Peer-reviewed version available at *Int. J. Mol. Sci.* **2019**, *20*, 5460; doi:10.3390/ijms20215460

Article

Sequence and structure properties uncover the natural classification of protein complexes formed by intrinsically disordered proteins via mutual synergistic folding

Bálint Mészáros[1,2,3,*], László Dobson[4,5], Erzsébet Fichó[3] 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, Hungarian Academy of Sciences, Magyar Tudósok krt 2, Budapest, H-1117 Hungary

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

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

*To whom correspondence should be addressed: bmeszaros@caesar.elte.hu, simon.istvan@ttk.mta.hu

Abstract

Intrinsically disordered proteins mediate crucial biological functions through their interactions with other proteins. Mutual synergistic folding (MSF) occurs when all interacting proteins are disordered, folding into a stable structure in the course of the complex formation. In these cases, the folding and binding processes occur in parallel, lending the resulting structures uniquely heterogeneous features. Currently there are no dedicated classification approaches that would take into account the particular biological and biophysical properties of MSF complexes. Here we present a scalable clustering-based classification scheme, built on redundancy-filtered features that describe the sequence and structure properties of the complexes, and the role of the interaction, which is directly responsible for structure formation. Using this approach, we define six major types of MSF complexes, corresponding to biologically meaningful groups. Hence, the presented method also shows that differences in binding strength, subcellular localization, and regulation are encoded in the sequence and structural properties of proteins. While current structure classification methods can also handle complex structures, we show that the developed scheme is fundamentally different, and since it takes into account defining features of MSF complexes, it serves as a better representation of structures arising through this specific interaction mode.

Keywords

Intrinsically disordered protein, IDP, protein-protein interaction, mutual synergistic folding, coupled folding and binding, structural analysis, structure-based classification, fold recognition

1. Introduction

Intrinsically disordered proteins (IDPs) are crucial elements of the molecular machinery indispensable for complex life [1]. IDPs are parts of regulatory pathways [2], control the cell cycle [3], function as chaperones [4] and regulate protein degradation [5,6], amongst others. In accord, IDPs are typically under tight regulation at several levels [7]. While some IDPs fulfill their functions directly through their lack of structure, such as spring-like entropic chains, the majority of disordered proteins interact with other macromolecules, most often other proteins [8]. IDP mediated interactions are essential for many hub proteins [9], and several IDPs serve as interaction scaffolds/platforms for macromolecular assembly [10]. Mounting evidence also shows that protein disorder plays a crucial role in the assembly of liquid-liquid phase separated non-membrane bounded organelles [11].

Depending on the partner protein and the specifics of the interaction, IDPs can bind through several mechanisms. Several IDPs recognize and bind to ordered protein domains, usually through a linear sequence motif [12]. While some IDPs retain their inherent flexibility in the bound form as well [13], in most known cases the complex structure lends itself to standard structure determination methods, such as X-ray crystallography or NMR. These cases of coupled-folding-and-binding have been studied intensively [14–16]. However, IDPs can utilize a fundamentally different molecular mechanism for interaction, through which they reach a folded state as well. Complexes that contain only IDPs as constituent protein chains, without the presence of a previously folded domain, are formed via a process called mutual synergistic folding (MSF) [17] – a much less understood way in which protein folding and binding can merge into a single biophysical process.

A major advancement in the field of IDP interactions in recent years were the development of specialized interaction databases for various mechanisms including coupled folding and binding [18,19], fuzzy complexes [20], generic databases encompassing interactions with non-protein partners [21], mutual synergistic folding [22], and proteins driving liquid-liquid phase separation [23]. Out of these aspects, possibly the most understudied one is mutual synergistic folding, owing to the fact that these are the only interactions where none of the partner proteins have a well-defined structure outside of the complex, forcing us to revise our current approaches used for describing protein structures and complexes. The biological and biophysical properties of these interactions are markedly different from those mediated by other types of proteins. While in other interaction types a stable, folded hydrophobic core is already present in at least one partner, here the folding and binding happen at the same for all partners. Comparative analysis has not only shown that MSF complexes are a separate biologically meaningful class, but also highlighted that these complexes are highly heterogeneous in terms of sequence and structure [24].

We now have knowledge of over 140,000 protein structures deposited in the PDB [25], a major part of which contains several proteins. In each of these cases, the proteins achieve stability either before, or upon interacting. A major question is how is stability achieved? Can

Peer-reviewed version available at Int. J. Mol. Sci. 2019, 20, 5460; doi:10.3390/ijms20215460

this be a basis of the definition of biologically meaningful classification? In the case of ordered proteins, current hierarchical classification schemes are rooted in the tertiary protein structures, such as in the case of SCOP [26] and CATH [27]. While these methods are extended to classify protein complexes as well, they do not explicitly factor in parameters that describe the interactions or the differences in sequence composition between complexes of similar overall structures. However, in the case of MSF complexes, these differences are defining features, as the interaction is the primary reason for the emergence of the structure itself, and this interaction requires highly specialized residue compositions [24]. While other classification methods were developed specifically for protein-protein interactions, they only aim to describe the interface, without taking the overall resulting structure into account [28].

Here we present the first classification method designed to identify biologically relevant types of protein complexes formed via mutual synergistic folding. Our work aims to answer specific questions about the types of MSF complexes based on the currently known more than 200 examples. Are there intrinsic classes of MSF complexes or are all known examples basically unique in terms of sequence and structure? If meaningful groups are definable in an objective way, what are the characteristics of each group in terms of sequence composition and adopted structure? In addition, how is the formation of MSF complexes regulated? Are mechanisms known to be important for other molecular interactions relevant to these complexes as well? If so, are there differences between various MSF groups regarding these regulatory mechanisms and other biologically relevant properties, such as binding strength and subcellular localization?

2. Results

2.1 Sequence-based properties define four clusters of complexes

Complexes formed by mutual synergistic folding were taken from the MFIB database [22], and each complex has been assigned a feature vector describing the sequence composition of its constituent protein chains. To represent the sequence composition, we use the amino acid grouping previously used for investigating protein-protein complexes involving IDPs [24] (see Data and Methods and Table 1 for definitions, and Supplementary Table S1 for exact values for all complexes). These vectors were used as input for hierarchical clustering (Supplementary Figure S1) to quantify the sequence-based relationship between various complexes. k-means clustering (Supplementary Figure S2) indicates four as the most appropriate number of clusters, and therefore we use four sequence-based clusters in all subsequent analyses. The main features of the four clusters are shown in Table 1, while cluster numbers for each complex are shown in Supplementary Table 1.

Table 1 shows the average sequence compositions of each of the four sequence-based clusters. While clusters were defined based on sequence compositions only, Table 1 also shows the average heterogeneity of the four clusters, meaning the average normalized difference in sequence composition between the interacting proteins of the complexes (see Data and Methods). Complexes in clusters 1 and 2 are both largely devoid of special

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 October 2019 doi:10.20944/preprints201910.0124.v1

Peer-reviewed version available at Int. J. Mol. Sci. 2019, 20, 5460; doi:10.3390/ijms20215460

residues, including Gly (flexible), Pro (rigid) and Cys (cysteine). Members of these two clusters contain an average fraction of hydrophobic residues; however are slightly depleted in aromatic residues, indicating that π-π interactions aren't the dominant source of stability. The most characteristic difference between clusters 1 and 2 is that members of cluster 1 typically contain a high fraction of polar residues, while members of cluster 2 are enriched in charged residues. Also, cluster 1 members are typically formed by proteins with highly different compositions (high heterogeneity values), while cluster 2 members are formed by proteins of very similar compositions.

In contrast, members of clusters 3 and 4 are typically enriched in Gly and Pro, and contain a higher-than-average fraction of aromatic residues. Again, polar/charged residue balance is a distinguishing feature, with cluster 3 and 4 showing preferences for polar and charged residues, respectively. Also, similarly to cluster 1 and 2, there is a notable difference in heterogeneity values between clusters 3 and 4: members of clusters 3 and 4 are typically composed of proteins with very similar and different residue compositions, respectively.

		Cluster 1	Cluster 2	Cluster 3	Cluster 4	Average
Number of members		25	60	38	84	-
Average amino acid composition	Aromatic (FWY)	0.029	0.049	0.089	0.072	0.064
	Hydrophobic (AILMV)	0.381	0.324	0.287	0.357	0.336
	Flexible (G)	0.024	0.021	0.076	0.056	0.046
	Rigid (P)	0.018	0.005	0.029	0.040	0.026
	Charged (HKRDE)	0.258	0.392	0.239	0.290	0.308
	Polar (NQST)	0.281	0.205	0.254	0.179	0.211
	Cysteine (C)	0.009	0.005	0.027	0.006	0.010
Heterogeneity (average dissimilarity between subunits)		0.162	0.098	0.057	0.117	0.111

Comparison to average:

+10%
+20%
+30%
-10%
-20%
-30%

Table 1: Average values of sequence features for the four sequence-based clusters. Blue and orange shadings mark values that are over- or under-represented compared to the average of all MSF complexes. Heterogeneity values weren't used for cluster definitions.

2.2 Structure-based properties offer a different means of defining complex types

The structural properties of the studied complexes were quantified using various features describing the secondary structure compositions, various molecular surfaces, incorporating hydrophobicity measures, and atomic contacts (see Supplementary Table 1 and Data and Methods). These structural features were used to describe each complex in the form of a feature vector, and similarly to the analysis of sequence properties, these vectors were input to hierarchical clustering; however, structural features were filtered and only those were kept that share a modest degree of correlation (see Supplementary Table 2 and Data and Methods for specifics) to avoid bias. The resulting tree is shown in Supplementary Figure S3. In contrast to the sequence-based clustering, k-means within cluster sum of squares analysis does not indicate any given number of clusters as optimal (Supplementary Figure S4). In order to have a medium number of clusters, we cut the hierarchical tree at a linkage distance

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 October 2019 doi:10.20944/preprints201910.0124.v1

Peer-reviewed version available at *Int. J. Mol. Sci.* **2019**, *20*, 5460; doi:10.3390/ijms20215460

that defines five clusters (Supplementary Figure S3). The average values of structural parameters for all five structure classes are shown in Table 2.

The obtained clusters show distinguishing structural features. Members of cluster 1 incorporate the highest amount of non-helical secondary structure elements. These complexes heavily rely on a large number of buried hydrophobic residues for stability and most stabilizing atomic contacts are formed between residues of the same protein, relying less on intermolecular interactions, which tend to be mostly polar in nature.

In contrast, members of cluster 2 adopt mainly helical structures. The stability of these complexes seems to rely more on the interactions formed between the subunits, mostly formed between side chains. The importance of interchain interactions is also reflected in the large relative interface and small relative buried surface areas.

