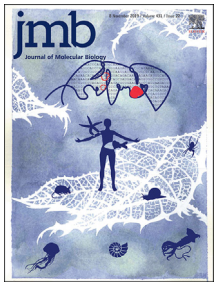


Sequential, Structural and Functional Properties of Protein Complexes Are Defined by How Folding and Binding Intertwine

Bálint Mészáros [1,2,3], László Dobson [4,5], Erzsébet Fichó [3], Gábor E. Tusnády [4], Zsuzsanna Dosztányi [1] and István Simon [3]

1 - MTA-ELTE Momentum Bioinformatics Research Group, *Department of Biochemistry, Eötvös Loránd University, Pázmány Péter stny 1/c, Budapest H-1117, Hungary*

2 - Structural and Computational Biology Unit, *European Molecular Biology Laboratory, Meyerhofstraße 1, 69117 Heidelberg, Germany*

3 - Protein Structure Research Group, *Institute of Enzymology, RCNS, HAS, PO Box 7, Budapest H-1518, Hungary*

4 - Membrane Protein Bioinformatics Research Group, *Institute of Enzymology, RCNS, HAS, PO Box 7, Budapest H-1518, Hungary*

5 - Faculty of Information Technology and Bionics, *Pázmány Péter Catholic University, Práter u. 50A, H-1083 Budapest, Hungary*

Correspondence to Bálint Mészáros: *MTA-ELTE Momentum Bioinformatics Research Group, Department of Biochemistry, Eötvös Loránd University, Pázmány Péter stny 1/c, Budapest H-1117, Hungary.* bmeszaros@caesar.elte.hu

https://doi.org/10.1016/j.jmb.2019.07.034

Edited by Anna Panchenko

Abstract

Intrinsically disordered proteins (IDPs) fulfill critical biological roles without having the potential to fold on their own. While lacking inherent structure, the majority of IDPs do reach a folded state via interaction with a protein partner, presenting a deep entanglement of the folding and binding processes. Protein disorder has been recognized as a major determinant in several properties of proteins, such as sequence, adopted structure upon binding and function. However, the way the binding process is reflected in these features in general lacks a detailed description. Here, we defined three categories of protein complexes depending on the unbound structural state of the interactors and analyzed them in detail. We found that strikingly, the properties of interactors in terms of sequence and adopted structure are defined not only by the intrinsic structural state of the protein itself but also to a comparable extent by the structural state of the binding partner. The three different types of interactions are also regulated through divergent molecular tactics of post-translational modifications. This not only widens the range of biologically relevant sequence and structure spaces defined by ordered proteins but also presents distinct molecular mechanisms compatible with specific biological processes, separately for each interaction type. The distinct attributes of different binding modes identified in this study can help to understand how various types of interactions serve as building blocks for the assembly of tightly regulated and highly intertwined regulatory networks.

© 2019 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Introduction

Proteins deliver the basic machinery for life, providing functions indispensable for all living organisms. The foundation of the molecular understanding of how proteins function was hallmarked by the determination of the first protein structures. The resulting dogma, called structure–function paradigm [1], delineated the central thesis of structural biology: protein function is born from structure, and the prerequisite of a functional protein is prior successful and complete folding.

In the late 1990s, mounting evidence led to the realization that ordered proteins, which conform to the dogma, represent only part of the protein world [2]. There are several other functional proteins that lack a stable tertiary structure in their isolated form. Although they defy the previous dogma, intrinsically

0022-2836/© 2019 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

disordered proteins (IDPs) are critically important functionally, especially in signaling and regulation [3,4]. IDPs have advanced our understanding of how proteins can function by showing that protein regions incompatible with autonomous folding can still convey biological processes. With this birth of "unstructural biology" [5], the protein world was divided into two major regions: ordered and disordered proteins. This binary view is deeply embedded in us at the conceptual level, exemplified by current disorder prediction methods [6,7] and their evaluation schemes [8]. As an extension of this binary representation, it has been long recognized that protein flexibility is rather a continuous spectrum, ranging from (almost) true random coils [9], through molten globules [10] and proteins that are marginally stable [11], to stable domains.

A more complete description of IDPs should consider not only their structural properties in their free state but also their interactions. While some IDPs stay disordered during exerting their function—some even when bound to a protein partner [12]—the vast majority of known IDP interactions result in the partial or full ordering of the disordered partner. In these cases, the folding happens at the same time as the binding, and the two processes, governed by the same biophysical forces [13], are deeply intertwined.

The entanglement of folding and binding for IDPs results in binding modes clearly distinct from the interactions of ordered proteins. The interaction between an IDP and an ordered protein partner—termed coupled folding and binding [14]—holds the potential for forming weaker, transient interactions [15] due to the loss of conformational entropy decreasing the binding strength [16]. The description and study of this binding mode of IDPs were approached from several different angles. Most known coupled folding and binding IDP regions are short, often exhibiting a single well-defined secondary structure character in the bound form. Identifying such short segments in bound structures is the basis of the definition of molecular recognition features (MoRFs) [17–19]. The study of the specific structural properties of such interactions gave rise to their better understanding [20] and also enabled a more detailed classification of IDPs [21], together with targeted sequence-based prediction method development [22–24]. Several such peptide-domain motifs have also been studied in structural detail [25,26] leading to the development of specialized structure-based predictions [27] and generic docking methods capable of accounting for the flexibility of one of the interacting partners [28]. These advances have the potential to ultimately lead to successful development of novel ways of pharmaceutical modulation through the identification of targetable epitopes on domains [29,30], possible inhibitory peptides [31], and the development of small molecules [32].

In contrast, complexes formed exclusively by IDPs—through a process termed mutual synergistic folding—are far less understood. Most of our knowledge stems from individually studied cases [33] and analyses of relatively small data sets [34,35]. While several related classes of protein complexes have been analyzed (such as intertwined complexes [36]), these works define their focus based on the properties of the bound structure instead of the structural states of the unbound proteins. However, this lack of targeted analysis of mutual synergistic folding is primarily due to the lack of data. Recently, two new databases focusing on various types of IDP interactions in structural detail were established, paving the way for the systematic analyses of their specific properties [37,38].

In this work, we examine the basic types of interactions between proteins, considering both ordered and IDP interactors. We classify binding scenarios purely based on the unbound structural states of the constituent proteins. While several analyses comparing the binding of ordered and disordered proteins have been conducted [17,20,39,40] using the analysis of protein sequences and structures adopted after binding, we assess how, and to what extent, the sequence and structure spaces compatible with biological functions are extended by IDPs. Also, to what extent does the presence and the structural state of a binding partner influence various IDP properties, such as sequence composition, adopted structure and cellular function? Considering the biological functions mediated by protein interactions, the main question is if there are certain biological functions that prefer a specific type of interaction. In addition, when considering higher-level biological and cellular processes, do various types of interactions segregate, or do they cooperate? Finally, we analyze whether the regulation of an interaction reflects the structural states of its constituent proteins, and propose that the disordered state and the transition to an ordered state themselves present additional regulatory switches for protein–protein interactions. We also explore how this mechanism differs for coupled folding and binding and mutual synergistic folding, and the possible mechanism of action for protein regions participating in both processes in the context of cell regulatory subnetworks.

Results

Interplay between folding and binding is reflected in the amino acid composition

Three protein interaction categories were considered in the analysis based on how constituent protein chains reach a structured state. These include proteins going through autonomous folding

and independent binding (where all proteins involved are ordered, marked as ord/ord complexes), coupled folding and binding (where one partner is an IDP binding to an ordered partner, marked as dis/ord complexes), and mutual synergistic folding (where all interacting proteins are disordered, marked as dis/dis interactions). A fourth category was also considered, composed of IDPs that presumably do not adopt a structure via interactions.

Fig. 1 and Table S1 show the calculated sequence properties for each group. As similar residues often substitute each other in homologous proteins with little

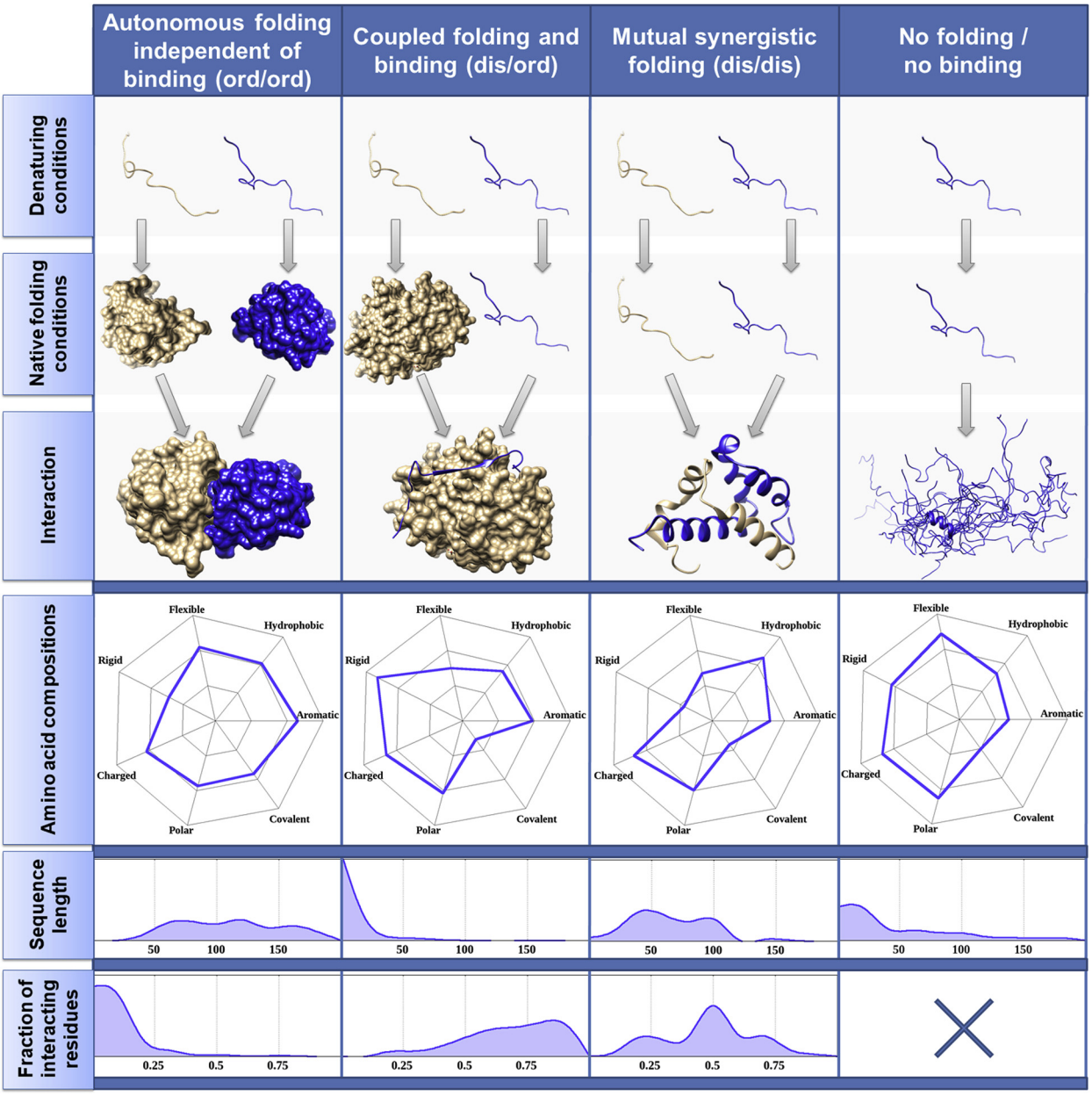


Fig. 1. Sequence properties of proteins based on the relationship between their folding and binding. Columns mark the four basic ways a protein can reach a structured state. Radar charts show the relative amino acid content compared to the human proteome (gray), where each line represents a 2-fold change (log2 radar plots). Amino acids are grouped according to their biochemical/structural properties: hydrophobic (A, I, L, M, V), aromatic (F, W, Y), polar (N, Q, S, T), charged (H, K, R, D, E), rigid (P), flexible (G), and covalently interacting (C). Amino acid groups on the right and left sides of the radars represent residues commonly considered to be stabilizing and structure breaking, respectively [41]. *p*-values quantifying the significance of the difference between value distributions for all four groups are given in Table S1. The lowermost two panels contain the sequence length distribution and the fraction of residues directly involved in the interaction, respectively.

