

Orders of Disorder: Computational Analysis of the Interactions of Intrinsically Disordered Proteins

PhD Thesis

Erzsébet Fichó

Structural Biochemistry Program, Doctoral School of Biology
Faculty of Science, Eötvös Loránd University

Head of the Doctoral School: Prof. Anna Erdei, DSc

Doctoral Program Leader: Prof. Mihály Kovács, DSc

Supervisor:

Dr. Bálint Mészáros, PhD

MTA-ELTE “Momentum” Bioinformatics Research Group, Eötvös Loránd University

Protein Structure Research Group, Institute of Enzymology
Research Centre for Natural Sciences, Hungarian Academy of Sciences



Budapest, Hungary
2018

Preface

This dissertation describes my work done between September 2014 and December 2018 in the Structural Biochemistry Program of the Doctoral School of Biology of the Eötvös Loránd University in the Protein Structure Research Group of the Institute of Enzymology, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Budapest under the guidance of Prof. István Simon and the supervision of Dr. Bálint Mészáros.

The dissertation is based on my results published (or submitted) in the following papers:

- Fichó E, Reményi I, Simon I, Mészáros B. *MFIB: a repository of protein complexes with mutual folding induced by binding*. *Bioinformatics*. 2017;33(22):3682-3684.
- 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*. 2018;34(3):535-537.
- Mészáros B, Dobson L, Fichó E, Tusnády EG, Dosztányi Z, Simon I: *How folding and binding intertwine during protein complex formation provides an additional layer of functional regulation*. *BioRxiv*, 2017 <https://doi.org/10.1101/211524>
- Magyar C, Mentés A, Fichó E, Cserző M, Simon I. *Physical Background of the Disordered Nature of "Mutual Synergetic Folding" Proteins*. *Int J Mol Sci*. 2018;19(11)

Other publications:

- Mészáros B, Dosztányi Z, Fichó E, Magyar C, Simon I: *Bioinformatical Approaches to Unstructured/Disordered Proteins and Their Complexes*. In Liwo A. *Computational Methods to Study the Structure and Dynamics of Biomolecules and Biomolecular Processes, From Bioinformatics to Molecular Quantum Mechanics*. Springer; 2018. Edited by Adam Liwo.
- Fichó E*, Shád E*, Simon I, Tompa P, Pancsa R: *Structural principles of primate positively selected residues*. In preparation – 2019
*Contributed equally

Acknowledgments

First and foremost, I would like to express my special appreciation and thanks to my group leader, Prof. Simon István and my supervisor, Dr. Bálint Mészáros, they have been tremendous mentors for me. I appreciate all their contributions of time, ideas, and funding to make my PhD experience productive and stimulating.

Nevertheless, I am grateful to all my past and present co-workers at the Protein Structure Research Group of the Institute of Enzymology. I would like to thank Prof. László Buday, the head of the Institute of Enzymology for providing me with the opportunity of working in the Institute.

I want to thank all my friends for their encouragement and coffee breaks.

Lastly, I want to thank my beloved parents and my little sister for supporting me spiritually throughout doing my PhD, writing this thesis and my life in general. And I would like to thank István Reményi whose faithful support during the final stages of my PhD is so appreciated.

Summary

Intrinsically Disordered Proteins (IDPs) exist without a well-defined tertiary structure, often playing roles in critical biological functions such as transcription, cell cycle regulation, and stress response. IDPs are abundant in eukaryotes, providing crucial elements of protein-protein interaction networks often playing vital roles in signal transduction, and hence many IDPs are involved in disease development as well. While IDP interactions are key to understanding these processes, their targeted and systematic analyses are lacking. In order to enable these analyses, first of all, large-scale manually curated datasets are needed.

Developing a combined automated and manual annotation pipeline, I assembled two databases - together with their dedicated web-servers for their dissemination - that focus on the two principal interaction mechanisms of IDPs. The Mutual Folding Induced by Binding (MFIB) database - available at <http://mfib.enzim.ttk.mta.hu/> - is the repository for protein complexes formed exclusively by IDPs through a process termed mutual synergistic folding (MSF). The Disordered Binding Sites (DIBS) database - available at <http://dibs.enzim.ttk.mta.hu/> - collects protein complexes that are formed between an IDP and ordered protein partners via coupled folding and binding. These databases are the first structure-focused, large scale, publicly available IDP interaction resources that can serve as the foundations of detailed systematic IDP interaction studies.

My comparative analysis between complexes from MFIB, DIBS and those formed by globular proteins, show that the presence and the structural characteristics of a binding partner profoundly affects the sequence, bound structure, subcellular localization, function, and regulation of proteins. This highlights the interplay between folding and binding, and how it can affect the function and regulation of the resulting protein complexes.

The analysis of features of proteins from different interaction classes not only enabled the delineation of the characteristic differences between them, but can also be used to lay the foundation of the currently missing classification system of MSF complexes. Developing a sequence/structure centric clustering approach, I recognized 6 separate biologically meaningful classes of MSF complexes. The defined groups have marked differences in multiple biological properties, and the proposed classification system can be used to hint at possible common functional roles of various MSF interactions in both health and disease states.

Összefoglalás

A rendezetlen fehérjék (Intrinsically Disordered Proteins - IDPs) nem rendelkeznek időben állandó térszerkezettel. Rendezetlenségük révén azonban számos létfontosságú biológiai funkció ellátására képesek: kiemelkedő szerepet játszanak a transzkripcióban, a sejtválaszban, különböző regulációs és jelátviteli folyamatokban, melyek során gyakran más fehérjékkel hatnak kölcsön. A rendezetlen fehérjék igen gyakoriak az eukarióta szervezetekben, a fehérje-fehérje kölcsönhatási hálózatok létfontosságú kulcsszereplői, az elmúlt évek során pedig a különböző betegségekben betöltött szerepük is világossá vált. A rendezett fehérjék által kialakított kölcsönhatásokat viszonylag jól ismerjük szerkezeti és funkcionális szempontból, azonban azok a kölcsönhatások, amelyeket rendezetlen fehérjék alakítanak ki egymással, vagy más rendezett fehérje-partnerekkel, jelenleg kevésbé jól tanulmányozottak. Ennek ellenére az eddig ismert néhány példa is azt bizonyítja, hogy a rendezetlen fehérjék által kialakított kölcsönhatások rendkívül fontosak, hiszen kapcsolatba hozhatók esszenciális gének transzkripciós szabályozásával, különböző daganatok kialakulásával vagy hoszt-patogén kölcsönhatásokkal.

A rendezetlen fehérjék kölcsönhatásainak átfogó vizsgálatának egyik hiányzó eleme a megfelelő minőségű és méretű, annotált adathalmaz volt. Doktori munkám fókuszpontjában a rendezetlen fehérjék és az általuk kialakított kölcsönhatások állnak, különös tekintettel azon esetekre, ahol a résztvevő két vagy több fehérje monomer formában rendezetlen, és egymást rendezik a kölcsönhatásuk során (Mutual Synergistic Folding - MSF).

Első lépésként automatizált és manuális szűrési eljárások eredményeként két adatbázist hoztam létre: a Mutual Folding Induced by Binding (MFIB – elérhető: <http://mfib.enzim.ttk.mta.hu/>) adatbázis az egymással kölcsönhatásba lépő és egymást kölcsönösen rendező rendezetlen fehérjéket tartalmazza. A Disordered Binding Sites (DIBS – elérhető: <http://dibs.enzim.ttk.mta.hu/>) olyan fehérje komplexeket tartalmaz, ahol egyetlen lánc rendezetlen, és az rendezett láncokhoz kötődve rendeződik (coupled folding and binding). Mindkét adatbázis tartalmazza a fehérjék szerkezeti és funkcionális annotációját is, mindezt felhasználóbarát módon prezentálva, akár letölthető formában is. A szisztematikus gyűjtések eredményeképpen az MFIB 205 egyedi, térszerkezettel rendelkező fehérjét listáz, míg „testvéradatbázisa”, a DIBS 773 fehérjét tartalmaz.

Az így összegyűjtött adatok és a rendezett fehérjékről eddig rendelkezésre álló ismereteim immár elegendőek voltak ahhoz, hogy megtegyem az első lépéseket a rendezetlen fehérjék által kialakított kölcsönhatások szekvenciális, szerkezeti és funkcionális elemzése felé. Az általam végzett vizsgálatok rámutattak arra, hogy milyen mély összefonódás van a fehérjék rendezetlenségi állapota és kölcsönhatásai között, milyen hatása van a rendezetlenségnek és a kölcsönható partnereknek a fehérjék szekvenciális és szerkezeti jellemzőire, hogyan befolyásolják a kialakított komplexek sejten belüli lokalizációját, funkcióját, és a fehérje-fehérje kölcsönhatási hálózatokban betöltött szerepeiket, valamint elvezetnek a különböző jelátviteli mechanizmusok mélyebb megértéséhez is.

Az analízis során tett megállapítások arra is rávilágítottak, hogy az MSF-komplexeket is osztályozni tudjuk: összesen hat különböző, biológiailag releváns alcsoportot tudtam bennünk megkülönböztetni, mindegyik klaszter markáns szekvenciális, szerkezeti, funkcionális és szabályozási jellemzőkkel rendelkezik. Az általam használt csoportosítás magában rejti az egymáshoz képest hasonló szerkezetek közötti kapcsolat mélyebb megértésének lehetőségét, jövőbeli farmakológia vizsgálatok kiindulópontjaként szolgálhat, illetve más fehérjeszerkezetek újfajta osztályozási lehetőségét is megteremti.

Table of contents

Preface	i
Acknowledgments	i
Summary	ii
Összefoglalás	iii
Table of contents	v
1. Scientific Background	1
1.1. Introduction to proteins.....	1
1.1.1. Levels of protein structure	1
1.1.2. Dominant forces in protein folding and stability	3
1.1.3. The energy landscape view of protein folding.....	4
1.2. Intrinsically disordered proteins	6
1.2.1. Re-assessing the structure-function paradigm	6
1.2.2. Basic properties of IDPs	6
1.2.3. Biological functions of IDPs.....	8
1.3. Interactions of intrinsically disordered proteins	9
1.3.1. Coupled folding and binding	9
1.3.2. Mutual synergistic folding	11
1.3.3. Interactions with non-protein molecules.....	12
1.3.4. Role of IDPs in protein-protein interactions networks	13
1.4. Detecting and predicting protein disorder	14
1.4.1. Experimental techniques to identify IDPs and their interactions	14
1.4.2. Overview of protein disorder prediction techniques.....	16
1.4.3. Prediction of disordered binding regions	17
1.5. Repositories of IDPs	18
1.5.1. Databases of experimentally validated IDPs	18
1.5.2. Databases based on predictions	20
1.5.3. MobiDB: the central source for IDPs	20
2. Scientific Aims	22
3. Data and Methods	24
3.1. Databases	24
3.1.1. PDB.....	24
3.1.2. UniProt and UniRef	24
3.1.3. DisProt	25
3.1.4. IDEAL	25
3.1.5. Pfam.....	25

3.1.6. Functional annotations	26
3.1.7. Other databases	29
3.2. Algorithms	30
3.2.1. BLAST.....	30
3.2.2. DSSP.....	31
3.2.3. Naccess	31
3.2.4. FoldX	31
3.2.4. Other calculations	32
3.3. Development of web servers.....	33
3.4. Other programs and programming environments	34
4. Results and Discussion.....	35
4.1. Assembly of the MFIB and DIBS databases	35
4.1.1. Collecting potential data for MFIB and DIBS.....	35
4.1.2. Annotating protein complexes for MFIB and DIBS.....	36
4.1.3. Web interface for the “twin-databases”	41
4.1.4. Statistics of MFIB entries	45
4.1.5. Distribution of data in DIBS	48
4.2. Analysis of sequence, structure and function relationships in different protein interaction classes	52
4.2.1. Amino acid composition mirrors the connection between folding and binding	53
4.2.2. The presence of IDPs affects the structural properties of the resulting complexes	54
4.2.3. Closer to the DNA, closer to the IDP-mediated interactions.....	57
4.2.4. Interactions of IDPs mediate distinct biological functions in the cell	58
4.2.5. Protein disorder extends the biologically relevant sequence, structure, and functional spaces of proteins.....	60
4.2.6. IDPs are heavily regulated via post-translational modifications	62
4.2.7. Cooperation between ordered proteins and IDPs in different interaction modes	66
4.3. Classification of MSF complexes	69
4.3.1. Sequence and structure features of MSF complexes	69
4.3.2. Energetic properties of interactions	74
4.3.3. Regulatory mechanisms of MSF complexes	76
4.3.4. Catalog of MSF complexes.....	79
5. Conclusions and Future Directions	87
References.....	90

1. Scientific Background

1.1. Introduction to proteins

Proteins are fundamental biomolecules in living cells, fulfilling a great variety of functions: amongst others, they play a role in repair and maintenance, transport and storage processes, can act as enzymes, hormones or antibodies. The functional and structural fate of proteins often depends on their interactions; most of them bind to other molecules: proteins, nucleic acids, small molecules or ions. Proteins adopt structures in three dimensions, and to understand and describe their different functions, we need structural interpretation. For example to understand proteolysis, a cleavage site usually only requires an amino acid sequence, but to understand the function of the corresponding enzyme, we need a spatial representation. In general, to describe protein structure, we distinguish four different, hierarchical levels (see Figure 1).

1.1.1. Levels of protein structure

Proteins are linear polymers of amino acid building blocks, generally built up by the 20 standard amino acids (although after synthesis, the polypeptide chain may undergo additional post-translational chemical modifications). A condensation reaction can link two amino acids, and this creates a peptide bond between the carbon atom in the carboxyl group of one, and the nitrogen atom in the amine group of the other. During the polymerization, a chain of amino acids can be built up with theoretically unlimited number of residues. The linear sequence of residues is termed the primary structure, the simplest level of the protein structure, and can generally be inferred from DNA/mRNA sequences that encode them [1].

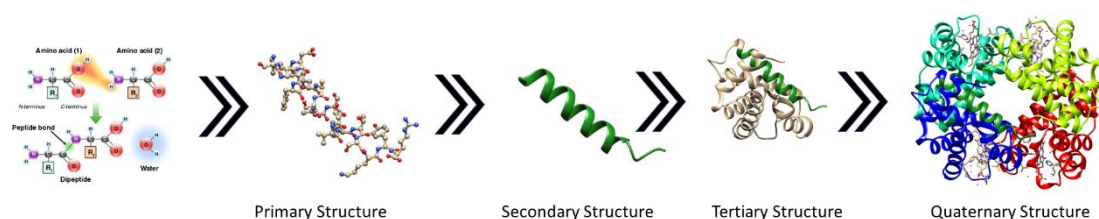


Figure 1: Hierarchy of protein structure.

For secondary, tertiary and quaternary structures only the backbone of the protein polymer is represented, without the amino acid sidechains.

The second level of protein structure, termed secondary structure, describes the local spatial conformation of backbone atoms in consecutive residues of the polypeptide chain, which can be characterized by three torsion angles for each residue. The rotations of the polypeptide backbone around the peptide bonds between N-C α (ϕ) and C α -C (ψ) are the two dihedral angles used in the description of secondary structures. There is a third possible torsion angle within the protein backbone (ω) which is mostly flat and fixed at 180 degrees. Ramachandran plots are the most common tools to visualize the secondary structure content of a protein, plotting the distribution of dihedral angles ϕ against ψ of every residue in the protein (termed Ramachandran plot) [2].

The two most common periodic secondary structural elements are α -helices and β -sheets, although, there are other local features, such as regions of the polypeptide where a change of chain direction occurs, including beta turns and hairpins, and other types of helices. α -helices are stabilized by H-bonds between the main chain atoms of the i^{th} and $i+4^{\text{th}}$ residues, and can be characterized with dihedral angles around -60° , -45° . Other less frequent types of helical structures include 3^{10} helices, which are stabilized by hydrogen bonds of the kind (i, i+3) and the π -helix, which is stabilized by hydrogen bonds of the type (i, i+5). β -sheets consist of several β -strands, stretched segments of the polypeptide chain kept together by a network of hydrogen bonds. From the point of view of experiments, Circular Dichroism (CD) measurements can be used to determine the relative amount of different secondary elements, because they exhibit distinctly different CD spectra [3].

The tertiary structure defines the full, 3D conformation of the protein chain. The third level of protein structure is defined by the coordinates of atoms of the protein, usually given in a Cartesian coordinate system. For simplified description, the topology of the secondary structural elements of the protein can be used instead. The adoption of a stable tertiary structure happens through a physical process called the folding of the protein. Knowing the 3D structure of proteins is essential for any attempt to understand how they work and how they interact with each other. Different experimental methods can be used to discover the details of a protein structure. The X-ray crystallography method can be used in the crystal state, while Nuclear Magnetic Resonance (NMR) spectroscopy can be employed in aqueous environment, and therefore can provide additional information about flexibility in the native solution state. Electron microscopy is also capable of determining the structure of various (usually above 200 kDa)

biomolecules. Nowadays, cryo-electron microscopy can overcome this limitation and plays an increasingly important part in the determination of protein structures.

Proteins are made up of a single polypeptide chain and have only three levels of structure. However, a native functional state of a protein is often not a single chain but an assembly of several chains. Protein complexes are made up of multiple polypeptide chains (known as subunits). The quaternary structure describes the spatial orientation of the subunits after forming their interactions. The oligomer may be composed of different (heteromultimer) or identical (homomultimer) subunits.

1.1.2. Dominant forces in protein folding and stability

The folded structure of a protein depends on the interplay of a vast number of interactions and linkage between its atoms. The covalent bonds define the connectivity of atoms in the primary sequence, but are not sufficient to define the fine details of a three-dimensional structure. The denaturation of a protein (the loss of the tertiary structure) is, therefore, the consequence of breaking non-covalent bonds that stabilize the native state. The main non-covalent driving forces behind the organization of secondary and tertiary structures are diverse, and mainly encompass the hydrophobic effect, electrostatic interactions, H-bonds, and van der Waals interactions.

The structural characteristics of a protein heavily depend on the interactions of its residues with the solvent water, affecting protein dynamics as well. The effect of the solvent is termed the hydrophobic effect, and is the dominant force behind the folding of proteins, being responsible for the burial of the hydrophobic residues in the core of the protein, while charged residues and to a lesser extent polar residues are disfavored at buried sites.

Electrostatic interactions arise both from ionizable amino acids and from polar groups that contain permanent dipoles. A salt bridge is a non-covalent interaction between two ionized sites within 5Å range. Furthermore, attraction can occur between the partial charges of polar groups. As a distinct mechanism, hydrogen bonding can occur between an acceptor and a donor group, partially exchanging a proton. Hydrogen bonding between the peptide backbone atoms represents the dominant stabilizing force of the secondary structure elements.

The Van der Waals forces are the relatively weak forces that arise between non-charged atoms. Apart from the dipole-dipole interactions, Van der Waals attraction arises from transient, random fluctuations of induced dipoles in the electron cloud surrounding

an atom. The strength of these interactions depends strongly on the distance. Although Van der Waals interactions are relatively weak, they can provide a non-negligible component of stability because of their large number.

1.1.3. The energy landscape view of protein folding

In the 1950s, Christian Anfinsen wanted to show that the information for protein folding resides entirely within the amino acid sequence of the protein. The postulate (also known as a thermodynamic hypothesis or Anfinsen's dogma) and its experimental validation was awarded a chemistry Nobel prize in 1972. In 1969, Cyrus Levinthal noted that it would take a nearly infinite amount of time for an unfolded protein to search through its full conformational space by using a random walk - the contradiction between the number of possible conformations and fast folding rates (the problem termed as Levinthal's paradox) [4].

The use of energy landscape theory of protein folding provides an alternative description of the transitions between the possible molecular conformations [5], that solves Levinthal's paradox. The energy landscape of a well-folding protein resembles a partially rough funnel [6]. The funnel-like energy surface of a protein ensures that proteins reach their native state from any unfolded conformation through specific intermediate states (folding intermediates). In the folding funnel theory, it is clear that the folding is not a random search, there are different folding pathways, the free energy gradient directs the transition from one conformation to the other.

A simplified schematic cartoon is typically used to illustrate the folding process using an internal free energy landscape view (see Figure 2). A point on the multidimensional surface represents a unique conformation of the protein.

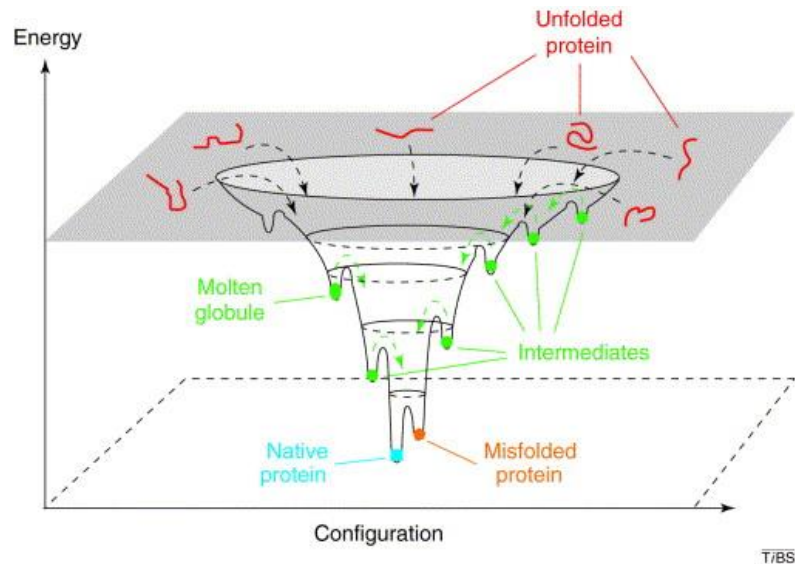


Figure 2: The funnel-like energy surface of protein folding [7].

The vertical axis represents the internal free energy, and the x and y-axes represent the conformational space. Therefore, each point in the graph is a possible conformation, having an assigned internal free energy value. This energy contains the enthalpic term; and it includes the contributions from hydrogen bonds, ion-pairs, torsion angle energies, and the hydrophobic and solvation free energies by averaging over the conformational space of water molecules. The funnel-shaped energy landscape of well-folding proteins ensures that the bottom of the funnel - the only global energy minimum - is the native state. The native state is kinetically accessible for the protein and has lower energy than any other single conformation at a local minimum. The picture of a folding funnel is able to capture that there might be multiple parallel pathways that are all channeled towards the unique native structure.

The protein folding process illustrated by folding landscapes can be quantified using basic thermodynamics equations. For this we distinguish between two well-defined states: a folded and denatured state. The conformational stability of a protein then can be determined using the equation

$$\Delta G_{\text{protein}} = -R T \ln K$$

where R is the gas constant, T is the temperature, K is the equilibrium constant between the unfolded and the folded state and $\Delta G_{\text{protein}}$ determines the total change in the Gibbs free energy between the denatured and the folded state of the protein. For a protein to remain folded, the ΔG has to be larger than the value of thermal fluctuations, and also the Gibbs free energy difference between the global and other, local minima.

1.2. Intrinsically disordered proteins

A major challenge in the current post-genome era will be the determination of the functions of the proteins encoded in genomic DNA sequences. The first decades of classical structural biology were driven by the idea that a protein needs a stable structure to function. As Francis Crick remarked: “If you want to understand the function, study the structure.” According to the structure-function paradigm, the function of the protein directly depends on its well-folded 3D structure [5, 8]. The past decade has witnessed major conceptual advances in our understanding of protein structure-function relationships regarding the ubiquitous existence of intrinsically disordered proteins, which defy the classical structure-function paradigm.

1.2.1. Re-assessing the structure-function paradigm

For a long time, the classical structure-function paradigm was the dominant view of protein research. The recent extension of the paradigm tries to encompass proteins that do not necessarily require a stable, 3D structure - even under physiological conditions - to fulfill their biological role [9]. These Intrinsically Unstructured/Disordered Proteins (IUPs/IDPs) lack a well-defined, stable structure in isolation; instead, they exist as an ensemble of different conformations and can still carry out biological functions.

Using bioinformatics predictors, it was estimated that more than 50% of eukaryotic proteins contain at least one long disordered segment (more than 30 residues long) [10], marking the importance of IDP research. Also, based on bioinformatics studies of available fully sequenced species, the frequency of IDPs is generally thought to be much higher in eukaryotes compared to prokaryotes [11]. Eukaryotic proteins are generally more complex than prokaryotic proteins, they are on average longer, contain more amino acid repeat patterns, have more Intrinsically Disordered Regions (IDRs), and they have an increased need for regulation in their more complex cellular environments. In accordance, IDPs are typically involved in many critical high-level processes such as transcription, translation, regulation, signal transduction, and stress response [12, 13].

1.2.2. Basic properties of IDPs

IDPs were characterized to have significantly different amino acid compositions compared to globular proteins. The hallmarks of IDPs are general depletion of amino acids with low flexibility indices, high net charge, and low net hydrophobicity [14].

Moreover, IDPs often exhibit low sequence complexity, as they often contain regions with little diversity in their amino acid composition [15].

Regarding the secondary and tertiary structure levels, IDPs fall onto a structural continuum (see Figure 3). There are different levels of flexibility in their isolated states, e.g. molten globule (fairly stable secondary structures without a tertiary structure, such as the nuclear receptor coactivator binding domain from the CREB-binding protein CBP [16]) or random coil (with no or almost no structural elements with the vast majority of residues being highly mobile, such as the Arc-repressor [17]). Functional sites of IDPs often display transient secondary structure elements (such as the transactivation N-terminal segment of p53 [18]), have a preference for hydrophobic residues, and have higher contact per residue ratio [19, 20]. In several modular proteins IDRs can also connect various independently folded globular domains as flexible linker sequences. These regions are highly variable according to their amino acid compositions and length. In addition, linker IDRs can act like structural strings, to attach distant spatial domains, and enable their binding.

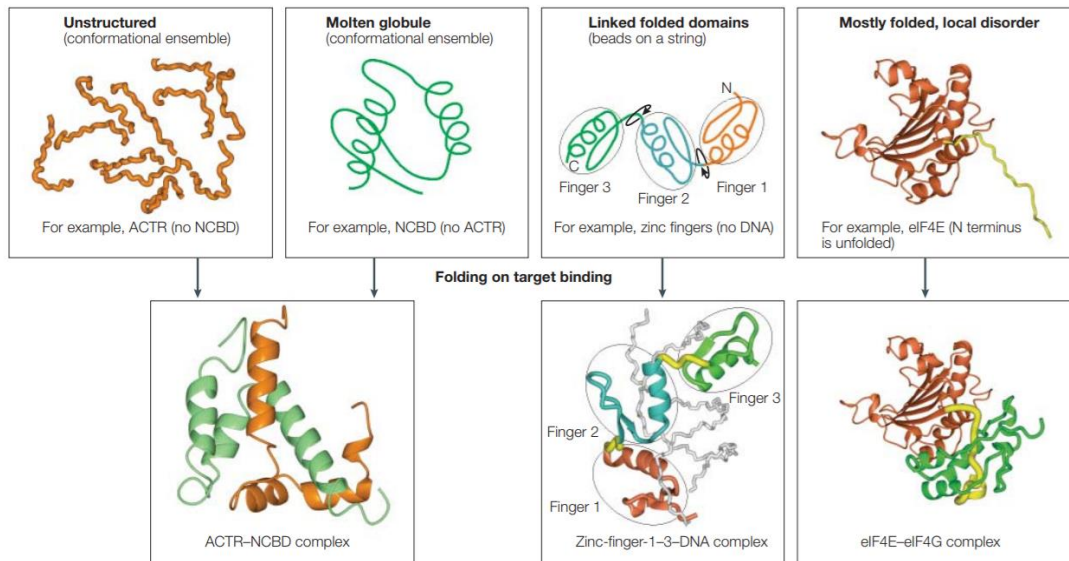


Figure 3: Proteins span a broad continuum of structural states [12].

According to the energy landscape view, for globular proteins, the global energy minimum represents a well-defined native conformation. In contrast, IDPs are characterized by the absence of a global minimum, as the differences between the lowest energy state and other local minima are in the range of thermal fluctuations [21, 22].

1.2.3. Biological functions of IDPs

In general, IDPs are involved in many biological processes. Their functional repertoire allows us to classify IDPs, and they can be largely categorized in six distinct functional groups [23].

IDPs are recognized to play a role in molecular recognition, acting as display sites. Post-Translational Modifications (PTMs) - such as phosphorylation or limited proteolysis - can regulate the interaction affinity of IDPs, their turnover, and their localization within the cell. For example, the activity of the transactivation domain of the cAMP Response Element Binding Protein (CREB), which binds TATA-box-associated factors, is regulated by phosphorylation. Chaperones have evolved to assist nascent proteins in reaching their native fold or to unfold these molecules if they are misfolded, to provide a chance for correct refolding. Chaperons are also capable of preventing the aggregation of their targets, and in some cases dispersing aggregates as well.

A different function of IDPs is acting as entropic chains, flexible linkers or spacers, which can connect functional regions, ordered domains or disordered segments, without a specific interaction. These IDPs carry out functions that benefit directly from their conformational disorder, from the ability of the polypeptide chain to fluctuate between a high number of different conformational states, eg. MAP2, RA70. A closely related function of IDPs is the entropic clock, which can mediate timer functions, as exemplified by the inactivation mechanism of Shaker channel of nerve axons [24]. Entropic springs, such as titin can contract reversibly after stretching.

Effector IDPs interact with other proteins and modify their activity, which involves activator and inhibitory roles during binding. When the disordered p27 protein is bound to the CyclinA-CDK2 complex, it leads to the inhibition of cell cycle. If an effector IDP has both activator and inhibitory functions, it is termed as a moonlighting.

Due to their frequent involvement in protein-protein interactions, IDPs often act as assemblers, that bring together multiple binding partners to promote the formation of higher-order protein complexes, such as activated T-cell receptor complexes.