Cluster 3 and 4 complexes exhibit similar features, including a balanced ratio of various secondary structure elements, and polar/hydrophobic balance of various molecular surfaces and contacts. For both clusters, interchain contacts rely mostly on side chain-side chain and backbone-backbone contacts. The main difference between the two clusters is the relative role of the interface between the participating proteins. Cluster 3 members have a larger than average interface, both in terms of molecular surface and number of contacts, meanwhile cluster 4 complexes have a very restricted interface size, incorporating only a few atomic contacts.

Members of cluster 5 are the most similar to the average in most structural features. There are only weak distinguishing features, including a slightly increased helical content at the expense of extended structural elements, a moderate increase in the role of backbone-side chain interactions in interchain contacts, and the increased ratio of interchain contacts. However, these increases in average parameter values are modest and - with the exception of the decreased extended structure content - none of them reaches 20% compared to the average values calculated for all complexes.

		Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Average
Number of members		47	58	23	28	52	-
Secondary structures	alpha *	0.153	0.879	0.565	0.691	0.672	0.603
	extended	0.398	0.008	0.208	0.043	0.040	0.131
	coil	0.448	0.113	0.227	0.267	0.288	0.266
Molecular surfaces	SASA – hydrophobic *	0.559	0.570	0.527	0.558	0.547	0.555
	SASA – polar	0.441	0.430	0.473	0.442	0.453	0.445
	Interface – hydrophobic *	0.681	0.783	0.706	0.797	0.768	0.750
	Interface – polar	0.319	0.217	0.294	0.203	0.232	0.251
	Buried surface – hydrophobic *	0.590	0.393	0.515	0.529	0.509	0.498
	Buried surface – polar	0.410	0.607	0.485	0.471	0.491	0.502
	Interface / total *	0.079	0.204	0.217	0.120	0.167	0.157
	Buried / total	0.606	0.400	0.438	0.519	0.474	0.485
Atomic contacts	Interchain hydro:hydro *	0.541	0.656	0.587	0.699	0.650	0.627
	Interchain hydro:polar	0.380	0.290	0.338	0.260	0.300	0.314
	Interchain polar:polar	0.080	0.054	0.074	0.041	0.050	0.059
	Interchain backbone:backbone	0.166	0.012	0.148	0.014	0.063	0.075
	Interchain backbone:side-chain *	0.373	0.309	0.295	0.305	0.379	0.339
	Interchain side-chain:side-chain *	0.462	0.678	0.557	0.682	0.558	0.586
	Intrachain hydro:hydro *	0.459	0.306	0.384	0.400	0.384	0.381
	Intrachain hydro:polar	0.438	0.548	0.491	0.480	0.491	0.494
	Intrachain polar:polar	0.103	0.145	0.125	0.120	0.125	0.125
	Intrachain backbone:backbone	0.284	0.475	0.398	0.361	0.379	0.384
	Intrachain backbone:side-chain *	0.383	0.417	0.379	0.410	0.413	0.403
	Intrachain side-chain:side-chain	0.333	0.107	0.223	0.229	0.209	0.213
	Interchain / total *	0.122	0.222	0.313	0.134	0.234	0.201

Comparison to average:
+10%
+20%
+30%
-10%
-20%
-30%

Table 2: Average values for structure features for the five structure-based clusters. Blue and orange shadings mark values that are over- or under-represented compared to the average of all MSF

Peer-reviewed version available at Int. J. Mol. Sci. 2019, 20, 5460; doi:10.3390/ijms20215460

complexes. SASA - solvent accessible surface area, hydro:hydro - fraction of contacts that are formed between two hydrophobic atoms. Asterisks mark features that were included in the clustering.

2.3 Defining interaction types based on sequence and structure clusters

Considering together the previously established sequence- and structure-based clusters, in total 20 types of complexes can be defined (Figure 1). The number of known complexes in possible types shows large variations, with some highly favoured ones (e.g. type 2[sequence]/2[structure]) and ones with a single known example (e.g. type 2/1), showing that not all sequence compositions are compatible with all types of adopted structures. In order to arrive at a reasonable number of basic complex types, types with 10 or fewer complexes were either merged with the adjacent sequence clusters, or were omitted. As structural differences in general are larger between clusters, types belonging to different structure clusters were never merged. This approach yielded five main interaction types, each of which has over 20 complexes. In order to include all known MSF complexes, a 6th pseudo-type was introduced, which contains all structures not compatible with any of the previously described five types (see Supplementary Table 1 for an exhaustive list).

The complex types defined so far are based on structure and sequence features. However, if these types represent biologically meaningful classes, there should be other relevant differences between them in terms of the energetics of the interaction, binding strength, subcellular localization or the biological regulation of the interaction. In the next chapters, we describe each complex type with biologically important characteristics, and assess the potential differences between the members of each class.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 October 2019 doi:10.20944/preprints201910.0124.v1

Peer-reviewed version available at Int. J. Mol. Sci. 2019, 20, 5460; doi:10.3390/ijms20215460

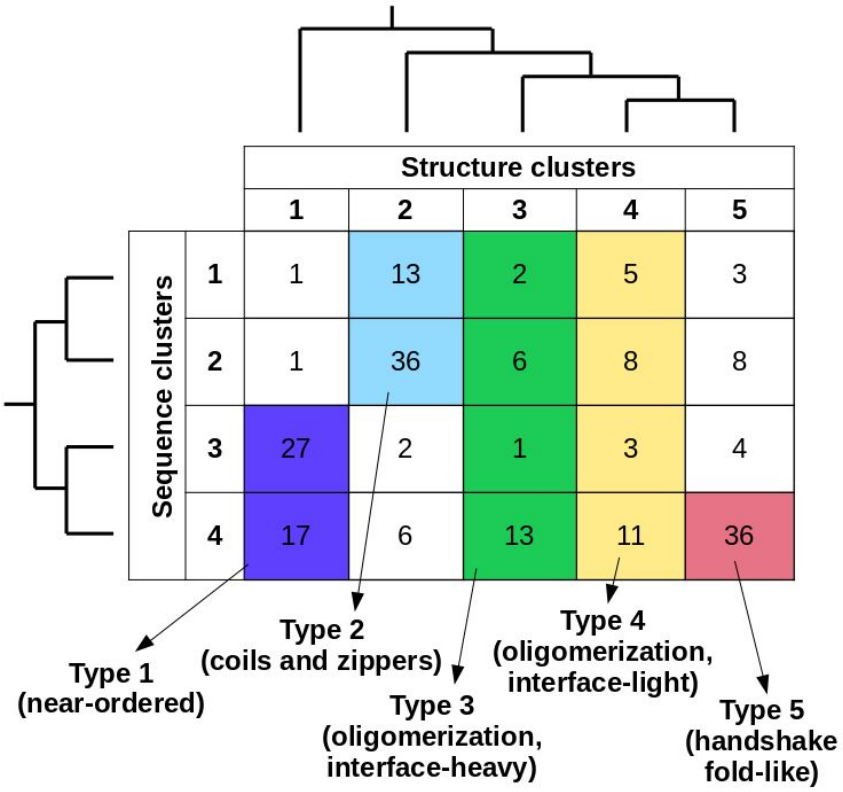


Figure 1: MSF complex types. Coloured regions mark separate interaction types considering sequence- and structure-based clusters (vertical and horizontal axes, respectively). The relationship of each sequence-and structure-based cluster taken from the hierarchical clustering (Supplementary Figures S1 and S3) is shown on the corresponding side of the table.

2.4 Complex types show characteristic energetic properties

From a biological perspective, the strength of association between interacting protein chains and the stability of the resulting complex is of utmost importance. Unfortunately, complexes formed exclusively by IDPs via MSF generally lack targeted measurements concerning thermodynamic and stability parameters. However, low-resolution energy calculations and prediction algorithms can give an indication about the characteristic energetics properties of the uncovered complex types in general. While these methods might have a fairly large error in individual cases, they are well equipped for comparative studies between groups of complexes.

In order to assess the energetic properties of complexes, we employed an energy calculation scheme using low-resolution force fields based on statistical potentials (see Data and Methods). As a reference, energetic properties were calculated for complexes formed exclusively by ordered proteins and complexes formed by an IDP binding to an ordered partner via coupled folding and binding (CFB) (see Data and Methods and Supplementary Table S3 and S4). Figure 2 shows two types of calculated energies for each complex. On one hand, we calculated the total energy per residue in the whole complex, which reflects the

Peer-reviewed version available at Int. J. Mol. Sci. 2019, 20, 5460; doi:10.3390/ijms20215460

overall stability. On the other hand, we also calculated the fraction of this stabilizing energy coming from intermolecular interactions, i.e. how important the interaction is for stability. In accordance with our expectations, complexes formed by ordered proteins feature strongly bound overall structures, with fairly low stabilizing energy/residue. In contrast, CFB complexes in general have less favourable per residue energies hinting at their comparatively weakly bound overall structures. However, the energetic feature providing the most recognizable difference between ordered and CFB complexes is the energy contribution of interchain contacts to the overall stability. In the case of ordered complexes, this contribution is fairly limited, as individual subunits have a stable structure on their own. In contrast, if the complex features an IDP, the interaction energy becomes a major contributor to stability (Figure 2a).

While ordered and CFB complexes tend to segregate in this energy space, complexes formed by MSF seem to be more heterogeneous, covering the whole available range of energetic values (Figure 2b). In the case of near-ordered proteins (Type 1), the energies resemble that of ordered complexes, hinting at the borderline ordered nature of the constituent IDPs, with the interaction between subunits playing a minor role. In contrast, coiled-coil like structures (Type 2) on average have a much less stable complex structure, with interaction playing a substantial role in stability. These complexes resemble IDPs bound to ordered domains, and are expected to include several transient interactions. Other types fall largely between these two extreme cases. Energetics properties of the two types of oligomerization modules (Types 3 and 4) reflect the differences in interface surface area and contact numbers shown in Table 2. While the overall stability for both types varies in a very wide range, on average, the contribution of the interaction is higher for interface-heavy complexes (Type 3) than for interface-light ones (Type 4). Handshake-like folds (Type 5) show interesting properties: these complexes are quite stable with only limited variation in the per-residue energies. Yet they achieve this high stability by relying heavily on the interaction between subunits of the dimer.