or no effect on protein structure and function, we used a reduced alphabet, grouping aromatic, hydrophobic, polar and charged residues, to concentrate on differences arising from other than this substitution effect. In terms of sequence composition, interacting ordered proteins on average resemble closely the reference residue composition of the human proteome, with a marked decrease in prolines, incompatible with ordered secondary structures. Intrinsically disordered regions (IDRs) not involved in protein–protein interactions (PPIs) conform to the generic view of the typical residue composition of disordered proteins—that is, depleted in stabilizing residues (hydrophobic, aromatic and cysteine residues) and enriched in charged and structure breaking residues, such as prolines and glycines [41]. In contrast, the average sequence compositions of various types of interacting IDRs show distinct differences, strongly reflecting the structural state of their binding partner. IDRs in dis/ord complexes are often highly charged, lack hydrophobic residues, and often contain prolines, possibly to decrease the loss of entropy upon binding to increase binding strength. On the other hand, IDRs in dis/dis complexes are typically more hydrophobic (on par with ordered proteins) and contain very few prolines/glycines (even less than the average for ordered proteins). They are also often highly charged and are devoid of cysteines and aromatic residues, highlighting the diminished role of disulfide bridges and π-stacking in their structure formation. When comparing the distribution of residue types between any two groups, nearly all differences are statistically significant (see Data and Methods and Table S1). The only exception is charge content—the charge distribution of proteins in ord/ord complexes is significantly different from that of all IDPs; however, charge seems to be the weakest distinguishing feature between the two types of interacting IDP segments, as well as non-interacting IDPs. Therefore, the well-known notion of the charge-enrichment of IDPs holds true for all disordered proteins to a similar extent, and other sequence features define subtypes. Although there are clear trends that can discriminate the four groups, variances within residue groups are high, reflecting the heterogeneous nature of proteins in all structural types.

Regarding sequence lengths, ord/ord proteins on average contain a relatively large number of residues, as the presence of a folded domain is incompatible with extremely short sequences. Non-interacting and dis/ord IDRs tend to be significantly shorter, while dis/dis proteins are similar to ordered proteins in terms of sequence length. Taking into consideration the fraction of residues directly involved in the interactions with the partner reveals a new layer of characteristic features. Ordered proteins use only a low fraction of their residues in the interaction, as a large fraction of their residues is buried in their hydrophobic cores. In contrast, IDPs tend to donate a larger relative number of residues to the binding, with several dis/ord IDRs consisting entirely of interacting residues, in line with recent similar analyses [42].

The uncovered characteristic differences in terms of sequence properties imply different binding modes for the three studied interacting groups. The differences between the length and the interacting fraction of the affected protein regions hint at basic structural differences, motivating a deeper structural analysis of the bound structures.

The bound conformation reflects the structural state of the interactors

The structures IDPs and ordered proteins adopt upon binding to a partner were analyzed (see Data and Methods), with a focus on secondary structures, molecular surface areas, atomic contacts and predicted interaction energies (Fig. 2, Table S2).

Ordered proteins show a relatively balanced composition of helical and extended secondary structures, and residues outside periodical secondary structures connecting them. Compared to this balanced structural makeup, structures of bound IDPs show characteristic differences. Dis/ord proteins generally lack periodic secondary structures and adopt irregular conformations. In contrast, dis/dis IDRs show a pronounced preference for helical structures, partially attributable to coiled-coil structures, while exhibiting a low β-structure preference, similarly to dis/ord proteins. It is worth noting that the use of DSSP for secondary structure determination might give a lower estimate for helices and β structures for NMR structures, primarily affecting the apparent secondary structure content of dis/dis and dis/ord complexes (with these data sets containing 25%–26% NMR structures) as opposed to ord/ord complexes (with only 7% of the structures being NMR). However, this bias cannot be responsible for the huge overrepresentation of irregular structures in dis/ord IDPs and only strengthens the reliability of the marked overrepresentation of helical structures in dis/dis IDPs.

The composition of molecular surfaces is primarily dictated by the hydrophobic effect arising from interactions with the surrounding solvent [43]. In the bound form, all three types of proteins have nearly identical hydrophobic/polar (H/P) ratio of solvent-accessible surface areas (SASA)—they all exist in the same aqueous environment. This is also reflected in the relatively high *p*-values obtained from the assessment of the statistical difference of SASA value distributions (Table S2). However, their interfaces show marked differences with hydrophobicity playing a more important role for the binding of IDPs. In contrast, buried surfaces are typically more polar; hence, hydrophobic and polar surfaces are generally made inaccessible due to inter- and intramolecular interactions, especially for dis/dis complexes.

The relative sizes of different molecular surfaces are also highly distinctive. IDPs on average utilize a

			Ord/ord	Dis/ord	Dis/dis
Secondary structures	Helical		0.329	0.178	0.618
	Extended		0.253	0.104	0.113
	Irregular		0.418	0.718	0.268
Molecular surfaces	Accessible surface area	H	0.565	0.565	0.557
		P	0.435	0.435	0.443
	Interface	H	0.618	0.671	0.752
		P	0.382	0.329	0.248
	Buried surface	H	0.591	0.391	0.490
		P	0.409	0.649	0.510
	Interface/total surface		0.042	0.357	0.170
	Buried/total surface		0.644	0.149	0.474
Atomic contacts	Interchain contacts	H-H	0.492	0.529	0.624
		H-P	0.408	0.385	0.315
		P-P	0.100	0.086	0.060
		B-B	0.087	0.096	0.071
		B-Sc	0.384	0.400	0.345
		Sc-Sc	0.529	0.504	0.585
	Intrachain contacts	H-H	0.458	0.364	0.377
		H-P	0.439	0.491	0.496
		P-P	0.103	0.145	0.127
		B-B	0.276	0.310	0.390
		B-Sc	0.393	0.430	0.406
		Sc-Sc	0.331	0.259	0.204
	Ratio of inter/intrachain contacts		0.102	0.822	0.352
Interaction energy	Total interaction energy/residue		-0.534	-0.352	-0.697
	Fraction of energy from interchain interactions		0.059	0.716	0.638

Fig. 2. Normalized average structural properties of proteins as a function of their folding and binding process. Columns mark the different interaction groups. Abbreviations: H, hydrophobic; P, polar; Bb, backbone; Sc, sidechain. Interaction energies are expressed in dimensionless arbitrary energy units derived from statistical potentials. For values quantifying the significance of the difference of value distributions for each feature between every pair of interaction groups, see Table S2.

much higher fraction of their molecular surfaces for interactions compared to that of ordered proteins, and this trend is most pronounced for dis/ord IDRs. Buried surfaces show an inverted trend, with IDPs burying only a small fraction of their surface when bound to ordered partners, in contrast to synergistically folding IDPs and ordered proteins. Apart from the comparison of SASA compositions, the differences in molecular surface parameters between the three groups are highly significant (Table S2).

In terms of atomic contacts, interactions between hydrophobic atoms aid interchain interactions, while hydrophobic–polar contacts play a major role in intrachain interactions and this trend is more pronounced for IDPs. Interchain interactions are primarily mediated through side chains, and the statistical differences between the ratio of backbone/sidechain-mediated interchain interactions are relatively modest (Table S2). Intrachain interactions, however, are evenly formed by sidechain and backbone atoms in the case of ordered proteins. In contrast, backbone atoms play a clearly more important role for IDPs. The ratio of interchain and intrachain contacts clearly shows that IDPs utilize their residues more efficiently for binding the partner protein. As dis/ord IDRs are usually shorter (see Fig. 1), a biologically meaningful stability has to be established by a small number of interacting residues. Although dis/dis IDRs display a larger number of intrachain interactions, they are still more heavily dominated by the interaction with the partner, compared to ordered proteins.

The properties and relative extents of calculated surfaces and contacts all contribute to the overall stability of the resulting complex. In order to assess this stability, interaction energies were calculated based on residue-level statistical potentials (see Data and Methods). According to the energy calculations, dis/dis protein complexes are the most tightly bound systems on average. Ordered complexes display comparable stabilizing per residue energies; however, the per residue stabilizing energy of dis/ord IDRs is significantly weaker, possibly corresponding to the prevalence of more transient interactions. The relative energetic weight of the interaction between subunits in the overall stability is low for ord/ord complexes, but over 10- and 12-fold higher for dis/dis and dis/ord complexes, respectively. The statistically most distinctive features between the three groups are the ratios of inter- and intrachain interactions, and interaction energies, as the *p*-values obtained from the comparison of the distribution of any of these values between any two groups are highly significant ($p < 2.2 \times 10^{-16}$, Table S2).

Various interactions mediate different biological functions with differential localization

It is known that the functional repertoire of IDPs, in general, is characteristically different from that of ordered proteins [4]. We extended the study of protein functions by analyzing the characteristic processes conveyed by the three types of interactions. Functional annotations were based on Gene Ontology (GO) terms describing biological processes (see Data and Methods).

In order to make GO annotations directly comparable, a reduced set of manually chosen terms were selected that describe the most important biological processes (see Data and Methods and Table S3). All GO annotations collected for the interactions were mapped to this reduced set termed PPI GO Slim (see Data and Methods). Terms in PPI GO Slim range from high-level, generic processes (such as development), to more specific, cellular processes through which the high-level functions are realized (such as regulation of gene expression). The most commonly occurring PPI GO Slim terms for all three classes of interactions are shown in Fig. 3, also marking the statistical differences between occurrence for each term between the three classes, with exact *p*-values shown in Table S3. Generic high-level processes, such as cell communication, morphogenesis and development, are typically executed via a large number of carefully coordinated interactions from all three interaction classes, and there are usually no strong preferences for any interaction type. However, certain processes show specificity toward specific interaction types. For example, the maintenance of homeostasis or transport processes are primarily executed through ordered protein interactions. In contrast, DNA damage response, cell cycle regulation or membrane organization is more heavily dependent on interactions mediated by IDPs.

The analysis of molecular-level specific subprocesses shows more pronounced distinctions between different interaction types. Protein-related processes, such as the control of catalytic activity, proteolysis, folding and phosphorylation, are dominated by ordered interactions. In contrast, DNA-related processes are actuated mostly through IDP-mediated interactions. In addition, there is a separation of functions depending on the structural state of the interaction partner. Functions connected to the information storage function of DNA (often involving direct DNA contact), such as transcription and gene expression, are more characteristic of dis/dis interactions. On the other hand, processes pertaining to the regulation of DNA as a macromolecule, such as DNA repair or chromosome organization, are dominated by dis/ord interactions.

GO annotations were also used to assess the typical subcellular localization of various interaction types (Fig. 4) via a restricted set of cellular component GO terms, termed CellLoc GO Slim (see Data and Methods and Table S3), and the statistically significant differences are marked with arrows and detailed in Table S3. Ordered interactions dominate the extracellular space, cellular interactions and receptors embedded in membranes. In addition, ordered interactions are more often found in various membrane-bounded organelles, such as the ER or mitochondria. The cytosol in general harbors all three types of interactions in large numbers. In contrast, localizations closer to the DNA are progressively more dominated by IDPs: the nucleoplasm is the characteristic location for dis/ord protein interactions, and localizations directly connected to the DNA, such as the DNA packaging complex or the chromatin, are most commonly associated with dis/dis interactions. Other common characteristic locations of IDP-mediated interactions are non-membrane-bounded organelles, such as stress granules or the centrosome, falling in line with the recently recognized importance of IDPs in the organization of such cell constituents [44,45].