As scavengers, IDPs can store and neutralize small ligands, such as ions or organic compounds, such as Chromogranin A or casein. Casein is a member of the family of proteins in the milk of mammals, and was one of the first recognized IDPs. It is traditionally thought to serve as nutrient for breastfed newborns, and functions by binding and neutralizing calcium phosphate.

The previous grouping illustrates the functional repertoire of IDPs, playing diverse and essential roles in the maintenance of life. In accord, the misregulation of IDPs are associated with several diseases, including various types of cancer, cardiovascular disease and diabetes [25-28]. For example, the tumor suppressor protein, p53, acts as a "cellular gatekeeper" playing crucial roles in the regulation of apoptosis, cell cycle, DNA-repair and senescence. The loss of p53 activity occurs in about 50% of human cancers and has a profound impact on the associated pathways. IDPs can also have a high potential to aggregate, a phenomenon common in neurodegeneration [29-31].

1.3. Interactions of intrinsically disordered proteins

As apparent from the previous functional overview, IDP functions often heavily rely on their interactions. In order to understand the wide range of functions of IDPs, we need to understand the intermolecular interactions formed by IDPs with other ordered or disordered proteins, or with other non-protein molecules.

The high chain flexibility and conformational disorder of IDPs and IDRs enable them to form complementary binding interfaces with their targets more easily than an ordered domain. Combining high specificity with low affinity is another widely-mentioned advantage of IDPs. High plasticity of IDPs allows them to bind to multiple partners more readily by changing conformations or interaction regions according to the templates provided by different target molecules.

1.3.1. Coupled folding and binding

IDPs are capable of binding to and folding upon the surface of ordered protein partners, in a process termed coupled folding and binding [32, 33]. The flexibility of the disordered partner decreases due to the binding, and the loss of entropy during the folding of the disordered partner results in a weaker overall binding compared to globular proteins. This type of interaction thus holds the potential to form transient interactions. Moreover, this way the specificity, which is independent of the entropic terms, is uncoupled from binding strength [32, 34].

There are two main concepts to describe the interacting regions in IDPs undergoing coupled folding and binding [35]. One is the definition of the structural transition between the disordered to the ordered state, initiated by the binding event. As a complementary approach, linear motifs - also referred to as short linear motifs or minimotifs - are short

functional sites typically found in disordered protein regions [36], that are defined in a sequence-centric way. Linear motif definitions specify common residues (constituting a "patch" as a motif), that mediate the binding largely independent of the other regions of the protein they are embedded in, functioning autonomously. The majority of protein-protein interaction-mediating linear motifs were described in eukaryotes.

Figure 4 shows a protein complex involving three proteins, the previously formed complex between the ordered cyclinA and Cyclin-Dependent Kinase 2 (CDK2) proteins, inhibited by the disordered p27 protein. The interaction between cyclinA and CDK2 plays a crucial role in the control of the transition between the S and G2 cell-cycle phases of eukaryotic cells. The segment of p27 involved in the binding shows only little helical preferences in the unbound form. However, some regions adopt a well-defined α -helix upon binding. The interacting residues of p27 are dominated by hydrophobic and aromatic residues that fit into hydrophobic clefts and grooves on the surface of the cyclinA-CDK2 complex. This interaction has also been described using linear motifs, as the disordered interacting region of p27 contains a cyclin docking motif [37].

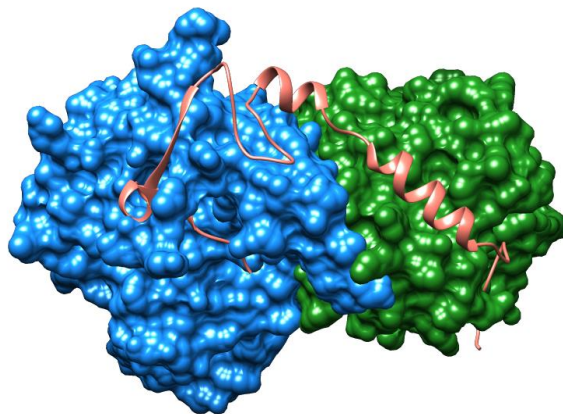


Figure 4: Example of interfaces between two ordered proteins and a disordered protein. The ordered CDK2 and cyclinA are shown in blue and green surface representations, respectively. The disordered p27 is shown in salmon cartoon representation. The figure was generated from the 1jsu PDB file.

Studies show that not only the interface of p27, but the interface of IDPs in general are more hydrophobic than typical ordered protein interfaces, in the case of coupled folding and binding [20]. Thus, typically the main preferred contacts are hydrophobic-hydrophobic interactions between the partner proteins, accounting for the majority of the high number of intermolecular contacts formed by these IDRs [34].

Exempt from the general trends, not all IDPs undergo folding upon binding to a partner protein. There are a few known cases where a single conformation cannot describe the complex structure, because the IDP partner adopts multiple conformations in the complex even after the binding. So-called "fuzziness" can retain the inherent dynamics of the IDP/IDR in the bound state [38]. Fuzzy protein complexes typically have weak, transient interactions.

1.3.2. Mutual synergistic folding

During coupled folding and binding, the IDP partner reaches the ordered state using the ordered protein partner(s) as a template that drives the folding process during the interaction. However, several IDPs are able to adopt stable structures during interactions without a folded partner. The folding of all participating protein partners happens at the same time: coupled with the interaction in a synergistic manner through a process called Mutual Synergistic Folding (MSF). In an MSF complex, all interacting partners are IDPs without stable tertiary structures outside of the complex.

MSF proteins are very sporadic in the literature, and we have only a handful of collected cases [19, 39-41]. The interaction between ACTR and CBP proteins was one of the first well-documented examples [16]. During the folding, ACTR and CBP form an ordered complex where a near random coil protein and a molten globule protein stabilize each other (see Figure 5).

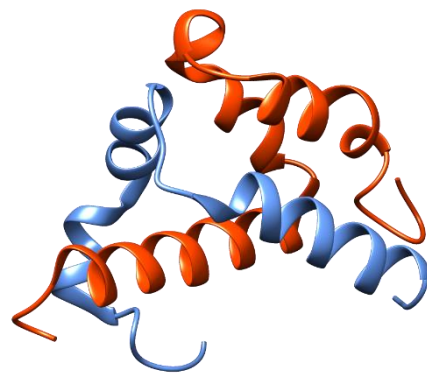


Figure 5: Example of a protein complex formed via mutual synergistic folding.

ACTR and CBP are shown in blue and red cartoon representations, respectively. The figure was generated from the 1kbh PDB file.

1.3.3. Interactions with non-protein molecules

IDPs are most commonly involved in protein-protein interactions. Nevertheless, to participate in various biological processes, they often bind to other non-protein molecules as well, such as nucleic acids, lipids, and ions.

IDPs often interact with DNA/RNA molecules. Permanent versions of such interactions are key to the organization of large nucleic acid-protein complexes, such as the chromatin and the ribosome, both of which are the essential organelles of the cell. In the case of ribosomal proteins, it has been shown that their inherent flexibility is most likely critical in the very assembly of the ribosome [42, 43]. Apart from tight and obligate interactions formed by core chromatin and ribosomal IDPs, a wide range of mRNA-binding proteins also often have a high level of disorder. Their interactions regulate all stages of gene expression from transcription, through mRNA processing and folding, to the regulation of translation. Histones and chromatin organizing proteins, together with proteins involved in the recruitment and assembly of the transcription machinery, function through the recognition of specific or aspecific DNA sequences, and many of these protein regions also appear disordered in solution [42]. HES-1 is a transcriptional repressor belonging to the basic helix-loop-helix family, and is one of the main downstream effectors in Notch signaling, which functions as a mediator of short-range cell-cell communication.

Several examples of unfolded domains or IDPs binding to membranes have been reported. These lipid-IDP interactions are frequent in eukaryotic membrane proteins and play a role in transmembrane signal transduction [44]. Multichain Immune Recognition Receptors (MIRRs) are found on the surface of T cells and B cells, and the cytoplasmic domains of MIRR signaling subunits are intrinsically unstructured in both monomeric and oligomeric states [45].

Structural disorder is also frequently present in metal-binding proteins. The interaction of d-block metal ions (Cu, Zn, Fe) with IDPs is weaker and more flexible than with a structured metalloprotein [46]. Several IDPs have been noted for their ability to bind metal ions with low affinity but high capacity, and some of these proteins are implicated in several diseases, mainly neurodegenerative. Prothymosin- α is a natively unfolded, and highly conserved protein located mostly in the nucleus of eukaryotic cells. The exact biological role of the protein is unknown, nevertheless, it has been shown to be

involved in cell proliferation, chromatin remodeling, antiapoptotic activity, and is able to bind metal cations [47].

Calcium binding IDPs are mostly extracellular (in a folded form), or are involved in secretion or calcium sequestering, and thus, they reside in locations where Ca^{2+} concentration is relatively high. Binding of the divalent calcium ion has a great potential to force structural change, including protein folding [48]. Calmodulin is an ubiquitous and highly conserved Ca^{2+} sensor that interacts with a wide variety of eukaryotic proteins and enzymes, controlling their activities in response to calcium [49].

The above examples are to illustrate the heterogeneity of the interactions of IDPs, which are prevalent in the cell, performing a wide range of signaling and regulatory pathways. This central role can be better captured with higher-level description of protein interactions, using the analysis of protein-protein interactions networks.

1.3.4. Role of IDPs in protein-protein interactions networks

IDRs/IDPs most often function through Protein-Protein Interactions (PPI), when they permanently or transiently bind partner molecules with diverse functional consequences. Disorder predictors, when applied to whole proteomes, indicated that the fraction of proteins with substantial amounts of disorder is predicted to be higher in eukaryotes than archaea or eubacteria. Structural flexibility and plasticity represent major functional advantages in the case of IDPs. They can interact with a broad range of binding partners, and their disordered nature enables an increased speed of interaction [50].

In one-to-many binding, a single disordered binding site binds to two or more different partners individually [51]. Disordered regions can bind partners with both high specificity and low affinity, suggesting that disorder-based signaling and regulatory interactions can be highly specific but easily reversed [52]. IDPs are sensitive to combinatorial post-translational modifications and alternative splicing, adding greater complexity to PPI regulatory networks and providing a mechanism for tissue-specific signaling as well.

Protein interaction networks display approximate scale-free topology, in which the highly connected proteins of PPIs, commonly referred to as hubs or hub proteins that interact with a vast number of other proteins, determine the overall organization of the PPI network. Different studies show that central hub proteins are enriched in IDPs [52-54]. The disordered state is allowing the same polypeptide to undertake different interactions, often with different functional outcomes. The more highly connected a

protein node is, the more physically interacting partners it has, the more critical it is for normal cellular function, and the more likely its removal will be detrimental to a cell [55].

Several hub proteins have been shown to be completely or almost entirely disordered in solution. A prime example is the high mobility group HMGA1 protein, which is a relatively small IDP of 107 residues, with random coil-like structure, participating in a variety of cellular processes including gene transcription, integration of retroviruses into chromatin, induction of neoplastic transformation, and promotion of metastatic progression [56]. The intrinsic flexibility of HMGA1 allows it to undergo reversible disorder-to-order structural transitions upon binding to its partners, and it can induce conformational changes in the bound DNA and protein substrates. HMGA1 serves as a hub for nuclear function (The IntAct database lists more than 50 binary interactions for the human HMGA1, including nucleic acids and proteins; the BioGrid database contains 108 HGMA1 interactions) [42, 52].

As IDP-mediated interactions are central to protein-protein interaction networks and cellular functions, they need to be tightly regulated to ensure precise signaling in time and space. Mutations in IDPs or changes in their cellular abundance are implicated in various diseases. Because IDPs are also involved in many disease pathways, they are also increasingly considered as potential drug targets [57, 58].

1.4. Detecting and predicting protein disorder

Although IDPs are abundant, their recognition has taken many years. During classical structural experiments, IDRs were either overlooked, unreported, or intentionally removed to enable crystallization, often being thought of as "unimportant" for function. Evidence of recent years point out how IDPs are common in the cell, and we need various tools to identify them. We know just a limited number of them, and emerging techniques for IDP identification can take scientific findings in a new direction or can provide a broader context for the reported discoveries about IDPs

1.4.1. Experimental techniques to identify IDPs and their interactions

The physical and chemical properties of IDPs display marked differences compared to globular proteins. IDPs have unusual behavior as they often resist heat, remain soluble under extreme conditions, have unusual SDS-page mobility, moving more slowly through the gel than globular proteins. The reason for these phenomena stems in their unique

amino acid composition being highly charged and having a low content of hydrophobic amino acids [23]. IDPs display enhanced proteolytic sensitivity [59], and are mostly insensitive to chemical denaturation due to the lack of a folded structure [33, 60].

Electron and X-ray crystallography have traditionally excelled at determining the structure of a single folded/ordered protein. IDPs, however, are not directly amenable for these two static structural determination methods [12], given that IDPs do not adopt a stable structure (and hence do not crystallize) unless bound to other partners. However, these methods can still provide - albeit sometimes indirect - information about protein disorder. Missing residues from the electron-density map can indicate the presence of disordered regions in X-ray crystallography [61]. To characterize IDPs at atomic resolutions and determine their structural properties, NMR spectroscopy is the most potent method [44, 62]. IDPs are highly flexible, fluctuating between a large number of alternative conformations, and various NMR methods are capable of detecting molecular motions over a wide range of timescales [63]. Apart from the previous techniques, CD measurements have been widely used to assess the secondary structure content of a protein, and are particularly useful for detecting the presence of transient secondary structural elements in IDPs [3, 64].

These direct experimental techniques can be complemented with various indirect experimental approaches. Hydrodynamic techniques, such as gel filtration and dynamic light scattering, can also aid IDP identification as they report on the radius of the protein, which is often larger for an IDP or denatured protein than for a folded protein of the same mass [65]. Isothermal titration calorimetry is one of the few methods available for completely evaluating the thermodynamic parameters describing coupled folding and binding and MSF in IDP-interactions [66]. Differential scanning calorimetry is another thermal analytic technique that is often used in characterizing IDPs. The transition to a more ordered/disordered state appears as a heat-absorption curve, and can provide information indirectly on protein disorder. During heat denaturation, the characteristics of the melting curves representing phase transitions can indicate the structural states of the monomers [67].

Small-Angle X-ray Scattering (SAXS) enables a low-resolution structural characterization of biological macromolecules in solution. SAXS not only provides shapes, oligomeric states and quaternary structures of folded proteins or protein complexes, but also allows quantitative analysis of flexible systems, such as IDPs. The

technique can be highly complementary to the high-resolution methods of X-ray crystallography and NMR.

In addition, complementing solution NMR with Magic-Angle Spinning (MAS) solid-state techniques allows us to study higher molecular mass aggregates that IDPs/IDRs may form, which are not accessible through solution NMR alone [62]. In MAS techniques, spinning the sample at the magic angle with respect to the direction of the magnetic field increases the resolution, enabling better identification and analysis of the spectrum [68].

1.4.2. Overview of protein disorder prediction techniques

Most of the above mentioned and other currently existing experimental procedures are costly and time-consuming, require a large amount of protein, and some provide only indirect information about the disorder. Because of experimental difficulties, bioinformatics tools that target the prediction of protein disorder from the sequence, play an important role in the identification and characterization of IDPs. Currently only these bioinformatic tools can give us large-scale information about the basic properties, evolution, and functions of IDPs [69].

In an algorithmic sense, the prediction of protein disorder can be viewed as a classic binary classification problem: the whole protein or its residues can be labeled disordered or ordered. Most prediction methods rely on the biased sequence features characterizing disordered segments and provide predictions at the per-residue level. The performance of disorder predictors can be evaluated with different metrics, such as balanced accuracy or the area under the ROC curve. In machine learning techniques, a Matthews correlation coefficient is also accepted as a measure of the quality of binary (two-class) classifications. Since 2002, the performance of various disorder prediction methods has been critically assessed at the CASP experiments [70].

The first prediction methods were simply direct implementations of physical principles governing the process of protein folding [14, 33, 71, 72]. The simplest methods, such as FoldIndex [73], rely on amino acid composition and mostly use charge and the scale of hydrophobicity or flexibility to discriminate various amino acid properties. Another physicochemical-based approach is the IUPred algorithm, which predicts disorder from a single amino acid sequence using a dedicated statistical potential, which is optimized to estimate the pairwise interaction energies between residues [71, 74].

In contrast to biophysics-based methods, machine learning approaches can automatically distill relationships between the input sequence features and the output properties (the ordered/disordered state of residues). As an input, usually the amino acid sequence within a local sequence window is used. The most commonly used machine learning techniques are support vector machines (SVMs) and neural networks.

The most widely used class of standard machine learning algorithms are SVMs, such as PONDR VSL2b [75, 76]. They have several advantages over neural networks, as they are less prone to overfitting, can be trained more efficiently, and handle noisy datasets better. Members of the PONDR family [77, 78], such as PONDR VL-XT [15] and ESpritz [79] are typically feedforward neural networks that use sequence attributes.

Meta approaches integrate the results of several pre-existing prediction methods [70, 80] and achieve improved performance by decreasing the noise of individual predictors. Meta-predictors dominated the last round of CASP experiment. DisCoP [81] uses a regression model to produce a new disorder prediction from seven methods (DisProt and X-ray versions of ESpritz, CSpritz [82], SPINE-D [83], DISOPRED2 [84], MD [85] and DISOclust [86, 87]).

1.4.3. Prediction of disordered binding regions

Interacting linear motifs are usually short and evolutionarily variable segments, which in most cases, fall into locally disordered regions [36]. The characteristic feature of disordered binding regions is that in isolation they exist as ensembles of rapidly interconverting conformations. During interactions, these sequence regions can undergo coupled folding and binding [35]. Due to their specific functional and structural properties, disordered binding regions have distinct properties compared to both globular proteins and IDPs in general, and these properties enable the construction of prediction algorithms to recognize them from the protein sequence. Predicting binding regions in IDPs is a challenging task, and as of yet we can only estimate such interacting regions involved in coupled folding and binding, while sites in MSF remain with no dedicated prediction algorithms.

The ANCHOR method can capture the basic biophysical properties of disordered binding segments from their amino acid sequences [88, 89]. The method's performance is largely independent from the adopted secondary structures of the binding regions. ANCHOR uses the IUPred energy predictor matrix, combined with the pairwise

interaction energies of the given residue may gain by forming intrachain contacts, and when interacting with a globular protein represented by an average amino acid composition.

In contrast to the biophysics modeling approach behind ANCHOR, the DISOPRED3 method has an SMV technique with the nearest neighbor classifier, capable of predicting disordered binding sites, also from the sequence of the protein [90].

MoRFpred can predict disordered binding sites adopting an α -helical conformation in their bound form. The algorithm uses an SVM model, which is built from evolutionary profiles, predicted disorder, relative solvent accessibility, and physicochemical properties [91]. MoRFpred-Plus is the improved version of MoRFpred, and it combines the previous release with Hidden Markov Models [92]. The MoRFchibi method is also a significantly improved version of MoRFpred that can be easily integrated into custom bioinformatics analysis pipelines. The suite also offers MoRFchibiWeb [93], as part of the MoRFchibi SYSTEM. The OPAL method is a result of PROMIS and MoRFchibi. It is evaluated using the same test sets that were previously used to evaluate MoRFpred [94]. OPAL accepts a single protein sequence of length greater than 26 residues as an input, and has online web services as well.

1.5. Repositories of IDPs

1.5.1. Databases of experimentally validated IDPs

There are several databases which contain experimentally verified IDPs, serving as a starting point for future analysis of the properties of IDPs. One of the first such repositories was the DisProt database (available at <http://www.disprot.org>). The DisProt repository aims to collect manually curated IDPs and protein regions characterized by various experimental techniques [95]. After a major update in 2017, the current 7.0 release of the database holds information on 2,167 IDP regions in 803 proteins. Entries come from the literature identified by text-mining in PubMed abstracts or manual annotations done by curators, from the previous version of DisProt (with several entries re-annotated), and from novel cases identified as PDB entries with long regions of missing electron density. Every entry contains at least one experimentally verified disordered region. Detection methods include X-ray crystallography, NMR spectroscopy, CD spectroscopy (both far and near UV) and protease sensitivity, in addition to several other, less frequently used experimental techniques.

Beside DisProt, the other largest IDP resource is the IDEAL database (Intrinsically Disordered proteins with Extensive Annotations and Literature, available at <http://www.ideal.force.cs.is.nagoya-u.ac.jp/IDEAL/>). Apart from disorder annotations, IDEAL also contains experimentally verified IDP interaction regions from 913 proteins by manual curation. The current version also contains PTM sites, references, and structural domain assignments as well [96]. Moreover, IDEAL explicitly annotates regions undergoing coupled folding and binding, when the disordered and ordered information are both available in the region.

DisBind (Disorder Binding Sites, available at <http://biophy.dzu.edu.cn/DisBind>) is dedicated to the classification of functional binding sites of IDPs from the DisProt database. DisBind is a residue-level collection of experimentally supported binding sites in IDPs, according to their binding partners, including proteins, RNA, DNA and metal ions [97]. The current version of DisBind contains 226 IDPs with functional site annotations.

While several of the above datasets incorporate IDPs in bound structures, IDPs can also bind in fuzzy structures, folding into several alternative structured conformations or remaining mostly disordered, exhibiting a fast exchange of conformations in their bound form. Fuzzy protein complexes can be found in FuzDB (available at <http://protdyn-database.org/>) with their structural and biochemical evidence for disorder. The database also includes higher-order assemblies, and presents a detailed analysis of the structural and functional data [98]. FuzDB currently contains over 100 fuzzy complexes.

While these databases define their target scope based on structural states (the binding protein segment has to be disordered in isolation), the largely overlapping sequence-based interaction definition of linear motifs also has dedicated databases. The largest motif database, ELM (Eukaryotic Linear Motifs, available at <http://elm.eu.org/>) [99], contains motif definitions together with several bound IDP structures. ELM is a hub for collecting and classifying short linear motif instances in a curated way from the experimental literature. The database contains over 275 motif classes and over 3,000 motif instances, and a pipeline to discover candidate short linear motifs in protein sequences.

1.5.2. Databases based on predictions

Besides the experiment-based databases, there is a repository for IDPs identified using prediction algorithms. The Database of Disordered Protein Prediction (D²P² - available at <http://d2p2.pro/>) stores predicted annotations of intrinsic disorder [100]. The primary purpose of the database is to enable disorder analysis of complete proteomes. It provides previously computed disorder prediction outputs for more than 1,500 whole proteomes of more than 1,200 species. D²P² includes the following predictors: PONDR VL-XT [15], PONDR VSL2b [75, 76], PrDOS [101], PV2 [102], ESpritz (all variants), IUPred (all variants) along with ANCHOR. These predictors have different run times and have different approaches to IDPs. Slower methods can not analyze whole proteomes in a realistic time, therefore D²P² can enable computational analysis without unrealistic computing capacity on behalf of users.

1.5.3. MobiDB: the central source for IDPs

Due to the difficulty of experimental characterization of IDPs, their databases usually collect a few hundred proteins, however, there are millions of known protein sequences. Some IDP-focused repositories include automatically generated predictions of intrinsic disorder. After a significant update, MobiDB 3.0 is a centralized resource for extensive disorder annotation for all protein sequences in the UniProt database. Another motivation behind the major update of MobiDB is to give end users a convenient user interface and user experience, and also give advanced stable programmatic access through web services and easily accommodate new annotations. The database features three levels of annotation: manually curated, indirect and predicted. It also includes a consensus annotation for long disordered regions. MobiDB 3.0 is organized by both the type of disorder annotation and the quality of disorder evidence. MobiDB includes experimental annotations of disorder taken from DisProt and the PDB, disorder information derived from NMR chemical shift data, and also integrates information from other specific disorder databases [103]. For predicted IDPs, a consensus annotation is provided including various predictors: ESpritz, DisEMBL, IUPred, GlobPlot [104], VSL2b [76]. These predictors enable MobiDB to provide disorder annotations for every protein, even when no indirect or curated data is available. MobiDB 3.0 contains information for the complete UniProt protein set. IDP annotations by MobiDB 3.0 were included in the core data of UniProt.

MobiDB 3.0 also offers access to the structural characterization of the putative disorder in MobiDB-Lite, to help interpret its role and function [105]. The MobiDB-Lite prediction algorithm was developed to recognize long IDRs, thus the method has been benchmarked on the largest possible dataset based on X-ray missing residues. MobiDB-Lite uses eight different predictors to derive a consensus, which is then filtered for false, short predictions and optimized to predict long IDRs. MobiDB-Lite is fully integrated into the MobiDB database, and also has a downloadable version.

2. Scientific Aims

IDPs exist as conformational ensembles without adopting a 3D structure in isolation. In the past two decades, experimental and bioinformatics studies have shown that IDPs play a central role in various signaling and regulatory processes. These proteins can adopt a stable structure upon interacting with other, ordered proteins via coupled folding and binding. However, some IDPs can also form an ordered complex via mutual synergistic folding. The known instances of IDP-containing complexes lag far behind their expected numbers, not exclusively due to the difficulty of experimental analysis of IDPs. Moreover, there is no comprehensive database that collects them; they are very sporadic in the literature [106].

The central biological roles of IDPs motivate us to investigate the interactions formed by them to understand their molecular functions. However, analysis of IDPs is currently hindered by the lack of structured, accessible data concerning their interactions. To help overcome this hindrance, I aimed to build databases to collect specific IDP-interactions, and using this information to perform a detailed analysis of the interacting parts of IDPs, in order to understand the interplay between folding and binding, and its connection to biological function.

The specific aims of my PhD work are the following:

1. In order to perform a comprehensive analysis of IDPs, we need large-scale datasets. There were no systematically collected, structure-based repositories for MSF complexes or coupled folding and binding complexes before. As a starting point, I aimed to build disorder-specific databases to collect these types of interactions.
2. The analyses of the IDP interactions in these databases serve as a cornerstone of a more profound understanding of IDP-mediated interactions, answering the following questions: What are the main differences between various ways IDPs adopt a stable structure through interactions, concerning their sequence, structure, and function? What are the significant sequential differences between complexes formed by only ordered proteins, an IDP and ordered protein partners, and IDP-only complexes? What are the main differences in structural parameters? What are the typical functions mediated by the three types of interactions?

3. Are there intrinsic classes of protein complexes formed by mutual synergistic folding? Is there an objective way to define these groups? What are the main characteristics of these groups at the sequence, structure and function levels? How are these proteins and their interactions regulated?

3. Data and Methods

3.1. Databases

3.1.1. PDB

The Protein Data Bank (PDB) contains experimentally determined three-dimensional structural data of biological molecules, such as nucleic acids and proteins [107]. The structures were determined by X-ray crystallography, NMR spectroscopy, or - increasingly - cryo-electron microscopy. The PDB contains annotations and cross-references to other databases, such as UniProtKB for sequences, Pfam for domain assignments, SCOP for structural classifications, Phosphosite for PTMs, or ExPasy for genomic information. Its website (available at <https://www.rcsb.org/>) offers multiple tools for structure query, browsing, analysis, and molecular visualization. Currently (as of October 2, 2018), around 144,000 structures have been deposited into the database. For the database assemblies, version of March 28, 2017 was used.

3.1.2. UniProt and UniRef

The Universal Protein Knowledgebase (UniProtKB, available at <https://www.uniprot.org/>) is the central resource of protein sequences and associated detailed annotations [108, 109]. UniProt unifies SwissProt, which contains manually curated protein entries, and TrEMBL, the automatically annotated extension of SwissProt. The annotations cover high-quality information about the sequence included, amongst others, biologically relevant domains and sites, functions, possible variants, post-translational modifications, structure details, diseases associated with deficiencies or abnormalities of the protein, and cross-references to other databases.

For the analysis of regulatory mechanisms of MSF complexes in chapter 4.3.3, protein isoforms were taken from UniProt. To determine alternative binding partners for IDPs, all oligomer PDB structure containing the same UniProt (using BLAST) region were selected.

The UniProt Reference Cluster (UniRef) is a clustered version of UniProt, containing representative sequences selected from UniProt in order to reduce redundancy, while maintaining full coverage of the sequence space [110]. UniRef100 provides a comprehensive non-redundant sequence collection clustered by sequence identity and

taxonomy with source attribution. UniREF90 and UniREF50 are built from UniRef100 to provide non-redundant sequence collections. All records from all source organisms with a bilateral sequence identity of >90% or >50%, respectively are merged into a single record. During the construction of MFIB and DIBS, Uniref90 was used to transfer disorder annotations and to reduce redundancy.

3.1.3. DisProt

DisProt (available at <http://www.disprot.org/>) is a curated database of disordered proteins and protein regions characterized by various experimental techniques [95]. The current version contains 803 proteins and 2,167 disordered protein regions with experimental and functional annotations and crosslinks to other databases.

3.1.4. IDEAL

Intrinsically Disordered proteins with Extensive Annotations and Literature (IDEAL, available at <http://www.ideal.force.cs.is.nagoya-u.ac.jp/IDEAL/>) is a repository for disordered regions and disordered binding segments with their manual annotations [96, 111]. The version of November 2017 contains 913 entries (proteins).

3.1.5. Pfam

The Pfam database (available at <https://pfam.xfam.org/>) is an extensive collection of protein families, each family represented by multiple sequence alignments and Hidden Markov Models (HMMs) [112]. The general purpose of the Pfam database is to provide a complete and accurate identification and classification of conserved protein regions based on seed alignments. Pfam contains six different types of entries: family, domain, repeat, motif, disordered, and coiled coil. The used version, 31.0 (March 2017) contains 16,712 entries. Currently the latest version of Pfam is 32.0 (as of September 2018).