The transient or obligate nature of interactions provides clues about their roles in biological systems. This is at least partially describable through K_d dissociation constants. While there is ample data about K_d values of IDPs binding via CFB to ordered domains [18], these values are largely missing for MSF complexes. In accord, we calculated estimated K_d values for MSF complexes (Supplementary Table S1) with Figure 3 showing the K_d distributions for the 6 previously defined complex types. The lowest average K_d values were calculated for complexes with a handshake-like fold (Type 5). The next two types with low K_ds are the near-ordered complexes (Type 1), and interface-heavy oligomerization modules (Type 3). These three types together possibly cover most cases of the interactions where the complex needs to stay stable for an extended period of time, such as histone dimes (Type 5), complexes with enzymatic activity (Type 1) and several transcription factors (Type 3). Coiled-coil-like structures and oligomerization modules with small interfaces in general have a higher K_d, indicating that several transiently bound complexes belong to these types.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 October 2019 doi:10.20944/preprints201910.0124.v1

Peer-reviewed version available at Int. J. Mol. Sci. 2019, 20, 5460; doi:10.3390/ijms20215460

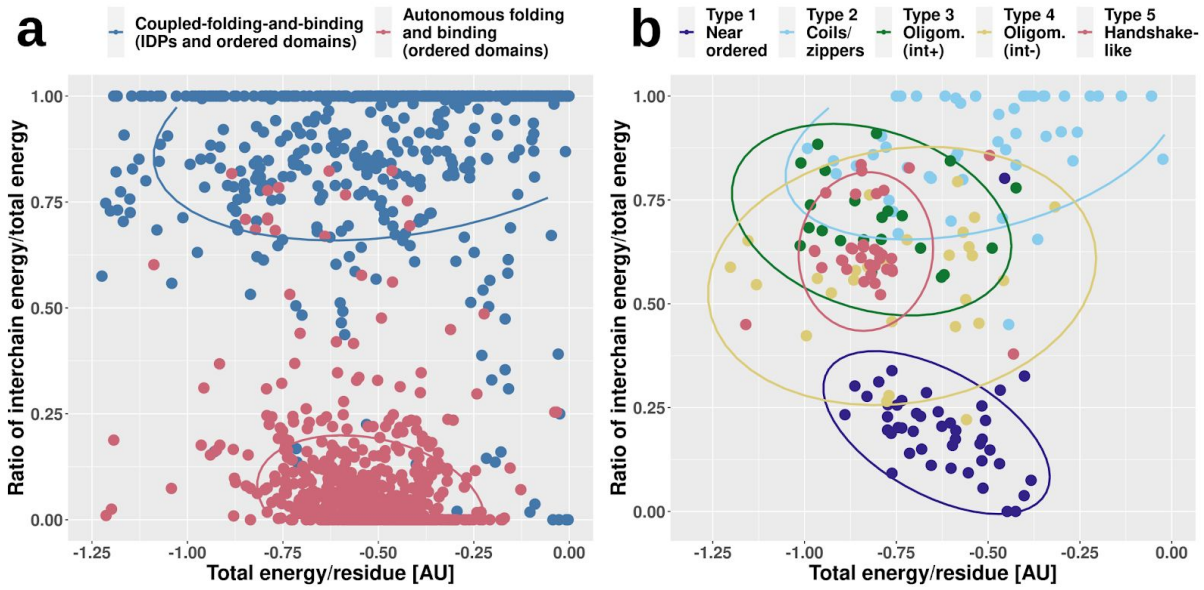


Figure 2: Energetic parameters of various interaction classes. The relative energetic weight of intersubunit interactions in the overall stability (y-axis) as a function of the overall energy per residue (x-axis, measured in arbitrary units, AU) for ordered complexes and complexes formed by coupled folding and binding (**a**), and the five well-defined types of MSF complexes (**b**).

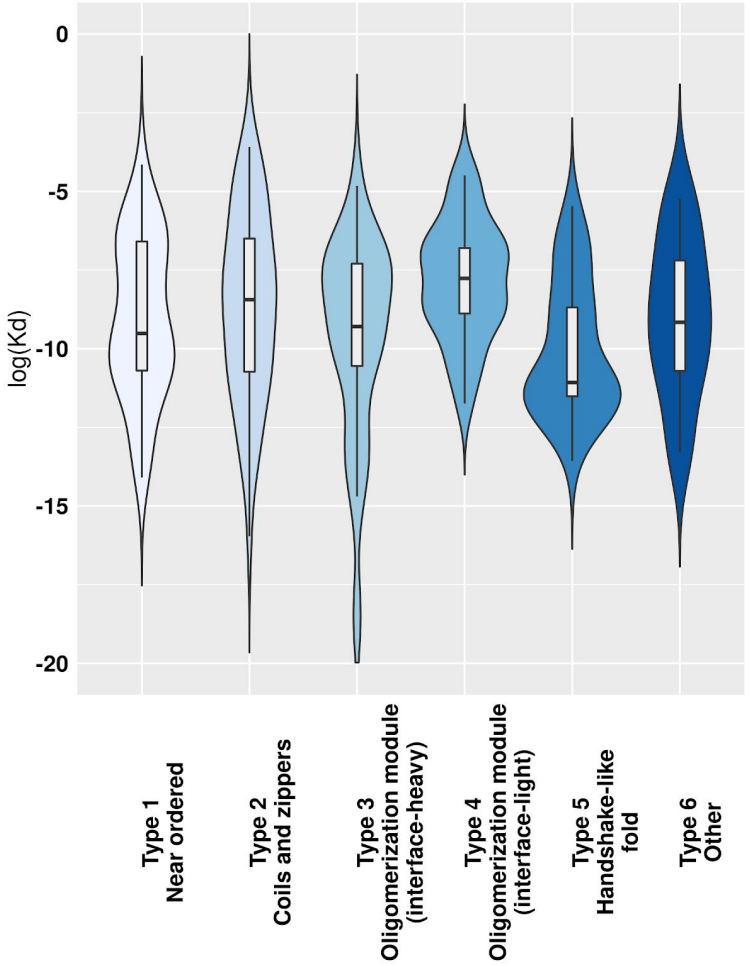


Figure 3: Predicted Kd value distributions for the six types of MSF complexes.

2.5 Interactions are heavily regulated by several mechanisms

While the energetics of various interactions can provide clues about their transient/obligatory nature, the regulatory mechanisms can give more direct evidence. For example, while most IDP enzymes (belonging to Type 1) form particularly stable oligomers, indicating an obligate interaction, the oligomeric state of superoxide dismutase (SOD1) is known to be controlled by the post-translational modification (PTM) serving as an on/off switch [29]; meaning that despite a strong interaction, it is reversible and the disordered state of the monomers is biologically relevant (Figure 4a). Figure 4a shows additional examples of various regulatory mechanisms of MSF interactions via PTMs. These regulatory steps have already been described in the case of IDPs that bind to ordered domains [30], but have not been studied in the context of IDPs participating in MSF interactions. Apart from the on/off switch exemplified by SOD1, PTMs can control the partner selection of synergistically folding IDPs, such as in the case of another tightly bound complex, formed by H3/H4 histones (Type 5) [31]. PTMs can also tune the affinity of certain interactions, as is the case for the activating p53/CBP interaction (Type 4) [32]. Apart from these mechanisms that directly control the interaction between IDPs, PTMs can have a more indirect effect modulating the activity of the dimer itself. In the case of the Max dimeric transcription factor, phosphorylation at the N-terminus of the binding region controls the dimers (Type 4) interaction capacity towards DNA [33]. An even more indirect modulation of function is displayed for the retinoblastoma protein Rb, which in complex with E2F1/DP1 (Type 3) has a strong transcriptional repression activity. Upon methylation, Rb recruits L3MBTL1 [34], which is a direct repressor of transcription via chromatin compaction, augmenting the effect of Rb through a related but separate mechanism extrinsic to the Rb/E2F1/DP1 complex.

To have a more systematic picture of the extent of regulatory mechanisms in MSF interactions, Figure 4B shows the fraction of known MSF complexes with experimentally verified PTM sites (Supplementary Table S5). In total, nearly 30% of studied complexes feature at least one PTM that was experimentally verified in a low-throughput experiment, presenting a regulatory mechanism that is able to directly or indirectly modulate either the interaction itself, or the activity of the resulting complex. The most prevalent PTM is phosphorylation affecting 22% of complexes, but 10%, 15% and 5% of MSF complexes contain methylation, acetylation and ubiquitination sites as well (Figure 4b).

In addition, complex formation can also be regulated through the availability of the subunits participating in the interaction. This availability can depend on the alternative mRNA splicing of the corresponding genes, where certain isoforms lack the binding site (Supplementary Table S6). Also, even if the translated isoform has the binding site, the protein itself can be sequestered by competing interactions with other protein partners (Supplementary Table S7). These mechanisms are present for 11% (alternative splicing) and 16% (competing interactions) of complexes, and together with PTMs, in total 36% of MSF complexes have at least one known regulatory mechanism for modulating the interaction. Furthermore, these

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 October 2019 doi:10.20944/preprints201910.0124.v1

Peer-reviewed version available at Int. J. Mol. Sci. 2019, 20, 5460; doi:10.3390/ijms20215460

regulatory mechanisms often act in cooperation, with 7 interactions known to employ PTMs, alternative splicing and competing interactions as well (Figure 4c).