Protein disorder extends the biologically relevant sequence, structure and function spaces

The previous analyses have shown that interacting IDPs have distinctly different residue compositions compared to interacting ordered proteins, and this composition reflects the structural state of the binding partner (Fig. 1). In addition, the presence of IDPs in protein interactions is reflected in the bound structure of the resulting complexes, and certain specific functions these interactions mediate. However, these analyses only considered the average

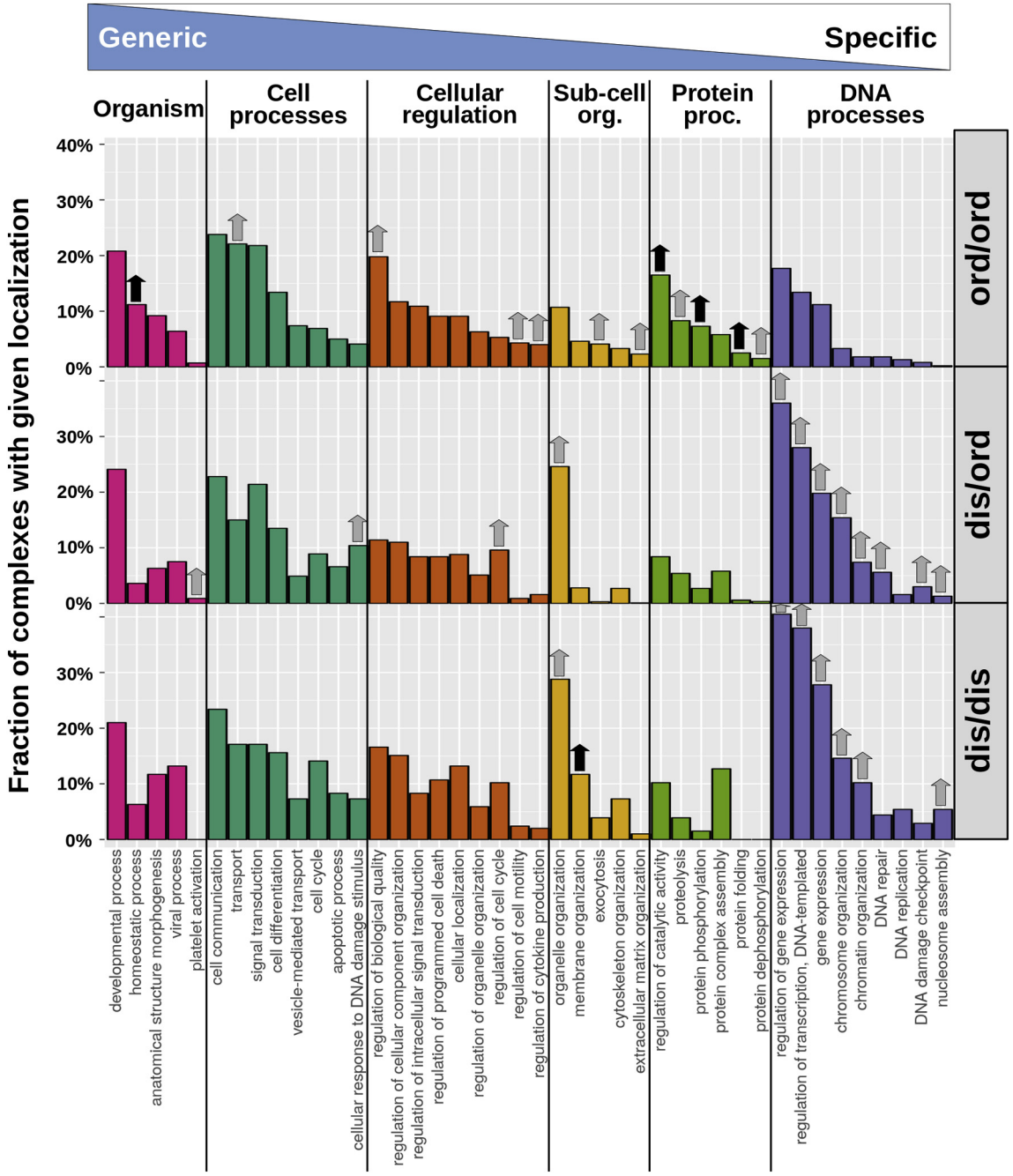


Fig. 3. The frequency of occurrence for PPI GO Slim terms for the three classes of interactions. Left: generic terms, right: specific terms. Only those terms are shown that occur in at least 5% of complexes for any group, or where there is a significant difference between any two pairs of groups. Black arrows mark significant over-representation compared to both other groups, and gray arrows mark significant overrepresentation compared to one of the other groups. *p*-values are reported in Table S3.

values of features, without quantifying the sequential, structural and functional heterogeneity of each interaction class.

To visualize how proteins and interactions from the three interaction classes are distributed in the sequence/structure/function spaces, principal component analysis (PCA) was employed. The three aspects were evaluated separately for all three interaction classes, using the annotations of proteins described in the previous chapters (see Data and Methods). Sequence parameters used are the same as presented in Fig. 1. Structural parameters encompass a redundancy-filtered set of features shown in Fig. 2 (for the filtering criteria, see Data and

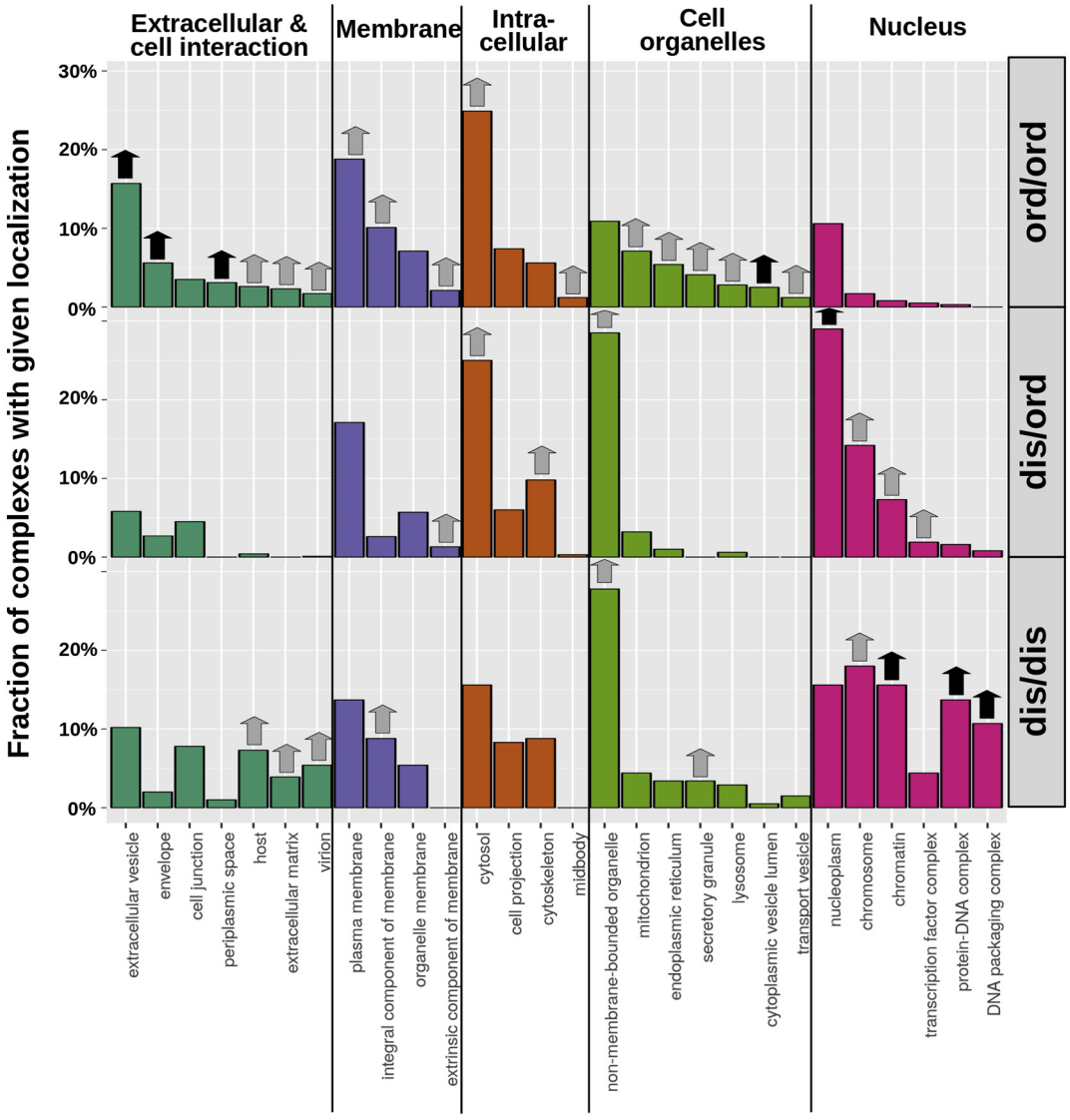
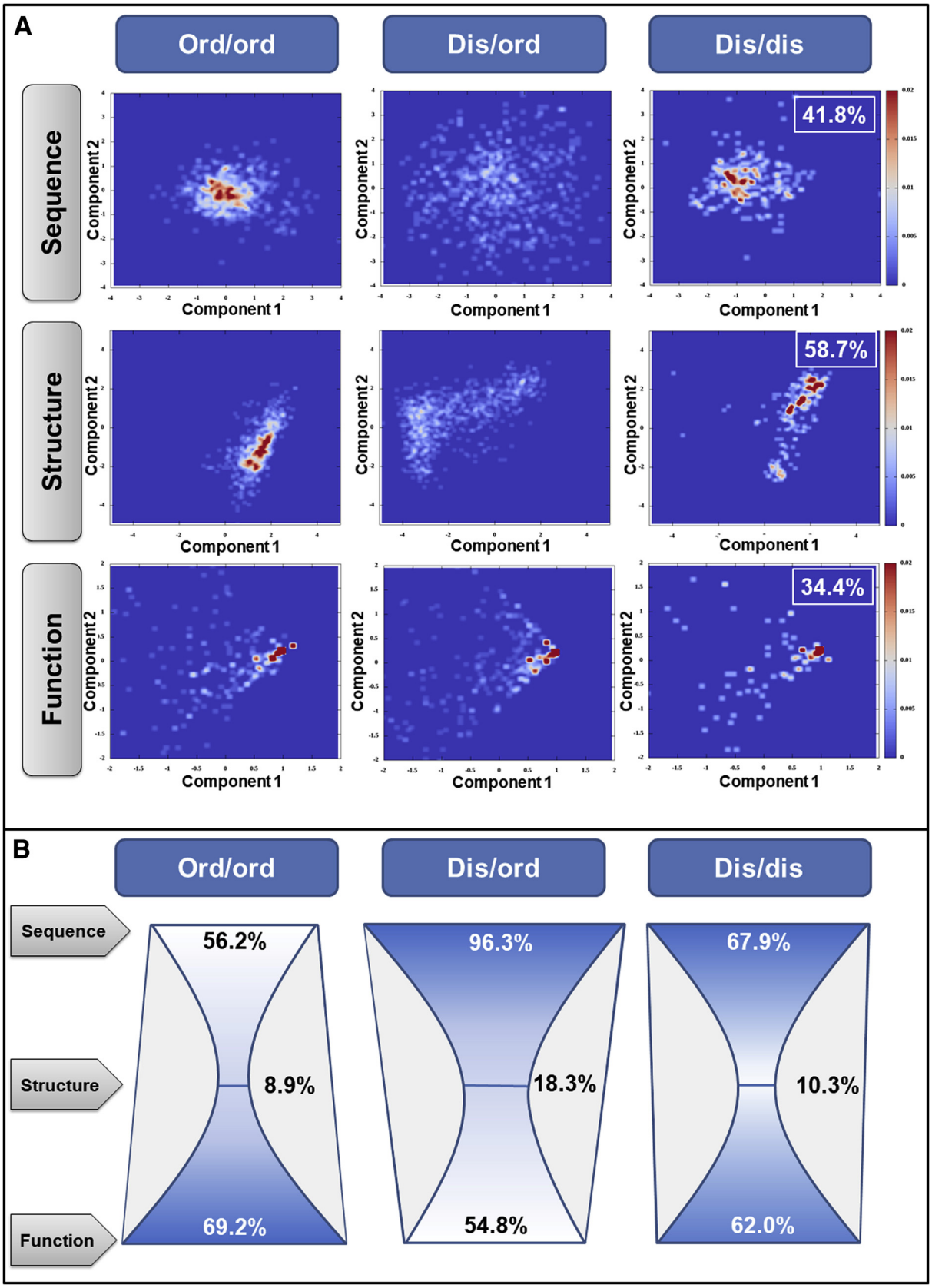


Fig. 4. Characteristic sub-cellular localizations of the three classes of interactions. Only those terms are shown that occur in at least 5% of complexes for any group, or where there is a significant difference between any two pairs of groups. Black arrows mark significant over-representation compared to both other groups, and gray arrows mark significant overrepresentation compared to one of the other groups. *p*-values are reported in Table S3, similarly to the representation in Fig. 3.