Pfam consisted of two parts, Pfam-A and Pfam-B. Pfam-A is the manually curated part of Pfam, and contains well-characterized protein domain families with high-quality alignments, which are maintained by using manually checked seed alignments and HMMs to find and align all members. Because the entries in Pfam-A do not cover all known proteins, an automatically generated supplement was provided, called Pfam-B. Pfam-B was no longer supported after Pfam version 28.0, and thus in all analyses I used only Pfam-A.

3.1.6. Functional annotations

For the functional annotations, terms from the Gene Ontology (GO, available at <http://www.geneontology.org/>) were used [113, 114]. The GO database is a bioinformatics initiative to unify the representation of gene and gene product attributes across all species. Each term of the ontology can represent one of three basic properties of a gene product: cellular component, biological process, or molecular function. The sets of terms are designed to be species-neutral, and structured as a directed acyclic graph connected by various relationships (such as ‘is a’ or ‘is a part of’), constituting the ontology.

The CellLoc GO Slim was created manually from the ‘cellular localization’ namespaces of GO, by selecting terms (see Table 1) that are either assigned to studied complexes or are ancestors of such terms. Terms were filtered for redundancy, and if two terms are in child/parent relationships, only one was kept.

Localizations			
ID	Term	ID	Term
GO:0000151	ubiquitin ligase complex	GO:0019867	outer membrane
GO:0000785	chromatin	GO:0019897	extrinsic component of plasma membrane
GO:0000786	nucleosome	GO:0019898	extrinsic component of membrane
GO:0005618	cell wall	GO:0030054	cell junction
GO:0005635	nuclear envelope	GO:0030133	transport vesicle
GO:0005654	nucleoplasm	GO:0030139	endocytic vesicle
GO:0005667	transcription factor complex	GO:0030141	secretory granule
GO:0005694	chromosome	GO:0030312	external encapsulating structure
GO:0005730	nucleolus	GO:0030496	midbody
GO:0005739	mitochondrion	GO:0031012	extracellular matrix
GO:0005764	lysosome	GO:0031090	organelle membrane
GO:0005768	endosome	GO:0031252	cell leading edge
GO:0005773	vacuole	GO:0031254	cell trailing edge
GO:0005776	autophagosome	GO:0031975	envelope
GO:0005777	peroxisome	GO:0032993	protein-DNA complex
GO:0005783	endoplasmic reticulum	GO:0036477	somatodendritic compartment
GO:0005794	Golgi apparatus	GO:0042597	periplasmic space
GO:0005802	trans-Golgi network	GO:0042995	cell projection
GO:0005813	centrosome	GO:0043209	myelin sheath
GO:0005819	spindle	GO:0043228	non-membrane-bounded organelle
GO:0005829	cytosol	GO:0043292	contractile fiber
GO:0005840	ribosome	GO:0044215	other organism
GO:0005856	cytoskeleton	GO:0044297	cell body
GO:0005874	microtubule	GO:0044815	DNA packaging complex
GO:0005884	actin filament	GO:0045121	membrane raft
GO:0005886	plasma membrane	GO:0045177	apical part of cell
GO:0005903	brush border	GO:0045178	basal part of cell
GO:0005912	adherens junction	GO:0045202	synapse
GO:0005938	cell cortex	GO:0048770	pigment granule
GO:0009536	plastid	GO:0060205	cytoplasmic vesicle lumen
GO:0010008	endosome membrane	GO:0070062	extracellular exosome
GO:0012506	vesicle membrane	GO:0097136	Bcl-2 family protein complex
GO:0015629	actin cytoskeleton	GO:0097458	neuron part
GO:0015630	microtubule cytoskeleton	GO:0098687	chromosomal region
GO:0016021	integral component of membrane	GO:0098796	membrane protein complex
GO:0016528	sarcoplasm	GO:0098797	plasma membrane protein complex
GO:0016604	nuclear body	GO:0099080	supramolecular complex
GO:0016605	PML body	GO:0099503	secretory vesicle
GO:0016607	nuclear speck	GO:1903561	extracellular vesicle
GO:0017053	transcriptional repressor complex	GO:1990234	transferase complex
GO:0018995	host	GO:1990391	DNA repair complex
GO:0019012	virion	GO:1990904	ribonucleoprotein complex
GO:0019867	outer membrane	GO:0043235	receptor complex

Table 1: The selected subcellular localization annotations for CellLoc GO Slim.

The PPI GO Slim was also created manually from the ‘biological process’ namespaces of GO used in DIBS and MFIB, in the same fashion as in CellLoc Slim, meaning that terms were selected that are either assigned to studied complexes or are ancestors of such terms (the terms are listed in Table 2 and 3).

PPI GO Slim was assembled to cover a wide range of possible biological functions and was partitioned into two levels:

- 'Generic' part contains high-level cellular/organismal processes (see Table 2).
- 'Specific' part of PPI GO Slim contains terms describing specific biological subprocesses through which generic processes are executed (see Table 3).

Again, terms from both the generic and specific parts were filtered for redundancy with respect to child/parent relationships, similarly to the construction of CellLoc GO Slim.

Generic biological processes			
ID	Term	ID	Term
GO:0001775	cell activation	GO:0009628	response to abiotic stimulus
GO:0006810	transport	GO:0009653	anatomical structure morphogenesis
GO:0006915	apoptotic process	GO:0016032	viral process
GO:0006950	response to stress	GO:0016049	cell growth
GO:0006955	immune response	GO:0016477	cell migration
GO:0007049	cell cycle	GO:0030154	cell differentiation
GO:0007154	cell communication	GO:0032502	developmental process
GO:0007155	cell adhesion	GO:0042221	response to chemical
GO:0007165	signal transduction	GO:0042592	homeostatic process
GO:0008283	cell proliferation	GO:0048511	rhythmic process
GO:0009405	pathogenesis	GO:0051301	cell division
GO:0009605	response to external stimulus		

Table 2: The selected generic biological processes for PPI GO Slim / ‘Generic’.

Specific biological processes			
ID	Term	ID	Term
GO:0000077	DNA damage checkpoint	GO:0008104	protein localization
GO:0001817	regulation of cytokine production	GO:0010467	gene expression
GO:0001932	regulation of protein phosphorylation	GO:0010468	regulation of gene expression
GO:0006260	DNA replication	GO:0015031	protein transport
GO:0006275	regulation of DNA replication	GO:0016192	vesicle-mediated transport
GO:0006281	DNA repair	GO:0016579	protein deubiquitination
GO:0006310	DNA recombination	GO:0019062	virion attachment to host cell
GO:0006325	chromatin organization	GO:0030168	platelet activation
GO:0006334	nucleosome assembly	GO:0030198	extracellular matrix organization
GO:0006355	regulation of transcription, DNA-templated	GO:0032259	methylation
GO:0006412	translation	GO:0033043	regulation of organelle organization
GO:0006457	protein folding	GO:0043067	regulation of programmed cell death
GO:0006461	protein complex assembly	GO:0043687	post-translational protein modification
GO:0006468	protein phosphorylation	GO:0050790	regulation of catalytic activity
GO:0006470	protein dephosphorylation	GO:0050821	protein stabilization
GO:0006508	proteolysis	GO:0051128	regulation of cellular component organization
GO:0006887	exocytosis	GO:0051276	chromosome organization
GO:0006897	endocytosis	GO:0051641	cellular localization
GO:0006914	autophagy	GO:0051726	regulation of cell cycle
GO:0006974	cellular response to DNA damage stimulus	GO:0061024	membrane organization
GO:0006996	organelle organization	GO:0065008	regulation of biological quality
GO:0007010	cytoskeleton organization	GO:1902531	regulation of intracellular signal transduction
GO:0007050	cell cycle arrest	GO:2000145	regulation of cell motility
GO:0007059	chromosome segregation		

Table 3: The selected generic biological processes for PPI GO Slim / ‘Specific’.

3.1.7. Other databases

During the annotation and analysis steps described in the Results and Discussion section, I used other already existing databases as well, such as CATH, ELM, and PhoshoSitePlus, complemented with my own collection of data and annotations, such as specific sets of globular/ordered structures.

- CATH

The CATH Protein Structure Classification database (available at <http://www.cathdb.info/>) is a resource for the evolutionary relationships of protein domain [115]. The domains were classified in a hierarchical manner, and the four primary levels of CATH classification are protein class (C), architecture (A), topology (T) and homologous superfamily (H).

According to CATH, protein domains are identified using multiple automatic methods and manual curation, and treated as an autonomous structural unit. Apart from domain definitions, CATH also includes fragments, as typically small protein regions outside of identified domains.

- ELM

The Eukaryotic Linear Motif database (ELM, available at <http://elm.eu.org/>) focuses on gathering, storing and providing information about short linear motifs in eukaryotic proteomes, which can also occur in viral proteins as well. ELM is integrated with a number of other databases, such as Reactome, PDB and UniProt [116]. After the latest major update, ELM also provides an Application Programmatic Interface (API) to the ELM exploration pipeline, and includes motif instances in bacterial cells as well. Currently, there are 3,069 experimentally validated ELM instances in 196 taxons (as of August 2018).

- PhosphoSitePlus

The PhosphoSitePlus database (available at <http://www.phosphosite.org>) collects manually curated and experimentally identified post-translational modifications, primarily of mammals, especially in human and mouse proteins [117]. The high-throughput (HTP) data is complemented with manually curated low-throughput (LTP) sources (however, only 4.5% of the data comes from LTP measurements). This highly interactive PTM resource contains nearly 445,000 non-redundant PTMs (1,8% has both

LTP and HTP measurements) from more than 20,000 non-redundant proteins, including different phosphorylation, acetylation, ubiquitination, methylation and sumoylation sites. In total, PhosphoSitePlus contains 27,175 PTMs from human sequences (the version of 2 October 2017).

Only LTP PTMs were used during the analysis of regulatory mechanisms of MSF in chapter 4.3.3. PTMs were mapped to complex structures using BLAST between UniProt and PDB sequences. A PTM was considered to be on an exposed surface region, if the Solvent Accessible Surface Area (SASA) of the target residue is at least 70% of the standard SASA for the corresponding residue type. A PTM was considered to affect a residue in contact, if the SASA of the residue calculated from the complex structure increased by at least 30% of the standard SASA of the corresponding residue type upon the deletion of all partner proteins from the structure.

- Ordered/ordered protein complexes

For the comparative structural analyses presented in the Results and Discussion section, I need a dataset containing protein complexes that are formed exclusively by ordered domains. These complexes were taken from the PDB, by selecting structures, where the interaction is formed by only two chains (considering biomatrix transformations, PISA records, and manual assignments as well). These structures did not contain any non-protein or other fragments, and the whole interacting chains can be considered to be single globular domains, according to CATH, without any fragments present.

3.2. Algorithms

3.2.1. BLAST

The Basic Local Alignment Search (BLAST) identifies local regions of similarity between two input sequences [118]. The BLAST algorithms compare nucleotide or protein sequences, and calculate the statistical significance of matches using a scoring matrix. BLAST is typically used to query sequence databases for sequences showing a pre-defined degree of similarity to the input sequence. Similarity is quantified using substitution matrices, and the BLOSUM62 matrix is the default for protein searches, which gives a positive score for each amino acid identity, or a penalty for substitutions/mismatches between two aligned sequences. BLAST can handle partial

sequence identity and gaps as well. During the annotation steps, BLAST was also used to transfer direct annotations from various databases to candidate chains.

3.2.2. DSSP

DSSP (Define Secondary Structure of Proteins) is a standard tool for assigning secondary structures to the residues of a protein, using the atomic-resolution structure as an input (i.e. a PDB-file) [119]. DSSP calculates the most likely secondary structure for each residue based on the input coordinates. This assignment was primarily based on identifying hydrogen bonds between main chain carbonyl and amide groups (as hydrogen atoms themselves are most often not present in PDB structures, this constitutes a prediction method), taking some geometric constraints into account. DSSP assigns to every residue one of eight possible states: ‘H’ for α -helix, ‘B’ for a residue in isolated β -bridge, ‘E’ extended strand, which participates in a β ladder structure, ‘G’ for 3-helix (3^{10} helices), ‘I’ for 5 helix (π -helix), ‘T’ for hydrogen bonded turn, and ‘S’ for bend, as a secondary structure element.

3.2.3. Naccess

Naccess is a stand-alone program that can calculate the solvent accessible surface area of a molecule from its PDB file. It can calculate the atomic-level and residue-level accessibilities for both proteins and nucleic acids [120]. Naccess can determine the available atomic surface using a "rolling ball" algorithm. The centre of a probe (the “ball”) defines the accessible surface when it is rolled around the macromolecule in a way, that its surface stays exactly at the Van der Waals distance from the nearest atom.

3.2.4. FoldX

FoldX can estimate the contribution of a point mutation to the overall stability of a protein structure [121]. The algorithm uses the so-called FoldX force field to calculate the free energy and to predict the effect of a mutation on the stability of proteins and nucleic acids. The effect on protein stability of introducing a PTM was assessed by switching the original residue with a mimetic one in the structure. Phosphorylations were mimicked with Asp; Lys and Arg methylations were mimicked with Leu and Met, respectively; 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.

3.2.4. Other calculations

Principal Component Analysis (PCA), hierarchical clustering and k-means clustering methods were used to assess heterogeneity in the analysis presented in chapter 4.2.5. The k-means clustering was also used to calculate the ideal number of clusters in chapter 4.3.1.

In the case of hierarchical clustering using sequence and structural features, the ‘Ward.2’ R implementation of Ward’s method was used with Euclidean distance in chapter 4.3.1.

Dissimilarity and heterogeneity values in chapters 4.2.5 and 4.3.4 we calculated as follows. Dissimilarity of two proteins i and j is defined as: $d_{ij} = \frac{L_{ij}}{L_{max}}$, where L_{ij} is the linkage distance given by the clustering (of all proteins from all three classes) between the two proteins from the same interaction class, and L_{max} is the maximal linkage distance. Heterogeneity values are defined as the geometrical averages of dissimilarity values between all protein pairs from a given class.

In the case of functional heterogeneity, the hierarchical cluster tree was replaced by the GO ontology tree. Distances between terms that are in a parent/child relationship was 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 is minimal. I.e. 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 , i.e. 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.

For sequential features in chapters 4.2.1 and 4.3.1, seven amino acid groups were used in quantifying sequence composition of proteins: 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). In chapter 4.2.1, compositions were calculated for a single protein chain, whereas in chapter 4.3.1, compositions were calculated for the entire complex. For complex-based calculations an 8th sequence parameter was used to quantify the compositional difference between subunits, and was defined as:

$\Delta_{total} = \sum_{i=1}^7 \Delta_i$, where Δ_i is the largest composition difference of residue group i between constituent chains.

During the analyses, interaction energies for residues were calculated using the statistical potentials described in [71]. In the classification of complexes of MSF, in chapter 4.3.2, these were calculated for the entire complex.

Dissociation constant (K_d) values for MSF complexes in chapter 4.3.2 were calculated using the PRODIGY binding affinity prediction tool, using standard parameters [122].

3.3. Development of web servers

- Backends of MFIB and DIBS

The servers behind the MFIB and DIBS databases are implemented in PHP (version 5.0) using an Apache HTTP server (version 2.4) on Ubuntu 14.04. The information of each entry is stored in a MySQL database (MariaDB 5.7), and in XML-files (version 1.0).

- Frontends of MFIB and DIBS

The interfaces were built with HTML5 and CSS3. For the structure viewer, an open source JSmol library (<http://jmol.sourceforge.net/>) was used for DIBS and LiteMol (<https://webchemdev.ncbr.muni.cz/LiteMol/>) for MFIB. For interactivity in the TreeMap section, a Bootstrap 3.0 library (<https://getbootstrap.com/>) was used, and to implement dynamic charts, an open-source JavaScript library, named JsChart (<https://www.chartjs.org/>) was used. For the sequence viewers in each entry pages, an open-source javascript-based library, named Feature Viewer (<https://github.com/caliphosib/feature-viewer>) was used. This is part of the NetProX project, and was modified by me for the specific requirements of the developed servers.

3.4. Other programs and programming environments

All other tasks were programmed in the Perl5 or Python 3.6 languages (except for the literature data mining part, which was done using the Ruby 2.2 language), in Windows 10 and Ubuntu 14.04/16.04 environments, using ATOM/Geany/Sublime and native text editors.

PCA and clustering calculations were performed using the R statistical computing environment (version 3.3.1).

Plots were generated mostly using Microsoft PowerPoint 2016 and Excel 2016, and the matplotlib library (version 3.0.0) of Python.

The protein figures (mostly with the ribbon and cartoon representation) were generated from the corresponding original or modified PDB-files, using the UCSF Chimera 11.2 visualization tool.

For a visual representation of GO annotation enrichments, a word cloud technique (based on Wordle, <http://www.wordle.net/>) was used in Figures 15 and 16. Font size for a given localization or biological generic and specific process represents the relative frequency of occurrence of the given GO term among the studied interaction class. Colour depth represents the specificity of the term for the given interaction class. The primary chosen colors (blue - ordered/ordered complexes, green - disordered/ordered complexes, red - disordered/disordered) only support the visual representations, do not have a direct meaning.

4. Results and Discussion

Protein complexes formed by ordered proteins are relatively well studied; however, the growing number of known disordered proteins and their crucial biological functions typically arising from their interactions require detailed analyses to understand. To perform such analyses, we need large-scale datasets. The lack of well-organized and accessible data for IDP-mediated interactions in structural detail was the primary motivation behind the establishment of two databases: the DIBS database (DIordered Binding Sites), for IDPs undergoing coupled folding and binding, and MFIB (Mutual Folding Induced by Binding), for complexes formed exclusively by IDPs. Targeted databases often prove to be not only beneficial but vital for the development of research areas in biology [123].

4.1. Assembly of the MFIB and DIBS databases

4.1.1. Collecting potential data for MFIB and DIBS

In the assembly of both DIBS and MFIB, structures from the PDB were taken as a starting point: the solved complex structure was a prerequisite for the inclusion of an interaction in the databases. The solved structure is proof that the proteins involved adopt a stable structure upon interacting, and is also a verification of the interaction. Accordingly, all 127,801 PDB structures, serving as candidates, were inspected during the analysis.

The main step in collecting IDP-mediated interactions was to identify IDPs in structures. As this is a time-consuming step involving manual annotation, first all structures from the PDB were scanned for potential protein-protein interactions. I discarded structures due to various reasons, as they are not part of my primary focus of interest. During the first filtering steps, structures containing any non-proteins, typically RNA or DNA, were discarded because I was interested in only protein-protein interactions, which are markedly different from protein-DNA or protein-RNA interactions.

To further filter candidate structures, I used various keywords to find chimeras, structures featuring other non-biological peptides containing a large number of non-standard residues, and artificial/synthetic proteins. These entries were also dropped because there is only a slight chance they exist in nature, and I opted to focus on biologically relevant native interactions. I also filtered structures based on the quality of

determination, keeping only NMR, and X-ray determined structures with a resolution better than 5Å. I discarded complexes solved by electron microscopy, because only the backbone conformation is recognizable with its resolution.

In order to identify all protein-protein interactions in the remaining candidate structures, I needed to generate possible missing chains in the PDB-files, where these are represented using biomatrices. After the generation of these protein chains, I considered only those structures that have at least two protein chains in interaction, and ignored monomer structures. I defined two chains as being in interaction if they have at least five interchain heavy-atom (non-hydrogen) contacts between them, meaning that the Euclidean distance between two heavy atoms of amino acids from different chains is less than the sum of their Van der Waals radii plus 1Å. I used the standard Van der Waals radii of the elements: 1.70Å for carbon, 1.55Å for nitrogen, 1.52Å for oxygen and 1.80Å for sulfur.

These filtering steps reduced the number of candidate structures by over 85%. In order to aid structural annotations, protein chains of the remaining candidates were mapped to UniProt sequences using BLAST, or if applicable, the DBREF record was checked for the correct UniProt identifiers. This enables the biologically relevant sequences to be directly compared for further annotation.

4.1.2. Annotating protein complexes for MFIB and DIBS

Figure 6 shows a simplified workflow for the filtering and annotation steps during the construction of the databases. After the filtering step of PDB for possible complexes, the protein chains constituting the candidate set of interactions were mapped to UniProt and UniRef90 sequences to enable their annotation with disorder or order information. Sequences from all candidate structures were grouped into sequence clusters if they show a high enough similarity, meaning that the two sequences belong to the same UniRef90 cluster, and they overlap with each other in at least 70% of their respective lengths.

As a next step, proofs for the structural state were sought for members of each sequence cluster. "Disorder proof" means there is experimental proof of disorder in DisProt/IDEAL for at least one of the proteins in the cluster, or there is a known linear motif in the interacting region (from ELM, UniProt, or Pfam). In addition, disorder proof were collected from the missing coordinates in X-ray determined structures, NMR structures showing highly flexible structures, or from manual literature searches.

As linear motifs are described to mediate interactions with ordered domains, these annotations from ELM, UniProt, Pfam or the literature were only accepted as disorder proof in the case of DIBS. Linear motif-based disorder proof is labeled as ‘Inferred from motif’, to reflect the less specific assignation of the disordered status.

„Order proof” means there is a monomeric solved structure in the PDB for at least one of the proteins in the cluster according to CATH or manual curation. Order annotations mostly came from the PDB, as it is rescanned for stable monomer structures, taking biomatrices into account as well. These chains were checked in the CATH database to only use structures describing a compact, single domain without fragments. The resulting monomer set contains 16,381 protein structures.

Using the above established disorder and order annotations, all sequence clusters can be annotated four different ways:

- disordered, when at least one of all sequences in that cluster has disorder proof, and the other proteins of the cluster lack structural annotations
- ordered, when at least one of the sequences has order evidence
- unknown, when there are no available disorder or order annotations for any of the sequences
- not clear ("conflict"), when there is conflicting information, e.g. one protein is annotated to be disordered, and another has a monomeric solved structure. These annotations had to be checked manually to resolve the conflict, if possible.

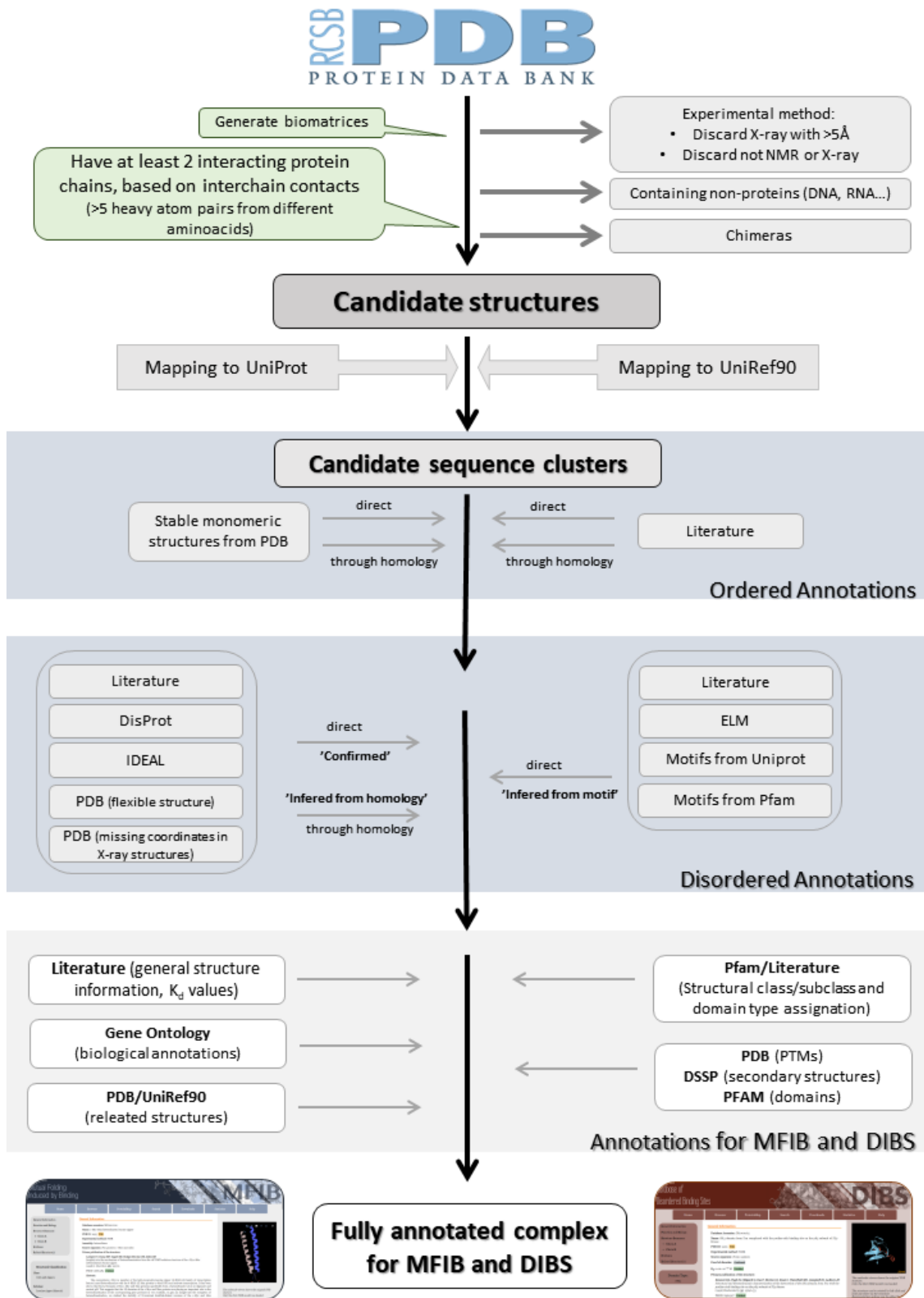


Figure 6: Workflow of the construction of the databases.
The filtering and annotation steps are detailed in chapters 4.1.1. and 4.1.2.

In order to be able to annotate the largest possible number of protein sequences as disordered, the annotations were transferred in two different ways. All information from source annotations (databases or the literature) were mapped to UniProt sequences to be directly comparable, and this information represents the direct evidence for being disordered. In the case of DIBS, direct evidence can either be “confirmed” or “inferred from motif” (as described above).

Apart from using direct evidence, disorder annotations were expanded to sequences that belong to the same UniRef90 cluster. The rationale behind this annotation transfer (yielding indirect annotations) is that it has been shown that in case of ordered proteins, a 30% sequence identity means probable homology, and in the case of sufficiently long alignments, the adoption of the same fold [124]. While there is no similar study conducted concerning protein disorder, it is safe to assume that if 30% identity is generally sufficient for two ordered proteins to share the same fold, the significantly higher level of sequence identity guaranteed by belonging to the same UniRef90 cluster, or bearing the same Pfam object, should be sufficient for belonging to the same structural class (ordered or disordered). These annotations are labeled as ‘inferred from homology’, as the candidate chains and all annotations were mapped to UniRef90 sequences, therefore, the found evidence is most likely transferred through homology between sequences.

Considering all types of available annotations, the constituent chains from candidate protein complexes were classified as ‘ordered’ or ‘disordered’ if either type of annotations covered at least 70% of their sequence, and ‘unknown’ otherwise. This yielded fully and partially annotated complexes, where all or some of the chains were bearing some kind of structural annotations. Complexes where exactly one chain was annotated as disordered, with the rest being annotated as ordered, were included in DIBS. Complexes where all chains were annotated as disordered, were included in MFIB. Promising partially annotated complexes were attempted to be fully annotated based on literature, to maximize the number of complexes in each database.

As a next step, in order to reduce the number of similar protein complexes, annotated interactions in MFIB and DIBS were clustered to entries in order to attenuate redundancy based on UniRef90 clusters. The main reason behind the redundancy is that there are several interactions that have been extensively studied due to their important biological functions, leading to more determined structures of the same or nearly same interactions (e.g., there are nearly 500 structures that contain p53). This type of redundancy becomes crucial to handle at the step of creating biologically relevant

database entries by clustering highly similar complexes. To cluster interactions, two complexes are deemed related (or highly similar) if they contain the same number of proteins, and the proteins from the two structures show a sufficient degree of pairwise similarity, i.e. they belong to the same UniRef90 cluster (the full proteins exhibit at least 90% sequence identity) and convey roughly the same region to their respective interactions (the two regions from the two proteins share a minimum of 70% overlap). Related complexes were grouped together into clusters forming the entries in MFIB and in DIBS. This clustering step reduced the number of MFIB entries from 1,405 to 205 for MFIB, and from 1,577 to 773 for DIBS.

From each of the resulting interaction clusters, only one structure was selected as a representative of the interaction, based on structure determination methods, structure quality, and source organisms (NMR structures were preferred over X-ray, better resolution structures and human proteins were favored over others). Each entry in MFIB is assigned a class and a subclass during the manual annotation and curation step. Table 4 shows the currently manually defined 88 classes and 33 subclasses.

Bulb-type lectin domain	Coils and zippers	
<ul style="list-style-type: none"> • Homodimeric lectin • Heterodimeric lectin 	<ul style="list-style-type: none"> • Coiled coil (dimeric) • Coiled coil (dimeric, forming a 4-helix bundle) • Coiled coil (hexameric) • Coiled coil (pentameric) • Coiled coil (tetrameric) • Coiled coil (tetrameric, 4-helix bundle) • Coiled coil (trimeric) • Alanine zipper (trimeric) • Leucine zipper (dimeric) • Leucine zipper (tetrameric) • Phenylalanine zipper (dimeric, forming a 4-helix bundle) 	
Histone-like interactions		
<ul style="list-style-type: none"> • Histones • Histone-like complexes 		
Homooligomeric enzymes		
<ul style="list-style-type: none"> • Homodimeric enzymes • Homotetrameric enzymes • Homohexameric enzymes 		
L27 domains		
<ul style="list-style-type: none"> • L27_1 type • L27_2/N type 		
NGF-like proteins		Other
<ul style="list-style-type: none"> • Homodimeric NGF-like proteins • Heterodimeric NGF-like proteins 		<ul style="list-style-type: none"> • Basic helix-loop-helix (bHLH) • E2 dimer • p53 tetramerization • Phd antitoxin • Ribbon-helix-helix (RHH) • Trp repressor-like • Other
Transthyretin-like folds		
<ul style="list-style-type: none"> • Transthyretin • HIUase 		

Table 4: Classes (in grey boxes) and subclasses (as bullet points) currently defined in MFIB.