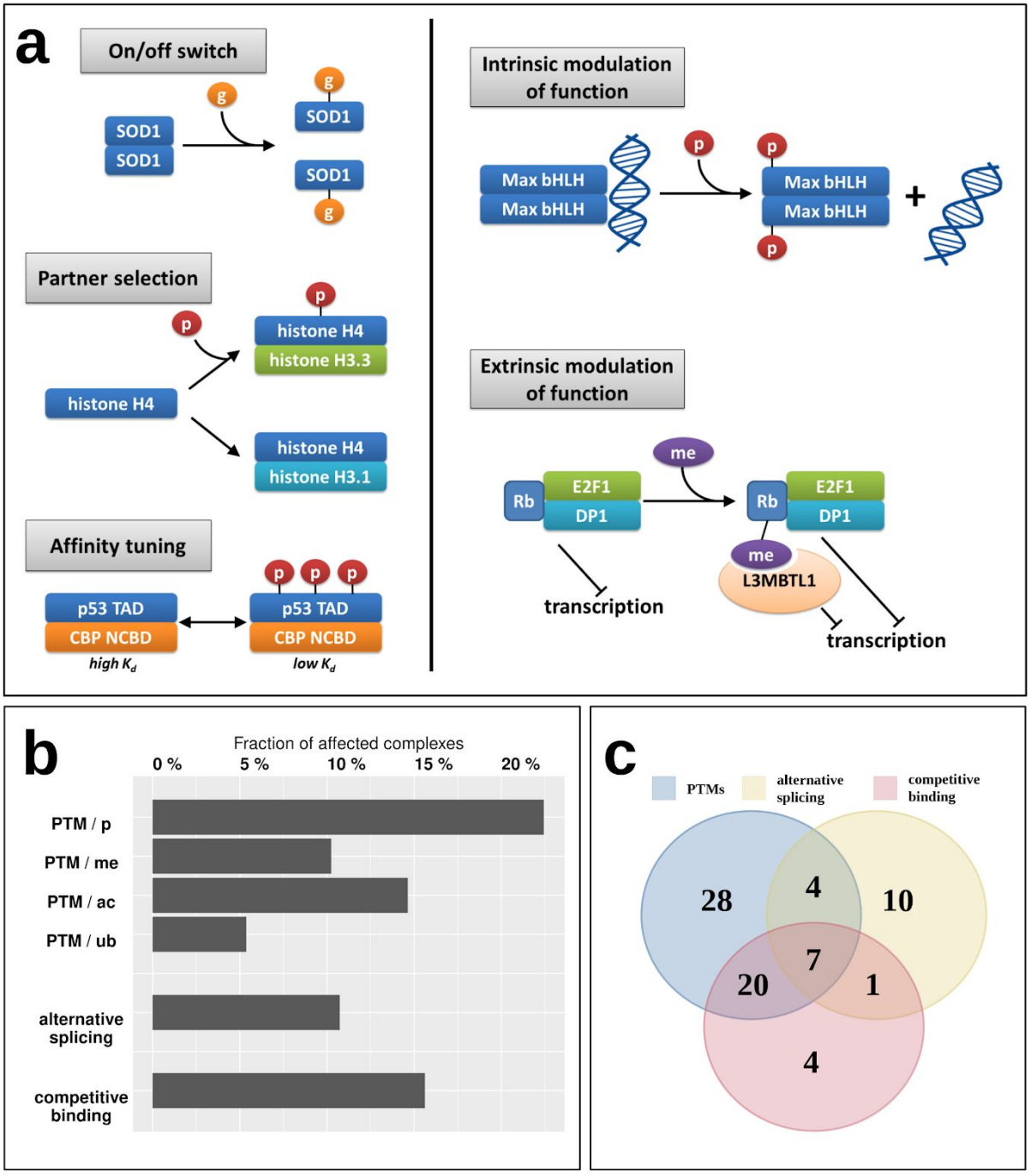


Figure 4: Regulatory mechanisms of MSF complexes. (**a**) examples of regulation and modulation of function through post-translational modifications. p - phosphorylation, g - glutathionylation, me - methylation, SOD1 - superoxide dismutase, CBP - CREB-binding protein, Rb - retinoblastoma associated protein. Coloured boxes represent interacting chains forming the MSF complexes. (**b**) the fraction of complexes with verified PTM sites, and the fraction of complexes where at least one interactor is regulated via alternative splicing or by competing interactions. (**c**) Number and overlap of MSF complexes affected by the three types of regulatory mechanisms.

Peer-reviewed version available at Int. J. Mol. Sci. 2019, 20, 5460; doi:10.3390/ijms20215460

2.6 Various complex types show differential subcellular localization

In addition to regulatory mechanisms detailed in the previous chapter, a crucial element in the spatio-temporal control of protein function is subcellular localization [35]. In order to assess this aspect of MSF complexes, and to understand if the defined interaction types have different properties in terms of cellular localization, we used ‘cellular component’ terms from GeneOntology (GO) [36] (see Data and Methods). Various GO terms were condensed into 5 categories including ‘Extracellular’, ‘Intracellular’, ‘Membrane’, ‘Nucleus’ and ‘Other’ to enable an overview of the differences in localization between the 6 complex types (Figure 5) (for exact GO terms for each complex see Supplementary Table S8).

The least amount of information is available for Type 1, near ordered complexes. Albeit GO terms are lacking for most complexes, even the limited annotations highlight that these complexes are able to efficiently function in the extracellular space, which in general is fairly uncommon for IDPs. Coil and zipper type helical complexes (Type 2) are somewhat more often attached to the membrane or function in the intracellular space, or in non-nuclear environments, such as the lysosome. In contrast, oligomerization modules (Types 3 and 4) are most prevalent in the nucleus and the intracellular space, which is in line with the function of the high number of transcription factors in these groups. However, modules with a large interface (Type 3) are relatively often found in other compartments, while modules with smaller interfaces (Type 4) also function in the extracellular space. Complexes adopting a handshake-like fold are enriched in histones, which is reflected in their enrichment in the nucleus and the chromatin (classified as ‘other’ in Figure 5). Type 6 complexes are heterogeneous in terms of localization as well, and hence members can be found in all studied localizations to a comparable degree. These preferences in subcellular localization for different complex types reinforce our notion that even though our classification scheme relies on sequence and structure properties alone, the obtained interaction types also have biological meaning.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 October 2019 doi:10.20944/preprints201910.0124.v1

Peer-reviewed version available at Int. J. Mol. Sci. 2019, 20, 5460; doi:10.3390/ijms20215460

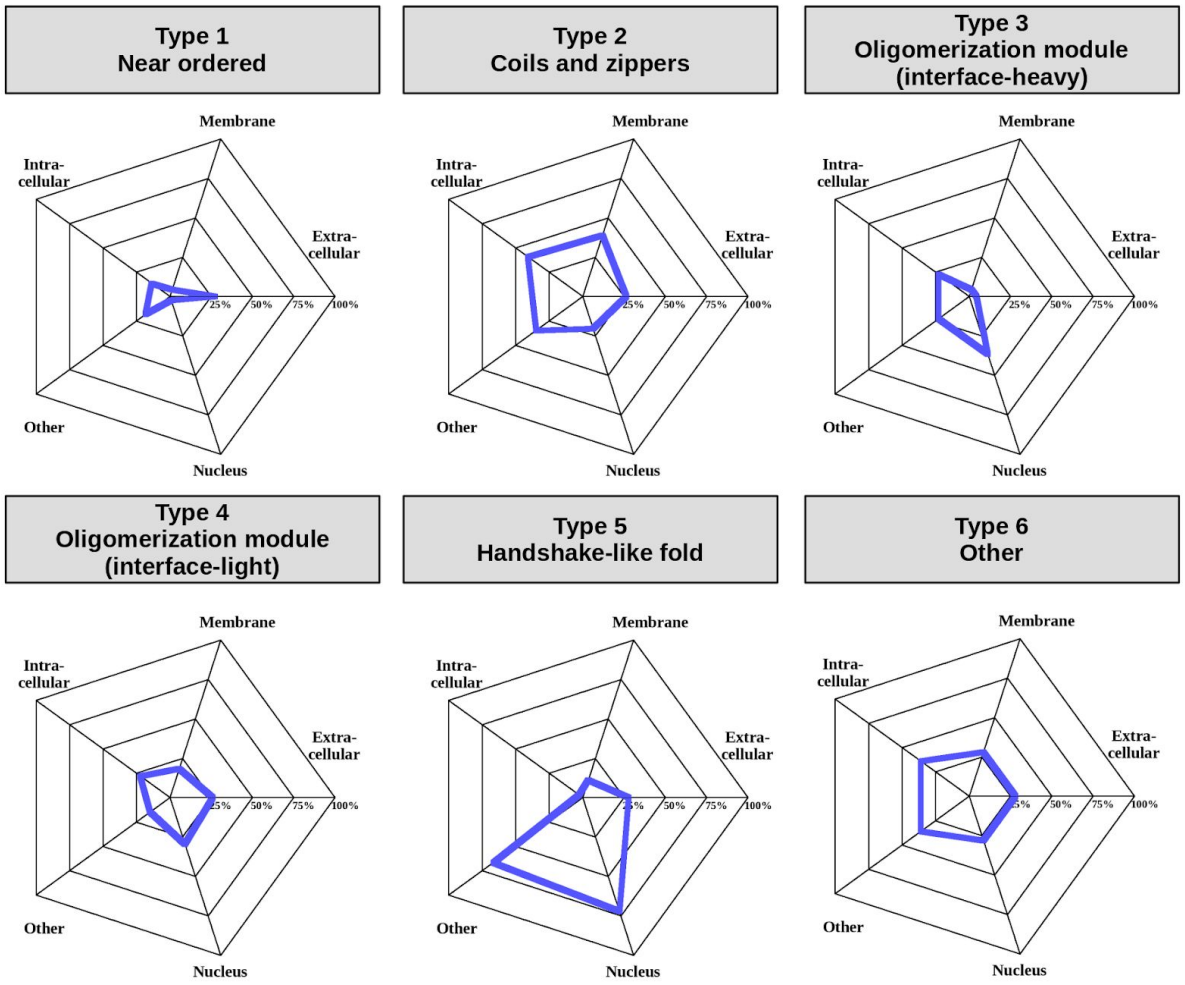
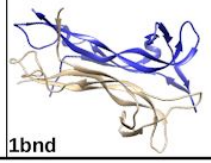
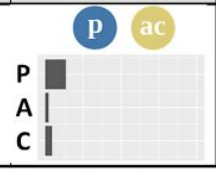
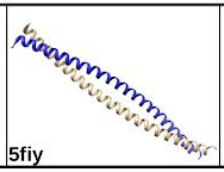
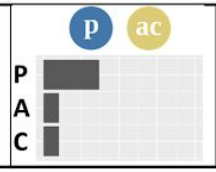
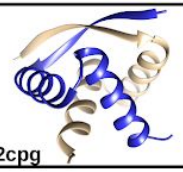
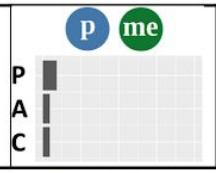
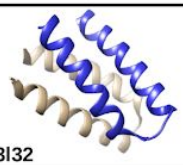
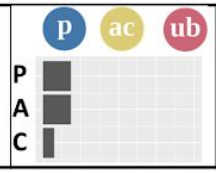
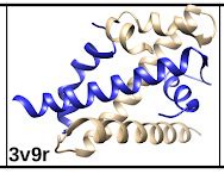
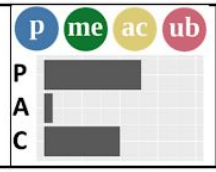
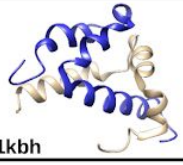
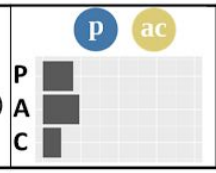


Figure 5: Subcellular localization of MSF complexes belonging to the 6 types. 'Other' contains the 'non-membrane bounded organelle', 'secretory granule', 'lysosome', 'cytoplasmic vesicle lumen' and 'transport vesicle' GeneOntology terms. Red circles mark the most dominant localization(s) for each class.

2.7 The annotated catalogue of complexes formed via mutual synergistic folding

Considering the previously analysed features of complexes and averaging calculated features for the six established interaction types provides the annotated catalogue of MSF interactions (Figure 6). Apart from the main sequential and structural features, figure 6 also shows example structures, energetic properties, subcellular localization and the main regulatory mechanisms for each complex type.