Methods and Table S2), while functional features are the generic high-level PPI GO Slim terms, shown in Table S4. Fig. 5A shows the best two-dimensional representation of these spaces, using the first two principal components that carry the highest variances of the data, while Table S4 shows the coordinates and variances explained by all principal components. While the total variance explained by the first two components is fairly low, even the distribution of interactions in this restricted space can highlight basic differences of interaction classes. Furthermore, the fact that a large number of features are needed to represent the full variance of the data (see Fig. S2) justifies previous sequence and structure feature selections, as well as including the fairly high number of GO terms in the PPI GO Slim.

Considering sequence-space distributions, ord/ord and dis/dis complexes show a moderate demarcation, hinting at (at least partially) mutually exclusive, but not radically different residue compositions. The first principal component mainly encodes the presence of charged and hydrophobic residues, while the second component mainly represents the charge and polar content (Table S4). IDPs capable of binding to ordered proteins show a much wider distribution of compatible compositions in these dimensions, overlapping with both dis/dis and ord/ord complexes. Notably, these dis/ord interacting proteins have the shortest


A
Ord/ord
Dis/ord
Dis/dis
Sequence
Structure
Function
41.8%
58.7%
34.4%
Component 1
Component 2
B
Ord/ord
Dis/ord
Dis/dis
Sequence
Structure
Function
56.2%
8.9%
69.2%
96.3%
18.3%
54.8%
67.9%
10.3%
62.0%

length, and therefore, their sequence compositions can be very extreme, often being comprised oft only a handful of types of residues. However, even these cases are biologically functional as evidenced by the interactions they mediate. The shortest such protein segments often harbor linear sequence motifs that can mediate transient interactions with several crucially important domain types, such as SH2, PDZ or 14−3−3 [19]. The distributions in the sequence space show that the biologically useful space of protein sequences is largely extended by IDPs, especially by IDPs involved in dis/ord interactions binding through linear motifs, as these proteins can be functional with highly biased compositions that are incompatible with other protein classes.

Considering structural properties, all three classes occupy a distinct subregion in the space of possible structures, even based on the first two principal components. Both components encode a relatively evenly weighted mix of secondary structure composition, molecular surfaces and atomic interactions (Table S4). Considering both sequence and structure properties, protein regions differ widely based on the intrinsic structure and the structural state of the binding partner. These differences are comparable to the differences between ordered regions and IDRs in general.

Distribution of the three interaction types in the functional space shows a high degree of overlap. The first two components—similarly to the structural ones—cannot be simply interpreted as certain functional subsets, rather they represent overall functional profiles, slightly weighted to discriminate processes characteristic of IDPs (Fig. 3, Table S4). This reinforces the notion of the previous section stating that most high-level biological processes rely on both ordered proteins and IDPs as well, utilizing an interconnected network of their interactions.

We also introduced a quantitative measure to characterize heterogeneity within each interaction class in terms of sequence, structure and functional spaces. The introduced heterogeneity values were defined as the average dissimilarity between two randomly chosen complexes from the same class. Dissimilarity between two complexes was defined based on the hierarchical clustering using various feature vectors as input (see Data and Methods). The so-calculated heterogeneity values lie between 0% and 100%, with 0% corresponding to all complexes being identical (i.e., having identical features) and 100% corresponding to all pairs of complexes being as different as possible. Heterogeneity values are calculated as the geometric mean of these dissimilarities, serving as a conservative measure for the true variance of the studied proteins, largely insensitive to extreme cases. A graphical interpretation of these heterogeneity values is the area covered by each interaction type in Fig. 5A with respect to the maximally available area, but calculated from all principal components.

Calculated heterogeneity values (shown in Fig. 5B) outline a general trend. Regardless of structural state, proteins in general utilize highly variable sequence compositions, to realize a comparatively much narrower set of structures. This bottleneck in the structural space is in agreement with the data presented in Fig. 5A, as the three interaction groups overlap heavily in the sequence and function spaces, but their overlap is quite limited in the structure space; therefore, only restricted areas are available at the structural level for each interaction class. This reduction in complexity at the structure level, however, does not limit functional roles, as the functional heterogeneity of interacting proteins is comparable to their sequence heterogeneity. Apart from general trends, the three classes of complexes show characteristic differences as well. Ord/ord complexes fulfill a wide range of functions using proteins with more restricted sequence compositions. Dis/ord IDRs represent the opposite, using wide variations in composition, but conveying a more restricted range of functions in comparison. In contrast to both classes, dis/dis IDRs show a strikingly similar level of heterogeneity both at the level of sequence compositions used and at the level of biological functions they mediate.

Complexes of IDPs are tightly regulated by post-translational modifications

As shown in our previous functional analyses, all three studied classes of interactions play roles in crucial biological processes. In order for these processes to function correctly, the interactions on which they are built must be precisely regulated. These regulatory mechanisms include control of expression levels, subcellular localization, post-translational modifications, and competing interactions, among others. Previous studies have shown that IDPs are under exceptionally tight regulation [46], yet the interconnection between the structural state of the partner and the mechanisms of regulation is largely unknown. We analyzed post-

Fig. 5. Variability of sequences, structures and functions for complexes from the three interaction classes. (A) Distribution of various complexes considering the first two principal components of the sequence, structure and functional spaces. In each row, the components are the same linear combination of features, and hence, distributions are directly comparable. Insets show the total variance of the data carried by the plotted components. (B) Sequence, structure and functional heterogeneity values calculated for all three classes of interactions.

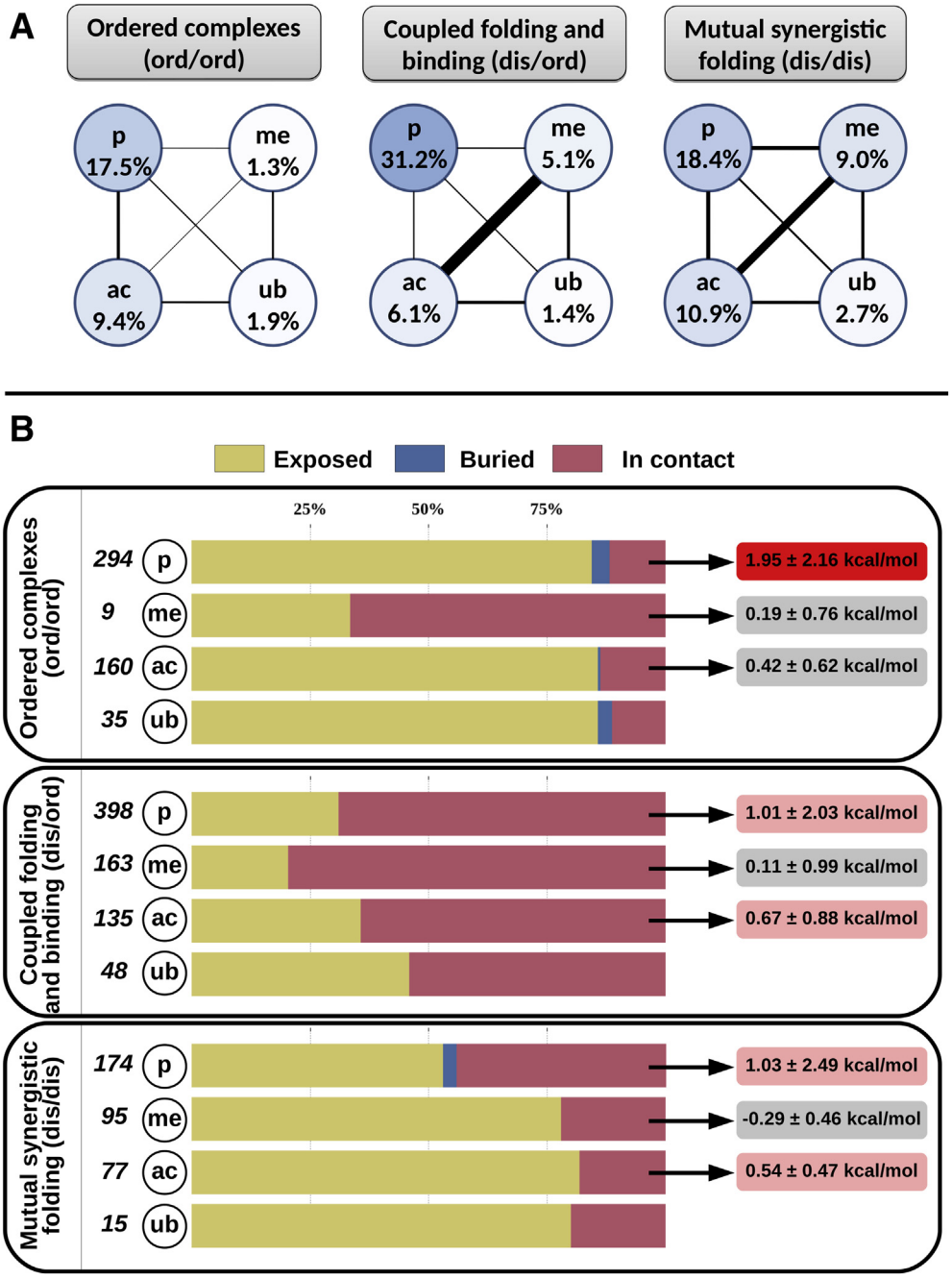


Fig. 6. PTM occurrences in proteins implicated in different interactions. (A) the occurrence of PTMs in interacting proteins (p, phosphorylation; me, methylation; ac, acetylation; ub, ubiquitination). Color depth and percentage values represent the fraction of proteins affected. The width of connecting lines shows the mutual information between the occurrences of PTM pairs (see Table S5 for exact values). (B) Location of PTM sites in the complex structure. Colored bars represent occurrences in the three types of structural configurations (solvent exposed, buried or interface). Values next to circles indicate the amount of PTMs found. Energy values show the mean and the standard deviation of estimated $\Delta\Delta G$ values of introducing the PTMs to interface residues, evaluated with the use of mimetic residues in FoldX (see Data and Methods). Colors indicate the destabilizing effect of the average value (gray, neutral; light red, slightly destabilizing; deep red, strongly destabilizing). As there are no usable mimetic substitutions for ubiquitination, free energy calculations have not been performed for these PTMs.

translational modifications (PTMs), which are the most prevalent regulatory mechanism for interactions, and for which the largest amount of data is available in PhosphoSitePlus [47] and Phospho.ELM [48].

Occurrences of four types of PTMs (phosphorylation, methylation, acetylation and ubiquitination—see Data and Methods and Table S5)—were studied for the three classes of interactions (see Fig. 6A). All four types of PTMs are present on both ordered and disordered interacting proteins, with a pronounced accumulation of known PTM sites in IDRs. In addition, these IDPs not only harbor more PTMs, but the occurrence of these modifications is far more coordinated, showing a higher level of mutual information than in ordered proteins. For dis/ord IDRs, methylation and acetylation occurrences show an extremely high cooperation. This is mainly due to the large number of N-terminal histone tails in this group, harboring the acetylation and methylation sites for epigenetic markers [49], bound to ordered domains. In contrast, for dis/dis interactions, all four PTMs seem to be highly cooperative in comparison to other interaction classes.