In contrast to MFIB, DIBS uses domain type assignments according to Pfam and the literature taking only the ordered partner(s) into account. Currently, there are 185 defined domain types in DIBS, such as the 14-3-3 domain (15 entries) or bromodomain (29 entries). The database also collects information about the measured binding strengths of the interactions (K_d values) from the literature, where possible.

4.1.3. Web interface for the “twin-databases”

DIBS and MFIB are freely accessible through their dedicated websites: MFIB is made available at <http://mfib.enzim.ttk.mta.hu/> and DIBS is made available at <http://dibs.enzim.ttk.mta.hu/>.

The websites consist of two components: (i) the servers are hosted in Apache web servers and are implemented in PHP5 with a MySQL backend, which processes the requests, fetches data from the database, provides search functionality and serves web pages and (ii) a front-end which includes various interactive visualizations based on charts.js, jquery.js, bootstrap.min.js and d3.js libraries (see Figure 7).

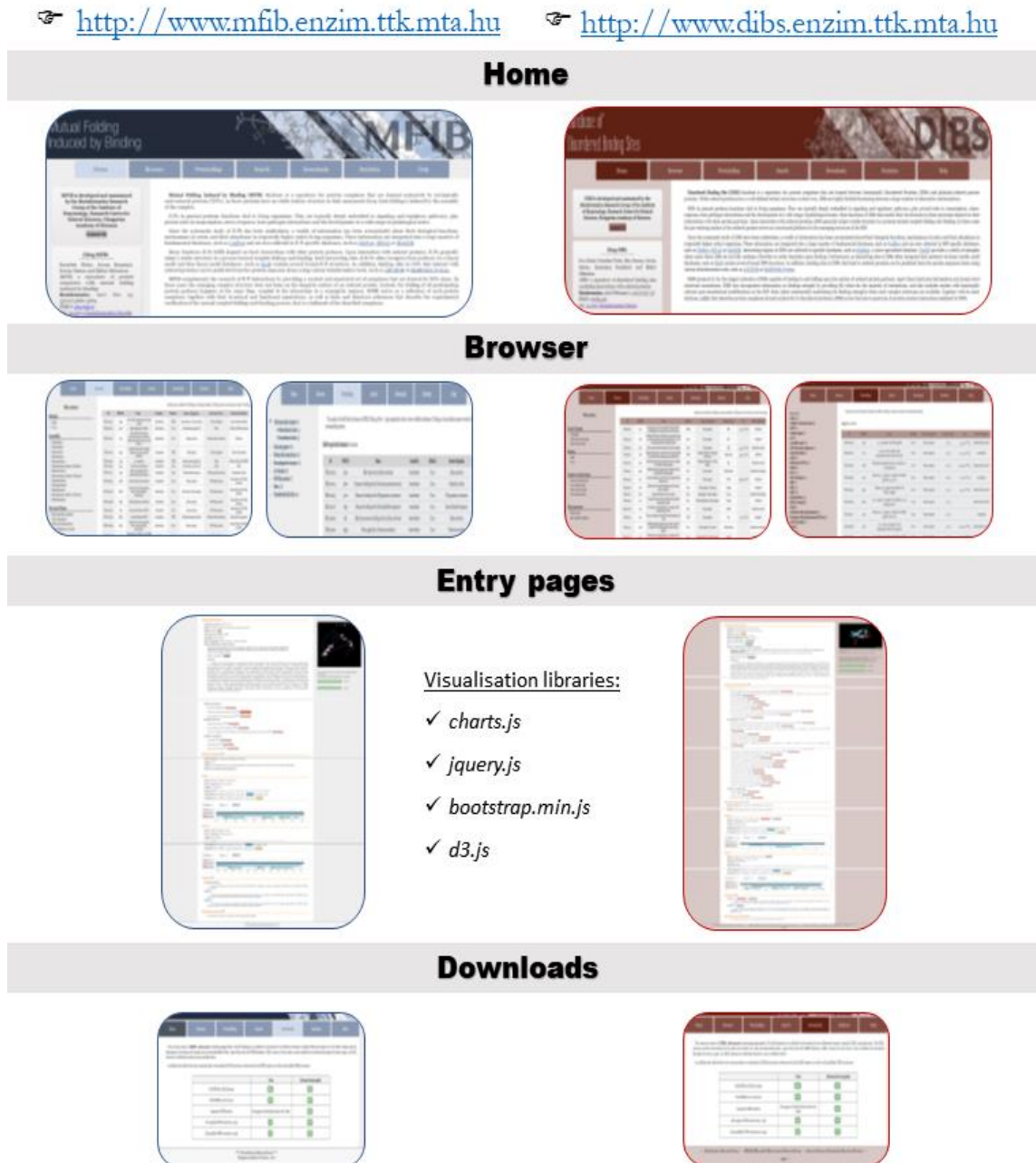


Figure 7: Web interfaces of MFIB and DIBS.

Beyond the primary purpose of disseminating the raw data, an additional purpose of the websites is helping the user to navigate easier. Each MFIB/DIBS entry is assigned a unique accession, which is composed of the letters 'MF'/'DI' at the beginning according to their parent databases, followed by seven digits. In MFIB, the first two digits mark the oligomeric state of the complex: (i) the first digit being equal to the total number of interacting protein chains in the complex, and (ii) the second digit indicating the number of unique proteins in the complex. For example, accessions for all homodimer complexes start with MF21, accessions for heterodimers start with MF22. For DIBS, only the first

digit marks the oligomeric state of the ordered chains in the complex. For dimers, where the interaction is between a single IDP and a single ordered protein, the accession starts with DI1.

In MFIB, the third and fourth, and in DIBS, the second and third digits contain information about the taxonomic group(s) from which the interacting chains originate. The first of these two digits shows the highest taxonomic group of all chains with '0' corresponding to human, '1' corresponding to all other eukaryotes, '2' meaning bacteria, '3' meaning archaea and '4' denoting viral proteins. The fourth digit shows the heterogeneity of the origin species of the interacting chains. It is '0' if all interacting proteins are from the same species, '1' if they cover more than one species but all are from the same taxonomic domain, and '2' if the proteins in the structure cover more than one taxonomic domain. The last three digits in MFIB and the last four digits in DIBS form a randomly assigned number that guarantees the uniqueness of the accession for each entry.

Each entry has a separate page that details information about a specific interaction. These pages are organized into various sections for easier navigation. The first section is the "General Information" which itemizes the primary information about the complex. In the beginning, it lists the assigned accession of the entry, the name (which is not necessarily the same as the name of the structure in the PDB database), displays references to the corresponding PDB structure, and other details about the solved structure. It includes the type of the structure determination method used and the primary publication of the structure, if applicable (authors, title, journals, abstract and PubMed IDs with links to the PubMed website), using the PDBe REST API.

The general information section also defines the biological oligomeric state of the complex; however, it is not always the same as the assembly state in the raw PDB structure. That is because in some cases the original PDB structure does not or does not only show the biologically relevant interactions. To remedy this, in these cases a modified PDB file is generated and displayed in the embedded structure viewer, which loads the structure of the complex and visualizes it in cartoon representation. If a modified PDB file is generated, a description of the transformations can be found below the viewer window in the right column. These modifications can be the generation of new protein chains based on the biomatrices in the PDB file, or the omission of a protein chain or chains to reduce possible duplicity present in the structure, or truncations of protein chains to highlight the relevant interacting regions. Links to download the original, and if

applicable, the modified PDB file are provided under the structure viewer. Also, it includes a link to the entry's XML file, which contains all the presented information.

The “Function and Biology” section features the known biological annotations of the protein complex. These annotations were taken from the Gene Ontology (GO) using the Gene Ontology Annotation Database via the API service provided by EMBL-EBI. Only terms that fit at least two of the interacting chains are shown (in the case of DIBS, one of these has to be the disordered protein), as individual proteins may bear biological functions related to regions not participating in the interaction. This section categorizes the annotations as molecular functions, biological process, and cellular component, as can be found in the GO.

The “Structure Summary” lists the total number of interacting protein chains in the complex, as well as the number of different chains. Each chain is assigned an identifier in the form of a capital letter, and they are taken from the PDB. Chains that were generated using biomatrices are assigned the first letter that was not used in the original PDB file. The description of the transformation used to generate these chains is also specified. Annotations of each chain in the complex are detailed in its own subsection. These subsections describe the sequence of the protein region in the PDB file, the corresponding protein regions from UniProt and UniRef90, and display the segments inside this protein region that have atomic coordinates in the PDB file. Moreover, basic secondary structural elements (helices and beta structures) are also shown, together with Pfam objects (if applicable). This information is displayed in an embedded scalable NetProX sequence viewer. The Feature Viewer was modified to meet my special requirements (the use of different colors and the addition of extra functions).

While all sections described so far hold the same type of information in both DIBS and MFIB, the “Evidence” section has slightly different content in the case of the two servers. In MFIB, it displays experimental evidence demonstrating that all of the participating proteins are disordered prior to the interaction. This section either shows evidence for the intrinsically unstructured nature of all participating protein chains separately (with cross-links to other disorder databases and literature), or shows evidence for the structured complex itself to arise directly from the interaction of disordered monomers (“Complex Evidence”). In some rare cases, both types of evidence are available for a complex. In DIBS, there is also experimental support proving the disordered nature of one of the interacting chains in its unbound form. Other quoted evidence proves that all other participating chains are ordered in their monomeric form,

or form an ordered oligomer together, existing in a folded structure without the disordered chain.

The “Related Structures” section displays links to other structures in the PDB that belong to the same redundancy cluster.

Together with the individual entry pages, the webpages have several auxiliary pages that aid easier navigation. The “Home” page describes the basis and purpose of the database for users unfamiliar with MFIB or DIBS. The “Statistics” page shows basic statistics about the databases, such as the number of entries belonging to various oligomeric states (assemblies) or the distribution of entries in various taxonomic groups, using interactive charts to illustrate them (using the JsChart library). The “Help” page functions as a FAQ for the databases, answers frequently asked questions connected to the design and usability of the database and the server.

MFIB and DIBS offer three ways of structured access to the data they contain. In the ‘Browser’ section, all entries in the database are listed. The list is sortable by all displayed information (complex name, source organism, etc.) and can be filtered by various options (oligomeric state, type of experimental methods etc.). The “Search” page offers a simple search engine (implemented in MySQL) able to return matching hits to various queries and subqueries in names, various IDs (database IDs, PDB, UniProt or UniRef90), as well as other information describing the entries (the type of assembly, etc.). The ‘ProteinMap’ can display entries belonging to selected classes and subclasses in case of MFIB, and the defined domain types in the case of DIBS.

Apart from online access, MFIB and DIBS offer multiple ways of downloading data in the database. The ‘Downloads’ section includes links to download the full databases in XML or text format, and all original and all modified PDB-structures in a zip archive. Furthermore, each entry page includes links to download the entry’s PDB structure(s) and XML format file.

4.1.4. Statistics of MFIB entries

The establishment of MFIB achieves the first systematic collection of data concerning MSF complexes, and provides comprehensive coverage of possible IDP-only interactions. The current version of MFIB contains 1,406 structures grouped into 205 entries, representing the core of MFIB. Typically, 2-6 protein chains form the

interactions, and thus MFIB entries cover the majority of biologically relevant oligomeric compositions (from dimers to hexamers). The database contains 147 dimers, 10 trimer entries, 40 tetramers, 3 pentameric interactions, and 5 hexamer complexes (see Figure 8).

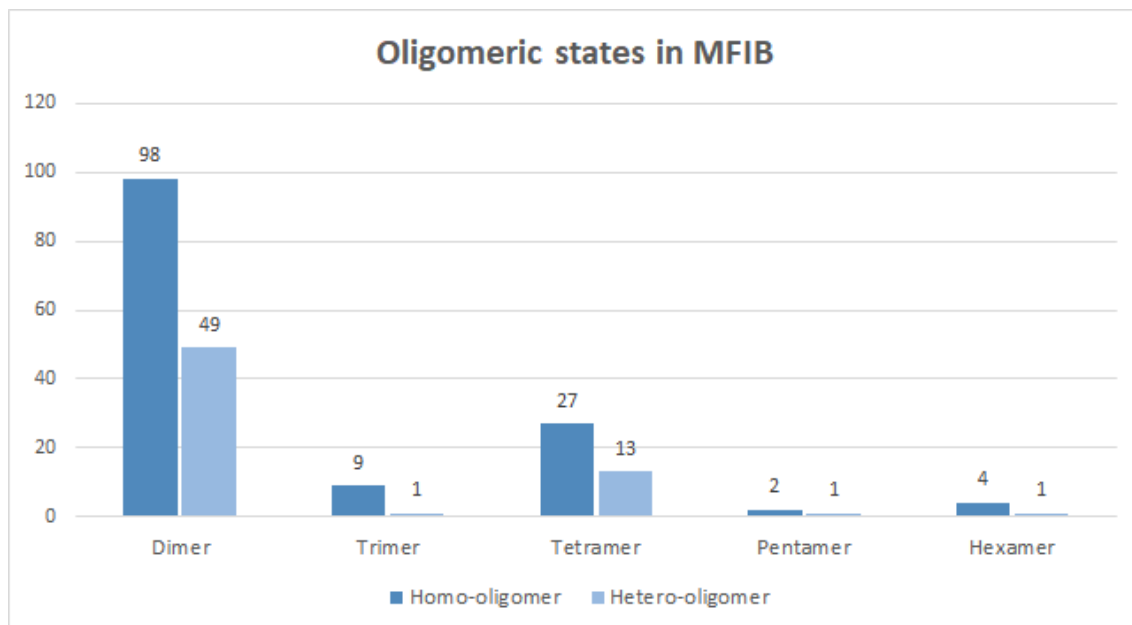


Figure 8: Statistics of oligomeric states in MFIB.

Entries in MFIB come from a wide range of taxonomic groups, covering all three domains of life, with most of the complexes being formed by human proteins. However, MFIB also includes complexes from viral proteins, shedding light on the importance of mutual synergistic folding in host-pathogen interactions. Interactions formed by proteins from different taxonomic domains are classified as “cross-domain” interactions (see Figure 9, top).

The number of related structures in entries can vary widely, with some interactions being unique (98 complexes have no solved related structures) and others having a large number of available similar structures (with the maximum number of related structures being 273 for human transthyretin). The average number of related structures is 14 (see Figure 9 bottom). Each structure detailing the interactions were determined by either X-ray (153 out of 205, accounting for 74.6%) or NMR (52 out of 205, accounting for 25.4%).

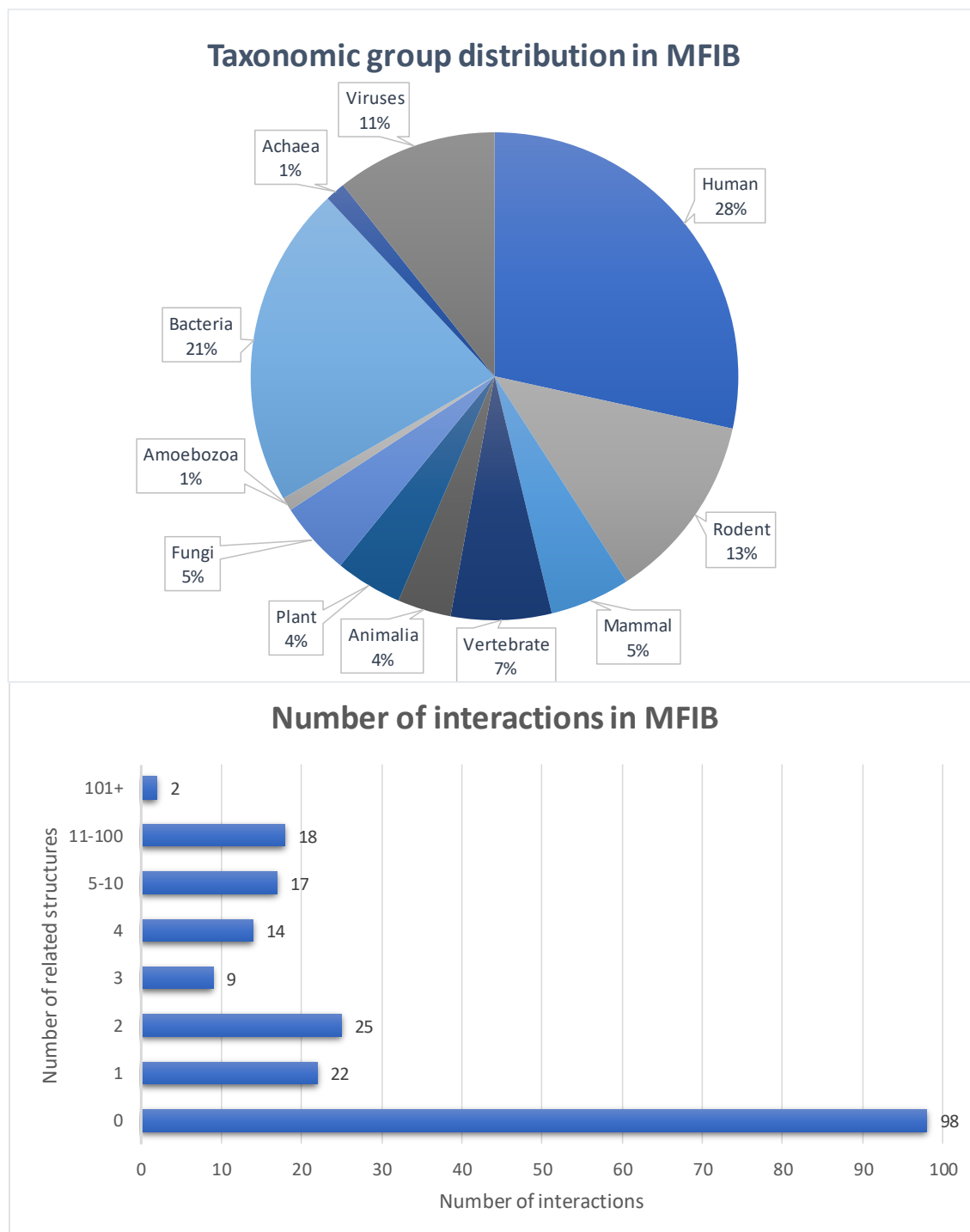


Figure 9: Distribution of source organisms (top) and number of related structures (bottom) in MFIB.

Taxonomic groups are mutually exclusive, i.e. for example “Mammal” represents all mammalian organisms except for rodents and human.

Complexes in MFIB also cover the known spectrum of protein disorder, illustrating that disorder is more like a continuum than a binary property. The database contains complexes of IDP regions from near random coil proteins (CBP-ACTR, PDB: 1kbh, MFIB: MF2201001) [16], through molten globules (Arc repressor) [17] to near-ordered

structures, where a monomeric IDP protein can be stabilized with a limited number of mutations (nucleoside diphosphate kinase, PDB: 1nkp, MFIB: MF6110001) [125].

4.1.5. Distribution of data in DIBS

The current version of DIBS contains 1,577 complex structures clustered into 773 entries. IDP-mediated interactions are the most abundant in eukaryotes, and in accord, entries in DIBS cover a wide range of eukaryotic taxonomic groups. In addition, DIBS also includes a fair number of bacterial and cross-domain interactions, where the interacting protein chains come from organisms of different taxonomic domains (see Figure 10, top).

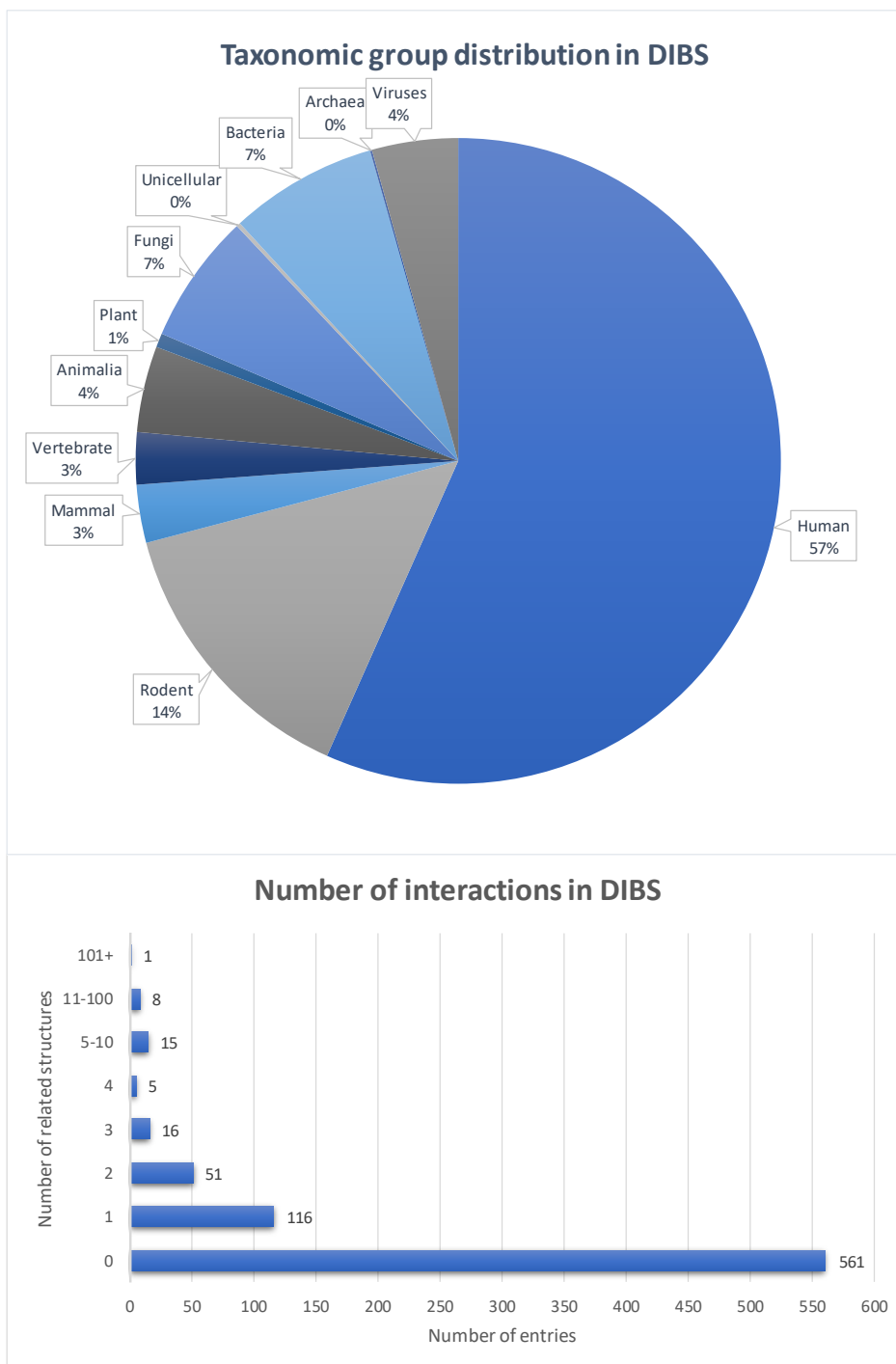


Figure 10: Distribution of source organism (top) and number of related structures (bottom) in DIBS.

Most of the interactions in DIBS are unique regarding the number of related structures (561 complexes have no related structures), while others have a large number of available similar structures (the maximum number of related structures being 162 for "Human estrogen receptor α ligand-binding domain in complex with NCOA2 peptide", PDB: 1gwq, DIBS: DI2000015). The average number of related structures is 5 (see Figure 10, bottom).

The majority of DIBS complexes feature known K_d values for the interactions (488 out of 773, accounting for 63.1%). Figure 11 shows the distribution of K_d values, covering a wide range between approximately 10^{-3} M and 10^{-11} M. Figure 11 also presents three example interactions with various K_d values. The first complex is formed by anophelin - a blood-clotting inhibitor from mosquito - and α -thrombin (PDB: 4e05, DIBS: DI2010010) with a relatively low K_d value, indicating a remarkably tight interaction. The other two examples both involve the disordered tail of integrin β 2, bound to 14-3-3 ζ (PDB: 2v7d, DIBS: DI2010013) and bound to filamin A (PDB: 2jf1, DIBS: DI1000145). Both interactions are transient, in accord with their signaling roles, yet there is still three orders of magnitude difference between the two K_d values. However, there is no direct competition between the two interactions, as they are coordinated via a PTMs. Integrin β 2 bound to 14-3-3 ζ requires a phosphorylation at T758, while the other interaction requires an unmodified integrin tail.

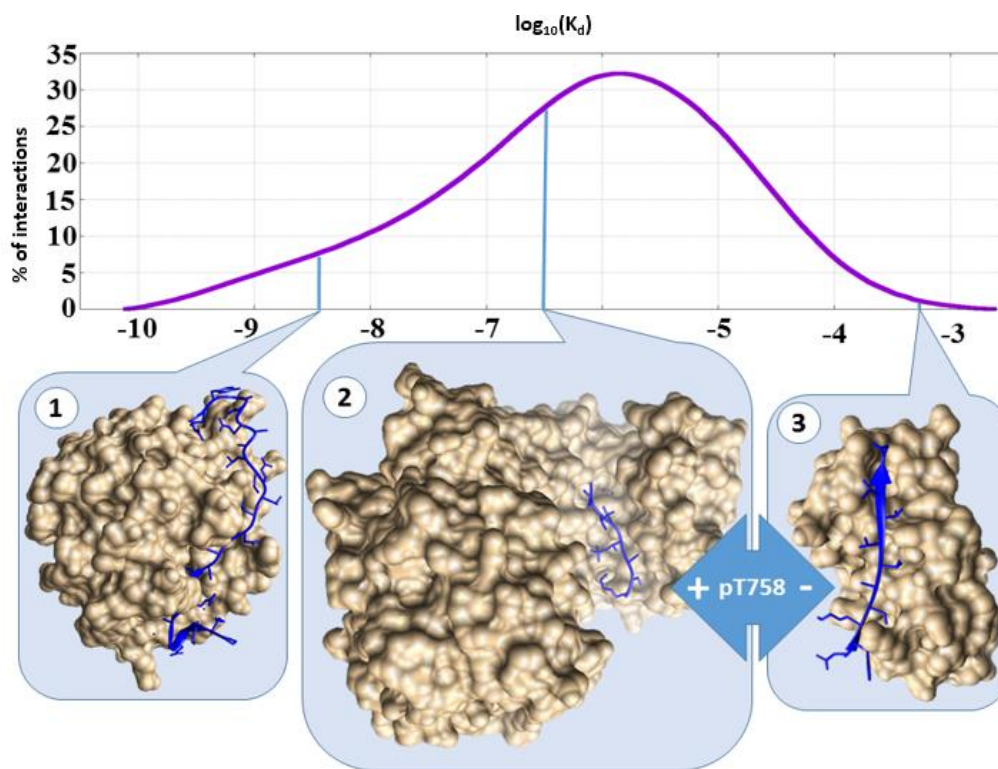


Figure 11: Distribution of K_d values in DIBS, together with three selected examples with differing biological functions [126].

1 - complex between anophelin and α thrombin, 2 - complex between integrin β 2 and 14-3-3 ζ , 3 - complex between integrin β 2 and filamin A.

DIBS shares only limited overlap with other existing disorder databases (see Figure 12). ELM contains known linear motifs, DisBIND is compiled from interacting regions of IDPs with automated annotations from the DisProt database, complemented with annotations from the PDB and UniProt. IDEAL and DisProt contain disordered protein sequences, assembled from manual curation. The highest overlap is only nearly 50% (ELM), though, these datasets focus on different aspect of IDPs.

Fraction of DIBS proteins also included in other disorder-specific databases	
DisProt	0.226
IDEAL	0.420
DisBIND	0.052
ELM	0.492

Figure 12: Overlap with related disorder-specific databases.

4.2. Analysis of sequence, structure and function relationships in different protein interaction classes

Using information from MFIB and DIBS, for the first time, we can get a full view of the entire spectrum of the IDP interactome, and we can uncover the differences between various types of interactions mediated by IDPs - concerning their sequences, structures, functions and regulation. Three protein interaction groups were considered in the analysis (see the first three columns in Figure 13).

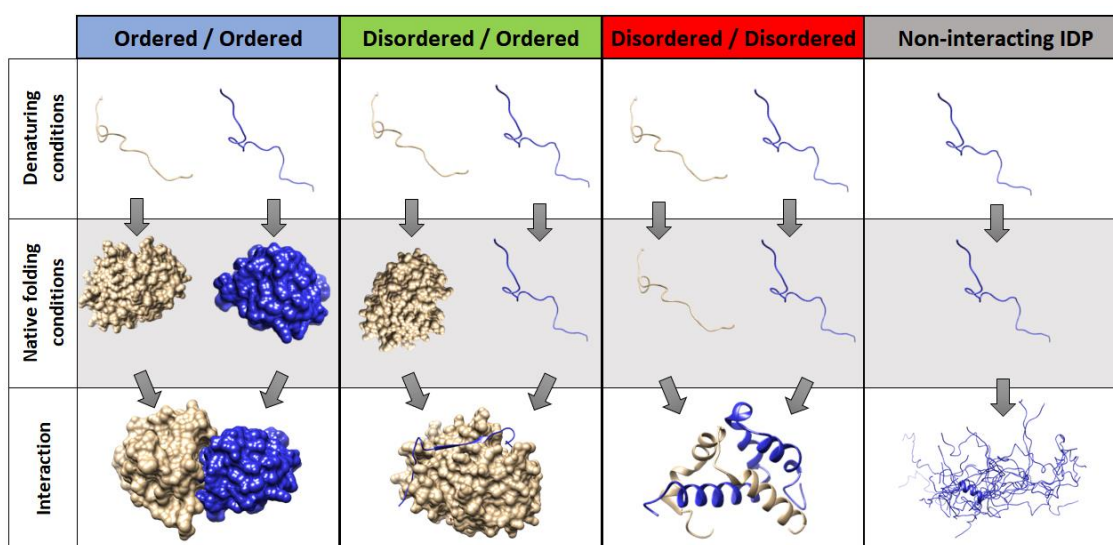


Figure 13: Protein interaction classes based on a various way a protein can reach a structured state.