Type ID	Type name	# of complexes	Example structure	Sequence		Structure				Strength		Regulation	Dominant subcellular localization
				Preferred	Depleted	Secondary structures	Interface	Buried surface	Dominant atomic contacts	Energy	Role of interaction		
1	Near ordered	44	 1bnd	Pro/Gly aromatic	charged	extended + coil	small, polar	hydroph.	intrachain	- -	+		Extracellular
2	Coils and zippers	49	 5fiy	hydrophobic charged OR polar	Pro/Gly Cys aromatic	highly helical	large	polar	balanced	- (variable)	+++		Membrane Intracellular Other
3	Oligo-merization module (interface heavy)	22	 2cpg	highly variable		balanced	large, slightly polar	average	interchain	- - -	+++		Nucleus Intracellular Other
4	Oligo-merization module (interface light)	27	 3l32	highly variable		mainly helical	small, hydroph.	average	intrachain	- - (variable)	++		Nucleus Extracellular Intracellular
5	Handshake-like fold	36	 3v9r	Pro/Gly aromatic	Cys polar	mainly helical	average	average	balanced	- - -	++		Nucleus Other
6	Other	25	 1kbh	highly variable		mixed	-	-	-	- - - (variable)	+ + (variable)		Heterogeneous

Peer-reviewed version available at Int. J. Mol. Sci. 2019, 20, 5460; doi:10.3390/ijms20215460

Figure 6: Annotated types of complexes formed by IDPs, based on sequence and structure features. Horizontal bars in the regulation column show the fraction of complexes in a given group involved in various types of regulatory mechanisms (P - post-translational modifications, A - alternative splicing affecting binding regions, C - competing interactions). Color circles mark the dominant post-translational modification(s) for the group (p - phosphorylation, me - methylation, ac - acetylation, ub - ubiquitination).

The first type of complexes bears a high similarity to ordered protein complexes, and hence are named *near ordered*. The constituent chains are usually similar, in many cases corresponding to homooligomers, with a high proline/glycine content and typically only a few charges. The main difference compared to protein complexes formed by ordered proteins is that in comparison, near-ordered subunits are depleted in α-helices [24]. For reaching a stable structure through the interaction, they utilize a large number of intrachain contacts, with inter-subunit interactions through a small polar interface playing only a secondary role in the stability of the complex. This group contains a large number of enzymes, transport proteins and nerve growth factors, where the exact structure is of utmost importance; however, in contrast to monomeric proteins, the presence of this structure relies on the interaction. This interaction type is mostly regulated through phosphorylation and acetylation of binding site residues. These proteins resemble ordered proteins in their localization as well, with extracellular regions being highly representative.

The second type of complexes contains structures with a high overall similarity, mostly consisting of *coiled-coils and zippers*, structures composed of parallel interacting helical structures, often stabilized by a restricted set of residues, such as leucines, alanines or tryptophans. In general, constituent proteins are depleted in residues incompatible with alpha helix formation, such as Pro and Gly, and also in aromatic residues. In turn, they are abundant in hydrophobic residues and show an enrichment for either polar or charged residues. The constituent helices usually for a fairly weakly bound system, where the interchain interactions via the relatively large interfaces play a major role. Constituent proteins are able to bury only a small fraction of their polar surfaces. Coiled-coil interactions are often regulated, typically via various types of PTMs, most often through phosphorylation or to a lesser degree acetylation. Despite their highly similar structures, complexes in this group convey a large variety of functions, mainly pertaining to regulating transcription and performing membrane-associated biological roles, such as organelle and membrane organization.

The third and the fourth type of complexes are both generic *oligomerization modules*, that can be split according to the importance for the interchain interactions, grouping them as either *interface-heavy* (Type 3) or *interface-light* (Type 4) complexes. In both cases, the sequences can be highly variable, and the unifying features are mostly structural. Both types typically have an average-sized relative buried area with balanced hydrophobic/polar composition. However, interface-heavy complexes have a large, slightly polar interface, that play a major role in achieving the tightly bound structures. In contrast, interface-light complexes form a more helical structure, and have smaller hydrophobic interfaces that play a more diminished role in achieving the stability of a less tightly bound system. This hints at

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 October 2019 doi:10.20944/preprints201910.0124.v1

Peer-reviewed version available at Int. J. Mol. Sci. 2019, 20, 5460; doi:10.3390/ijms20215460

interface-light complexes being more transient, also supported by the fact that these complexes have a higher number of known regulatory PTMs, and are also modulated by alternative splicing. Both type 3 and type 4 complexes preferentially occur in nuclear and intracellular processes, as several of them are ribbon-helix-helix (interface-heavy) or basic helix-loop-helix (interface-light) transcription factors, able to shuttle between the nuclear and the intracellular spaces. In addition to the similarities in subcellular localization, type 4 complexes preferentially occur in the extracellular space, and type 3 complexes in other cell compartments, as well.

The fifth type of complexes typically adopts a *handshake-like fold*, characteristic of histones and homologous proteins. While these structures are usually largely helical, the interacting proteins usually contain a relatively high ratio of prolines and glycines, in addition to the enrichment of aromatic residues. While they are depleted in polar residues, both the interface and the buried surface have a fairly balanced hydrophobic/polar makeup. The complexes are relatively tightly bound, and interchain interactions play a fairly large role in stabilizing the interaction. This type of complex has the highest ratio of both PTMs and competitive interactions, providing a large amount of regulation. In addition, PTMs are highly heterogeneous, containing phosphorylations, acetylations, methylations and ubiquitinations as well. Members of this cluster primarily serve DNA/chromosome-related functions, and hence are usually located in the nucleus.

While types 1-5 represent well-defined groups with members of clear unifying similarities, the final group serves as an umbrella term for complexes that are not members of any previous structural/sequential classes. In accord, these complexes cannot be described by simple characteristic features, and are the most sequentially and structurally heterogeneous group. This group contains highly specialized interactions that present unique protein complexes, which are regulated through all three control mechanisms and occur in all studied subcellular localizations.

2.8 Interaction types present a novel classification of protein complexes

The described MSF classification method bears similarity to the approach employed in CATH, as both approaches use a hierarchical classification of PDB structures. However, CATH does not consider interactions and simply relies on the secondary structure elements and their connectivity and arrangement, in contrast to the presented analysis taking into account adjacent protein chains and their interactions too, together with sequence composition features.

Table 3 shows the studied MSF complexes in both our MSF classification system and in CATH considering the top 2 levels (‘Class’ and ‘Architecture’). The highest level CATH definitions, corresponding to ‘Class’, reflect the overall secondary structure element distribution of the structures. In this framework, Type 1 near-ordered complexes mostly occupy the ‘Mainly Beta’ CATH class, while complexes from the other 5 types mostly fall into the ‘Mainly Alpha’ class or the ‘Other’ class. At the next CATH level, ‘Architecture’, certain

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 October 2019 doi:10.20944/preprints201910.0124.v1

Peer-reviewed version available at *Int. J. Mol. Sci.* **2019**, *20*, 5460; doi:10.3390/ijms20215460

MSF type complexes (such as type 2 coils and zippers) are segregated into further subclasses. Using more specific CATH levels (such as 'Topologies') shows little correlation with MSF types, as complexes tend to be distributed all over CATH definitions (data not shown).

Considering 'Class' and 'Architecture' definitions, there is very little correspondence between the CATH and the new MSF classification. If the two schemes showed a high degree of similarity, the matrix in Table 3 should be close to a diagonal matrix. In reality, however, off-diagonal elements are large, confirming the novelty of the presented MSF classification scheme.

		CATH classes and architectures							
		Mainly Alpha (1)		Mainly Beta (2)		Alpha Beta (3)		Few Secondary Structures (4)	Other
		Orthogonal bundle (1.10)	Up-down bundle (1.20)	Sandwich (2.60)	Orthogonal prism (2.90)	2-layer sandwich (3.30)	3-layer(aba) sandwich (3.40)	Irregular (4.10)	-
MSF classification	Near-ordered (Type 1)	2	0	11	9	8	4	0	10
	Coils and Zippers (Type 2)	0	31	0	0	0	0	0	18
	Oligomerization (interface-heavy) (Type 3)	8	1	0	0	1	4	1	7
	Oligomerization (interface-light) (Type 4)	11	3	0	0	3	0	3	7
	Hand-shake fold like (Type 5)	29	2	0	0	0	1	0	4
	Other (Type 6)	8	7	0	0	1	0	2	7

Table 3: Overlap between CATH and MSF classification.

3. Discussion

Here we present the first approach aiming at the classification of complex structures formed exclusively by disordered proteins via mutual synergistic folding. We developed and applied an method that can classify these complexes into various types based on sequence- and structure-based properties. The classification scheme takes into account on the one hand, the overall sequence and structure properties of the complex, and on the other hand, the interaction itself, quantifying the role of intra- and intermolecular interactions in relation to the overall contact/surface properties of the structure. As the classification protocol is based on hierarchical clustering, it is freely scalable. Tuning the resolution via changing the number of sequence-based or structure-based clusters, the method can be used to yield any number of types and subtypes. The presented classification is a top-level one highlighting the major types of MSF classes.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 October 2019 doi:10.20944/preprints201910.0124.v1

Peer-reviewed version available at Int. J. Mol. Sci. 2019, 20, 5460; doi:10.3390/ijms20215460

While both sequence- and structure-based parameters are taken into account when defining the final complex types, the two sets of descriptors have different roles in the scalability of the method. In our presented approach to defining complex types, the main features are structural properties, while sequence parameters are more descriptive in the sense that they highlight the sequential features needed to be able to fold into a complex of given structural properties (Figure 1). However, sequence features can be used to distinguish subtypes of structure-defined complex types. For example, type 1 near ordered complexes come in two flavours according to the two sequence clusters they cover (Figure 1 and Table 1): polar-driven interactions between mostly homodimers, and charge/hydrophobic driven interactions between mostly heterodimers. Also, type 2 complexes (coils and zippers) come in two varieties: relying on polar-driven interactions for heterodimers and charge-driven for homodimers.

In addition to providing a scalable classification scheme, the described method and the defined complex types have biological relevance. The presented complex types have different biological properties; although only information describing the sequence and structure properties were put in, the resulting types show different properties in terms of the energetics and strength of the interactions (Figures 2 and 3), the relevant regulatory processes (Figure 4) and subcellular localization (Figure 5).

The analysis of the energetics properties of the interactions can provide a glimpse into the biophysical details of the binding and folding. The use of low-resolution statistical force fields has proved to be a suitable approach to discriminate complexes based on the structural features of constituent chains [24], and to describe the binding of IDPs [37,38]. While complexes of ordered proteins and domain-recognition IDP binding sites have a fairly narrow range in energetics parameters (Figure 2a), complexes formed exclusively by IDPs are more heterogeneous, basically covering the whole range of the energy spectrum (Figure 2b). Furthermore, based on predictions, MSF complexes cover at least 10 orders of magnitude in K_d values (Figure 3). Hence, in terms of binding strength and stability, these complexes have the potential to cover a very wide range of biological functions, overlapping with those of ordered complexes and domain-binding IDPs as well, in agreement with the previous comparative functional analysis of a wide range of interactions [24].