The structural locations of PTMs offer insights into the mechanistic effects they have on the binding event (Fig. 6B). Most of the PTMs located within ordered proteins are enriched on the solvent-accessible surface of domains, outside of the interface (except for methylations, where the low amount of data may result in high noise). These PTMs are not expected to directly modulate the binding, although they might have an indirect effect on the interaction (e.g., through controlling the availability of the protein via localization or degradation signals). In addition, methylations and acetylations found in the ordered interfaces seem to have a moderate estimated change in the free energy of the complex structure when modeling the PTM with mimetic residue substitutions, that is, calculating the estimated change in stability when introducing a mutation that mimics the presence of a modified residue. In contrast, ordered interface phosphorylations seem to have a large destabilizing role, possibly capable of switching the interaction on or off.

In contrast, PTMs in dis/ord type IDRs generally seem to influence the binding event in a more direct fashion. In these cases, all studied types of PTMs preferentially cluster in residues involved in the binding. While the estimated free energy change of the introduction of these PTMs is rather moderate, the sheer number of these PTMs offers a large amount of control over the binding event.

PTM sites in dis/dis complexes seem to utilize a different molecular technique. In these proteins, targeting residues that are buried in the complex structure would be a logical approach, as these modifications would probably heavily disrupt the binding. Incidentally, this is the only interaction class where this would be a feasible approach, as buried residues in ordered proteins are generally not accessible, and dis/ord IDRs generally contain no buried residues. Interestingly, buried PTM sites seem to hardly occur in dis/dis proteins either. Instead, these proteins feature a moderate fraction of interface PTM sites, roughly halfway between those in ord/ord and dis/ord proteins, with the estimated free energy change introduced by these PTMs being on par with values calculated for ord/ord interactions.

These results show that PTMs play a major role in the modulation of IDP-mediated interactions. PTMs in solvent-accessible residues of IDPs—as opposed to ordered proteins—can heavily affect the binding through the tuning of local flexibility and predisposition for adopting a stable structure [50]. Considering interface PTMs, the structural state of the partner has a huge bearing on the basic molecular technique of this regulatory mechanism. IDPs binding to ordered partners are in general regulated through a high number of interface PTMs, while IDPs interacting with other IDPs are targeted by a more restricted number of PTMs, albeit occurring in a more coordinated fashion between different types of PTMs.

Discussion

According to most current descriptions, protein order and disorder are treated as binary structural features. In the past 15 years, a more refined view of protein disorder started to emerge, emphasizing the existence of different flavors of disorder and a continuous spectrum of protein flexibility. However, a detailed understanding of how IDP sequence characteristics are related to IDP structure and function remained elusive. Recent analyses have begun the systematic exploration of the relationship between IDP sequence properties and various types of disorder [42]. In this work, we focused on various types of protein complexes taking advantage of the emergence of recent IDP interaction databases. We analyzed the properties ord/ord, dis/ord and dis/dis complexes (Fig. 1). We found that the three categories exhibit markedly different characteristics in terms of sequence, structure and—to a slightly lower degree—function. As the three interaction categories represent three different scenarios of how folding and binding intertwine, the observed differences mirror the different strategies employed to balance the various biophysical factors driving folding and binding. Ordered protein complexes complete their folding before their interaction, IDPs bind to their ordered partner by adapting to the steric constraints presented by the structured domain,

while IDPs involved in synergistic folding form the core together and rely significantly on intrachain interactions for mutual stability.

Our results also highlighted that there is a pronounced difference in the sequence composition of proteins involved in various interaction types. One of the most well-known distinguishing features of IDPs is their low hydrophobic content and high net charge, and this single observation opened the way for the construction of early disorder prediction methods [51]. However, sequence composition is a function of IDR length [52] and experimental determination method [53], or can be highly biased in certain functional sites, such as histone tails [54] or polyQ regions [55]. Here we showed that the hydrophobicity and proline content are also highly dependent not only on the involvement in protein interactions but also on the presence and structural state of the partner. IDPs that bind to ordered partners utilize the highest fraction of prolines on average to reduce the entropic penalty of the binding. In contrast, for IDPs forming complexes via mutual synergistic folding, prolines are restricted to terminal segments of secondary structural elements. These IDPs are also prime exceptions to the low-hydrophobicity IDP rule: the hydrophobic core of the complex must be encoded in these sequences, as hydrophobic collapse happens during the binding event.

The dependence of IDP features on the binding partner is also represented in their bound structures. Although helical binding [56] and β-augmentation [57] represent possibly the two most well-known binding modes in coupled folding and binding, analysis of the complete available set of such complexes shows that these IDP segments overwhelmingly prefer irregular conformations in their bound forms. Altogether, the uncovered structural differences are even more pronounced than the sequential characteristics of the respective interaction classes; while certain sequence compositions are compatible with multiple modes of binding, the three types of interactions are very clearly separated in the structural space.

This indicates that IDPs vastly widen the range of not only protein sequences but also protein structures that are compatible with a biological function. This notion also indicates that various modes of binding might be favored or required for certain processes and molecular mechanisms. In general, interactions from the three studied classes often contribute to the same high-level biological functions (Fig. 3). However, considering more specific processes highlights the increased importance of IDP interactions in mediating functions related to DNA. In addition, this trend also becomes apparent when focusing on subcellular localization (Fig. 4), with ordered interactions dominating the extracellular space and the cytosol, while IDP interactions are enriched in the nucleus. Furthermore, this "disorder attraction" of the DNA also differentiates between dis/ord and dis/dis interactions, with the former pertaining to DNA regulation and the latter to DNA information content. This discrimination is also reflected in the increased importance of methylation in the regulation of synergistically folding complexes (Fig. 6), with methylation being primarily connected to the information access control of DNA. The deep connection between biological function and sequence/structure properties can be best demonstrated through select examples.

Fig. 7 shows representative examples of the three classes, where the sequence compositions are the closest to the group averages. These examples highlight not only the interdependence of the sequential and structural characteristics but also illustrate how various modes of binding might be favored or required for certain processes and molecular mechanisms. The properties of ordered complexes are demonstrated through nuclease A (NucA) forming a tightly bound complex with its inhibitor, NuiA (Fig. 7A). The stable monomeric form for NucA is a prerequisite of the enzymatic function. NuiA is able to specifically recognize and to tightly bind to NucA owing to its native structure that is very close to the bound conformation, increasing affinity to the picomolar range [58]. The prime example for an IDP undergoing coupled folding is the LC3-interacting region (LIR) of FYCO1 (FYVE and coiled-coil domain-containing protein 1) interacting with the ordered ubiquitin-like domain of autophagy-related protein LC3 A (Fig. 7B). As this interaction directly links autophagosomes with the microtubule-based kinesin motor, it has to be fast and reversible. This is made possible by the ordering of the LIR upon interaction, facilitating a highly specific interaction with a weaker binding (with K_d being in the low micromolar range) [59]. These structural properties allow the unique molecular binding mechanism resulting in transient, yet specific binding. The third interaction scenario is represented by the DNA-binding domain of trp repressor that forms homodimers composed of two identical IDRs, forming a structurally malleable complex [60] (Fig. 7C). This malleability enables trp repressor to adapt and bind to three different operator sites—a function largely incompatible with the presence of protein order. The molecular properties of complexes formed exclusively by IDPs resemble those of both other two interaction classes. On one hand, the constituent proteins feature a large fraction of hydrophobic residues in order to collaboratively form a stabilizing core. On the other hand, they also have to exhibit extreme plasticity to mutually adapt to each other structurally.

Interestingly, these interaction types are not always segregated. In the case of the transactivation domain of p53, the corresponding IDR can function

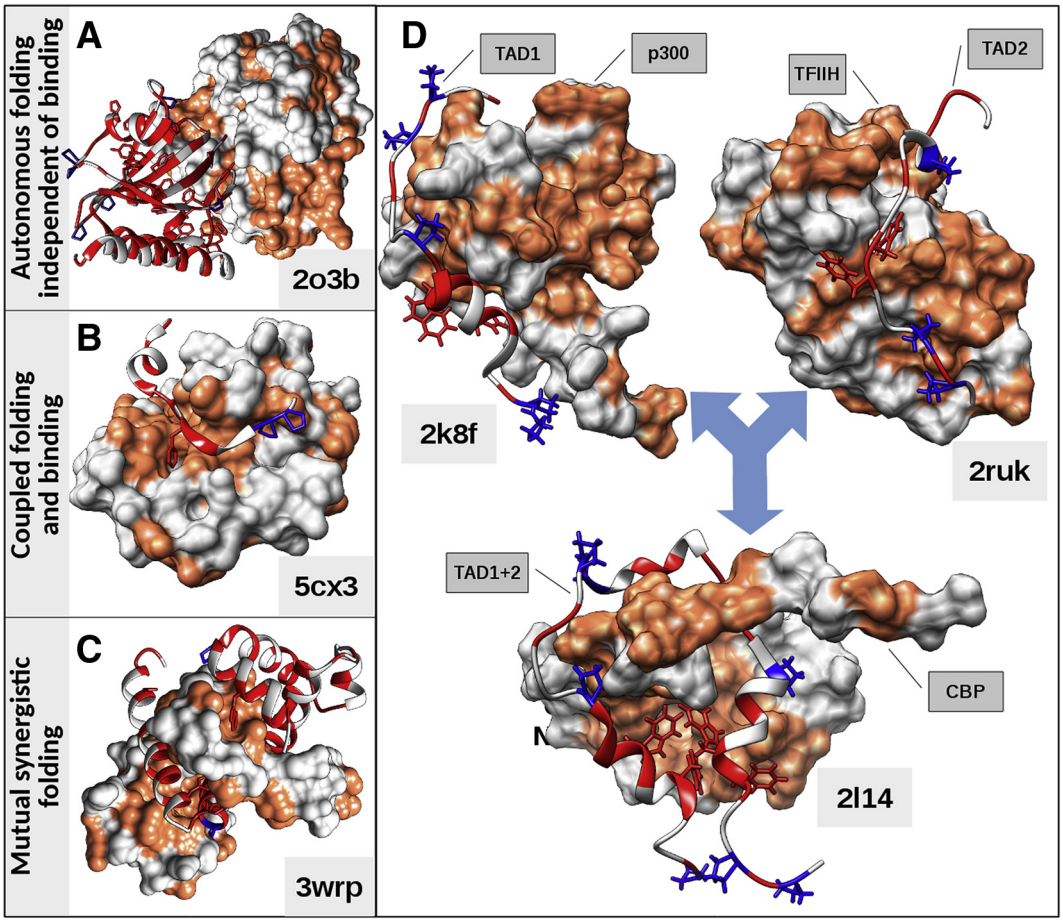


Fig. 7. Representative examples of the three classes of interaction mechanisms. The analyzed protein is shown in ribbon representation, while the partner is shown as a surface. Hydrophobic residues are colored red (orange for surfaces), and aromatic sidechains are shown in stick representation. Prolines are shown in blue sticks. (A) Interaction between the nuclease NucA and its inhibitor NuiA. (B) The LC3-interacting region of FYCO1 bound to the ubiquitin-like domain of autophagy-related protein LC3 A. (C) Homodimer of the DNA-binding IDR region of trp repressor. (D) The two transactivator regions (TAD1 and TAD2) bound to ordered domains from p300 and TFIIH, respectively (top), and TAD1 + TAD2 bound to an IDR of CBP (bottom).

as two independent domain-recognition IDRs in tandem (TAD1 and TAD2), or can work together as a single binding site that recognizes a disordered region of the CREB-binding protein (CBP) (Fig. 7D). This is possible because both TAD1 and TAD2 interact with domains mainly via hydrophobic interactions, yet both binding sites are small enough to avoid mutual synergistic folding on their own due to their limited size. However, the two TADs working in synergy surpass this size boundary, and while they are still small enough not to fold on their own, they are large enough to be able to form a stable structure with a suitable IDP partner.