Under physiological conditions, disordered proteins are shown in cartoon and ordered proteins are shown in surface representation.

The interaction categories are based on how constituent proteins reach their structured states with regard to the binding process:

- 691 structures (henceforth ordered/ordered complexes), where all proteins involved are ordered, going through autonomous folding prior to the binding event without the presence of the partner protein (see Data and Methods).
- 773 complexes from the DIBS dataset (henceforth disordered/ordered complexes), where coupled folding and binding happens, with one IDP bound to ordered partner proteins.
- 205 complexes from the MFIB dataset (henceforth disordered/disordered complexes), where the complexes are formed exclusively by IDPs through mutual synergistic folding.

4.2.1. Amino acid composition mirrors the connection between folding and binding

For the sequence analysis, a fourth category was taken into account, IDPs that presumably do not participate in interactions at all (see the fourth column in Figure 13). This category does not overlap with the other groups, contains 1,045 sequence regions, and was taken from DisProt after discarding sequence regions present in MFIB and DIBS.

For the classification of amino acids, the following categories were considered: aromatic (F, W, Y), charged (H, K, R, D, E), covalently interacting (C), flexible (G), hydrophobic (A, I, L, M, V), polar (N, Q, S, T) and rigid (P). Amino acid compositions were calculated for single constituent protein chains in interaction, using this reduced amino acid alphabet. For reference, the human proteome was used with 71,576 sequences (version 11 August 2017). The various calculated sequence properties are shown in Figure 14:

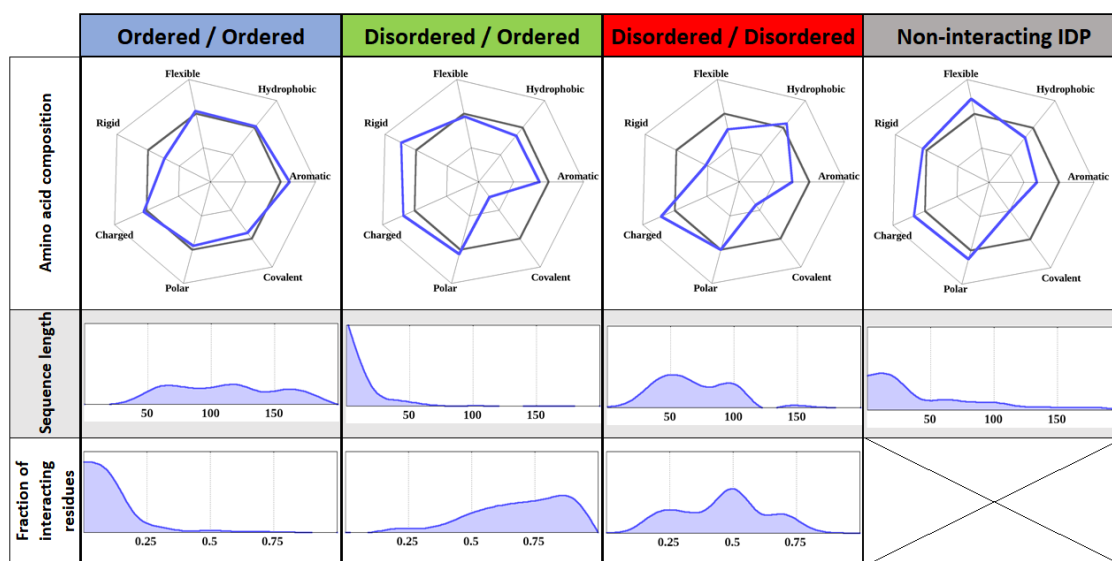


Figure 14: Calculated sequence properties for the four interactions classes.

Amino acid compositions are presented as fold changes compared to the human proteome composition (bold line). Each line represents a fold change.

Non-interacting IDPs reflect the general view of disordered proteins, they lack hydrophobic residues, have a high net charge and are enriched in prolines [14]. On average, interacting ordered/ordered proteins closely resemble the amino acid composition of the human reference proteome. The most notable difference is that they have low proline content, because most of the ordered secondary structure elements are incompatible with proline, as it is a structure breaking residue [127].

In contrast, IDRs in disordered/ordered complexes are usually highly charged and are also depleted in hydrophobic residues. They also often contain prolines, possibly to decrease the loss of entropy upon binding to increase binding strength. IDRs in disordered/disordered complexes are typically more hydrophobic, presenting an exception to the general view of IDPs [128], they contain very few prolines (even less than ordered/ordered proteins) and glycines, and are also highly charged.

Besides the amino acid compositions, other parameters also display differences between the various protein classes. Ordered/ordered complexes contain more residues than the other three categories, because remarkably short sequences can not fold into a globular domain. Disordered sequences undergoing coupled folding and binding and non-interacting IDPs tend to be shorter on average than sequences from disordered/disordered complexes. The fraction of directly interacting residues can be a new angle of distinctive features. IDPs in general use a more significant fraction of their residues for binding than ordered proteins, and this tendency is the most pronounced for IDPs in disordered/ordered complexes.

The exhibited fundamental sequential differences uncover the fact that amino acid compositions of interacting IDPs often do not correspond to the general IDP-view, instead, their sequences reflect the structural state of their binding partners. IDPs in MSF complexes lack prolines and they contain more hydrophobic residues than an average IDP. Ordered/ordered proteins on average contain a relatively large number of residues, as a short sequence is not able to fold into domains. Their large number of hydrophobic residues provide the hydrophobic core to stabilize the tertiary structure; therefore proteins forming ordered/ordered complexes use a low fraction of their residues in the interaction. In the case of disordered/ordered complexes, the hydrophobic core is already provided the ordered part of the complex, and they can donate most of their residues into the binding, while for disordered/disordered complexes the unusually high hydrophobic content serves to create the stabilizing core during the interaction.

4.2.2. The presence of IDPs affects the structural properties of the resulting complexes

The structural properties of interacting ordered proteins and IDPs were analyzed, with the focus on secondary structure elements, molecular surface areas, intramolecular and intermolecular atomic contacts, and predicted interactions energies (see Table 5). The structural features were calculated for one interacting protein chain in the bound form.

DSSP was used to assign secondary structure states to the residues in the structure, specifying residues as helical ('G', 'H', 'I'), extended ('B', 'E') and irregular elements ('S', 'T' or unassigned).

			Ordered / Ordered	Disordered / Ordered	Disordered / Disordered
Secondary structures	Helical		0.333	0.175	0.596
	Extended		0.249	0.103	0.110
	Irregular		0.418	0.722	0.294
Molecular surfaces	Accessible surface	H	0.567	0.565	0.558
		P	0.433	0.435	0.442
	Interface	H	0.623	0.671	0.750
		P	0.377	0.329	0.250
	Buried surface	H	0.589	0.378	0.489
		P	0.411	0.622	0.511
	Interface/total surface		0.105	0.401	0.297
Buried/total surface		1.907	0.263	0.948	
Atomic contacts	Interchain contacts	H-H	0.495	0.529	0.622
		H-P	0.406	0.385	0.317
		P-P	0.099	0.086	0.061
		B-B	0.086	0.095	0.071
		B-Sc	0.382	0.401	0.347
		Sc-Sc	0.531	0.504	0.582
	Intrachain contacts	H-H	0.457	0.364	0.377
		H-P	0.439	0.491	0.496
		P-P	0.104	0.145	0.127
		B-B	0.278	0.310	0.387
		B-Sc	0.393	0.431	0.407
		Sc-Sc	0.329	0.259	0.206
	Ratio of inter/intrachain contacts		0.104	0.820	0.348
Interaction energy	Total interaction energy/residue		-0.519	-0.191	-0.461
	Fraction of energy from interchain interactions		0.040	0.832	0.532

Table 5: Normalized average structural properties of the protein interactions classes.

H - hydrophobic, P - polar, B - backbone, Sc - sidechain. Depth of the shading represents the deviation from the average.

From the three interaction classes, ordered/ordered protein complexes have the most balanced composition between various secondary structure elements. IDPs undergoing coupled folding and binding largely lack periodic secondary structures, and often adopt irregular conformations. Disordered/disordered complexes strongly prefer helices, and in a few cases, coil structures as well.

The analysis also suggested that in bound form, all three classes have nearly the same hydrophobic/polar ratio of the solvent accessible surface area, due to the fact that the complexes, after formation, have to exist in the same aqueous environment. However, the interfaces that get buried during the interactions do not share the same hydrophobicity characters, reflecting the different binding modes. IDPs usually shield a larger fraction of hydrophobic residues by burying them, and this effect is most notable for disordered/disordered complexes. In addition, there is a reverse trend between the fraction of molecular surface a protein segment buries in the interface and by intramolecular shielding. Ordered proteins bury large fractions of surfaces during folding, and only donate a fairly small fraction to the binding. The reverse trend is true for IDPs in disordered/ordered complexes, with the majority of surface being utilized in the interface. Disordered/disordered IDPs fall between the two extreme cases on average.

Focusing on molecular interactions, IDPs in disordered/disordered complexes mostly depend on intrachain interactions, whereas during coupled folding and binding, interchain contacts are more involved in the final stability of the complex. Disordered/ordered interactions are primarily mediated by a larger number of interchain interactions, as these IDR segments are on average shorter (see Figure 14). The ratio of interchain and intrachain contacts supports the observation that bound IDPs use their residues more efficiently during binding, which is increasingly true for IDR undergoing coupled folding and binding.

In order to assess the overall stability of the complexes, interaction energies were calculated (based on the pairwise energy potentials used in IUPred [71]). Considering these energetics, ordered/ordered complexes are the most tightly bound systems on average, although the majority of this stabilizing energy come from the contacts providing the stable fold for the individual domains. Disordered/ordered complexes in contrast have the lowest overall per residue stabilizing energy, and the major contribution to this stability comes from the interaction itself. Disordered/disordered complexes have similar overall stability to ordered/ordered complexes, but the interchain interactions play a more dominant role.

4.2.3. Closer to the DNA, closer to the IDP-mediated interactions

The appropriate subcellular localization of proteins is critical because it provides the spatial context for their function, determining - amongst other factors - the range of possible interaction partners. It has been suggested that IDPs are specialized to various functions, but how is that reflected in their localization? To address this question, subcellular localizations were taken from the "cellular component" namespace of GO, as used in the MFIB and DIBS databases. For all complexes from the three interaction categories a GO term was considered if at least two of the constituent chains were annotated with that term. To make the selected GO terms comparable, a reduced set of higher-level terms of typical subcellular localizations was selected, termed CellLoc GO Slim (see Figure 15).

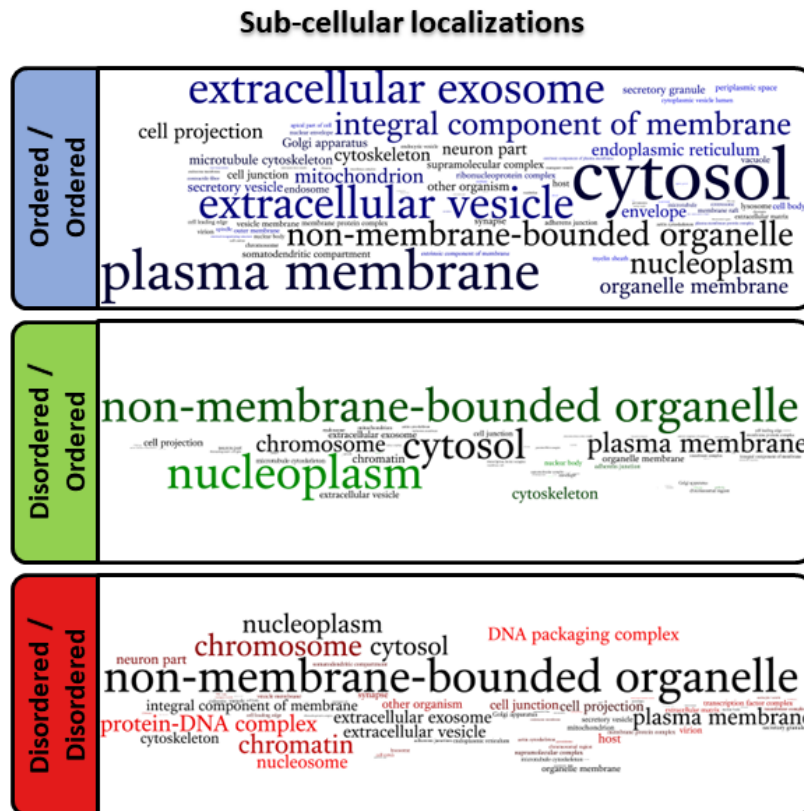


Figure 15: Characteristic sub-cellular localizations of the three interaction classes, visualized using a word cloud technique.

The relative frequency of occurrences for CellLoc GO terms is represented by the font sizes. Color depth represents the specificity of the given term for that protein class. Localization names in black show terms that ubiquitously occur in all types, terms in full color represent features that are unique to the given class. The chosen main colors for the classes were blue for ordered/ordered proteins, green for the disordered/ordered class and red for the disordered/disordered interactions.

The extracellular space is dominated by ordered/ordered interactions. Furthermore, we can often find ordered complexes embedded as receptors in the plasma membrane and the membrane of extracellular vesicles, as part of the connection between the extracellular and intracellular spaces in the cell. Moreover, the most characteristic localization of complexes formed by ordered proteins is the cytosol. Moving closer to the nucleus/nucleoplasm, more disordered/ordered complexes are present. In general, IDP-mediated functions are typically centered around the DNA. These interactions can be found in the nucleoplasm, and in non-membrane bounded organelles, such as ribosomes, stress granules, or centrosomes. Localizations close to the DNA are also enriched in disordered/disordered complexes, even more so than disordered/ordered complexes, as they often interact directly with the DNA, and typically occur in chromatin organization or DNA packaging. The dominance of disordered/disordered interactions around the DNA is in line with the need for transient interactions in transcription related biological processes.

4.2.4. Interactions of IDPs mediate distinct biological functions in the cell

Protein disorder is related to numerous biological processes and molecular functions, which have been addressed in several studies [86, 129]. It is also known that IDPs are associated with distinctively different functions than ordered proteins [130]. I assessed the typical biological functions of various IDP-mediated interactions using a reduced set of GO biological process terms. From PPI GO Slim (see Data and Methods), we can recognize the most characteristic biological functions for the interaction classes.

The analysis highlights that all three interaction types are involved in all major high-level/generic biological processes (such as transport or communication). However, lower-level processes show preferences between different interaction classes (see Figure 16).

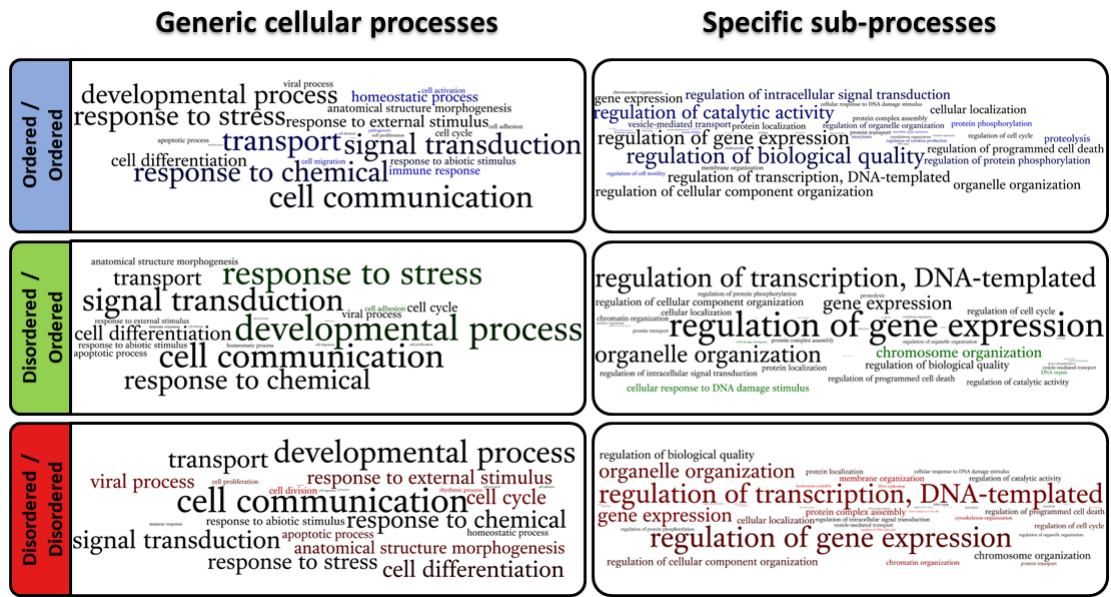


Figure 16: The relative frequency of occurrence for PPI GO terms for all interaction classes, represented by the font sizes.

The used font colors are similar to Figure 15, with font color depth marking the specificity of a cellular process for the given interaction type.

The most notable functions of ordered proteins are connected to the maintenance of homeostasis, including different regulatory processes, such as gene expression, control of catalysis of chemical reactions as enzymes, stress response, they also function as transport vehicles for other molecules in circulating blood, and they are heavily involved in replication and the repair of DNA to ensure the biological quality in the cell.

IDP-mediated functions are usually involved in DNA-related functions, though, there is a separation of processes depending on the structural state of the partner proteins. Chromosome organization and DNA repair pathways are enriched in disordered/ordered complexes. Disordered/disordered interactions often connect to the information storage function of DNA, including direct DNA contact. Thus, proteins with MSF often play a crucial role in transcription and gene regulation.

The analysis of different sequential, structural properties, together with various localization preferences strongly support the idea that certain biological processes have strong preferences for distinct interaction types. From the side of the formed protein complexes, the ordered or disordered state of the interacting partner of a protein has a strong effect on the sequential, structural and functional properties of the formed complex.

4.2.5. Protein disorder extends the biologically relevant sequence, structure, and functional spaces of proteins

The previous analyses only consider the average values of various features, without quantifying their heterogeneity across each interaction class. To assess this heterogeneity, we need to directly quantify and visualize the sequence, structure and function spaces covered by various interaction categories. The three levels (sequence, structure and function) were evaluated separately for all three interaction classes. The functional annotations were represented by a 23-element vector, as the GO PPI Slim contains 23 possible high level cellular/organismal processes. For visualization, Principal Component Analysis (PCA) was employed, and I used the first 2 components for the best 2D representation, as they are carrying the highest fraction of variation of data, and can highlight the main differences of the different interaction classes (see Figure 17). Ordered/ordered and disordered/disordered complexes dominate different sequential and structural spaces. Protein complexes with coupled folding and binding (disordered/ordered) overlap with the other two classes according to their sequential features, however, in the structural space they occupy a more distinct subregion. Distribution in the functional space shows high overlap in all three interactions classes, to support the fact, that at a higher functional level, the roles of different interaction classes are intertwined, instead of separately accomplishing various biological roles.

To more objectively quantify the extent of various spaces used by different interactions, sequence-, structure-, and functional heterogeneity values were calculated for all three types of interactions (see Data and Methods for full description and definitions). Heterogeneity values were defined as the average dissimilarity between two randomly chosen complexes from the same class. The so calculated heterogeneity values lie between 0% and 100%, with 0% corresponding to all complexes being identical and 100% corresponding to all pairs of complexes being as different as possible. Dissimilarity between two complexes from the same class was defined based on the hierarchical clustering of complexes (see Data and Methods), being defined as the Euclidean distance of complexes in the tree constructed by the clustering. Heterogeneity values are defined as the geometric averages of dissimilarity values between all protein pairs from a specific interaction class, and dissimilarity of two proteins is the quotient of the linkage distance given by the clustering of Ward algorithm and the maximal linkage distance.

For functional heterogeneity, no clustering was employed, instead the GO ontology tree was used. Distances between parent/child terms were defined as 1. The dissimilarity between two complexes was defined based on their most similar GO term pairs. For each complex from a pair, I chose the biological process selected for minimal distance in the ontology. Based on these dissimilarity values, heterogeneity values were calculated as described for other features. The calculated heterogeneity values are again in a range of 0-100%, 100% means that all pairs of complexes being as different as possible, and 0% corresponding to all complexes being identical.

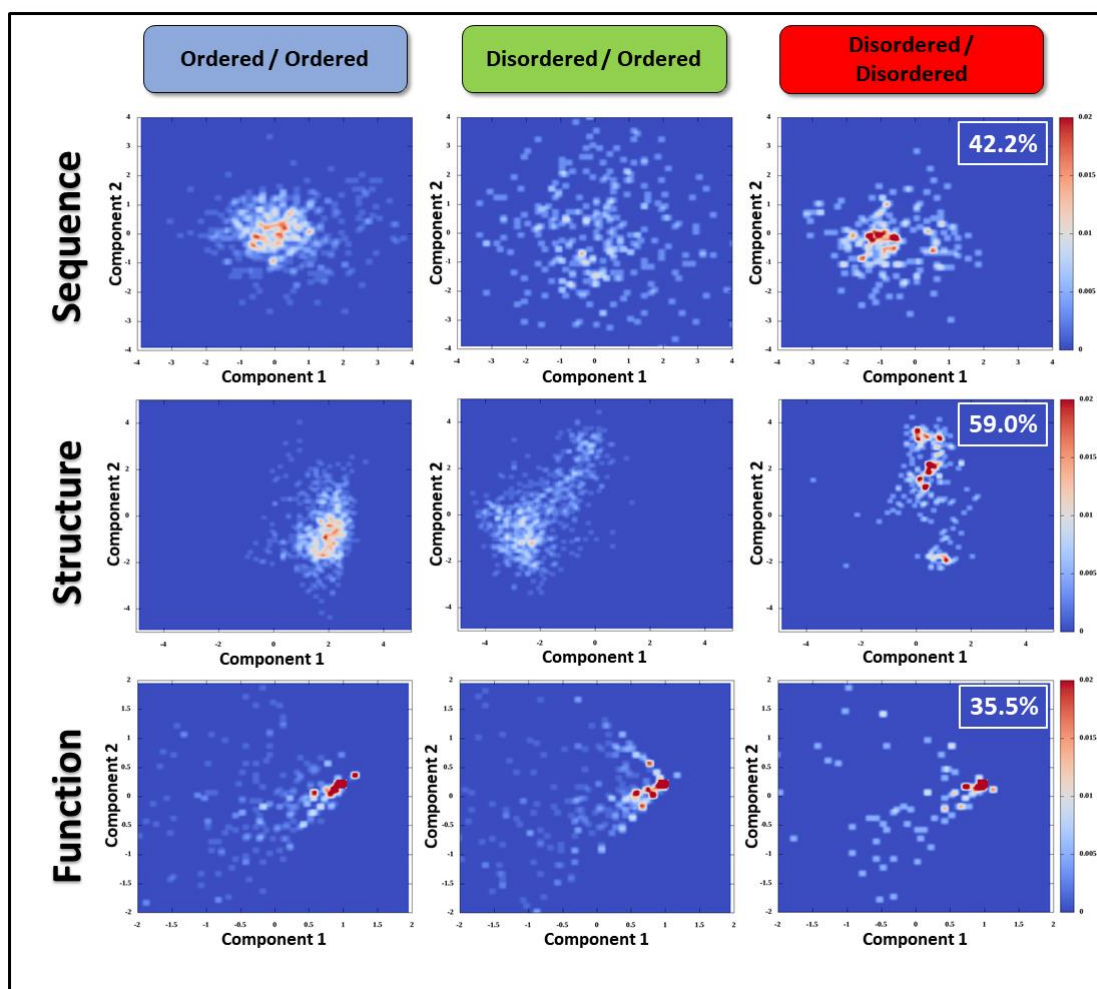


Figure 17: Distribution of various complexes considering the best two dimensional (the first two principal components) representation of the sequence-, structure- and functional spaces.

Insets show the total variance of the data carried by the plotted components.

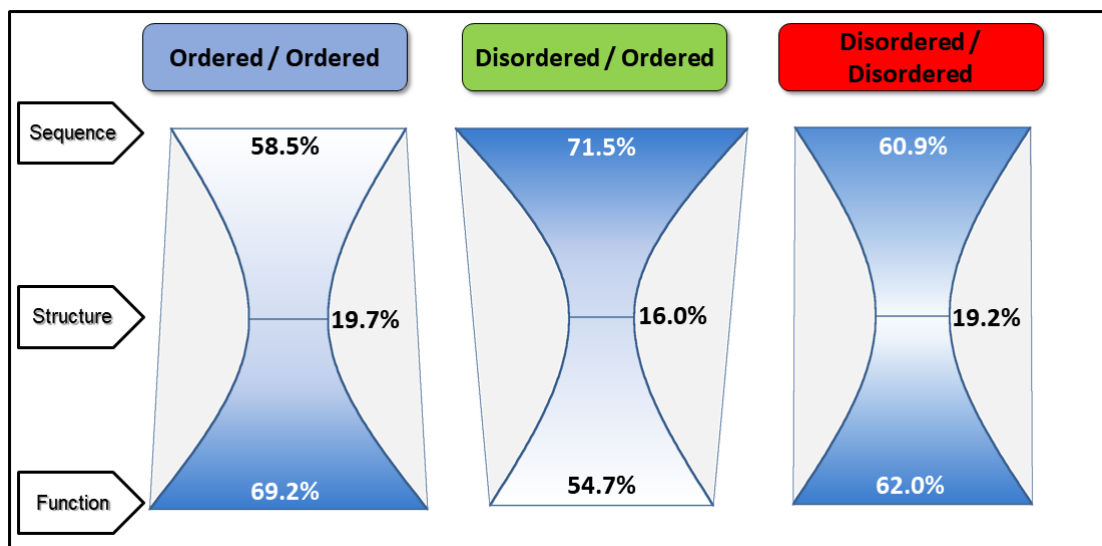


Figure 18: Variability of sequences, structures, and functions, using calculated heterogeneity values for all three interaction classes.

Figure 18 shows the calculated heterogeneity values for all three interaction classes. In general, the highly variable sequence compositions of proteins utilize a much narrower set of the structure space; nevertheless, it does not limit the functional possibilities of the complexes. Disordered/disordered proteins show a balance between the heterogeneity of sequences and functions. Ordered/ordered complexes have restricted sequence compositions to fulfill a wide range of functions, and in opposition to disordered/ordered interactions, the restricted range of functions are accomplished by a highly variable sequence space.

4.2.6. IDPs are heavily regulated via post-translational modifications

As different levels of GO-annotations suggested, the three studied interaction classes play essential roles in vital biological processes, and need to be precisely regulated. Many studies shed light on the general importance of various post-translational modifications that occur in response to a dynamic change in the external and internal environment of a given cell type. Works conducted in recent years highlighted that IDPs are under especially strong regulation [131].

In the presented analysis, I studied the four most commonly occurring types of PTMs (phosphorylation, acetylation, methylation, and ubiquitination) using the low-throughput, experimentally verified data from PhosphoSitePlus, Phospho.ELM and UniProt (see Figure 19).

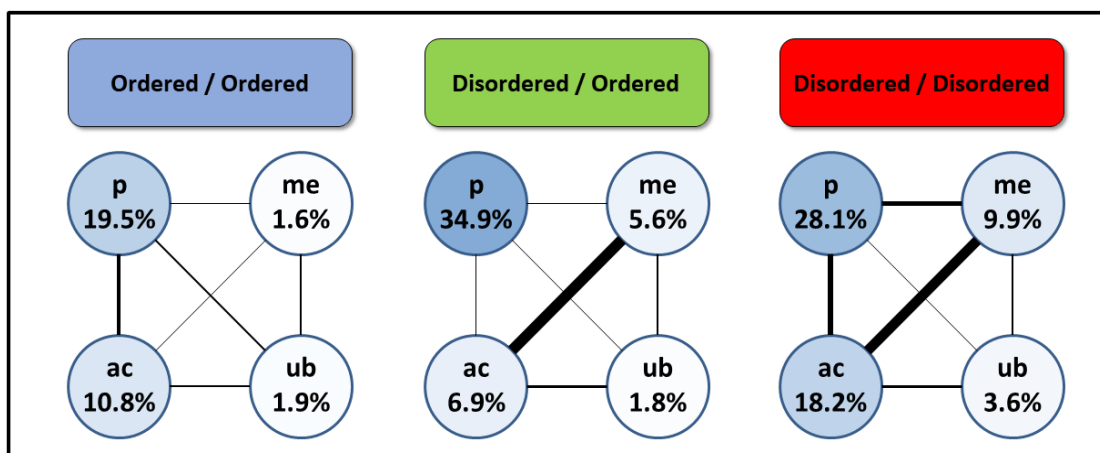


Figure 19: The occurrence of PTMs in interacting proteins from the three interaction classes.

ac - acetylation, me - methylation, p - phosphorylation, ub - ubiquitination. The width of connecting lines shows the amount of mutual information between the occurrence of PTM pairs over proteins in the given interaction class. The percentage values and color depth represents the fraction of proteins affected.