For most known MSF complexes, the resulting structure is instrumental for proper function, such as the coiled-coil structure for the SNARE complex in mediating membrane fusion [39], the dimeric structure for a wide range of transcription factors in precise DNA-binding [40–42], and the proper coordination of catalytic residues for oligomeric enzymes [43,44]. Therefore, for MSF complexes the interaction de facto switches on the protein function, and hence the precise regulation of the interaction strength is vital in the biological context of these complexes. While structure-based K_d value predictions are informative, in some cases they do not fully describe the interactions. Many MSF complexes are tightly bound, yet they are not necessarily obligate complexes, and their association/dissociation can be under heavy regulation. For example, solely based on K_d values and energetics, type 5 (handshake-like fold) interactions seem to form obligate complexes. However, there are several cases where these interactions do break up in a biological setting, most notably for histones. Histone H4 is able to form dimers with at least 8 different H3 variants [45], and it was described that in the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 October 2019 doi:10.20944/preprints201910.0124.v1

Peer-reviewed version available at Int. J. Mol. Sci. 2019, 20, 5460; doi:10.3390/ijms20215460

case of H3.1 and H3.3, it is governed by a H4 phosphorylation which H3 is preferred [31]. The post-translational modifications can enhance complex formation of dissociation in many other cases as well [29]. In addition, competition for the same binding partner and binding site availability as a function of alternative splicing are additional mechanisms for the regulation of the formation of MSF complexes (Figure 4).

Exploring the precise regulatory mechanisms for MSF complexes would be highly informative. Unfortunately, experimental K_d measurements are lacking for the majority of these interactions, and interactions in structural detail have usually been only analyzed in a single PTM state. Therefore, the molecular details and biologically relevant steps of the regulation of these interactions are difficult to assess; but from a biological sense, it is probable that even several low K_d complexes can dissociate rapidly in certain cases. At least some regulatory mechanisms are currently known for about 36% of studied MSF complexes, but the real numbers are bound to be higher. This means that most probably the majority of MSF complexes are not obligate complexes, where the disordered state is physiologically irrelevant, but can exist in both the bound stable state and the unbound disordered state as well, under native conditions. Thus, MSF complexes are integral parts or direct targets of regulatory networks, although the extent varies with the interaction type considered.

Apart from the studied regulatory mechanisms, additional layers of spatio-temporal regulation can play crucial roles for MSF complexes, similarly to other IDP interactions [35]. An emerging such regulatory mechanism is liquid-liquid phase separation (LLPS). A prime example is the Nck/neuronal Wiskott-Aldrich syndrome protein (N-WASP). N-WASP is known to undergo LLPS when interacting with Nck and nephrin [46], via linear motif mediated coupled folding and binding. Mutually synergistic folding between the secreted EspFu pathogen protein from enterohaemorrhagic *Escherichia coli* and the autoinhibitory GTPase binding domain (GBD) in host WASP proteins (MFIB ID:MF2202002, type 5 complex) hijacks the native LLPS-mediated cellular processes [47], showing that competing interactions are not always stoichiometric in nature, and the true extent of MSF regulation is likely to be even more complex than highlighted here.

The difference between complex types in various biological and biophysical properties shows that these type-definitions reflect true biological differences. Apart from being useful for complex classification, the presented method also shows that differences in binding strength, subcellular localization, and regulation are encoded in the sequence and structural properties of proteins. This can be the basis for developing future prediction methods, where these sequence- and structure-based parameters can be used as input for the prediction of biological features of complexes. In addition, the establishment of MSF complex types has direct implications, as knowledge present for a specific complex might be transferable to other complexes of the same type. For example, certain pathological conditions arise through the pathological aggregation of IDPs. A well known example is transthyretin (TTR) aggregation that can lead to various amyloid diseases, such as senile systemic amyloidosis [48]. Another example from the same near-ordered complex type is the superoxide dismutase SOD1, which is able to form aggregates in ALS [49]. While the localization and the biological function of TTR and SOD1 (hormone transport and enzymatic catalysis) are radically different, their potency of malfunctioning (often connected to various mutations)

Peer-reviewed version available at Int. J. Mol. Sci. 2019, 20, 5460; doi:10.3390/ijms20215460

share a high degree of resemblance. On one hand this marks other type 1 complexes as candidates for toxic aggregation, on the other hand, indicates that potential therapeutic techniques for one complex (e.g. CLR01 for TTR) can give clues about potential targeting of other interactions.

Such structural classification approaches can have a high impact of structure research, most importantly in the study of protein structure or evolution, in training and/or benchmarking algorithms, augmenting existing datasets with annotations, and examining the classification of a specific protein or a small set of proteins [50]. Up to date several structure-based classification approaches have been developed, such as SCOP [26] and CATH [27], which are extended to protein complexes as well. In this sense, previously existing methods are able to classify MSF complexes too. However, the approaches used do not take into account that these structures are only stable in the context of the interaction, and that a certain protein region can adopt fundamentally different structures depending on the interacting partner. The explicit encoding of parameters describing the properties and importance of the interaction into the classification scheme makes current methods unable to accurately describe the spectrum of MSF complexes, and to date, no such classification scheme has been proposed. In contrast to previously existing methods that largely encode the same information [51], the presented MSF classification scheme is highly independent (Table 3), and thus serves as an orthogonal approach capable of properly handling the specific properties of IDP-driven complex formation through mutual synergistic folding.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 October 2019 doi:10.20944/preprints201910.0124.v1

Peer-reviewed version available at Int. J. Mol. Sci. 2019, 20, 5460; doi:10.3390/ijms20215460

4. Data and Methods

Complexes formed through mutual synergistic folding (MSF)

MSF complexes were taken from the MFIB database [22]. Two entries, MF2100018 and MF5200001, from the 205 were discarded due to issues with the corresponding PDB structures 1ejp and 1vzj, as constituent chains have an unrealistically low number of interchain contacts. Problems with these two structures are apparent from the high outlier scores and clash scores provided in the PDB server. As the developed classification scheme relies heavily on structural parameters, we opted to leave these two entries out of the calculations. The final list of entries is given in Supplementary Table S1.

Other complexes of ordered and disordered proteins

As a reference, two other datasets of protein complexes were used. A set of complexes formed exclusively by ordered single domain protein interactors was taken from [24]. These 688 complexes (see Supplementary Table S3) are formed via autonomous folding followed by binding, i.e. both interacting protein chains adopt a stable structure in their monomeric forms, prior to the interaction. A set of 772 complexes with an IDP interacting with ordered domains was taken from the DIBS database [18]. These complexes (see Supplementary Table S4) are formed via coupled folding and binding, where the IDP adopts a stable structure in the context of the interaction.

Calculating sequence features

Similarly to the approach described in [24], the following amino acid groups were used in quantifying sequence composition of proteins: hydrophobic (containing A, I, L, M, V), aromatic (containing F, W, Y), polar (containing N, Q, S, T), charged (containing H, K, R, D, E), rigid (containing only P), flexible (containing only G), and covalently interacting (containing only C). This low-resolution sequence composition at least partially compensates for commonly occurring amino acid substitutions that in most cases do not affect protein structure and function. In all cases, compositions were calculated for the entire complex including all interacting protein chains. An 8th sequence parameter was used to quantify the compositional difference between subunits. This dissimilarity measure was defined as: $\Delta_{total} = \sum_{i=1}^{7} \Delta_i$, where Δ_i is the largest composition difference of residue group *i* between any pair of constituent chains. The average dissimilarities for various sequence-based clusters are shown in Table 1. For exact sequence composition values for all MSF entries see Supplementary Table S1.

Calculating structure features

Secondary structure assignment was performed by DSSP [52] using a three-state classification distinguishing helical ('H','G','I'), extended ('B','E') and irregular ('S','T', unassigned) residues.

Molecular surfaces were calculated using Naccess [53]. Solvent accessible surface area (SASA) was defined by the Nacces absolute surface column. Interface is defined as the increase in SASA as a result of removing interaction partners from the structure. Buried

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 October 2019 doi:10.20944/preprints201910.0124.v1

Peer-reviewed version available at Int. J. Mol. Sci. 2019, 20, 5460; doi:10.3390/ijms20215460

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 analysed 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 Angstrom. For exact structural feature values for all MSF entries see Supplementary Table S1.

Filtering features for clustering

Standard Pearson correlation values were calculated between all sequence and structure features calculated (Supplementary Table S2). If two features show a correlation with an absolute value above 0.7, only one was kept. In each case we discarded the feature that shows a high correlation with a higher number of other features, or the one with the lower standard deviation. In total, none of the 8 sequence parameters were discarded, but 13 out of the 24 structure parameters were omitted from subsequent clustering steps.

Clustering

Both sequence and filtered structure parameters were used as input for clustering, separately. First, hierarchical clustering was done using the scaled features as input, using Euclidean distance and Ward's method (Supplementary figures S1 and S3). Then, k-means clustering was employed, and the within groups sum of squares were plotted as a function of the number of clusters (Supplementary figures S2 and S4). Clustering was done using R with the Ward.D2 and k-means packages.

Energetic features

Interaction energies for residues were calculated using the statistical potentials described in [54]. These interaction potentials were demonstrated to well describe the energetic features of IDP interactions [37], and are the basis for recognizing them from the sequence [38]. These potentials yield dimensionless quantities in arbitrary units, and hence their absolute values bear no direct physical meaning. However, their signs are accurate, and values below 0 correspond to stabilizing interactions. Furthermore, they can be directly compared, and hence more negative values typically correspond to more stable structures. In each analyses, the total energies were calculated from the residue-level interactions from the entire complex. Two residues were considered to be in interaction if there is at least one heavy atom contact between them. Energetic values are given in Supplementary Tables S1 (for MSF complexes), S3 (for ordered complexes) and S4 (for complexes containing both IDPs and ordered domains).

Prediction of Kd values

Dissociation constants for MSF complexes were estimated using the method described in [55]. In each case, the modified PDB structures taken from the MFIB database [22] were used as input. For technical reasons, not all structures yield a K_d value prediction, and thus the number of values used in representing the average per complex type K_ds (Figure 3) are

calculated from fewer values than the actual number of complexes per type. K_d values are listed in Supplementary Table S1.

Post-translational modifications (PTMs), isoforms and competitive binding

Post-translational modifications were taken from the 2 October 2017 version of PhosphoSitePlus [56], PhosphoELM [57] and UniProt [58]. Only PTMs that were identified in low-throughput experiments were used. These were mapped to complex structures using BLAST between UniProt and PDB sequences (Supplementary Table S5). Protein isoforms were taken from the 4 October 2017 version of UniProt (Supplementary Table S6). To determine alternative binding partners for IDPs, all oligomer PDB structures containing the same UniProt region were selected. PDB structures listed as related in the corresponding MFIB entry were removed. Structures containing the same interaction partners as the original complex were also removed (Supplementary Table S7).