At the network level, the cooperation between various binding modes is a general feature of biological regulation. The dual nature of p53 TAD is the key to understanding the interaction and regulatory sub-network of N-terminal p53/CBP, which lies at the intersection of a range of critical regulatory and signaling processes (Fig. 8). p53 is the main tumor suppressor in multicellular organisms capable of initiating apoptosis upon irreparable DNA damage [61] and is activated via the interaction of both TADs with CBP in mutual synergistic folding. This interaction is mutually exclusive with a range of other interactions between one or both p53 TADs and ordered proteins. To add further complexity, CBP is also involved in interactions with other partners competing for the same binding site. As various interactions of CBP and p53 mediate different biological functions, these molecule-level interactions actually encode a switching mechanism for several cellular processes, such as apoptosis, autophagy, DNA repair or proliferation. Therefore, the switching of p53 from a mutual synergistic binding mode to coupled folding and binding

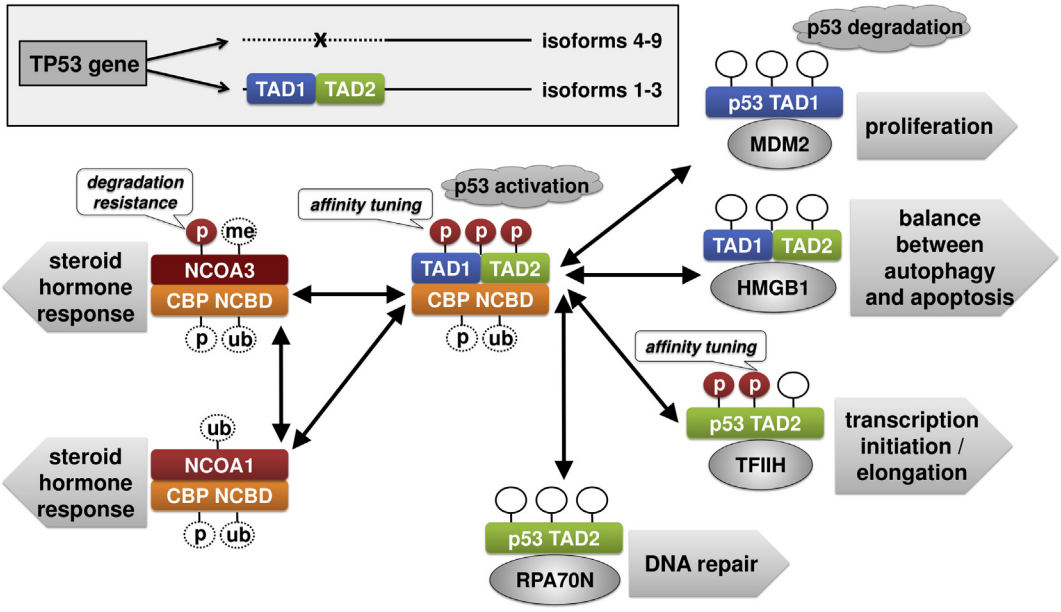


Fig. 8. Regulatory sub-network of the dual nature TAD1/2 segments of p53. Interactions and their regulation in the p53/CBP regulatory sub-network. Color boxes represent IDP binding sites, and gray ovals represent ordered proteins. Dashed circles represent PTMs with unknown functional effects identified in high-throughput measurements.

corresponds to switching between various biological processes. This is coupled with additional regulatory steps, such as PTMs affecting binding strength [62,63], and by alternative splicing with six of the nine existing p53 isoforms completely lacking the TAD. Furthermore, p53 interactions and hence function are heavily dependent on localization [64], providing an additional layer of regulation.

The uncovered differences between various types of interactions in terms of sequence, structure, function and regulation present the first step in the basic understanding of how the interplay between protein folding and interaction modulates critical properties of the resulting complexes. While the discussed examples represent only a limited fraction of the human interactome, they already outline how the inherent structural properties of interacting protein segments determine the sequence and structure properties pertaining to their binding events. Interacting regions embedded in IDPs are not only heavily regulated by PTMs and other molecular processes, but their intrinsic structural properties present an additional bona fide regulatory mechanism. Our present knowledge of this regulation is limited by the fact that while IDP-mediated interactions are expected to be present in the human interactome on the scale of hundreds of thousands [65], we only have detailed structural information of only a fraction of a percent. However, a recent large-scale experimental work has indicated the importance of the preferential modulation of IDP solubility and thermal stability in various cell cycle phases, and has highlighted phosphorylation as a major contributor [66], giving support to the idea that intertwined structure-based regulatory mechanisms play critical roles in the living cell. The further appreciation of this regulation will hopefully contribute to the ignition of the targeted research of this as-of-yet largely unexplored region of the protein interactome.

Data and Methods

Sequence and interaction data sets

Data for the three studied classes of protein–protein interactions were collected by selecting appropriate structures from the PDB. Hence, for every studied protein or protein complex, there is high-quality structural data. All of the interacting protein chains were annotated with reliable order or disorder information in order to assess the interaction class (ord/ord, dis/ord or dis/dis).

Complexes formed by coupled folding and binding (dis/ord complexes), and mutual synergistic folding (dis/dis complexes) were downloaded from the DIBS [37] and the MFIB [38] databases, containing 772 and 205 protein complexes, respectively. Entry DI1110004 from DIBS was removed as it refers to a PDB structure that has been marked as obsolete. All interacting protein chains were mapped to UniProt sequences. Only those sequence regions were kept that had corresponding ATOM coordinates in the PDB structure and in case of NMR structures, where the RMSD between locally aligned models do not exceeds 3 Å, assessed with CYRANGE [67]. In total, the 772 used DIBS entries contain 555 distinct disordered protein regions (Table S1) considered for

sequence analysis, and 772 bound IDR structures considered for structural analysis (Table S2). The 205 used MFIB entries contain a total of 256 distinct disordered protein regions (Table S1) considered for sequence analysis, and 283 bound IDR structures considered for structural analysis (Table S2).

Complexes formed by ordered proteins were taken from the PDB by selecting structures containing dimeric protein interactions, as evidenced by the number of proteins (considering biomatrix transformations), PISA records and the authors' manual assignments. Only structures solved by NMR or x-ray were considered, and in the latter case, the resolution had to be better than 5 Å. Two protein chains were considered to be in interaction if they have at least five atom pairs in contact. Only those structures were kept that did not contain any non-protein entities and where both interacting proteins consist of a single domain without any fragments, as defined by CATH [68]. Complexes were discarded if they contain more than 10 disordered residues, defined as non-terminal residues that have no corresponding ATOM records, and residues in NMR structures, where the locally aligned models show a larger than 3Å RMSD, assessed with CYRANGE. The resulting set of ordered domain interactions was subjected to the same sequence-based redundancy filtering, as the data in DIBS and MFIB. Complexes, where the constituent protein chains show a high degree of pairwise similarity (i.e., they belong to the same UniRef90 cluster, and their respective overlap is over 70%), were filtered, and only the complex with the best resolution was kept. Apart from the quality of the structures and sequence redundancy, no other filters were applied to achieve the best coverage of complexes with respect to size, sequence composition, taxonomy and other properties. The remaining 614 structures contain 629 distinct ordered protein regions (Table S1) considered for sequence analysis, and 688 bound structures considered for structural analysis (Table S2).

For all three interaction data sets, Table S6 contains all relevant IDs and accessions, including DIBS and MFIB IDs, PDB IDs with chain IDs, and UniProt accessions together with region boundaries.

Sequences of IDPs devoid of interacting regions were generated from DisProt [69] records by removing sequence regions that are present in either DIBS or MFIB. Remaining sequences shorter than five residues were removed. The resulting set of 1045 sequence regions is shown in Table S1. The human proteome containing 71,567 protein sequences was downloaded from UniProt [70] on August 11, 2017.

Sequence features

After considering various type of classifications, we found that the following amino acid categories are the most descriptive for distinguishing protein groups: hydrophobic (A, I, L, M, V), aromatic (F, W, Y), polar (N, Q, S, T), charged (H, K, R, D, E), rigid (P), flexible (G) and covalently interacting (C). Average content and standard variances for all 20 amino acids measured in various protein groups supporting this classification is shown in Fig. S1. We observed that the standard deviations calculated in the reduced alphabet are considerably lower than those calculated for the 20 individual residue types. This indicates that the used grouping probably removes a large portion of variance that is due to the trivial fact that residues with similar physicochemical properties often substitute each other in protein sequences without affecting structural and functional properties. All protein sequence compositions were calculated on the reduced alphabet. When comparing proteins from the three interaction classes, compositions were calculated for one protein alone.

Structure features

Structural features of proteins were calculated from their bound structures. NMR structures were assessed with the CYRANGE program [67], and residues outside the core regions were treated as disordered. Secondary structure assignment was performed by DSSP [71] using a three-state classification distinguishing helical (“H,” “G,” “I”), extended (“B,” “E”) and irregular (“S,” “T,” unassigned) residues. In the case of NMR structures, the first model was considered.

Molecular surfaces were calculated using Naccess [72]. SASA was defined by the Naccess absolute surface column. Interface is defined as the increase in SASA as a result of removing interaction partners from the structure. Buried surface was calculated by subtracting interface area and SASA from the sum of standard surfaces of residues in the protein chain. Thus, interface and buried surfaces represent the area that is made inaccessible to the solvent by the partner(s) or by the analyzed protein itself. All calculated areas were split into hydrophobic (H) and polar (P) contributions based on the polarity of the corresponding atom. Polar/hydrophobic assignations were taken from Naccess.

Contacts were defined at the atomic level. Two atoms were considered to be in contact if their distances are shorter than the sum of the two atoms' van der Waals radii plus 1 Å.

Interaction energies for residues were calculated using the statistical potentials described in Ref. [73]. These interaction potentials were demonstrated to adequately describe the energetic features of interacting proteins, including IDPs [20]. These dimensionless energy-like quantities behave like true

energies in being additive and negative/positive values corresponding to stabilizing/destabilizing contributions; however, their absolute values have no direct physical interpretation and should be interpreted in comparison to one another only.

When comparing proteins from the three interaction classes, structural parameters were always calculated for only a single protein. When using structural parameters as input for PCA and hierarchical clustering (used for calculating heterogeneity values), only a restricted set of poorly correlated structural features were used to reduce bias. These features encompass the fraction of all three secondary structural elements; the hydrophobic fractions of SASA, the interface area, and the buried area; the ratio of the interface and the total surface; the ratio of the buried surface and the total surface; and the fraction of backbone-backbone interactions for all inter- and intrachain interactions. Correlation values between structural parameters are given in Table S2.

Functional annotations

Biological functions and subcellular localizations were taken from the DIBS and MFIB databases in the forms of GO terms. Annotations for ordered complexes were generated from the GO annotations of constituent proteins (taken from UniProt-GOA) using the approach described in DIBS/MFIB (http://dibs.enzim.ttk.mta.hu/help.php). We assigned GO terms to complexes that are assigned to both interacting partners. In order to expand the number of annotations, "matches" between GO terms are defined permissively. Two terms are considered to be matching if they are the same, or if they are in children/ancestor relationship and their distance in the ontology is no more than two steps. This definition of complex GO terms is fully consistent with GO definitions for dis/dis and dis/ord complexes, making them directly comparable.

PPI GO Slim and CellLoc GO Slim were created manually from the "biological process" and "cellular localization" namespaces of GO, by selecting terms that are either assigned to studied complexes or are ancestors of such terms. PPI GO Slim was assembled to cover a wide range of possible biological functions from high-level organismal terms (such as developmental process) to low-level molecular terms (such as proteolysis). The terms contained in PPI GO Slim and CellLoc GO Slim are shown in Table S3.