All four types of PTMs occur in all three interaction classes, and they are present in both ordered and disordered interacting proteins. Disordered/disordered proteins have more known PTMs than the other two classes, with IDPs in general harboring more PTMs than ordered proteins. Furthermore, IDPs not only contain more PTM sites that can affect complex formation, but the occurrence of these PTMs seem to be more overlapping, suggesting a coordinated regulatory system between various PTM types. IDPs in disordered/disordered complexes show an extremely high cooperation between methylation and acetylation events, whereas IDPs in disordered/disordered complexes show a more general coordination among all four types of PTMs.

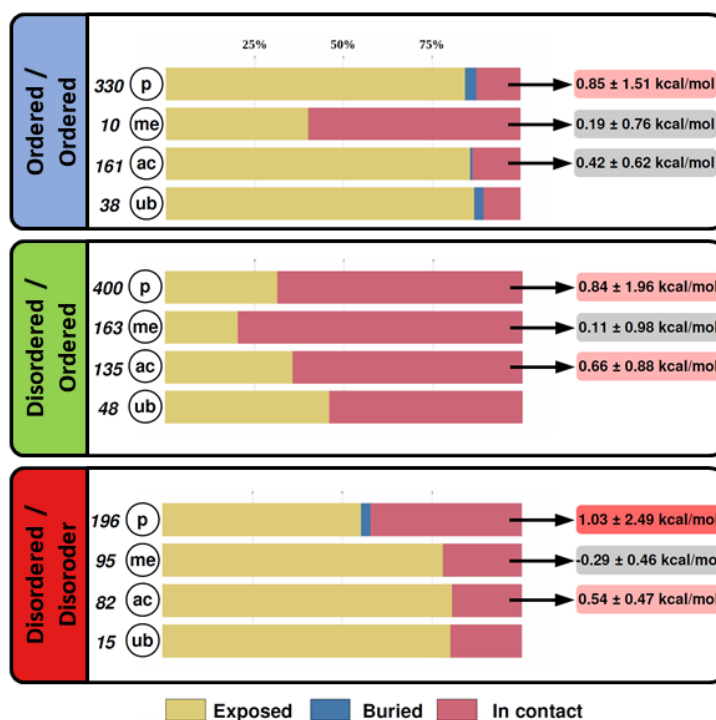


Figure 20: Location of PTM sites in the complex structure.

Values next to circles indicate the amount of PTMs found. Colored bars represent occurrences in the three types of structural elements (exposed, buried and interface/“in contact” residues). 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. Colors indicate the destabilizing effect of the average value (grey – 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.

The structural location of PTMs offers insights into the mechanistic effects they have on the binding event. As Figure 20 highlights, PTMs in ordered/ordered complexes are enriched in the solvent accessible surface of the domains (exposed surface) which probably do not have a direct influence on the binding, however, they can have an indirect impact. On the other hand, a smaller fraction of PTMs do affect residues in direct contact with the partner, and substitution calculations for these PTMs show (when modeling the PTM with mimetic residue substitutions, and calculating the estimated change in stability when introducing the “mutation”), they change the estimated free energy of the resulting complex moderately. The number of PTMs that have a direct effect on the binding (in contact residues) in case of disordered/ordered complexes is relatively higher, providing a larger possible control over the binding event, albeit the energy contribution of single PTMs remains fairly subtle. The change of free energy brought about by the introduction of in contact PTMs for disordered/disordered complexes is higher than for

disordered/ordered and globular proteins, but the number of PTMs is in the middle between the two other classes.

PTMs are capable of producing significant changes in the interaction properties of IDPs, and hence play a role in the modulation of IDP-mediated interactions. My analysis suggests that IDP-ordered complexes in general are regulated through a high number of PTMs each with subtle energetic changes, while a restricted number of PTMs have a more significant effect on stability in the case of disordered/disordered complexes. PTMs in ordered proteins might indirectly modulate the given interactions, but in IDPs they have a more direct role.

4.2.7. Cooperation between ordered proteins and IDPs in different interaction modes

Figure 21 shows representative interactions from each of all three analysed interaction classes, illustrating the intertwined connection between structural and sequential characteristics.

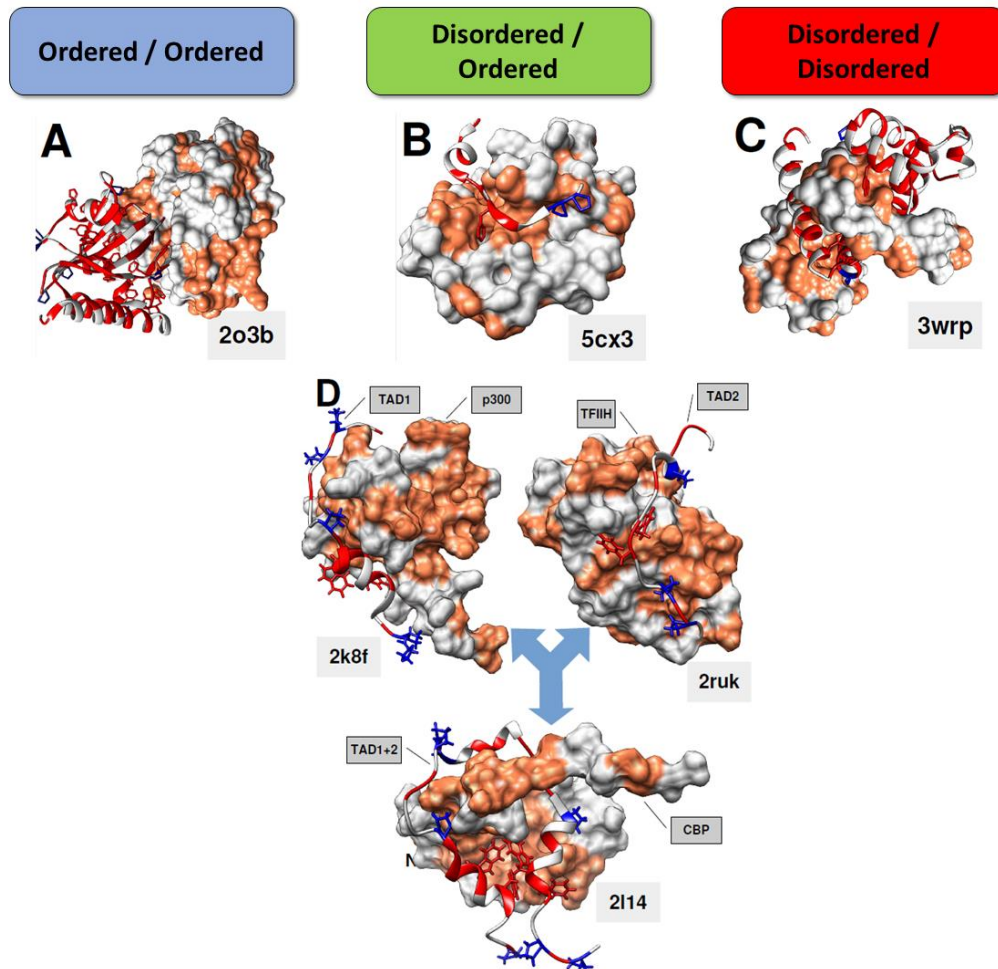


Figure 21: Representative examples of the three classes of interaction mechanisms. The analyzed partner is shown in ribbon representation, while the partner is shown as a surface. Hydrophobic residues are red (orange for surfaces), aromatic side chains are shown in stick representation. Prolines are 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 the Trp repressor. D: the two transactivator regions (TAD1 and TAD2) of human p53 bound to ordered domains from p300 and TFIIH, respectively (top), and TAD1+TAD2 bound to an IDR of CBP (bottom). Examples in panels A-C have both sequence and structure features very close to the averages of their respective interaction classes.

NucA inhibition by NuiA is an example from the ordered/ordered category (Figure 21A). Nuclease A is a small, extracellular monomeric enzyme. Its inhibitor is NuiA, which can specifically bind to it with a high affinity. Both proteins are ordered in the monomeric form. The stable monomeric form is required as a prior condition for the

enzymatic activity of NucA. Both proteins have hydrophobic cores on their own, enriched in aromatic residues. NuiA can specifically recognize and bind to NucA owing to its native structure that is very close to the bound conformation [132].

The disordered/ordered class is represented by the disordered LC3-interacting region (LIR) of FYCO1 and the ordered ubiquitin-like domain of autophagy-related LC3 (Figure 21B). In this case, the IDP undergoes coupled folding and binding. This interaction has to be reversible and fast with high specificity, as a microtubule-based kinesin motor and autophagosomes interact with each other. LIR adopts a largely irregular conformation, and lacks buried residues. The relatively high number of prolines is able to lower the need for structural adaptation.

The DNA-binding domain of Trp repressor is a structurally malleable homodimer from the disordered/disordered category (Figure 21C). The flexibility of the complex ensures that this protein can bind to three different operator sites. The chains are enriched in hydrophobic residues to form the hydrophobic core together in order to stabilize the resulting complex, and they exhibit extreme plasticity that enables them to mutually bind to each other.

These examples indicate that certain biological processes require one definite type of interactions. However, these interaction types are not clearly segregated, as they show cooperation on at least two different levels: individual binding site-level and the network level. The former scenario corresponds to specialized cases of IDRs that have evolved to satisfy the requirements of both ordered and disordered partners. In the case of the transactivation domain of p53, the corresponding IDR can function 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) (Figure 21D). The transition between the two scenarios is highly non-trivial and requires the structural adaptation and re-positioning of several residues, especially for TAD2. The key to this duality lies in the fact that for both TAD1 and TAD2, the main force being their interactions is hydrophobicity, resulting in a fairly high fraction of hydrophobic residues. Yet, both binding sites are small enough to avoid mutual synergistic folding on their own, due to the lack of a sufficient amount of hydrophobic amino acids. 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.

The uncovered differences and similarities between the types of interaction classes can provide a cornerstone for a basic understanding of how the interplay between protein folding and interaction modulates the various features of the complexes. Amongst these, the presented analysis also shows how sequential properties are recognized at the structural level, and thus intrinsic disorder means a bona fide regulation mechanism for the activity of the subunits and the complex. In the last few years, interactions between IDPs and the other proteins have been an exciting research topic, as their detailed analysis might open new opportunities for therapeutic targeting. Some of the studies based on IDP interactions have already led to successful pharmaceutical targeting [133], and the better understanding of the intricate details of these binding modes have the potential to serve with further, currently overlooked therapeutic options.

4.3. Classification of MSF complexes

The disordered and ordered nature of the proteins enables to characterize them in different ways. For ordered proteins, the typically used groupings are rooted in the three-dimensional structures expressed as various domain or fold types, collected for example in the CATH or SCOP databases. The fold classes can also be a starting point to the structural classification in the case of IDPs bound to domains via coupled folding and binding. This classification approach is employed in the DIBS database, where the assignment of the complexes primarily relies on the domain type of the ordered interacting partner(s). However, MSF complexes cannot be classified using this approach because they do not incorporate ordered proteins. However, the previous analyses focusing on sequence/structure heterogeneity (see section 4.2.5) suggested that disordered/disordered complexes have distinctive sequential and structural features, which enable their recognition as a separate interaction category. The MSF protein class also shows large enough variations to group them into subgroups, moreover the previous PCA calculations also suggested the existence of well separated subcategories as well (Figure 17). In the following chapters I use these features to propose the first hierarchical classification for MSF complexes.

4.3.1. Sequence and structure features of MSF complexes

Previously I analyzed ordered proteins and IDPs involved in the interactions, but in this scenario instead of analyzing proteins, I focused on full complexes formed exclusively by IDPs. I recalculated the previously mentioned sequence and structure features for these complex structures (see chapters 4.2.1 and 4.2.2). MSF complexes have been assigned a feature vector describing the sequence composition of the entire complex, with an additional parameter quantifying the compositional difference between subunits (see Data and Methods), and these vectors were used as input for hierarchical clustering (Figure 22). K-means clustering indicated 4 as the most appropriate number of clusters in case of sequence features.

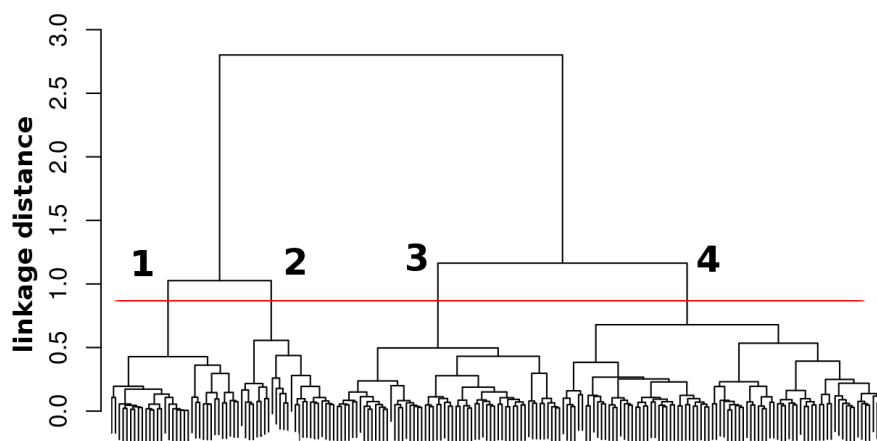


Figure 22: Sequence-based hierarchical clustering of complexes with mutual synergistic folding.

The red line indicates the optimal cut in linkage distance to define clusters, derived from k-means clustering (not shown).

		Cluster 1	Cluster 2	Cluster 3	Cluster 4	Average
Number of members		34	25	59	85	-
Average amino acid composition	Aromatic	0.0531	0.0436	0.0547	0.0724	0.0605
	Hydrophobic	0.3301	0.3239	0.3218	0.3227	0.3238
	Flexible	0.0661	0.0425	0.0319	0.0625	0.0517
	Rigid	0.0319	0.0376	0.0209	0.0338	0.0302
	Charged	0.3199	0.3049	0.3811	0.2587	0.3102
	Polar	0.1931	0.2399	0.1819	0.2366	0.2138
	Cysteine	0.0058	0.0076	0.0077	0.0133	0.0097
Heterogeneity (average dissimilarity between subunits)		0.2319	0.4146	0	0.0064	0.0926
The difference compared to disordered proteins undergoing coupled folding and binding		0.192	0.171	0.297	0.174	0.212
The difference compared to ordered proteins		0.154	0.186	0.260	0.100	0.166

Table 4: Sequence features calculated for MSF complexes.

Blue and red shadings mark values that are more than 20% lower or higher than the average, respectively.

Table 6 highlights the main average sequence features of the resulting clusters. Sequence clusters 1 and 4 mostly resemble ordered/ordered complexes, and cluster 4 contains more cysteines and aromatic residues, possibly to enhance stability. Moreover, protein chains of cluster 4 are almost identical, while cluster 1 is dominated by hetero-oligomeric complexes. The amino acid composition of the complexes in cluster 2 shows the highest similarity to disordered proteins undergoing coupled folding and binding. They are exclusively hetero-oligomers, enriched in polar residues and prolines. Sequence cluster 3 contains homo-oligomers, enriched in charged amino acids, depleted in glycines and prolines.

Structural properties of complexes were quantified using the secondary structure composition, atomic contacts, and various molecular surface parameters, as detailed in the section above (see chapter 4.2.2). I discarded the features with relatively high pairwise correlation values to avoid selection bias. The remaining parameters were used to represent the complexes in a feature vector, which - similarly to the sequence feature vectors - were used in hierarchical clustering (see the resulting tree in Figure 23). K-means clustering indicates again 4 as the optimal number of structural groups.

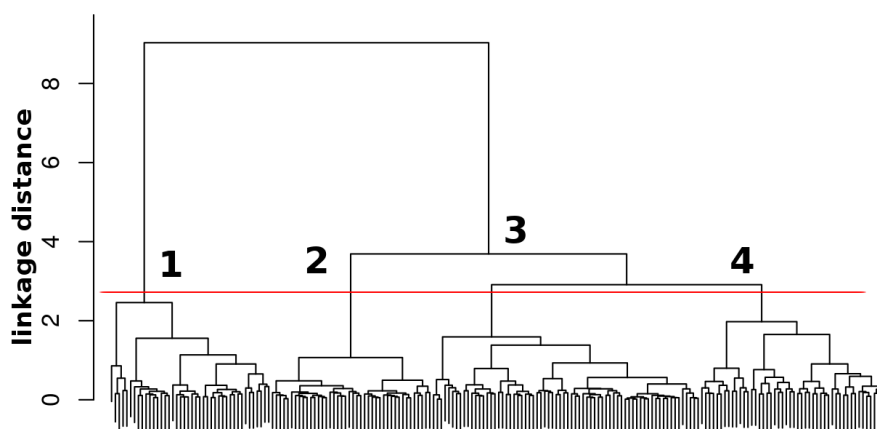


Figure 23: Structure-based hierarchical clustering of MSF complexes.

The red line indicates the optimal cut in linkage distance to define clusters, derived from k-means clustering (not shown).

		Cluster 1	Cluster 2	Cluster 3	Cluster 4	Average
Number of members		42	42	70	49	-
Secondary structures	alpha	0.127	0.908	0.699	0.509	0.578
	strand	0.417	0.001	0.029	0.131	0.128
	coil	0.457	0.091	0.272	0.360	0.294
Molecular surfaces	Hydrophobic SASA	0.560	0.563	0.558	0.553	0.558
	Hydrophobic interacting surface	0.680	0.783	0.781	0.726	0.747
	Hydrophobic buried surface	0.591	0.386	0.510	0.497	0.498
	Interface/ total	0.198	0.347	0.278	0.301	0.281
	Buried/total	1.632	0.656	0.997	0.720	0.990
Atomic contacts	Interchain, backbone- backbone	0.161	0.008	0.041	0.108	0.075
	Intrachain, backbone- backbone	0.280	0.483	0.376	0.394	0.382
	Interchain/ total	0.117	0.225	0.186	0.274	0.201

Table 7: The structure features calculated for complexes formed exclusively by IDPs.

Blue and red shadings mark values that are more than 20% lower or higher than the average, respectively.

All four groups exhibit distinct structural features (see Table 7). Complexes from cluster 1 contain the highest amount of non-helical secondary structure elements, in contrast to clusters 2 and 3, which adopt mainly helical structures (complexes in cluster 2 on average have over 90% of their residues in α -helical conformation), and have more intermolecular interactions between the chains to stabilize the resulting complex. Members of cluster 1 also have a large number of buried hydrophobic residues, and have a large number of intramolecular interactions for stability. Complexes from cluster 2 on average use a high fraction of their interchain backbone-backbone atomic contacts. Complexes from Cluster 4 seem to be the most balanced regarding the studied structural properties (such as the composition of secondary structures or various molecular surfaces, or the ratio of atomic contacts). They utilize more interchain contacts paired with low

fractions of buried surface, meaning the stability of these complexes typically comes from interchain interactions.

Based on the calculated sequence and structure features, and considering them together, 16 complex types can be defined, as shown in Figure 24. There are many strikingly not favored types, with very few or no known examples in them. Categories with 10 or fewer representatives were omitted or merged with the adjacent sequence clusters. This resulted in 5 main interaction types (near-ordered, multi-chain domains, coils/zippers, compacted coils and histone-like), each of which has over 30 members, and the 6th class is termed the “other” category to include the rest of the MSF complexes not classifiable using the 5 well-defined clusters.

This division only considers sequence and structure properties, however these categories can be biologically meaningful as well. If that is the case and they represent "true" classes in nature, there should be noticeable differences in their regulation or energetics, and they are expected to have characteristic functional roles in biological processes.

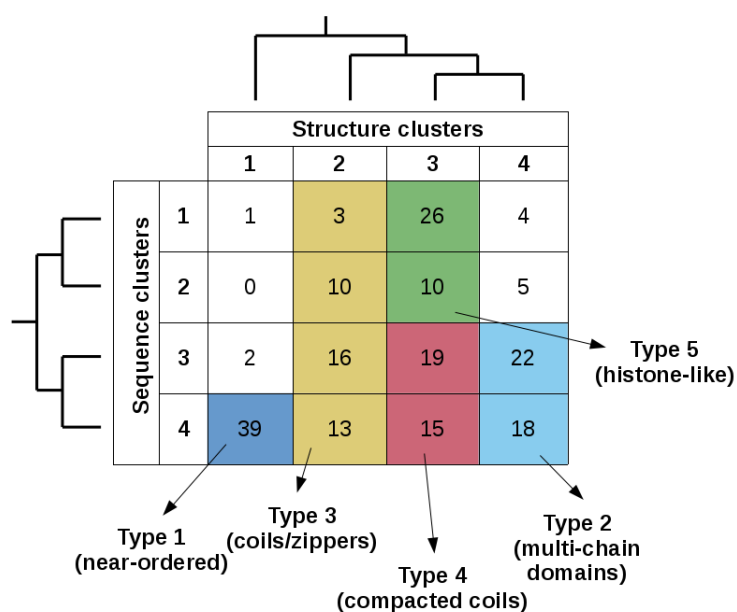


Figure 24: Definition of interaction types considering sequence and structure clusters.

4.3.2. Energetic properties of interactions

The protein interactome is dominated by a large number of weak interactions that are, nevertheless, more important for overall network topology than the smaller number of strong interactions [134]. The PPI network is critically dependent on the strengths of interactions between constituent chains of complexes. Unfortunately, we do not have enough targeted measurements for IDPs in MSF complexes to experimentally assess the overall binding properties. To circumvent this limitation and to assess the energetics of MSF interactions, I used low-resolution energy calculations based on statistical potentials described in [71]. Using this force field I was able to quantify the residue-level interaction energies for the entire complex. As a reference, I used the ordered/ordered and disordered/ordered sets of proteins from the previous analyses to give context to the results, as shown in Figure 25.

Based on these predictions, ordered/ordered complexes typically have strongly bound structures, with low stabilizing energies per residue, the constituent chains having stable structures on their own, gaining only a limited amount of stabilizing energy from interchain interactions. In contrast, disordered/ordered complexes are comparatively weakly bound, with interchain interactions playing a major role in the stability of the complex. In comparison, MSF complexes seem to cover the whole available range of energetic properties, with characteristic differences between the defined groups. While most groups are in-between the reference interaction classes, both in terms of overall stability, and the importance of the interaction. However, there are two extreme groups of MSF complexes. Type 3 complexes (coiled coils) on average are the least stable types of structures, and they gain virtually all of their stability from inter-subunit interactions. At the other extreme are type 1 (near-ordered) complexes, that are on average display the same level of stability as ordered/ordered complexes, with subunit interactions playing only a limited role in overall stability.

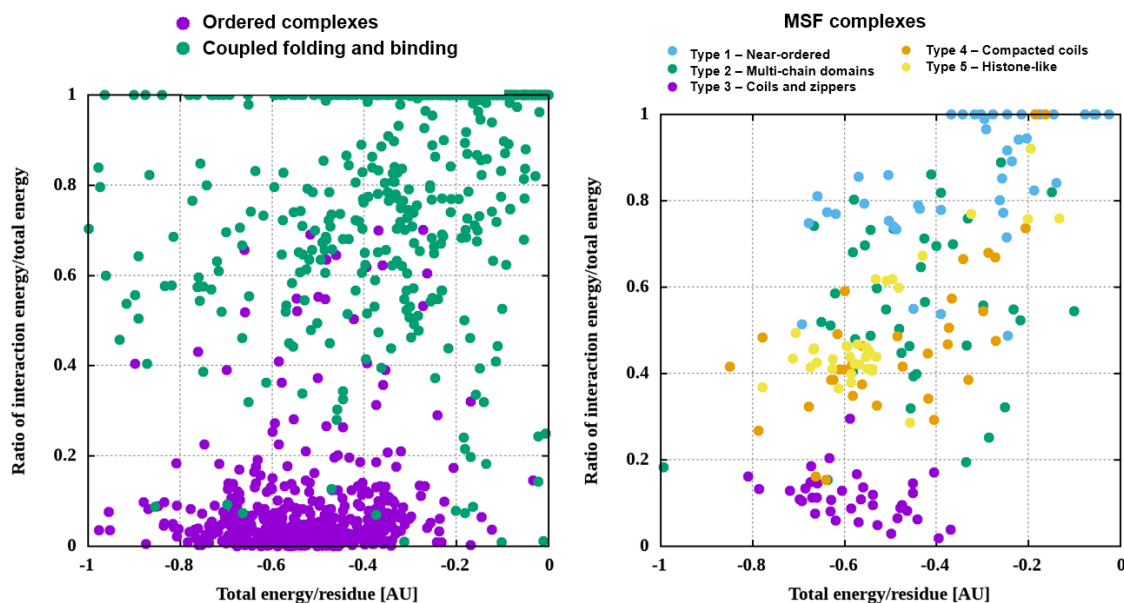


Figure 25: Energetic parameters of various interaction classes.

The relative energetic importance of intersubunit interactions as a function of the overall energy per residue for ordered complexes and complexes formed by coupled folding and binding (left), and the five well-defined types of MSF complexes (right).

The transient and reversible nature of certain interactions is a fundamental pillar of the interactome. In the case of IDPs with coupled folding and binding, there is a large number of K_d values in DIBS (see chapter 4.1.5), but these values are lacking in MFIB, as MSF interactions are surprisingly rarely studied in this regard. In order to further assess MSF complex binding properties, K_d dissociation constants were predicted based on the PDB structures (see Data and Methods). Figure 26 shows the distribution of estimated K_d values for the six previously described MSF complex groups. The lowest average K_d values were calculated for histone-like and near-ordered complexes, as they possibly include the highest fraction of obligate interactions. Non-classifiable complexes have the highest K_d values, presumably they contain the most transient interactions with unusual sequence and structure features.

These results suggest that MSF complexes are energetically the most diverse type of interactions from the three studied classes. They seem to cover the whole range of biologically relevant stability, including both highly transient and reversible, and obligate interactions, with a wide range of energetic properties and dissociation constants. Certain groups of MSF interactions form complexes that are highly similar to ones formed by ordered proteins, while others resemble IDPs stabilized via coupled folding and binding.

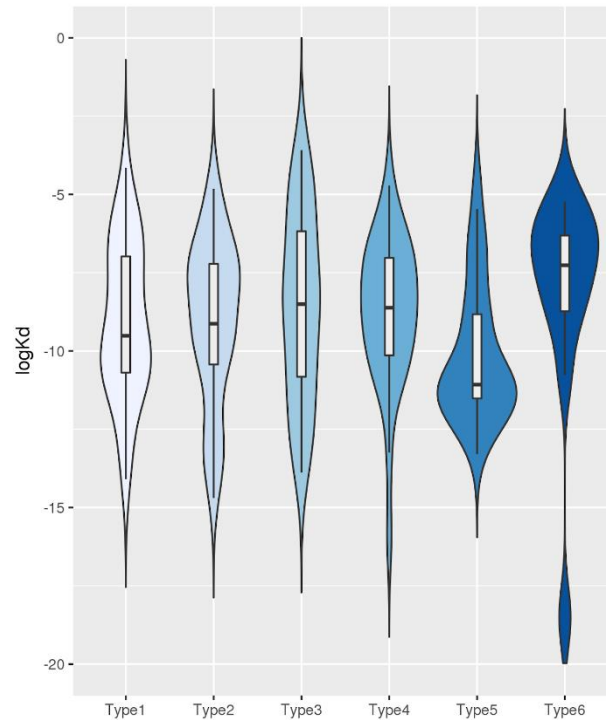


Figure 26: K_d value distributions for complexes formed via mutual synergistic folding.

4.3.3. Regulatory mechanisms of MSF complexes

The energetic properties fundamentally define the transient or obligate nature of MSF interactions. However, cellular processes typically require even more precise control over crucial interactions with regard to the presence of the right amounts of precisely activated IDPs, at the right localization, and at the right time. There are several regulatory mechanisms that are often used for the precise control of protein-protein interactions, including PTMs, competing interactions, and alternative splicing to control the presence or absence of protein binding regions. I studied the prevalence of these three regulatory mechanisms in proteins participating in MSF complexes. In chapter 4.2.6 I already established the role of various PTMs in IDP-mediated interactions. In addition, I collected alternative splicing events and the presence of alternative binding partners competing for the same interaction region present in the studied MSF complexes (see Data and Methods). About 40% of proteins participating in the studied MSF complexes incorporate at least one of the three studied regulatory mechanisms. As MSF complexes are generally lacking focused studies, these numbers most probably are conservative estimates of the true extent of MSF regulatory mechanisms.

Figure 27A shows the fraction of MSF complexes that are affected by various experimentally validated PTMs identified in low throughput experiments. The most prevalent PTM is phosphorylation, but acetylation and methylation can also often modulate MSF complexes. Apart from PTMs, alternative splicing also plays a role in the regulation of about 10% of MSF interactions. In these cases, the presence of the binding region of at least one of the interaction partners depends on the splicing of the mRNA of the corresponding gene(s) (see Figure 27B). In general, alternative splicing tends to avoid protein domains [135], and the spliced segments are enriched in IDPs and contain linear interaction motifs or PTM sites [136]. Analysis of MSF complexes shows that the existence of various isoforms and their regulation should be taken into account when studying MSF binding events. The relationship between alternative splicing and the binding event can be quantified by assessing the “precision” of the splicing event with respect to the binding site. I define this precision value as the fraction of residues affected by the splicing event that belong to the binding site. If the splicing removes a large protein region, that contains a lot of other functional sites apart from the MSF binding region, precision values will be low. However, for splicing events that only remove the MSF binding site and no other regions from the protein, precision will be 1. Figure 27B shows the precision values for each MSF binding site affected by alternative splicing (if for the same region multiple splice variants are known, only the one with the highest precision is taken into account). These data not only show that alternative splicing affects synergistically folding IDP sites, but that it specifically targets them, thus alternative splicing is likely a bona fide regulatory mechanism for MSF complexes. Apart from PTMs and alternative splicing events, about 15% of the studied interactions are subject to competition from other molecular partners. Furthermore, Figure 27C illustrates that the different regulation modes are intertwined for several interactions. Alternative splicing and competing interactions seem to present two mostly mutually exclusive alternative mechanisms for interaction control, however PTMs seem to readily coexist with both, with seven complexes featuring all three studied regulatory mechanisms. These studies show that MSF interactions are typically heavily regulated via independent, but often intertwined mechanisms at both the protein and mRNA level.