GeneOntology terms for assessing subcellular localization

Subcellular localization was represented using GeneOntology [36] terms from the cellular_component namespace. Terms attached to complexes in MFIB were mapped to a restricted set of terms, called CellLoc GO Slim, used in previous studies [24] to compare localization of protein-protein interactions. Terms in CellLoc GO Slim were split into 5 categories: extracellular, intracellular, membrane, nucleus, and other, encompassing other membrane-bounded cellular compartments, such as the lysosome, as well as non-membrane bounded compartments, such as the nucleus. For CellLoc GO terms attached to MSF complexes see Supplementary Table S8.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 October 2019 doi:10.20944/preprints201910.0124.v1

Peer-reviewed version available at *Int. J. Mol. Sci.* **2019**, *20*, 5460; doi:10.3390/ijms20215460

Supplementary Material

Figure S1

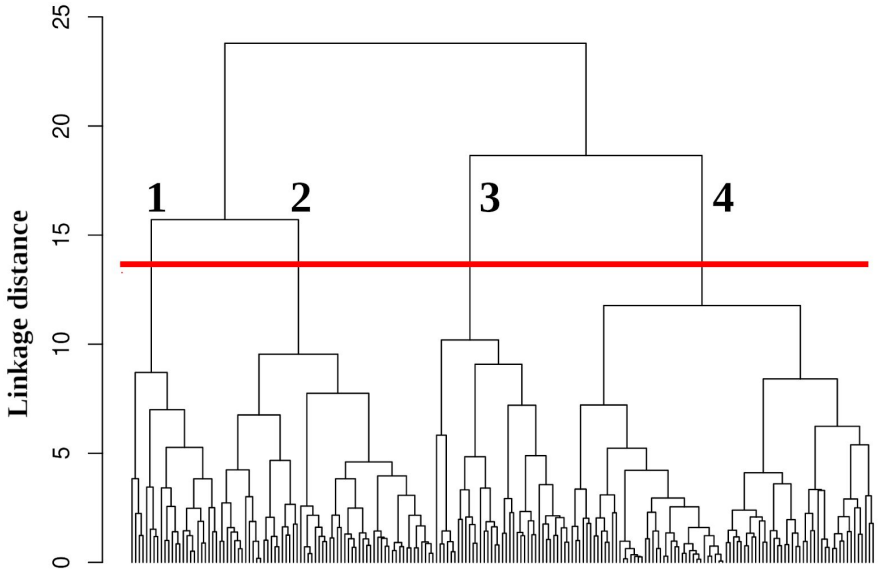


Figure S1: Sequence-based hierarchical clustering of complexes formed via MSF. Red line marks the cutoff used to define the four sequence-based clusters.

Figure S2

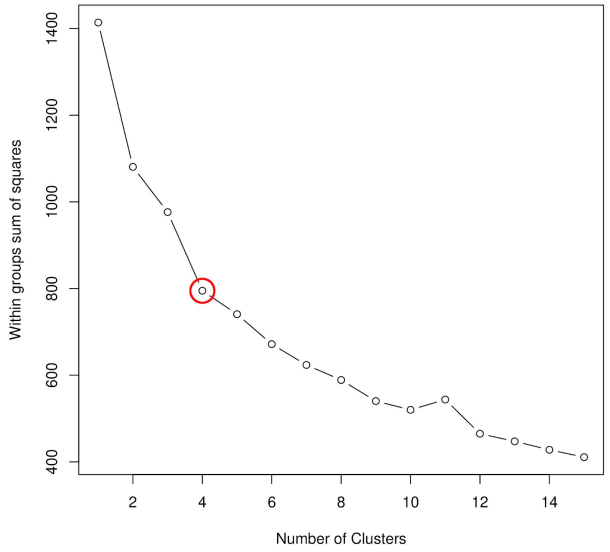


Figure S2: The within groups sum of squares as a function of the number of clusters. k-means clustering was done using sequence parameters as input.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 October 2019 doi:10.20944/preprints201910.0124.v1

Peer-reviewed version available at Int. J. Mol. Sci. 2019, 20, 5460; doi:10.3390/ijms20215460

Figure S3

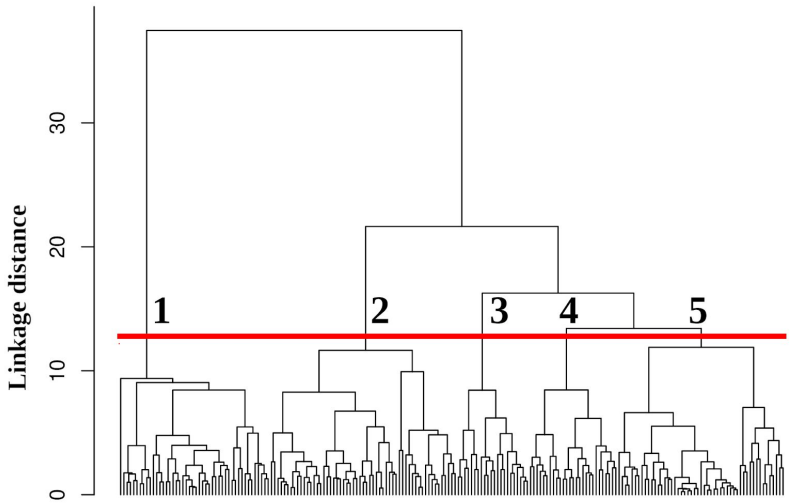


Figure S3: Structure-based hierarchical clustering of complexes formed via MSF. Red line marks the cutoff used to define the five structure-based clusters.

Figure S4

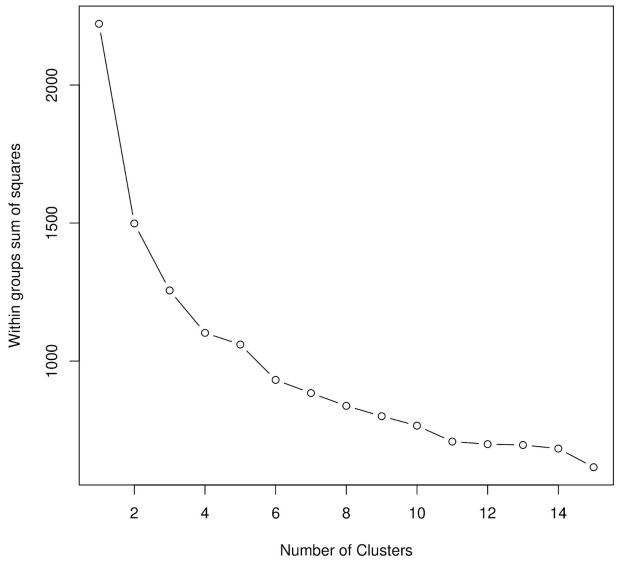


Figure S4: The within groups sum of squares as a function of the number of clusters. k-means clustering was done using structure parameters as input.

Author Contributions

Conceptualization, B.M., I.S.; Methodology, B.M.; Software, B.M., L.D.; Formal Analysis, B.M., L.D., E.F.; Investigation, B.M., I.S.; Resources, B.M., L.D.; Data Curation, B.M.; Writing – Original Draft Preparation, B.M.; Writing – Review & Editing, B.M., L.D., E.F.; Visualization, B.M.; Supervision, B.M., I.S.; Project Administration, B.M., I.S.; Funding Acquisition, B.M., L.D., I.S.

Funding

This research was funded by the EMBO|EuropaBio fellowship 7544 (B.M.), the UNKP-17-3 new national excellence program of the ministry of human capacities of Hungary (L.D.), the project no. FIEK_16-1-2016-0005 financed under the FIEK_16 funding scheme (National Research, Development and Innovation Fund of Hungary) (I.S.), Hungarian Research and Developments Fund OTKA K115698 (I.S.). Bioinformatic infrastructure was supported by ELIXIR Hungary.

Acknowledgements

The authors express gratitude to Dr. Zsuzsanna Dosztányi and Dr. Zoltán Gáspári for their comments on the project.

Conflicts of Interest

The authors declare no conflict of interest.

Abbreviations

IDP	Intrinsically Disordered Protein
MSF	Mutual Synergistic Folding
CFB	Coupled Folding and Binding
PTM	Post-Translational Modification
SOD1	Superoxide Dismutase
Rb	Retinoblastoma protein
SCOP	Structural Classification Of Proteins
CATH	Class/Architecture/Topology/Homologous superfamily
GO	GeneOntology

References

1. Dyson, H.J.; Wright, P.E. Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.* **2005**, *6*, 197–208.
2. Wright, P.E.; Dyson, H.J. Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.* **2015**, *16*, 18–29.
3. Galea, C.A.; Wang, Y.; Sivakolundu, S.G.; Kriwacki, R.W. Regulation of cell division by intrinsically unstructured proteins: intrinsic flexibility, modularity, and signaling conduits. *Biochemistry* **2008**, *47*, 7598–7609.
4. Tompa, P.; Kovacs, D. Intrinsically disordered chaperones in plants and animals. *Biochem. Cell Biol.* **2010**, *88*, 167–174.
5. He, J.; Chao, W.C.H.; Zhang, Z.; Yang, J.; Cronin, N.; Barford, D. Insights into degron recognition by APC/C coactivators from the structure of an Acm1-Cdh1 complex. *Mol. Cell* **2013**, *50*, 649–660.
6. Mészáros, B.; Kumar, M.; Gibson, T.J.; Uyar, B.; Dosztányi, Z. Degrons in cancer. *Sci. Signal.* **2017**, *10*.
7. Gsponer, J.; Futschik, M.E.; Teichmann, S.A.; Babu, M.M. Tight regulation of unstructured proteins: from transcript synthesis to protein degradation. *Science* **2008**, *322*, 1365–1368.
8. van der Lee, R.; Buljan, M.; Lang, B.; Weatheritt, R.J.; Daughdrill, G.W.; Dunker, A.K.; Fuxreiter, M.; Gough, J.; Gsponer, J.; Jones, D.T.; et al. Classification of intrinsically disordered regions and proteins. *Chem. Rev.* **2014**, *114*, 6589–6631.
9. Dosztányi, Z.; Chen, J.; Dunker, A.K.; Simon, I.; Tompa, P. Disorder and sequence repeats in hub proteins and their implications for network evolution. *J. Proteome Res.* **2006**, *5*, 2985–2995.
10. Cortese, M.S.; Uversky, V.N.; Dunker, A.K. Intrinsic disorder in scaffold proteins: getting more from less. *Prog. Biophys. Mol. Biol.* **2008**, *98*, 85–106.
11. Harmon, T.S.; Holehouse, A.S.; Pappu, R.V. Differential solvation of intrinsically disordered linkers drives the formation of spatially organized droplets in ternary systems of linear multivalent proteins. *New Journal of Physics* 2018, *20*, 045002.
12. Davey, N.E.; Van Roey, K.; Weatheritt, R.J.; Toedt, G.; Uyar, B.; Altenberg, B.; Budd, A.; Diella, F.; Dinkel, H.; Gibson, T.J. Attributes of short linear motifs. *Mol. Biosyst.* **2012**, *8*, 268–281.
13. Tompa, P.; Fuxreiter, M. Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. *Trends Biochem. Sci.* **2008**, *33*, 2–8.
14. Sugase, K.; Dyson, H.J.; Wright, P.E. Mechanism of coupled folding and binding of an intrinsically disordered protein. *Nature* **2007**, *447*, 1021–1025.
15. Wang, Y.; Chu, X.; Longhi, S.; Roche, P.; Han, W.; Wang, E.; Wang, J. Multiscaled exploration of coupled folding and binding of an intrinsically disordered molecular recognition element in measles virus nucleoprotein. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, E3743–52.
16. Shammas, S.L.; Crabtree, M.D.; Dahal, L.; Wicky, B.I.M.; Clarke, J. Insights into Coupled Folding and Binding Mechanisms from Kinetic Studies. *J. Biol. Chem.* **2016**, *291*, 6689–6695.
17. Demarest, S.J.; Martinez-Yamout, M.; Chung, J.; Chen, H.; Xu, W.; Dyson, H.J.; Evans, R.M.; Wright, P.E. Mutual synergistic folding in recruitment of CBP/p300 by p160 nuclear receptor coactivators. *Nature* **2002**, *415*, 549–553.
18. Schad, E.; Fichó, E.; Pancsa, R.; Simon, I.; Dosztányi, Z.; Mészáros, B. DIBS: a