PCA

PCA was used to map the multidimensional sequence (7 dimensions) and structure feature vectors (11 dimensions) of proteins (listed in Tables S1 and S2) into two dimensions for visualization (Fig. 5). PCA was performed using the princomp module in the R statistical computing environment (version 3.3.1) [74]. PCA was also used to visualize the studied complexes in terms of variance in functions. For this, each complex was assigned a 23-element vector, where each element marks the number of GO terms that can be mapped to each of the 23 selected high-level, generic cellular/organismal processes of the GO PPI Slim, which describe cellular/organismal level biological processes of the proteins (see Table S4). The number of GO PPI Slim terms was reduced to avoid having a drastically higher dimensionality for representing function compared to sequence and structure. In all three cases, PCA was performed by first merging the data points in the three data sets, and then calculating the principal components. Biplots for the 7 and 11 sequence/structure parameters are shown in Figs. S2 and S3. The calculated components, expressed as a linear combination of input features, are shown in Table S4.

Heterogeneity

Heterogeneity values aim to quantify the total variability of complexes in each class with respect to a given aspect. For example, the sequence heterogeneity of ordered complexes describes the average variability in the sequence composition of the constituent proteins, compared to the maximal possible variability. Large heterogeneity values indicate that several very different sequence compositions are compatible with this interaction type. Low values would indicate that this type of interaction requires a very specific composition. As the heterogeneity values are calculated using the same features for all three classes, they are directly comparable.

To obtain numerical heterogeneity values, two separate hierarchical clusterings were employed using the sequence and structure features (the 7 sequence and 11 structure features used for PCA) as input for the Ward.2 algorithm in R, using Euclidean distances. Heterogeneity values are determined using the linkage distances between various proteins, based on which we calculated dissimilarity values. Dissimilarity of two proteins *i* and *j* is defined as $d_{ij} = \frac{L_{ij}}{L_{\max}}$, where L_{ij} is the linkage distance between the two proteins in the obtained tree, and $L_{\max}$ is the maximal linkage distance (i.e., the linkage distance between the root and the leaves of the tree). Heterogeneity is defined as the geometrical mean of dissimilarity values between all protein pairs in a given class. This definition ensures that heterogeneity is a conservative measure in the sense that outliers cannot dominate the final value. On one hand, even if a protein has a sequence composition or structure very dissimilar to

all other proteins from the same group, its dissimilarity is maximized as 1, and the number of times this value is counted in the mean scales linearly with the number of proteins (N), while the other dissimilarity values in the group scale with N^2. On the other hand, calculating geometric mean instead of arithmetic mean suppresses the effect of extreme values.

In the case of functional heterogeneity, the tree obtained from the hierarchical clustering was replaced by the GO ontology tree. Distances between terms that are in a parent/child relationship were defined as 1. Dissimilarity between two complexes was defined based on their most similar GO term pairs. Let t_i be the GO biological process terms of complex A and t_j be the GO biological process terms of complex B. For each t_i, we choose a t_j pair, for which their distances in the ontology are minimal; that is, let t^* be the most specific (low level) term in the ontology that is the common parent of both t_i and t_j. The distance between t_i and t_j is the distance between t_i and t^*, plus the distance between t_j and t^*. Next, we normalize this distance with the maximal possible distance that could be between t_i and t_j, that is, the sum of the distances of the two terms and the ontology root ("biological_process"). The dissimilarity between two complexes in the functional sense is defined as the average normalized distance between their term pairs, selected for minimal distance. From these measures, heterogeneity values are derived in the same fashion as for sequence and structure, described above. This definition, similarly to the definition based on hierarchical clustering, is resistant to the effect of outliers, that is, proteins with very dissimilar functions.

Statistical tests

For comparing sequence compositions and structural features between various protein groups, two-sided Kolmogorov–Smirnov test was used, as implemented in R (ks.test function). This compares the distributions of the selected parameter in the two groups, assigning a *p*-value that describes the probability of the two sets of values originating from the same distribution. These *p*-values are reported in Tables S1 and S2.

For comparing occurrences of various GO terms in the three groups, Fisher's exact test was employed, using the implementation in the Text::NSP::Measures::2D::Fisher Perl module. For each GO term occurrence for each group, two sets of expected values were generated based on the occurrence of the same term in the other two groups. Based on these two sets of expected values, two *p*-values were obtained from using one-sided Fisher's exact test, quantifying the overrepresentation of the given term. These values are reported in Table S3 and are represented in Figs. 3 and 4.

Regulation

PTMs identified in low-throughput experiments were taken from the 2 October 2017 version of PhosphoSitePlus [47], Phospho.ELM [48] and UniProt [70], and were mapped to complex structures using BLAST between UniProt and PDB sequences.

The effect on protein stability of introducing a PTM was assessed by switching the original residue with a mimetic one in the structure. Ser and Thr phosphorylations were mimicked with Asp; Lys and Arg methylations were mimicked with Leu and Met, respectively; and Lys acetylation was mimicked with Gln. FoldX was used to calculate the $\Delta\Delta G$ values of the introduced mutation using the standard settings on an optimized structure. All reported values are averages of five runs.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmb.2019.07.034.

Acknowledgments

B.M. is the recipient of the postdoctoral fellowship of the Hungarian Academy of Sciences and the EMBO | EuropaBio fellowship 7544. L.D. is supported by the UNKP-17-3 new national excellence program of the Ministry of Human Capacities. G.E.T. and L.D. are recipients of grants K119287 and K125607 from the Hungarian Scientific Research Fund (OTKA) and "Momentum" Program of the Hungarian Academy of Sciences (LP2012/35). L.D., G.E.T. and I.S. are supported by project no. FIEK_16-1-2016-0005 financed under the FIEK_16 funding scheme (National Research, Development and Innovation Fund of Hungary). I.S. is the recipient of the Hungarian Research and Developments Fund OTKA K115698. Z.D. receives funding from the "Momentum" grant from the Hungarian Academy of Sciences (LP2014-18) and the OTKA grant (K108798).

Declaration of Competing Interest

None.

Received 28 February 2019;
Received in revised form 10 July 2019;
Accepted 29 July 2019
Available online 12 August 2019

Keywords:
intrinsically disordered proteins;
protein–protein interactions;
coupled folding and binding;
mutual synergistic folding;
regulatory networks

Abbreviations used:
IDP, intrinsically disordered protein; IDR, intrinsically disordered region; PPI, protein–protein interaction; GO, Gene Ontology; PCA, principal component analysis; PTM, post-translational modification; LIR, LC3-interacting region; TAD, transactivation domain; CBP, CREB-binding protein.

References

[1] O.C. Redfern, B. Dessailly, C.A. Orengo, Exploring the structure and function paradigm, Curr. Opin. Struct. Biol. 18 (2008) 394–402.

[2] P.E. Wright, H.J. Dyson, Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm, J. Mol. Biol. 293 (1999) 321–331.

[3] H.J. Dyson, P.E. Wright, Intrinsically unstructured proteins and their functions, Nat. Rev. Mol. Cell Biol. 6 (2005) 197–208.

[4] P.E. Wright, H.J. Dyson, Intrinsically disordered proteins in cellular signalling and regulation, Nat. Rev. Mol. Cell Biol. 16 (2015) 18–29.

[5] P. Tompa, Unstructural biology coming of age, Curr. Opin. Struct. Biol. 21 (2011) 419–425.

[6] X. Deng, J. Gumm, S. Karki, J. Eickholt, J. Cheng, An overview of practical applications of protein disorder prediction and drive for faster, more accurate predictions, Int. J. Mol. Sci. 16 (2015) 15384–15404.

[7] Z. Dosztányi, B. Mészáros, I. Simon, Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins, Brief. Bioinform. 11 (2010) 225–243.

[8] B. Monastyrskyy, A. Kryshtafovych, J. Moult, A. Tramontano, K. Fidelis, Assessment of protein disorder region predictions in CASP10, Proteins. 82 (Suppl. 2) (2014) 127–137.

[9] K. Gast, H. Damaschun, K. Eckert, K. Schulze-Forster, H.R. Maurer, M. Müller-Frohne, D. Zirwer, J. Czarnecki, G. Damaschun, Prothymosin alpha: a biologically active protein with random coil conformation, Biochemistry. 34 (1995) 13211–13218.

[10] H. Sutovsky, E. Gazit, The von Hippel–Lindau tumor suppressor protein is a molten globule under native conditions: implications for its physiological activities, J. Biol. Chem. 279 (2004) 17190–17196.

[11] N. Wang, C.Y. Majmudar, W.C. Pomerantz, J.K. Gagnon, J.D. Sadowsky, J.L. Meagher, T.K. Johnson, J.A. Stuckey, C.L. Brooks 3rd, J.A. Wells, A.K. Mapp, Ordering a dynamic protein via a small-molecule stabilizer, J. Am. Chem. Soc. 135 (2013) 3363–3366.

[12] P. Tompa, M. Fuxreiter, Fuzzy complexes: polymorphism and structural disorder in protein–protein interactions, Trends Biochem. Sci. 33 (2008) 2–8.

[13] C.J. Tsai, S. Kumar, B. Ma, R. Nussinov, Folding funnels, binding funnels, and protein function, Protein Sci. 8 (1999) 1181–1190.

[14] H.J. Dyson, P.E. Wright, Coupling of folding and binding for unstructured proteins, Curr. Opin. Struct. Biol. 12 (2002) 54–60.

[15] J.R. Perkins, I. Diboun, B.H. Dessailly, J.G. Lees, C. Orengo, Transient protein–protein interactions: structural, functional, and network properties, Structure. 18 (2010) 1233–1243.

[16] X. Chu, J. Wang, Specificity and affinity quantification of flexible recognition from underlying energy landscape topography, PLoS Comput. Biol. 10 (2014), e1003782.

[17] A. Mohan, C.J. Oldfield, P. Radivojac, V. Vacic, M.S. Cortese, A.K. Dunker, V.N. Uversky, Analysis of molecular recognition features (MoRFs), J. Mol. Biol. 362 (2006) 1043–1059.

[18] D. Shaji, T. Amemiya, R. Koike, M. Ota, Interface property responsible for effective interactions of protean segments: intrinsically disordered regions that undergo disorder-to-order transitions upon binding, Biochem. Biophys. Res. Commun. 478 (2016) 123–127.

[19] M. Gouw, S. Michael, H. Sámano-Sánchez, M. Kumar, A. Zeke, B. Lang, B. Bely, L.B. Chemes, N.E. Davey, Z. Deng, F. Diella, C.-M. Gürth, A.-K. Huber, S. Kleinsorg, L.S. Schlegel, N. Palopoli, K.V. Roey, B. Altenberg, A. Reményi, H. Dinkel, T.J. Gibson, The eukaryotic linear motif resource—2018 update, Nucleic Acids Res. 46 (2018) D428–D434.

[20] B. Mészáros, P. Tompa, I. Simon, Z. Dosztányi, Molecular principles of the interactions of disordered proteins, J. Mol. Biol. 372 (2007) 549–561.

[21] R. van der Lee, M. Buljan, B. Lang, R.J. Weatheritt, G.W. Daughdrill, A.K. Dunker, M. Fuxreiter, J. Gough, J. Gsponer, D.T. Jones, P.M. Kim, R.W. Kriwacki, C.J. Oldfield, R.V. Pappu, P. Tompa, V.N. Uversky, P.E. Wright, M.M. Babu, Classification of intrinsically disordered regions and proteins, Chem. Rev. 114 (2014) 6589–6631.

[22] N. Malhis, M. Jacobson, J. Gsponer, MoRFchibi SYSTEM: software tools for the identification of MoRFs in protein sequences, Nucleic Acids Res. 44 (2016) W488–W493.

[23] F. Meng, V.N. Uversky, L. Kurgan, Comprehensive review of methods for prediction of intrinsic disorder and its molecular functions, Cell. Mol. Life Sci. 74 (2017) 3069–3090.

[24] B. Mészáros, G. Erdos, Z. Dosztányi, IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding, Nucleic Acids Res (2018), https://doi.org/10.1093/nar/gky384.

[25] M. Slutzki, D. Reshef, Y. Barak, R. Haimovitz, S. Rotem-Bamberger, R. Lamed, E.A. Bayer, O. Schueler-Furman, Crucial roles of single residues in binding affinity, specificity, and promiscuity in the cellulosomal cohesin–dockerin interface, J. Biol. Chem. 290 (2015) 13654–13666.