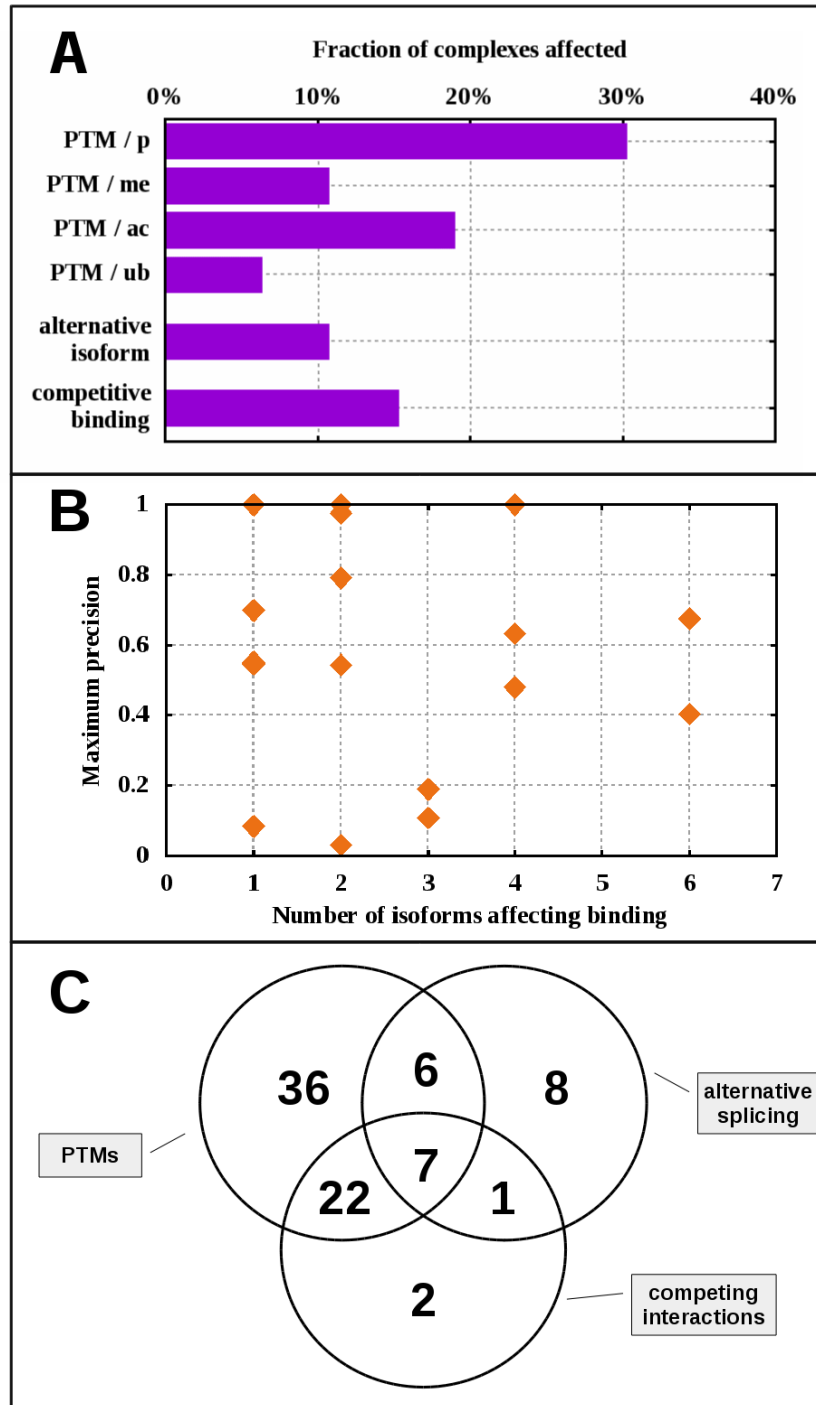


Figure 27: Regulatory mechanisms of MSF complexes.

A: 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. B: The number and specificity of isoforms affecting the binding regions of IDPs with MSF. Specificity represents the ratio of the spliced residues belonging to the binding site. For each protein, only the specificity of the most specific isoform is shown. C: Number of MSF complexes affected by the three types of regulatory mechanisms.

4.3.4. Catalog of MSF complexes

The previous subchapters present an outline of different biological aspects of MSF complexes, covering sequential, structural, and regulatory properties. To complement the described features, and to consider the previously calculated parameters, for each of the six defined MSF groups I also calculated and compiled the main subcellular localizations (see chapter 4.2.3), biological processes (chapter 4.2.4), and heterogeneity values (chapter 4.2.5). Taking all of these features into consideration, I propose the first systematic annotated classification of MSF complexes. While I defined the groups based on sequence and structure properties, this approach yields biologically meaningful groups, with markedly different associated functions, subcellular localizations, heterogeneity, and regulatory mechanisms. The following six figures detail these features for each of the six groups. Group IDs are consistent with the group definitions in chapter 4.3.1. Example structures and the associated protein classes were taken from entries in MFIB (chapter 4.1.2). The sequence and structure sections summarize the main distinguishing features of the group. Stability summarizes the average energetic parameters analyzed in chapter 4.3.2. Subcellular localizations and Functions are based on the CellLoc Slim and PPI Slim used in chapters 4.2.3 and 4.2.4. Regulatory mechanisms are taken from chapter 4.3.3 with the most prevalent PTMs shown in icons. Heterogeneity values were re-calculated using only the proteins involved in the complexes of the specific groups, using the same approach as presented in chapter 4.2.5.

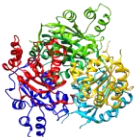
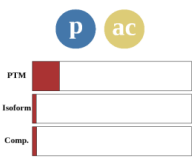
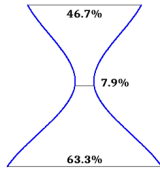
Group ID	Number of complexes	Example structure				Protein classes			
1	38	 1npk				<ul style="list-style-type: none"> ● NGFs ● Enzymes ● Bulb-type lectins ● Transthyretin-like ● E2 dimers 			
								Sequence	
		Composition	Subunit similarity	Secondary structures	Interface	Buried surface	Dominant atomic contacts	Energy	Role of inter-action
		Ordered-like	Near-identical	Mixed (extended and coil are pronounced)	small, polar	large, hydrophobic	Intrachain	+++	+
Functions		Regulation		Heterogeneity	Subcellular localization				
<ul style="list-style-type: none"> ● regulation of biological quality, catalytic activity and cellular component organization ● regulation of gene expression DNA replication ● regulation of programmed cell death 					<ul style="list-style-type: none"> ● host ● secretory vesicle ● secretory granule ● mitochondrion ● extracellular exosome 				

Figure 28: Basic properties of near-ordered type complexes (Group 1).

Figure 28 highlights the main features of **Type 1, near-ordered complexes**. Proteins forming this type of interactions show high similarity to the sequences of interacting ordered proteins, as they have a high number of polar and aromatic residues, together with a high cysteine content. The subunits are bearing near identical sequence compositions, and these (predominantly homo-)oligomers are stabilized via mostly interchain contacts. They are strongly bound structures with restricted interchain interaction energies. The complexes are mainly localized in the extracellular space and the mitochondria, including transporter proteins, nerve growth factors and enzymes. These complexes are mainly regulated by PTMs, mostly by phosphorylation and acetylation.

Complexes from Type 1 often have catalytic activities and play roles in cellular component organization, regulation of programmed cell death, and gene expression. A prime example of this type of interaction is the human glutathione S-transferase (PDB: 1k3y, MFIB: MF2100012), that functions by the addition of glutathione to target electrophilic compounds, including carcinogens, therapeutic drugs, environmental toxins, and products of oxidative stress. This action is an essential step in detoxification. The previously mentioned nucleoside diphosphate kinase (PDB: 1npk, MFIB: MF6110001) also belongs to this cluster, it is an enzyme required for the synthesis of nucleoside triphosphates other than ATP.

Near-order complexes can be considered to mark the edge of the spectrum of disorder, as these MSF proteins are disordered in isolation, but even the limited energetic

contribution from the subunit interactions are able to stabilize them. This near-ordered character is also reflected at the functional level, fulfilling similar biological functions as ordered proteins, such as enzymatic activity, and are localized mostly in the extracellular space, which is in general depleted in IDPs. However, one of the particular functional advantages of disorder manifests in their tight regulation - e.g. PTMs govern the formation of the SOD (superoxide dismutase) complex, which in turn governs the adoption of a stable structure, a prerequisite of catalytic activity. Overall, near-order complexes represent the fine balance between characteristics of ordered proteins and IDPs, as they exhibit the main features from both subgroups to fulfill their biological roles.


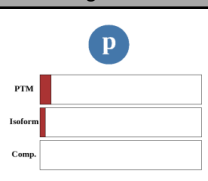
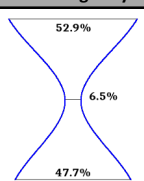
Group ID	Number of complexes	Example structure		Protein classes											
2	29	 2j0z		<ul style="list-style-type: none"> • RHHs • p53 tetramerization • Phd antitoxins 											
								Sequence		Structure		Stability			
								Composition	Subunit similarity	Secondary structures	Interface	Buried surface	Dominant atomic contacts	Energy	Role of inter-action
								Moderately variable	Near-identical	Helical & coil	large	small	Interchain	++	++
Functions		Regulation		Heterogeneity	Subcellular localization										
<ul style="list-style-type: none"> • regulation of transcription and gene expression • regulation of cell cycle 					<ul style="list-style-type: none"> • protein-DNA complex • NMBO • cytosol • chromatin • plasma membrane 										

Figure 29: Basic properties of multichain domain type complexes (Group 2).
 NMBO - non-membrane-bounded organelle.

Type 2, multi-chain domain structures are less tightly bound than the previous near-ordered complexes, and their stability depends more heavily on the interchain contacts between their highly similar subunits (see Figure 29). Their amino acid composition is highly variable, but usually markedly different from that of ordered proteins. These complexes prefer helical secondary structure elements, and they have relatively large hydrophobic interfaces. In terms of regulation, only a few of these interactions are under control via PTMs or alternative splicing.

Type 2 complexes play major roles in regulatory processes including transcription, gene expression and the cell cycle. This group includes ribbon-helix ribbon proteins and members of the p53 tetramerization family. The tumor suppressor protein p53 is crucial

in multicellular organisms, where it prevents cancer formation. Formation of a tetrameric structure of p53 is critical for its activation through DNA binding. A small destabilization of the tetrameric structure could result in dysfunction of tumor suppressor activity p53 (PDB: 2j0z, MFIB: MF4100003) [137]. Multi-chain domains are often involved in processes which are vital for the cell.

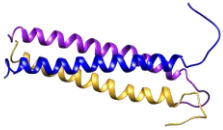
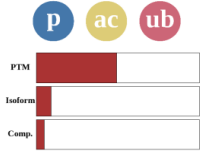
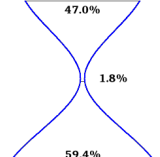
Group ID	Number of complexes	Example structure			Protein classes		
3	42	 1aq5			<ul style="list-style-type: none"> Regular coiled coils 		
Sequence		Structure			Stability		
Composition	Subunit similarity	Secondary structures	Interface	Buried surface	Dominant atomic contacts	Energy	Role of inter-action
Highly variable	Highly variable (oligomeric order correlates with subunit-dissimilarity)	Helical	large, hydrophobic	small, polar	Interchain	+ / ++	+++
Functions		Regulation	Heterogeneity	Subcellular localization			
<ul style="list-style-type: none"> regulation of gene expression and transcription organelle and membrane organization cellular localization 				<ul style="list-style-type: none"> plasma and organelle membrane <ul style="list-style-type: none"> NMBO cytosol cytoskeleton neuron part 			

Figure 30: Basic properties of coils and zippers type complexes (Group 3).
 NMBO - non-membrane-bounded organelle.

Type 3, coils and zippers is a structurally homogeneous group, as shown in Figure 30, composed entirely of coiled-coils. Despite the similar structures, their sequences are highly variable with no discernable unifying theme. They are enriched in helical and irregular secondary elements, and constituent protein chains bury only a small fraction of their polar surfaces upon interaction. The members of the group are weakly or moderately bound structures, where the interchain interactions are the most dominant. Proteins building up type 3 complexes mainly play role in membrane and organelle organization, and regulation of gene expression; and their interactions are often regulated by PTMs (mainly phosphorylation). A prime example, the cartilage oligomeric matrix protein may play a role in the structural integrity of cartilage via its interaction with other extracellular matrix proteins, such as collagen and fibronectin (PDB: 1aq5, MFIB: MF3110001). Despite their highly similar structures, coils and zippers convey a large variety of functions.

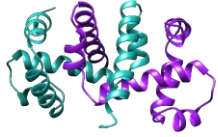
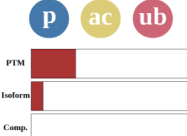
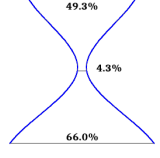
Group ID	Number of complexes	Example structure				Protein classes					
4	25	 <small>3wrp</small>				<ul style="list-style-type: none"> • Coiled coils with a twist • bHLHs • Trp repressors 					
		Sequence		Structure				Stability			
		Composition	Subunit similarity	Secondary structures	Interface			Buried surface	Dominant atomic contacts	Energy	Role of inter-action
		Moderately variable	Near-identical	Mainly helical	large, hydrophobic			average	-	+ / ++	+++
Functions		Regulation		Heterogeneity	Subcellular localization						
<ul style="list-style-type: none"> • regulation of transcription and gene expression • regulation of cell cycle 					<ul style="list-style-type: none"> • cytosol • nucleoplasm • non-membrane-bounded organelle • virion 						

Figure 31: Basic properties of compacted coils type complexes (Group 4).

Type 4, compacted coils are shown in Figure 31. Members of this group are sequentially similar to type 2, multi-chain domain complexes, while structurally being similar to type 3, coiled-coil complexes; however they contain additional structural elements, such as helix-loop-helices, resulting in a structural preference for both helices and irregular segments. Their interchain interactions between typically highly similar subunits are stronger, and their intrachain connections are relatively weak. A representative example is the Trp repressor protein (PDB: 3wrp, MFIB: MF2120008), which is an aporepressor. When complexed with L-tryptophan, it binds the operator region of the Trp operon, and prevents the initiation of transcription. Type 4 complexes in general resemble coils and zippers in terms of function as well, as they often occur in functions related to gene expression and transcription, and cell cycle regulation.

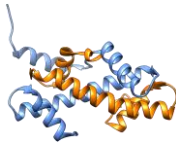
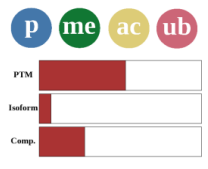
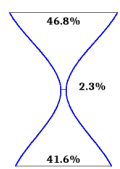
Group ID	Number of complexes	Example structure	Protein classes				
5	32	 3afa	<ul style="list-style-type: none"> • Histone-like interactions • L27 domains 				
Sequence		Structure			Stability		
Composition	Subunit similarity	Secondary structures	Interface	Buried surface	Dominant atomic contacts	Energy	Role of inter-action
Variable, but high G/P content	Moderately / highly variable	Mainly helical	large, hydrophobic	average	-	++	++
Functions		Regulation	Heterogeneity	Subcellular localization			
<ul style="list-style-type: none"> • chromosome organization • nucleosome assembly • DNA repair 				<ul style="list-style-type: none"> • NMBO • nucleosome • DNA packaging complex • nucleoplasm • extracellular vesicle 			

Figure 32: Basic properties of histone-like complexes (Group 5).

NMBO - non-membrane-bounded organelle.

Type 5, histone-like complexes include the L27 dimers, and complexes adopting the handshake fold of histone-like dimers (see Figure 32). They have variable sequences with a narrow structure space, depleted in extended structures. These complexes are formed by different subunits with large hydrophobic interfaces.

They have mostly chromosome and nucleosome-related functions, with direct contacts with the DNA. These interactions are often regulated by competitive binding, and to a lesser extent by alternative splicing, together with all of the four studied PTMs. Histones form parts of the nucleosome particle by dimerization and subsequent multimerization (PDB: 3afa, MFIB: MF2200005). The histone dimer contains both histone subunits in a highly intertwined conformation with a possible domain-swapped origin [138].


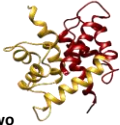
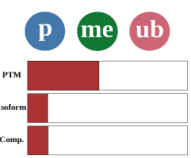
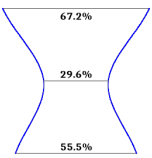
Group ID	Number of complexes	Example structures				Protein classes	
6	39					• Others	
Sequence		Structure				Stability	
Composition	Subunit similarity	Secondary structures	Interface	Buried surface	Dominant atomic contacts	Energy	Role of inter-action
Moderately variable	Highly variable	Mixed	-	-	-	Variable in wide range	Variable in wide range
Functions		Regulation	Heterogeneity		Subcellular localization		
<ul style="list-style-type: none"> regulation of gene expression and transcription regulation of biological quality and cellular component organization <ul style="list-style-type: none"> cellular localization regulation of programmed cell death 					<ul style="list-style-type: none"> NMBO nucleoplasm cytosol cell junction plasma membrane extracellular vesicle host 		

Figure 33: Basic properties of other type complexes (Group 6).

NMBO - non-membrane-bounded organelle.

Type 6 is the “other” group, consisting of complexes that are not members of any previous classes, containing highly unique complexes, most probably specialized to well-defined cellular functions. In accord, there is very little common among members of this group: their structure space has the widest range, often mixing helical, extended and coil elements to varying degrees (see Figure 33). These interactions are unique according to their sequence and structure features, they have specialized functions, including regulation of transcription and gene expression, cellular component organization, programmed cell death, and they occur in almost all compartments of the cell.

A well studied example is the S100B complex, which is a homodimeric member of the EF-hand calcium-binding protein superfamily (PDB: 1uwo, MFIB: MF2100013). It has been implicated in cellular processes such as cell differentiation and growth, and also plays a role in cytoskeletal structure and function. The previously mentioned interaction between nuclear receptor coactivators CBP and ACTR is also a member of this family (PDB: 1kbh, MFIB: MF220100).

This structure and sequence based classification of MSF complexes yields biologically meaningful complex groups. This not only enables a novel way for the classification and categorization of this previously understudied interaction class, but also has implications in biological inference. For example, the aggregation of transthyretin (PDB: 3a4d, MFIB: MF4100001) has been reported as the cause of the life-threatening

pathological conditions [139]. On the other hand, the aggregation of superoxide dismutase 1 (SOD1) (PDB: 2c9v, MFIB: MF2100014) often appears to accompany amyotrophic lateral sclerosis [140]. The localization and function of transthyretin and SOD1 are different, but their molecular pathogenesis is similar, and they both belong to the same complex type (type 1) in the proposed classification scheme. This suggests the possibility of other type 1 complexes to aggregate, and this can serve as a basis for future targeted biochemical characterization of other type 1 complexes, which in turn can serve as a lead to finding potential therapeutic targets. Similarly, certain pieces of biological and biomedical knowledge might be transferable within other complex types as well.

5. Conclusions and Future Directions

The growing number of known IDPs and their known crucial functions in signaling and regulatory pathways - typically mediated by their interactions with other proteins - motivate us to analyze protein complexes where either one or all participating partners are disordered. The structural and functional understanding of IDP-mediated interactions is essential for the understanding of the molecular basis of critical functions in the cell, the molecular basis for the architecture of protein-protein interaction networks, as well as the development of future approaches for therapeutic interventions. However, currently large-scale, systematic analyses focusing on the interactions of IDPs and the conveyed functions are lacking. This is not due to the lack of interest, but due to lack of data.

In order to understand and analyse the characteristics of the various binding modes of IDPs, I built the first, separate repositories for protein complexes that are formed between IDPs through mutual synergistic folding (MFIB), and for IDPs binding to ordered partner proteins via coupled folding and binding (DIBS). These databases present the first systematic and by far the most extensive collection of IDP complexes in structural detail, supported by high-quality manually curated annotations.

Reliable databases can serve as a stepping stone for a better understanding of IDP-mediated interactions, and in general, IDP functionality. Recent years have seen the development and major updates of several IDP-focused databases. DisProt, the central resource of experimentally verified instances of protein disorder was relocated to Europe and underwent a major overhaul. In parallel, MobiDB 3.0 became the central hub for intrinsic disorder sequence annotation, integrating data from DisProt, prediction methods and other sources. In turn, information from MobiDB (and thus from DisProt as well) is now part of UniProt annotations, serving as part of an ELIXIR core resource. Database integration is key in bioinformatics, and in line, both MFIB and DIBS are already integrated into MobiDB [103], and hence - at least indirectly - into UniProt as well. Further integration into DisProt has already commenced, and the majority of MFIB and DIBS data is expected to be included in the next major DisProt update, scheduled for late 2019. Regions described in MFIB and DIBS are also used to add experimentally verified annotations to the results of the newest incarnation of one of the most widely used disorder prediction methods, IUPred2A. Apart from data integration, a high quality dataset is always a reliable and robust resource for the development, training and testing of various bioinformatics methods, contributing to the development of improved

prediction algorithms. In the case of the newest version of ANCHOR, ANCHOR2, complexes from the DIBS database provided a reliable basis for recognizing disordered binding regions, with DIBS entries serving as training and test sets [89]. In parallel with serving with raw data for individual studies and bioinformatics development efforts, MFIB and DIBS also serve as a basis for a comprehensive analyses of interactions of IDPs. In my work, I uncovered the principal differences between the three fundamental types of interactions regarding sequence, structure, function, localization and regulation. This is the first major step in the basic understanding of how the interplay between protein folding and interaction modulates critical properties of the constituent chains and the resulting complexes. As a next step, further studies aiming at the delineation of the physical background of homooligomer MSF-proteins has already commenced in our research group [141].

The three studied interaction classes all have distinctive features that make them characteristically different from other types of interactions. However, none of the interaction classes are homogeneous, and all can realistically be divided into subgroups. In the case of ordered proteins, this grouping can be done by the tertiary structure classification of constituent domains, based on domain-focused databases (CATH [115] or SCOP [142]). These already defined fold classes can also be used as the basis for the definition of subgroups in case of IDPs that interact with ordered partner proteins, and this classification is already implemented in the DIBS server [126]. However, there is no already existing, evident way for the classification of MSF complexes, as these interactions have lacked focused structural studies, and have largely stayed in the uncharted territories of IDP-interactions.

Classification of MSF complexes based on structure and sequence properties yields biologically meaningful complex type definitions. The strength of the applied approach is that it is inherently scalable by choosing the appropriate number of clusters, and thus can be used to generate many more types or subtypes if the future accumulation of structural data warrants it. This classification system - apart from providing the first such effort for MSF complexes - can have direct applications as well, as knowledge gained for a specific complex - such as aggregation tendency, participation in the emergence of disease states, etc. - might be transferable to other complexes of the same type. Since its launch, MFIB has incorporated a grouping system for the contained complexes (see Table 4) to aid searches. However, this grouping system was created manually at the curation step of the database assembly (see chapter 4.1.2). The newly proposed automated

classification system will be implemented in the MFIB server, to offer a better way of searching, navigating, and comparing the presented MSF complexes.

References

- [1] S. J. Chan, S. O. Emdin, S. C. Kwok, J. M. Kramer, S. Falkmer, and D. F. Steiner, "Messenger RNA sequence and primary structure of preproinsulin in a primitive vertebrate, the Atlantic hagfish," *J. Biol. Chem.*, vol. 256, no. 14, pp. 7595-7602, 1981/7/25 1981.
- [2] G. N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan, "Stereochemistry of polypeptide chain configurations," *J. Mol. Biol.*, vol. 7, pp. 95-99, 1963/7 1963.
- [3] N. J. Greenfield, "Using circular dichroism spectra to estimate protein secondary structure," *Nat. Protoc.*, vol. 1, no. 6, pp. 2876-2890, 2007 2007.
- [4] M. Karplus, "The Levinthal paradox: yesterday and today," *Fold. Des.*, vol. 2, no. 4, pp. S69-75, 1997 1997.
- [5] C. J. Tsai, S. Kumar, B. Ma, and R. Nussinov, "Folding funnels, binding funnels, and protein function," *Protein Sci.*, vol. 8, no. 6, pp. 1181-1190, 1999/6 1999.
- [6] J. N. Onuchic, N. D. Socci, Z. Luthey-Schulten, and P. G. Wolynes, "Protein folding funnels: the nature of the transition state ensemble," *Fold. Des.*, vol. 1, no. 6, pp. 441-450, 1996 1996.
- [7] S. Radford, "Protein folding: progress made and promises ahead," *Trends Biochem. Sci.*, vol. 25, no. 12, pp. 611-618, 2000 2000.
- [8] O. C. Redfern, B. Dessailly, and C. A. Orengo, "Exploring the structure and function paradigm," *Curr. Opin. Struct. Biol.*, vol. 18, no. 3, pp. 394-402, 2008 2008.
- [9] P. E. Wright and H. Jane Dyson, "Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm," *J. Mol. Biol.*, vol. 293, no. 2, pp. 321-331, 1999 1999.
- [10] A. K. Dunker, Z. Obradovic, P. Romero, E. C. Garner, and C. J. Brown, "Intrinsic protein disorder in complete genomes," *Genome Inform. Ser. Workshop Genome Inform.*, vol. 11, pp. 161-171, 2000 2000.
- [11] R. Pancsa and P. Tompa, "Structural disorder in eukaryotes," *PLoS One*, vol. 7, no. 4, p. e34687, 2012/4/5 2012.
- [12] H. J. Dyson and P. E. Wright, "Intrinsically unstructured proteins and their functions," *Nat. Rev. Mol. Cell Biol.*, vol. 6, no. 3, pp. 197-208, 2005/3 2005.
- [13] H. Xie *et al.*, "Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions," *J. Proteome Res.*, vol. 6, no. 5, pp. 1882-1898, 2007/5 2007.
- [14] V. N. Uversky, J. R. Gillespie, and A. L. Fink, "Why are "natively unfolded" proteins unstructured under physiologic conditions?," *Proteins*, vol. 41, no. 3, pp. 415-427, 2000/11/15 2000.
- [15] P. Romero, Z. Obradovic, X. Li, E. C. Garner, C. J. Brown, and A. Keith Dunker, "Sequence complexity of disordered protein," *Proteins: Structure, Function, and Genetics*, vol. 42, no. 1, pp. 38-48, 2000 2000.
- [16] S. J. Demarest *et al.*, "Mutual synergistic folding in recruitment of CBP/p300 by p160 nuclear receptor coactivators," *Nature*, vol. 415, no. 6871, pp. 549-553, 2002/1/31 2002.
- [17] X. Peng, J. Jonas, and J. L. Silva, "Molten-globule conformation of Arc repressor monomers determined by high-pressure ¹H NMR spectroscopy," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 90, no. 5, pp. 1776-1780, 1993/3/1 1993.

- [18] M. Wells *et al.*, "Structure of tumor suppressor p53 and its intrinsically disordered N-terminal transactivation domain," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 105, no. 15, pp. 5762-5767, 2008/4/15 2008.
- [19] K. Gunasekaran, C.-J. Tsai, and R. Nussinov, "Analysis of ordered and disordered protein complexes reveals structural features discriminating between stable and unstable monomers," *J. Mol. Biol.*, vol. 341, no. 5, pp. 1327-1341, 2004/8/27 2004.
- [20] B. Mészáros, P. Tompa, I. Simon, and Z. Dosztányi, "Molecular principles of the interactions of disordered proteins," *J. Mol. Biol.*, vol. 372, no. 2, pp. 549-561, 2007/9/14 2007.
- [21] P. Csermely, R. Palotai, and R. Nussinov, "Induced fit, conformational selection and independent dynamic segments: an extended view of binding events," *Trends Biochem. Sci.*, vol. 35, no. 10, pp. 539-546, 2010/10 2010.
- [22] Y. Chebaro, A. J. Ballard, D. Chakraborty, and D. J. Wales, "Intrinsically disordered energy landscapes," *Sci. Rep.*, vol. 5, p. 10386, 2015/5/22 2015.
- [23] P. Tompa, "Intrinsically unstructured proteins," *Trends Biochem. Sci.*, vol. 27, no. 10, pp. 527-533, 2002/10 2002.
- [24] N. Zandany, L. Lewin, V. Nirenberg, I. Orr, and O. Yifrach, "Entropic clocks in the service of electrical signaling: 'Ball and chain' mechanisms for ion channel inactivation and clustering," *FEBS Lett.*, vol. 589, no. 19PartA, pp. 2441-2447, 2015 2015.
- [25] Y. Cheng, T. LeGall, C. J. Oldfield, A. K. Dunker, and V. N. Uversky, "Abundance of intrinsic disorder in protein associated with cardiovascular disease," *Biochemistry*, vol. 45, no. 35, pp. 10448-10460, 2006/9/5 2006.
- [26] J. H. Fong, B. A. Shoemaker, S. O. Garbuzynskiy, M. Y. Lobanov, O. V. Galzitskaya, and A. R. Panchenko, "Intrinsic disorder in protein interactions: insights from a comprehensive structural analysis," *PLoS Comput. Biol.*, vol. 5, no. 3, p. e1000316, 2009/3 2009.
- [27] V. N. Uversky *et al.*, "Unfoldomics of human diseases: linking protein intrinsic disorder with diseases," *BMC Genomics*, vol. 10 Suppl 1, p. S7, 2009/7/7 2009.
- [28] G. Mahmoudabadi, K. Rajagopalan, R. H. Getzenberg, S. Hannenhalli, G. Rangarajan, and P. Kulkarni, "Intrinsically disordered proteins and conformational noise: implications in cancer," *Cell Cycle*, vol. 12, no. 1, pp. 26-31, 2013/1/1 2013.
- [29] C. Dobson, "Protein Misfolding and Human Disease," *The Scientific World JOURNAL*, vol. 2, pp. 132-132, 2002 2002.
- [30] L. M. Luheshi, D. C. Crowther, and C. M. Dobson, "Protein misfolding and disease: from the test tube to the organism," *Curr. Opin. Chem. Biol.*, vol. 12, no. 1, pp. 25-31, 2008/2 2008.
- [31] V. N. Uversky, "Intrinsically disordered proteins and their (disordered) proteomes in neurodegenerative disorders," *Front. Aging Neurosci.*, vol. 7, p. 18, 2015/3/2 2015.
- [32] H. J. Dyson and P. E. Wright, "Coupling of folding and binding for unstructured proteins," *Curr. Opin. Struct. Biol.*, vol. 12, no. 1, pp. 54-60, 2002/2 2002.
- [33] V. N. Uversky, "Natively unfolded proteins: a point where biology waits for physics," *Protein Sci.*, vol. 11, no. 4, pp. 739-756, 2002/4 2002.
- [34] P. E. Wright and H. Jane Dyson, "Linking folding and binding," *Curr. Opin. Struct. Biol.*, vol. 19, no. 1, pp. 31-38, 2009 2009.