repository of disordered binding sites mediating interactions with ordered proteins. *Bioinformatics* **2017**.
19. Fukuchi, S.; Sakamoto, S.; Nobe, Y.; Murakami, S.D.; Amemiya, T.; Hosoda, K.; Koike, R.; Hiroaki, H.; Ota, M. IDEAL: Intrinsically Disordered proteins with Extensive Annotations and Literature. *Nucleic Acids Res.* **2012**, *40*, D507–11.
20. Miskei, M.; Antal, C.; Fuxreiter, M. FuzDB: database of fuzzy complexes, a tool to develop stochastic structure-function relationships for protein complexes and higher-order assemblies. *Nucleic Acids Res.* **2017**, *45*, D228–D235.
21. Yu, J.-F.; Dou, X.-H.; Sha, Y.-J.; Wang, C.-L.; Wang, H.-B.; Chen, Y.-T.; Zhang, F.; Zhou, Y.; Wang, J.-H. DisBind: A database of classified functional binding sites in disordered and structured regions of intrinsically disordered proteins. *BMC Bioinformatics* **2017**, *18*, 206.
22. Fichó, E.; Reményi, I.; Simon, I.; Mészáros, B. MFIB: a repository of protein complexes with mutual folding induced by binding. *Bioinformatics* **2017**, *33*, 3682–3684.
23. Mészáros, B.; Beáta, E.G.S.; Schád, E.; Agnes, T.; Abukhairan, R.; Horváth, T.; Nikoletta, M.; Kovács, O.P.; Kovács, M.; Tosatto, S.C.E.; et al. PhaSePro: the database of proteins driving liquid-liquid phase separation. *Nucleic Acids Res.* **2019**.
24. Mészáros, B.; Dobson, L.; Fichó, E.; Tusnády, G.E.; Dosztányi, Z.; Simon, I. Sequential, Structural and Functional Properties of Protein Complexes Are Defined by How Folding and Binding Intertwine. *J. Mol. Biol.* **2019**.
25. wwPDB consortium Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res.* **2019**, *47*, D520–D528.
26. Chandonia, J.-M.; Fox, N.K.; Brenner, S.E. SCOPe: classification of large macromolecular structures in the structural classification of proteins-extended database. *Nucleic Acids Res.* **2019**, *47*, D475–D481.
27. Sillitoe, I.; Dawson, N.; Lewis, T.E.; Das, S.; Lees, J.G.; Ashford, P.; Tolulope, A.; Scholes, H.M.; Senatorov, I.; Bujan, A.; et al. CATH: expanding the horizons of structure-based functional annotations for genome sequences. *Nucleic Acids Res.* **2019**, *47*, D280–D284.
28. Zhao, N.; Pang, B.; Shyu, C.-R.; Korkin, D. Structural similarity and classification of protein interaction interfaces. *PLoS One* **2011**, *6*, e19554.
29. Redler, R.L.; Wilcox, K.C.; Proctor, E.A.; Fee, L.; Caplow, M.; Dokholyan, N.V. Glutathionylation at Cys-111 induces dissociation of wild type and FALS mutant SOD1 dimers. *Biochemistry* **2011**, *50*, 7057–7066.
30. Van Roey, K.; Gibson, T.J.; Davey, N.E. Motif switches: decision-making in cell regulation. *Curr. Opin. Struct. Biol.* **2012**, *22*, 378–385.
31. Kang, B.; Pu, M.; Hu, G.; Wen, W.; Dong, Z.; Zhao, K.; Stillman, B.; Zhang, Z. Phosphorylation of H4 Ser 47 promotes HIRA-mediated nucleosome assembly. *Genes Dev.* **2011**, *25*, 1359–1364.
32. Lee, C.W.; Ferreon, J.C.; Ferreon, A.C.M.; Arai, M.; Wright, P.E. Graded enhancement of p53 binding to CREB-binding protein (CBP) by multisite phosphorylation. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 19290–19295.
33. Bousset, K.; Oelgeschläger, M.H.; Henriksson, M.; Schreek, S.; Burkhardt, H.; Litchfield, D.W.; Lüscher-Firzlaff, J.M.; Lüscher, B. Regulation of transcription factors c-Myc, Max, and c-Myb by casein kinase II. *Cell. Mol. Biol. Res.* **1994**, *40*, 501–511.
34. Saddic, L.A.; West, L.E.; Aslanian, A.; Yates, J.R., 3rd; Rubin, S.M.; Gozani, O.; Sage, J. Methylation of the retinoblastoma tumor suppressor by SMYD2. *J. Biol. Chem.* **2010**, *285*, 37733–37740.
35. Gibson, T.J. Cell regulation: determined to signal discrete cooperation. *Trends Biochem. Sci.* **2009**, *34*, 471–482.
36. The Gene Ontology Consortium The Gene Ontology Resource: 20 years and still GOing

strong. *Nucleic Acids Res.* **2019**, *47*, D330–D338.
37. Mészáros, B.; Tompa, P.; Simon, I.; Dosztányi, Z. Molecular principles of the interactions of disordered proteins. *J. Mol. Biol.* **2007**, *372*, 549–561.
38. Mészáros, B.; Erdos, G.; Dosztányi, Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.* **2018**, *46*, W329–W337.
39. Strop, P.; Kaiser, S.E.; Vrljic, M.; Brunger, A.T. The structure of the yeast plasma membrane SNARE complex reveals destabilizing water-filled cavities. *J. Biol. Chem.* **2008**, *283*, 1113–1119.
40. Bonvin, A.M.; Vis, H.; Breg, J.N.; Burgering, M.J.; Boelens, R.; Kaptein, R. Nuclear magnetic resonance solution structure of the Arc repressor using relaxation matrix calculations. *J. Mol. Biol.* **1994**, *236*, 328–341.
41. Madl, T.; Van Melderen, L.; Mine, N.; Respondek, M.; Oberer, M.; Keller, W.; Khatai, L.; Zangger, K. Structural basis for nucleic acid and toxin recognition of the bacterial antitoxin CcdA. *J. Mol. Biol.* **2006**, *364*, 170–185.
42. Sauvé, S.; Tremblay, L.; Lavigne, P. The NMR solution structure of a mutant of the Max b/HLH/LZ free of DNA: insights into the specific and reversible DNA binding mechanism of dimeric transcription factors. *J. Mol. Biol.* **2004**, *342*, 813–832.
43. Le Trong, I.; Stenkamp, R.E.; Ibarra, C.; Atkins, W.M.; Adman, E.T. 1.3-A resolution structure of human glutathione S-transferase with S-hexyl glutathione bound reveals possible extended ligandin binding site. *Proteins* **2002**, *48*, 618–627.
44. Dams, T.; Auerbach, G.; Bader, G.; Jacob, U.; Ploom, T.; Huber, R.; Jaenicke, R. The crystal structure of dihydrofolate reductase from Thermotoga maritima: molecular features of thermostability. *J. Mol. Biol.* **2000**, *297*, 659–672.
45. Tachiwana, H.; Osakabe, A.; Shiga, T.; Miya, Y.; Kimura, H.; Kagawa, W.; Kurumizaka, H. Structures of human nucleosomes containing major histone H3 variants. *Acta Crystallogr. D Biol. Crystallogr.* **2011**, *67*, 578–583.
46. Banjade, S.; Wu, Q.; Mittal, A.; Peeples, W.B.; Pappu, R.V.; Rosen, M.K. Conserved interdomain linker promotes phase separation of the multivalent adaptor protein Nck. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112*, E6426–35.
47. Cheng, H.-C.; Skehan, B.M.; Campellone, K.G.; Leong, J.M.; Rosen, M.K. Structural mechanism of WASP activation by the enterohaemorrhagic E. coli effector EspF(U). *Nature* **2008**, *454*, 1009–1013.
48. Westermark, P.; Sletten, K.; Johansson, B.; Cornwell, G.G., 3rd Fibril in senile systemic amyloidosis is derived from normal transthyretin. *Proc. Natl. Acad. Sci. U. S. A.* **1990**, *87*, 2843–2845.
49. Pansarasa, O.; Bordoni, M.; Diamanti, L.; Sproviero, D.; Gagliardi, S.; Cereda, C. SOD1 in Amyotrophic Lateral Sclerosis: “Ambivalent” Behavior Connected to the Disease. *Int. J. Mol. Sci.* **2018**, *19*.
50. Fox, N.K.; Brenner, S.E.; Chandonia, J.-M. The value of protein structure classification information-Surveying the scientific literature. *Proteins* **2015**, *83*, 2025–2038.
51. Hadley, C.; Jones, D.T. A systematic comparison of protein structure classifications: SCOP, CATH and FSSP. *Structure* **1999**, *7*, 1099–1112.
52. Touw, W.G.; Baakman, C.; Black, J.; te Beek, T.A.H.; Krieger, E.; Joosten, R.P.; Vriend, G. A series of PDB-related databanks for everyday needs. *Nucleic Acids Res.* **2015**, *43*, D364–8.
53. Hubbard, S.; Thornton, J. *Naccess*; 1992;.
54. 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.* **2005**, *347*, 827–839.
55. Vangone, A.; Bonvin, A.M. Contacts-based prediction of binding affinity in protein-protein

complexes. *Elife* **2015**, *4*, e07454.

56. Hornbeck, P.V.; Zhang, B.; Murray, B.; Kornhauser, J.M.; Latham, V.; Skrzypek, E. PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res.* **2015**, *43*, D512–20.
57. Dinkel, H.; Chica, C.; Via, A.; Gould, C.M.; Jensen, L.J.; Gibson, T.J.; Diella, F. Phospho.ELM: a database of phosphorylation sites--update 2011. *Nucleic Acids Res.* **2011**, *39*, D261–7.
58. UniProt Consortium UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **2019**, *47*, D506–D515.