[26] A. Dhulesia, J. Gsponer, M. Vendruscolo, Mapping of two networks of residues that exhibit structural and dynamical changes upon binding in a PDZ domain protein, J. Am. Chem. Soc. 130 (2008) 8931–8939.

[27] N. Alam, L. Zimmerman, N.A. Wolfson, C.G. Joseph, C.A. Fierke, O. Schueler-Furman, Structure-based identification of HDAC8 non-histone substrates, Structure. 24 (2016) 458–468.

[28] N. Alam, O. Goldstein, B. Xia, K.A. Porter, D. Kozakov, O. Schueler-Furman, High-resolution global peptide–protein docking using fragments-based PIPER–FlexPepDock, PLoS Comput. Biol. 13 (2017), e1005905.

[29] A. Lavi, C.H. Ngan, D. Movshovitz-Attias, T. Bohnuud, C. Yueh, D. Beglov, O. Schueler-Furman, D. Kozakov, Detection of peptide-binding sites on protein surfaces: the first step toward the modeling and targeting of peptide-mediated interactions, Proteins. 81 (2013) 2096–2105.

[30] N. London, B. Raveh, O. Schueler-Furman, Druggable protein–protein interactions—from hot spots to hot segments, Curr. Opin. Chem. Biol. 17 (2013) 952–959.

[31] Y. Sedan, O. Marcu, S. Lyskov, O. Schueler-Furman, Peptiderive server: derive peptide inhibitors from protein–protein interactions, Nucleic Acids Res. 44 (2016) W536–W541.

[32] H. Shen, C.G. Maki, Pharmacologic activation of p53 by small-molecule MDM2 antagonists, Curr. Pharm. Des. 17 (2011) 560–568.
[33] S.J. Demarest, M. Martinez-Yamout, J. Chung, H. Chen, W. Xu, H.J. Dyson, R.M. Evans, P.E. Wright, Mutual synergistic folding in recruitment of CBP/p300 by p160 nuclear receptor coactivators, Nature. 415 (2002) 549–553.
[34] K. Gunasekaran, C.-J. Tsai, R. Nussinov, Analysis of ordered and disordered protein complexes reveals structural features discriminating between stable and unstable monomers, J. Mol. Biol. 341 (2004) 1327–1341.
[35] J.A.O. Rumfeldt, C. Galvagnion, K.A. Vassall, E.M. Meiering, Conformational stability and folding mechanisms of dimeric proteins, Prog. Biophys. Mol. Biol. 98 (2008) 61–84.
[36] S.S. Mackinnon, A. Malevanets, S.J. Wodak, Intertwined associations in structures of homooligomeric proteins, Structure. 21 (2013) 638–649.
[37] E. Schad, E. Fichó, R. Pancsa, I. Simon, Z. Dosztányi, B. Mészáros, DIBS: a repository of disordered binding sites mediating interactions with ordered proteins, Bioinformatics (2017), https://doi.org/10.1093/bioinformatics/btx640.
[38] E. Fichó, I. Reményi, I. Simon, B. Mészáros, MFIB: a repository of protein complexes with mutual folding induced by binding, Bioinformatics (2017), https://doi.org/10.1093/bioinformatics/btx486.
[39] E.T.C. Wong, D. Na, J. Gsponer, On the importance of polar interactions for complexes containing intrinsically disordered proteins, PLoS Comput. Biol. 9 (2013), e1003192.
[40] N.E. Davey, K. Van Roey, R.J. Weatheritt, G. Toedt, B. Uyar, B. Altenberg, A. Budd, F. Diella, H. Dinkel, T.J. Gibson, Attributes of short linear motifs, Mol. BioSyst. 8 (2012) 268–281.
[41] A. Campen, R.M. Williams, C.J. Brown, J. Meng, V.N. Uversky, A.K. Dunker, TOP-IDP-scale: a new amino acid scale measuring propensity for intrinsic disorder, Protein Pept. Lett. 15 (2008) 956–963.
[42] M. Necci, D. Piovesan, S.C.E. Tosatto, Where differences resemble: sequence-feature analysis in curated databases of intrinsically disordered proteins, Database 2018 (2018), https://doi.org/10.1093/database/bay127.
[43] T.J. Richmond, Solvent accessible surface area and excluded volume in proteins. Analytical equations for overlapping spheres and implications for the hydrophobic effect, J. Mol. Biol 178 (1984) 63–89.
[44] S. Boeynaems, E. Bogaert, D. Kovacs, A. Konijnenberg, E. Timmerman, A. Volkov, M. Guharoy, M. De Decker, T. Jaspers, V.H. Ryan, A.M. Janke, P. Baatsen, T. Vercruysse, R.-M. Kolaitis, D. Daelemans, J.P. Taylor, N. Kedersha, P. Anderson, F. Impens, F. Sobott, J. Schymkowitz, F. Rousseau, N.L. Fawzi, W. Robberecht, P. Van Damme, P. Tompa, L. Van Den Bosch, Phase separation of C9orf72 dipeptide repeats perturbs stress granule dynamics, Mol. Cell 65 (2017) 1044–1055.e5.
[45] C.P. Brangwynne, P. Tompa, R.V. Pappu, Polymer physics of intracellular phase transitions, Nat. Phys. 11 (2015) 899–904.
[46] J. Gsponer, M.E. Futschik, S.A. Teichmann, M.M. Babu, Tight regulation of unstructured proteins: from transcript synthesis to protein degradation, Science. 322 (2008) 1365–1368.
[47] P.V. Hornbeck, B. Zhang, B. Murray, J.M. Kornhauser, V. Latham, E. Skrzypek, PhosphoSitePlus, 2014: mutations, PTMs and recalibrations, Nucleic Acids Res. 43 (2015) D512–D520.
[48] H. Dinkel, C. Chica, A. Via, C.M. Gould, L.J. Jensen, T.J. Gibson, F. Diella, Phospho.ELM: a database of phosphorylation sites—update, Nucleic Acids Res. 39 (2011) (2011) D261–D267.
[49] A.J. Bannister, T. Kouzarides, Regulation of chromatin by histone modifications, Cell Res. 21 (2011) 381–395.
[50] A. Bah, J.D. Forman-Kay, Modulation of intrinsically disordered protein function by post-translational modifications, J. Biol. Chem. 291 (2016) 6696–6705.
[51] V.N. Uversky, J.R. Gillespie, A.L. Fink, Why are "natively unfolded" proteins unstructured under physiologic conditions? Proteins. 41 (2000) 415–427.
[52] P. Radivojac, Protein flexibility and intrinsic disorder, Protein Sci. 13 (2004) 71–80.
[53] E. Garner, P. Cannon, P. Romero, Z. Obradovic, A.K. Dunker, Predicting disordered regions from amino acid sequence: common themes despite differing structural characterization, Genome Inform. Ser. Workshop Genome Inform. 9 (1998) 201–213.
[54] J.C. Hansen, X. Lu, E.D. Ross, R.W. Woody, Intrinsic protein disorder, amino acid composition, and histone terminal domains, J. Biol. Chem. 281 (2006) 1853–1856.
[55] F. Totzeck, M.A. Andrade-Navarro, P. Mier, The protein structure context of PolyQ regions, PLoS One 12 (2017), e0170801.
[56] Y. Cheng, C.J. Oldfield, J. Meng, P. Romero, V.N. Uversky, A. Keith Dunker, Mining α-helix-forming molecular recognition features with cross species sequence alignments, Biochemistry. 46 (2007) 13468–13477.
[57] H. Remaut, G. Waksman, Protein–protein interaction through beta-strand addition, Trends Biochem. Sci. 31 (2006) 436–444.
[58] T.W. Kirby, G.A. Mueller, E.F. DeRose, M.S. Lebetkin, G. Meiss, A. Pingoud, R.E. London, The nuclease A inhibitor represents a new variation of the rare PR-1 fold, J. Mol. Biol. 320 (2002) 771–782.
[59] X. Cheng, Y. Wang, Y. Gong, F. Li, Y. Guo, S. Hu, J. Liu, L. Pan, Structural basis of FYCO1 and MAP1LC3A interaction reveals a novel binding mode for Atg8-family proteins, Autophagy. 12 (2016) 1330–1339.
[60] M.R. Gryk, O. Jardetzky, L.S. Klig, C. Yanofsky, Flexibility of DNA binding domain of trp repressor required for recognition of different operator sequences, Protein Sci. 5 (1996) 1195–1197.
[61] K.T. Bieging, S.S. Mello, L.D. Attardi, Unravelling mechanisms of p53-mediated tumour suppression, Nat. Rev. Cancer 14 (2014) 359–370.
[62] L.M.M. Jenkins, S.R. Durell, S.J. Mazur, E. Appella, p53 N-terminal phosphorylation: a defining layer of complex regulation, Carcinogenesis. 33 (2012) 1441–1449.
[63] M. Okuda, Y. Nishimura, Extended string binding mode of the phosphorylated transactivation domain of tumor suppressor p53, J. Am. Chem. Soc. 136 (2014) 14143–14152.
[64] D.R. Green, G. Kroemer, Cytoplasmic functions of the tumour suppressor p53, Nature. 458 (2009) 1127–1130.
[65] P. Tompa, N.E. Davey, T.J. Gibson, M.M. Babu, A million peptide motifs for the molecular biologist, Mol. Cell 55 (2014) 161–169.
[66] I. Becher, A. Andrés-Pons, N. Romanov, F. Stein, M. Schramm, F. Baudin, D. Helm, N. Kurzawa, A. Mateus, M.-T. Mackmull, A. Typas, C.W. Müller, P. Bork, M. Beck, M.M. Savitski, Pervasive protein thermal stability variation during the cell cycle, Cell 173 (2018) 1495–1507.e18.

[67] D.K. Kirchner, P. Güntert, Objective identification of residue ranges for the superposition of protein structures, BMC Bioinformatics. 12 (2011) 170.

[68] F.M.G. Pearl, C.F. Bennett, J.E. Bray, A.P. Harrison, N. Martin, A. Shepherd, I. Sillitoe, J. Thornton, C.A. Orengo, The CATH database: an extended protein family resource for structural and functional genomics, Nucleic Acids Res. 31 (2003) 452–455.

[69] D. Piovesan, F. Tabaro, I. Mičetić, M. Necci, F. Quaglia, C.J. Oldfield, M.C. Aspromonte, N.E. Davey, R. Davidović, Z. Dosztányi, A. Elofsson, A. Gasparini, A. Hatos, A.V. Kajava, L. Kalmar, E. Leonardi, T. Lazar, S. Macedo-Ribeiro, M. Macossay-Castillo, A. Meszaros, G. Minervini, N. Murvai, J. Pujols, D.B. Roche, E. Salladini, E. Schad, A. Schramm, B. Szabo, A. Tantos, F. Tonello, K.D. Tsirigos, N. Veljković, S. Ventura, W. Vranken, P. Warholm, V.N. Uversky, A.K. Dunker, S. Longhi, P. Tompa, S.C.E. Tosatto, DisProt 7.0: a major update of the database of disordered proteins, Nucleic Acids Res 45 (2017) D1123–D1124.

[70] R. Apweiler, A. Bairoch, C.H. Wu, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M.J. Martin, D.A. Natale, C. O'Donovan, N. Redaschi, L.-S.L. Yeh, UniProt: the universal protein knowledgebase, Nucleic Acids Res. 32 (2004) D115–D119.

[71] W.G. Touw, C. Baakman, J. Black, T.A.H. te Beek, E. Krieger, R.P. Joosten, G. Vriend, A series of PDB-related databanks for everyday needs, Nucleic Acids Res. 43 (2015) D364–D368.

[72] S. Hubbard, J. Thornton, Naccess. http://www.bioinf.manchester.ac.uk/naccess/, 1992.

[73] Z. Dosztányi, V. Csizmók, P. Tompa, I. Simon, The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins, J. Mol. Biol. 347 (2005) 827–839.

[74] L. Tierney, The R statistical computing environment, in: Lecture Notes in Statistics, 2012, pp. 435–447.