- [35] B. Mészáros, Z. Dosztányi, and I. Simon, "Disordered Binding Regions and Linear Motifs—Bridging the Gap between Two Models of Molecular Recognition," *PLoS One*, vol. 7, no. 10, p. e46829, 2012 2012.
- [36] M. Fuxreiter, P. Tompa, and I. Simon, "Local structural disorder imparts plasticity on linear motifs," *Bioinformatics*, vol. 23, no. 8, pp. 950-956, 2007/4/15 2007.
- [37] G. M. Verkhivker, D. Bouzida, D. K. Gehlhaar, P. A. Rejto, S. T. Freer, and P. W. Rose, "Simulating disorder-order transitions in molecular recognition of unstructured proteins: where folding meets binding," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 100, no. 9, pp. 5148-5153, 2003/4/29 2003.
- [38] P. Tompa and M. Fuxreiter, "Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions," *Trends Biochem. Sci.*, vol. 33, no. 1, pp. 2-8, 2008/1 2008.
- [39] C. J. Tsai and R. Nussinov, "Hydrophobic folding units at protein-protein interfaces: implications to protein folding and to protein-protein association," *Protein Sci.*, vol. 6, no. 7, pp. 1426-1437, 1997/7 1997.
- [40] R. Nussinov, D. Xu, and C.-J. Tsai, "Mechanism and evolution of protein dimerization," *Protein Sci.*, vol. 7, no. 3, pp. 533-544, 1998 1998.
- [41] J. A. O. Rumfeldt, C. Galvagnion, K. A. Vassall, and E. M. Meiering, "Conformational stability and folding mechanisms of dimeric proteins," *Prog. Biophys. Mol. Biol.*, vol. 98, no. 1, pp. 61-84, 2008/9 2008.
- [42] J. Liu, N. B. Perumal, C. J. Oldfield, E. W. Su, V. N. Uversky, and A. K. Dunker, "Intrinsic disorder in transcription factors," *Biochemistry*, vol. 45, no. 22, pp. 6873-6888, 2006/6/6 2006.
- [43] H. J. Dyson, "Roles of intrinsic disorder in protein-nucleic acid interactions," *Mol. Biosyst.*, vol. 8, no. 1, pp. 97-104, 2012/1 2012.
- [44] M. Kjaergaard, "Can proteins be intrinsically disordered inside a membrane?," *Intrinsically Disord Proteins*, vol. 3, no. 1, p. e984570, 2015/3/2 2015.
- [45] A. B. Sigalov, D. A. Aivazian, V. N. Uversky, and L. J. Stern, "Lipid-binding activity of intrinsically unstructured cytoplasmic domains of multichain immune recognition receptor signaling subunits," *Biochemistry*, vol. 45, no. 51, pp. 15731-15739, 2006/12/26 2006.
- [46] P. Faller, C. Hureau, and G. La Penna, "Metal ions and intrinsically disordered proteins and peptides: from Cu/Zn amyloid- β to general principles," *Acc. Chem. Res.*, vol. 47, no. 8, pp. 2252-2259, 2014/8/19 2014.
- [47] N. V. Chichkova *et al.*, "Divalent metal cation binding properties of human prothymosin alpha," *Eur. J. Biochem.*, vol. 267, no. 15, pp. 4745-4752, 2000/8 2000.
- [48] E. A. Grzybowska, "Calcium-Binding Proteins with Disordered Structure and Their Role in Secretion, Storage, and Cellular Signaling," *Biomolecules*, vol. 8, no. 2, 2018/6/19 2018.
- [49] A. A. Maximciuc, J. A. Putkey, Y. Shamoo, and K. R. Mackenzie, "Complex of calmodulin with a ryanodine receptor target reveals a novel, flexible binding mode," *Structure*, vol. 14, no. 10, pp. 1547-1556, 2006/10 2006.
- [50] P. Tompa, "The interplay between structure and function in intrinsically unstructured proteins," *FEBS Lett.*, vol. 579, no. 15, pp. 3346-3354, 2005 2005.
- [51] W.-L. Hsu *et al.*, "Exploring the binding diversity of intrinsically disordered proteins involved in one-to-many binding," *Protein Sci.*, vol. 22, no. 3, pp. 258-273, 2013/3 2013.

- [52] A. K. Dunker, M. S. Cortese, P. Romero, L. M. Iakoucheva, and V. N. Uversky, "Flexible nets. The roles of intrinsic disorder in protein interaction networks," *FEBS J.*, vol. 272, no. 20, pp. 5129-5148, 2005/10 2005.
- [53] Z. Dosztányi, J. Chen, A. K. Dunker, I. Simon, and P. Tompa, "Disorder and sequence repeats in hub proteins and their implications for network evolution," *J. Proteome Res.*, vol. 5, no. 11, pp. 2985-2995, 2006/11 2006.
- [54] C. Haynes *et al.*, "Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes," *PLoS Comput. Biol.*, vol. 2, no. 8, p. e100, 2006/8/4 2006.
- [55] H. Jeong, S. P. Mason, A. L. Barabási, and Z. N. Oltvai, "Lethality and centrality in protein networks," *Nature*, vol. 411, no. 6833, pp. 41-42, 2001/5/3 2001.
- [56] T. F. Sumter *et al.*, "The High Mobility Group A1 (HMGA1) Transcriptome in Cancer and Development," *Curr. Mol. Med.*, vol. 16, no. 4, pp. 353-393, 2016 2016.
- [57] S. J. Metallo, "Intrinsically disordered proteins are potential drug targets," *Curr Opin Chem Biol*, vol. 14, no. 4, pp. 481-8, Aug 2010.
- [58] P. Kulkarni, "Intrinsically disordered proteins and prostate cancer: pouring new wine in an old bottle," *Asian J Androl*, vol. 18, no. 5, pp. 659-61, Sep-Oct 2016.
- [59] C. A. Galea, V. R. Pagala, J. C. Obenauer, C.-G. Park, C. A. Slaughter, and R. W. Kriwacki, "Proteomic studies of the intrinsically unstructured mammalian proteome," *J. Proteome Res.*, vol. 5, no. 10, pp. 2839-2848, 2006/10 2006.
- [60] V. N. Uversky, "A protein-chameleon: conformational plasticity of alpha-synuclein, a disordered protein involved in neurodegenerative disorders," *J. Biomol. Struct. Dyn.*, vol. 21, no. 2, pp. 211-234, 2003/10 2003.
- [61] A. K. Dunker, A. Keith Dunker, and Z. Obradovic, "The protein trinity—linking function and disorder," *Nat. Biotechnol.*, vol. 19, no. 9, pp. 805-806, 2001 2001.
- [62] I. C. Felli and R. Pierattelli, "Recent progress in NMR spectroscopy: toward the study of intrinsically disordered proteins of increasing size and complexity," *IUBMB Life*, vol. 64, no. 6, pp. 473-481, 2012/6 2012.
- [63] D. A. Torchia, "Dynamics of biomolecules from picoseconds to seconds at atomic resolution," *J. Magn. Reson.*, vol. 212, no. 1, pp. 1-10, 2011/9 2011.
- [64] L. B. Chemes, L. G. Alonso, M. G. Noval, and G. de Prat-Gay, "Circular dichroism techniques for the analysis of intrinsically disordered proteins and domains," *Methods Mol. Biol.*, vol. 895, pp. 387-404, 2012 2012.
- [65] K. A. Ball, D. E. Wemmer, and T. Head-Gordon, "Comparison of structure determination methods for intrinsically disordered amyloid- β peptides," *J. Phys. Chem. B*, vol. 118, no. 24, pp. 6405-6416, 2014/6/19 2014.
- [66] D. Sahu, M. Bastidas, C. W. Lawrence, W. G. Noid, and S. A. Showalter, "Assessing Coupled Protein Folding and Binding Through Temperature-Dependent Isothermal Titration Calorimetry," *Methods Enzymol.*, vol. 567, pp. 23-45, 2016 2016.
- [67] V. Receveur-Bréchet, J.-M. Bourhis, V. N. Uversky, B. Canard, and S. Longhi, "Assessing protein disorder and induced folding," *Proteins: Struct. Funct. Bioinf.*, vol. 62, no. 1, pp. 24-45, 2005 2005.
- [68] P. Schanda and M. Ernst, "Studying Dynamics by Magic-Angle Spinning Solid-State NMR Spectroscopy: Principles and Applications to Biomolecules," *Prog. Nucl. Magn. Reson. Spectrosc.*, vol. 96, pp. 1-46, 2016/8 2016.

- [69] C. Bracken, L. M. Iakoucheva, P. R. Romero, and A. K. Dunker, "Combining prediction, computation and experiment for the characterization of protein disorder," *Curr. Opin. Struct. Biol.*, vol. 14, no. 5, pp. 570-576, 2004/10 2004.
- [70] B. Monastyrskyy, A. Kryshtafovych, J. Moult, A. Tramontano, and K. Fidelis, "Assessment of protein disorder region predictions in CASP10," *Proteins*, vol. 82 Suppl 2, pp. 127-137, 2014/2 2014.
- [71] Z. Dosztányi, V. Csizmók, P. Tompa, and I. Simon, "The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins," *J. Mol. Biol.*, vol. 347, no. 4, pp. 827-839, 2005/4/8 2005.
- [72] O. V. Galzitskaya, S. O. Garbuzynskiy, and M. Y. Lobanov, "FoldUnfold: web server for the prediction of disordered regions in protein chain," *Bioinformatics*, vol. 22, no. 23, pp. 2948-2949, 2006/12/1 2006.
- [73] J. Prilusky *et al.*, "FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded," *Bioinformatics*, vol. 21, no. 16, pp. 3435-3438, 2005/8/15 2005.
- [74] Z. Dosztányi, "Prediction of protein disorder based on IUPred," *Protein Sci.*, vol. 27, no. 1, pp. 331-340, 2018/1 2018.
- [75] Z. Obradovic, K. Peng, S. Vucetic, P. Radivojac, and A. K. Dunker, "Exploiting heterogeneous sequence properties improves prediction of protein disorder," *Proteins*, vol. 61 Suppl 7, pp. 176-182, 2005 2005.
- [76] K. Peng, P. Radivojac, S. Vucetic, A. K. Dunker, and Z. Obradovic, "Length-dependent prediction of protein intrinsic disorder," *BMC Bioinformatics*, vol. 7, p. 208, 2006/4/17 2006.
- [77] Z. Obradovic, K. Peng, S. Vucetic, P. Radivojac, C. J. Brown, and A. K. Dunker, "Predicting intrinsic disorder from amino acid sequence," *Proteins*, vol. 53 Suppl 6, pp. 566-572, 2003 2003.
- [78] P. Radivojac, Z. Obradović, C. J. Brown, and A. K. Dunker, "Prediction of boundaries between intrinsically ordered and disordered protein regions," *Pac. Symp. Biocomput.*, pp. 216-227, 2003 2003.
- [79] I. Walsh, A. J. M. Martin, T. Di Domenico, and S. C. E. Tosatto, "ESpritz: accurate and fast prediction of protein disorder," *Bioinformatics*, vol. 28, no. 4, pp. 503-509, 2012/2/15 2012.
- [80] J. M. Bujnicki, A. Elofsson, D. Fischer, and L. Rychlewski, "LiveBench-2: large-scale automated evaluation of protein structure prediction servers," *Proteins*, vol. Suppl 5, pp. 184-191, 2001 2001.
- [81] X. Fan and L. Kurgan, "Accurate prediction of disorder in protein chains with a comprehensive and empirically designed consensus," *J. Biomol. Struct. Dyn.*, vol. 32, no. 3, pp. 448-464, 2013 2013.
- [82] I. Walsh, A. J. M. Martin, T. Di Domenico, A. Vullo, G. Pollastri, and S. C. E. Tosatto, "CSpritz: accurate prediction of protein disorder segments with annotation for homology, secondary structure and linear motifs," *Nucleic Acids Res.*, vol. 39, no. Web Server issue, pp. W190-6, 2011/7 2011.
- [83] T. Zhang, E. Faraggi, B. Xue, A. Keith Dunker, V. N. Uversky, and Y. Zhou, "SPINE-D: Accurate Prediction of Short and Long Disordered Regions by a Single Neural-Network Based Method," *J. Biomol. Struct. Dyn.*, vol. 29, no. 4, pp. 799-813, 2012 2012.
- [84] J. J. Ward, J. S. Sodhi, L. J. McGuffin, B. F. Buxton, and D. T. Jones, "Prediction and functional analysis of native disorder in proteins from the three kingdoms of life," *J Mol Biol*, vol. 337, no. 3, pp. 635-45, Mar 26 2004.

- [85] A. Schlessinger, M. Punta, G. Yachdav, L. Kajan, and B. Rost, "Improved disorder prediction by combination of orthogonal approaches," *PLoS One*, vol. 4, no. 2, p. e4433, 2009/2/11 2009.
- [86] J. J. Ward, J. S. Sodhi, L. J. McGuffin, B. F. Buxton, and D. T. Jones, "Prediction and functional analysis of native disorder in proteins from the three kingdoms of life," *J. Mol. Biol.*, vol. 337, no. 3, pp. 635-645, 2004/3/26 2004.
- [87] D. B. Roche, M. T. Buenavista, S. J. Tetchner, and L. J. McGuffin, "The IntFOLD server: an integrated web resource for protein fold recognition, 3D model quality assessment, intrinsic disorder prediction, domain prediction and ligand binding site prediction," *Nucleic Acids Res.*, vol. 39, no. suppl, pp. W171-W176, 2011 2011.
- [88] B. Mészáros, I. Simon, and Z. Dosztányi, "Prediction of Protein Binding Regions in Disordered Proteins," *PLoS Comput. Biol.*, vol. 5, no. 5, p. e1000376, 2009 2009.
- [89] B. Mészáros, G. Erdos, and Z. Dosztányi, "IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding," *Nucleic Acids Res.*, vol. 46, no. W1, pp. W329-W337, 2018/7/2 2018.
- [90] D. T. Jones and D. Cozzetto, "DISOPRED3: precise disordered region predictions with annotated protein-binding activity," *Bioinformatics*, vol. 31, no. 6, pp. 857-863, 2015/3/15 2015.
- [91] F. M. Disfani *et al.*, "MoRFPred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins," *Bioinformatics*, vol. 28, no. 12, pp. i75-83, 2012/6/15 2012.
- [92] R. Sharma, M. Bayarjargal, T. Tsunoda, A. Patil, and A. Sharma, "MoRFPred-plus: Computational Identification of MoRFs in Protein Sequences using Physicochemical Properties and HMM profiles," *J. Theor. Biol.*, vol. 437, pp. 9-16, 2018/1/21 2018.
- [93] N. Malhis, M. Jacobson, and J. Gsponer, "MoRFchibi SYSTEM: software tools for the identification of MoRFs in protein sequences," *Nucleic Acids Res.*, vol. 44, no. W1, pp. W488-93, 2016/7/8 2016.
- [94] R. Sharma, G. Raicar, T. Tsunoda, A. Patil, and A. Sharma, "OPAL: prediction of MoRF regions in intrinsically disordered protein sequences," *Bioinformatics*, vol. 34, no. 11, pp. 1850-1858, 2018/6/1 2018.
- [95] D. Piovesan *et al.*, "DisProt 7.0: a major update of the database of disordered proteins," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D219-D227, 2017/1/4 2017.
- [96] S. Fukuchi *et al.*, "IDEAL: Intrinsically Disordered proteins with Extensive Annotations and Literature," *Nucleic Acids Res.*, vol. 40, no. Database issue, pp. D507-11, 2012/1 2012.
- [97] J.-F. Yu *et al.*, "DisBind: A database of classified functional binding sites in disordered and structured regions of intrinsically disordered proteins," *BMC Bioinformatics*, vol. 18, no. 1, p. 206, 2017/4/5 2017.
- [98] M. Miskei, C. Antal, and M. Fuxreiter, "FuzDB: database of fuzzy complexes, a tool to develop stochastic structure-function relationships for protein complexes and higher-order assemblies," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D228-D235, 2017/1/4 2017.
- [99] H. Dinkel *et al.*, "ELM 2016--data update and new functionality of the eukaryotic linear motif resource," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D294-300, 2016/1/4 2016.
- [100] M. E. Oates *et al.*, "D²P²: database of disordered protein predictions," *Nucleic Acids Res.*, vol. 41, no. Database issue, pp. D508-16, 2013/1 2013.

- [101] T. Ishida and K. Kinoshita, "PrDOS: prediction of disordered protein regions from amino acid sequence," *Nucleic Acids Res.*, vol. 35, no. Web Server issue, pp. W460-4, 2007/7 2007.
- [102] M. F. Ghalwash, A. K. Dunker, and Z. Obradovic, "Uncertainty analysis in protein disorder prediction," *Mol Biosyst*, vol. 8, no. 1, pp. 381-91, Jan 2012.
- [103] D. Piovesan *et al.*, "MobiDB 3.0: more annotations for intrinsic disorder, conformational diversity and interactions in proteins," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D471-D476, 2017 2017.
- [104] R. Linding, R. B. Russell, V. Neduva, and T. J. Gibson, "GlobPlot: Exploring protein sequences for globularity and disorder," *Nucleic Acids Res.*, vol. 31, no. 13, pp. 3701-3708, 2003/7/1 2003.
- [105] M. Necci, D. Piovesan, Z. Dosztányi, and S. C. E. Tosatto, "MobiDB-lite: fast and highly specific consensus prediction of intrinsic disorder in proteins," *Bioinformatics*, vol. 33, no. 9, pp. 1402-1404, 2017/5/1 2017.
- [106] P. Tompa, N. E. Davey, T. J. Gibson, and M. M. Babu, "A million peptide motifs for the molecular biologist," *Mol. Cell*, vol. 55, no. 2, pp. 161-169, 2014/7/17 2014.
- [107] H. M. Berman *et al.*, "The Protein Data Bank," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 235-242, 2000/1/1 2000.
- [108] The UniProt Consortium, "UniProt: the universal protein knowledgebase," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D158-D169, 2017/1/4 2017.
- [109] T. UniProt Consortium, "UniProt: the universal protein knowledgebase," *Nucleic Acids Res.*, vol. 46, no. 5, p. 2699, 2018/3/16 2018.
- [110] B. E. Suzek, Y. Wang, H. Huang, P. B. McGarvey, C. H. Wu, and C. UniProt, "UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches," *Bioinformatics*, vol. 31, no. 6, pp. 926-932, 2015/3/15 2015.
- [111] S. Fukuchi *et al.*, "IDEAL in 2014 illustrates interaction networks composed of intrinsically disordered proteins and their binding partners," *Nucleic Acids Res.*, vol. 42, no. Database issue, pp. D320-5, 2014/1 2014.
- [112] R. D. Finn *et al.*, "The Pfam protein families database: towards a more sustainable future," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D279-85, 2016/1/4 2016.
- [113] M. Ashburner *et al.*, "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nat. Genet.*, vol. 25, no. 1, pp. 25-29, 2000/5 2000.
- [114] The Gene Ontology Consortium, "Expansion of the Gene Ontology knowledgebase and resources," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D331-D338, 2017/1/4 2017.
- [115] N. L. Dawson *et al.*, "CATH: an expanded resource to predict protein function through structure and sequence," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D289-D295, 2017/1/4 2017.
- [116] M. Gouw *et al.*, "The eukaryotic linear motif resource - 2018 update," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D428-D434, 2018/1/4 2018.
- [117] P. V. Hornbeck, B. Zhang, B. Murray, J. M. Kornhauser, V. Latham, and E. Skrzypek, "PhosphoSitePlus, 2014: mutations, PTMs and recalibrations," *Nucleic Acids Res.*, vol. 43, no. Database issue, pp. D512-20, 2015/1 2015.
- [118] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *J. Mol. Biol.*, vol. 215, no. 3, pp. 403-410, 1990/10/5 1990.

- [119] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers*, vol. 22, no. 12, pp. 2577-2637, 1983/12 1983.
- [120] S. Hubbard and J. Thornton, "Naccess," *Journal of molecular biology*, 1993 1993.
- [121] J. Schymkowitz, J. Borg, F. Stricher, R. Nys, F. Rousseau, and L. Serrano, "The FoldX web server: an online force field," *Nucleic Acids Res*, vol. 33, no. Web Server issue, pp. W382-8, Jul 1 2005.
- [122] L. C. Xue, J. P. Rodrigues, P. L. Kastiris, A. M. Bonvin, and A. Vangone, "PRODIGY: a web server for predicting the binding affinity of protein-protein complexes," *Bioinformatics*, vol. 32, no. 23, pp. 3676-3678, Dec 1 2016.
- [123] A. D. Baxevanis, "The Importance of Biological Databases in Biological Discovery," in *Current Protocols in Bioinformatics*, 2011.
- [124] B. Rost, "Twilight zone of protein sequence alignments," *Protein Eng. Des. Sel.*, vol. 12, no. 2, pp. 85-94, 1999 1999.
- [125] A. Giartosio *et al.*, "Thermal stability of hexameric and tetrameric nucleoside diphosphate kinases. Effect of subunit interaction," *J. Biol. Chem.*, vol. 271, no. 30, pp. 17845-17851, 1996/7/26 1996.
- [126] E. Schad, E. Fichó, R. Pancsa, I. Simon, Z. Dosztányi, and B. Mészáros, "DIBS: a repository of disordered binding sites mediating interactions with ordered proteins," *Bioinformatics*, vol. 34, no. 3, pp. 535-537, 2018/2/1 2018.
- [127] A. Campen, R. M. Williams, C. J. Brown, J. Meng, V. N. Uversky, and A. K. Dunker, "TOP-IDP-scale: a new amino acid scale measuring propensity for intrinsic disorder," *Protein Pept. Lett.*, vol. 15, no. 9, pp. 956-963, 2008 2008.
- [128] V. N. Uversky, "The multifaceted roles of intrinsic disorder in protein complexes," *FEBS Lett.*, vol. 589, no. 19 Pt A, pp. 2498-2506, 2015/9/14 2015.
- [129] H. Xie *et al.*, "Functional anthology of intrinsic disorder. 3. Ligands, post-translational modifications, and diseases associated with intrinsically disordered proteins," *J. Proteome Res.*, vol. 6, no. 5, pp. 1917-1932, 2007/5 2007.
- [130] P. E. Wright and H. J. Dyson, "Intrinsically disordered proteins in cellular signalling and regulation," *Nat. Rev. Mol. Cell Biol.*, vol. 16, no. 1, pp. 18-29, 2015/1 2015.
- [131] J. Gsponer, M. E. Futschik, S. A. Teichmann, and M. M. Babu, "Tight regulation of unstructured proteins: from transcript synthesis to protein degradation," *Science*, vol. 322, no. 5906, pp. 1365-1368, 2008/11/28 2008.
- [132] T. W. Kirby *et al.*, "The nuclease A inhibitor represents a new variation of the rare PR-1 fold," *J. Mol. Biol.*, vol. 320, no. 4, pp. 771-782, 2002/7/19 2002.
- [133] C. Corbi-Verge and P. M. Kim, "Motif mediated protein-protein interactions as drug targets," *Cell Commun. Signal.*, vol. 14, p. 8, 2016/3/2 2016.
- [134] M. Y. Hein *et al.*, "A human interactome in three quantitative dimensions organized by stoichiometries and abundances," *Cell*, vol. 163, no. 3, pp. 712-723, 2015/10/22 2015.
- [135] H. Hegyi, L. Kalmar, T. Horvath, and P. Tompa, "Verification of alternative splicing variants based on domain integrity, truncation length and intrinsic protein disorder," *Nucleic Acids Res.*, vol. 39, no. 4, pp. 1208-1219, 2011/3 2011.
- [136] M. Buljan *et al.*, "Alternative splicing of intrinsically disordered regions and rewiring of protein interactions," *Curr. Opin. Struct. Biol.*, vol. 23, no. 3, pp. 443-450, 2013/6 2013.
- [137] P. Chène, "The role of tetramerization in p53 function. - PubMed - NCBI."

- [138] V. Alva, M. Ammelburg, J. Söding, and A. N. Lupas, "On the origin of the histone fold," *BMC Struct. Biol.*, vol. 7, p. 17, 2007/3/28 2007.
- [139] R. J. Gasperini, D. W. Klaver, X. Hou, M. I. Aguilar, and D. H. Small, "Mechanisms of transthyretin aggregation and toxicity," *Subcell Biochem*, vol. 65, pp. 211-24, 2012.
- [140] L. Saelices *et al.*, "Uncovering the Mechanism of Aggregation of Human Transthyretin," *J. Biol. Chem.*, vol. 290, no. 48, pp. 28932-28943, 2015/11/27 2015.
- [141] C. Magyar, A. Mentés, E. Fichó, M. Cserző, and I. Simon, "Physical Background of the Disordered Nature of "Mutual Synergetic Folding" Proteins," *Int. J. Mol. Sci.*, vol. 19, no. 11, 2018/10/26 2018.
- [142] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, "SCOP: a structural classification of proteins database for the investigation of sequences and structures," *J. Mol. Biol.*, vol. 247, no. 4, pp. 536-540, 1995/4/7 1995.

ADATLAP

a doktori értekezés nyilvánosságra hozatalához*

I. A doktori értekezés adatai

A szerző neve: Fichó Erzsébet

MTMT-azonosító: 10061311

A doktori értekezés címe és alcíme: Orders of disorder: computational analysis of the interactions of intrinsically disordered proteins (Rendezett rendezetlenség: a rendezetlen fehérjék kölcsönhatásainak elméleti vizsgálata)

DOI-azonosító⁴⁶: 10.15476/ELTE.2018.256

A doktori iskola neve: ELTE TTK, Biológia Doktori Iskola

A doktori iskolán belüli doktori program neve: Szerkezeti Biokémia

A témavezető neve és tudományos fokozata: Dr. Mészáros Bálint, PhD

A témavezető munkahelye: MTA-ELTE Lendület Bioinformatika kutatócsoport

II. Nyilatkozatok

1. A doktori értekezés szerzőjeként

a) hozzájárulok, hogy a doktori fokozat megszerzését követően a doktori értekezésem és a tézisek nyilvánosságra kerüljenek az ELTE Digitális Intézményi Tudástárban. Felhatalmazom a Természettudományi kar Dékáni Hivatali Doktori, Habilitációs és Nemzetközi Ügyek Csoportjának ügyintézőjét, hogy az értekezést és a téziseket feltöltse az ELTE Digitális Intézményi Tudástárba, és ennek során kitöltse a feltöltéshez szükséges nyilatkozatokat.

b) kérem, hogy a mellékelt kérelemben részletezett szabadalmi, illetőleg oltalmi bejelentés közzétételéig a doktori értekezést ne bocsássák nyilvánosságra az Egyetemi Könyvtárban és az ELTE Digitális Intézményi Tudástárban;

c) kérem, hogy a nemzetbiztonsági okból minősített adatot tartalmazó doktori értekezést a minősítés (dátum)-ig tartó időtartama alatt ne bocsássák nyilvánosságra az Egyetemi Könyvtárban és az ELTE Digitális Intézményi Tudástárban;

d) kérem, hogy a mű kiadására vonatkozó mellékelt kiadó szerződésre tekintettel a doktori értekezést a könyv megjelenéséig ne bocsássák nyilvánosságra az Egyetemi Könyvtárban, és az ELTE Digitális Intézményi Tudástárban csak a könyv bibliográfiai adatait tegyék közzé. Ha a könyv a fokozatszerzést követően egy évig nem jelenik meg, hozzájárulok, hogy a doktori értekezésem és a tézisek nyilvánosságra kerüljenek az Egyetemi Könyvtárban és az ELTE Digitális Intézményi Tudástárban.

2. A doktori értekezés szerzőjeként kijelentem, hogy

a) az ELTE Digitális Intézményi Tudástárba feltöltendő doktori értekezés és a tézisek saját eredeti, önálló szellemi munkám és legjobb tudomásom szerint nem sértem vele senki szerzői jogait;

b) a doktori értekezés és a tézisek nyomtatott változatai és az elektronikus adathordozón benyújtott tartalmak (szöveg és ábrák) mindenben megegyeznek.

3. A doktori értekezés szerzőjeként hozzájárulok a doktori értekezés és a tézisek szövegének plágiumkereső adatbázisba helyezéséhez és plágiumellenőrző vizsgálatok lefuttatásához.

Kelt: Budapest, 2018. december 10.



.....
a doktori értekezés szerzőjének aláírása

*ELTE SZMSZ SZMR 12. sz. melléklet